CS2

Risk Modelling and Survival Analysis

Combined Materials Pack
for exams in 2019

The Actuarial Education Company
on behalf of the Institute and Faculty of Actuaries

Subject CS2

2019 Study Guide

Introduction

This Study Guide has been created to help guide you through Subject CS2. It contains all the information that you will need before starting to study Subject CS2 for the 2019 exams, and you may also find it useful to refer to throughout your Subject CS2 journey.

The guide is split into two parts:

- Part 1 contains general information about the Core Principles subjects
- Part 2 contains specific information about Subject CS2.

Please read this Study Guide carefully before reading the Course Notes, even if you have studied for some actuarial exams before.

Contents

1.1 Before you start

When studying for the UK actuarial exams, you will need:

- a copy of the **Formulae and Tables for Examinations of the Faculty of Actuaries and the Institute of Actuaries, 2nd Edition (2002)** – these are often referred to as simply the *Yellow Tables* or the *Tables*
- a 'permitted' **scientific calculator** – you will find the list of permitted calculators on the profession's website. Please check the list carefully, since it is reviewed each year.

These are both available from the Institute and Faculty of Actuaries' eShop. Please visit **www.actuaries.org.uk**.

1.2 Core study material

This section explains the role of the Syllabus, Core Reading and supplementary ActEd text. It also gives guidance on how to use these materials most effectively in order to pass the exam.

Some of the information below is also contained in the introduction to the Core Reading produced by the Institute and Faculty of Actuaries.

Syllabus

The Syllabus for Subject CS2 has been produced by the Institute and Faculty of Actuaries. The relevant individual Syllabus Objectives are included at the start of each course chapter and a complete copy of the Syllabus is included in Section 2.2 of this Study Guide. We recommend that you use the Syllabus as an important part of your study.

Core Reading

The Core Reading has been produced by the Institute and Faculty of Actuaries. The purpose of the Core Reading is to assist in ensuring that tutors, students and examiners have clear shared appreciation of the requirements of the syllabus for the qualification examinations for Fellowship of the Institute and Faculty of Actuaries.

The Core Reading supports coverage of the syllabus in helping to ensure that both depth and breadth are reinforced. It is therefore important that students have a good understanding of the concepts covered by the Core Reading.

The examinations require students to demonstrate their understanding of the concepts given in the syllabus and described in the Core Reading; this will be based on the legislation, professional guidance *etc* that are in force when the Core Reading is published, *ie* on 31 May in the year preceding the examinations.

Therefore the exams in April and September 2019 will be based on the Syllabus and Core Reading as at 31 May 2018. We recommend that you always use the up-to-date Core Reading to prepare for the exams.

Examiners will have this Core Reading when setting the papers. In preparing for examinations, students are advised to work through past examination questions and will find additional tuition helpful. The Core Reading will be updated each year to reflect changes in the syllabus, to reflect current practice, and in the interest of clarity.

Accreditation

The Institute and Faculty of Actuaries would like to thank the numerous people who have helped in the development of the material contained in this Core Reading.

ActEd text

Core Reading deals with each syllabus objective and covers what is needed to pass the exam. However, the tuition material that has been written by ActEd enhances it by giving examples and further explanation of key points. Here is an excerpt from some ActEd Course Notes to show you how to identify Core Reading and the ActEd material. **Core Reading is shown in this bold font.**

Note that in the example given above, the index *will* fall if the actual share price goes below the theoretical ex-rights share price. Again, this is consistent with what would happen to an underlying portfolio.

After allowing for chain-linking, **the formula for the investment index then becomes:**

$$I(t) = \frac{\sum_i N_{i,t}P_{i,t}}{B(t)}$$

where $N_{i,t}$ **is the number of shares issued for the** ***i*****th constituent at time** ***t*****;**

$B(t)$ **is the base value, or divisor, at time** ***t*****.**

This is ActEd text

This is Core Reading

Here is an excerpt from some ActEd Course Notes to show you how to identify Core Reading for R code.

The R code to draw a scatterplot for a bivariate data frame, `<data>`**, is:**

```
plot(<data>)
```

Further explanation on the use of R will not be provided in the Course Notes, but instead be picked up in the Paper B Online Resources (PBOR). We recommend that you refer to and use PBOR at the end of each chapter, or couple of chapters, that contains a significant number of R references.

Copyright

All study material produced by ActEd is copyright and is sold for the exclusive use of the purchaser. The copyright is owned by Institute and Faculty Education Limited, a subsidiary of the Institute and Faculty of Actuaries. Unless prior authority is granted by ActEd, you may not hire out, lend, give out, sell, store or transmit electronically or photocopy any part of the study material. You must take care of your study material to ensure that it is not used or copied by anybody else.

Legal action will be taken if these terms are infringed. In addition, we may seek to take disciplinary action through the Institute and Faculty of Actuaries or through your employer.

These conditions remain in force after you have finished using the course.

1.3 ActEd study support

This section gives a description of the products offered by ActEd.

Successful students tend to undertake three main study activities:

1. *Learning* – initial study and understanding of subject material
2. *Revision* – learning subject material and preparing to tackle exam-style questions
3. *Rehearsal* – answering exam-style questions, culminating in answering questions at exam speed without notes.

Different approaches suit different people. For example, you may like to learn material gradually over the months running up to the exams or you may do your revision in a shorter period just before the exams. Also, these three activities will almost certainly overlap.

We offer a flexible range of products to suit you and let you control your own learning and exam preparation. The following table shows the products that we produce. Note that not all products are available for all subjects.

LEARNING	**LEARNING & REVISION**	**REVISION**	**REVISION & REHEARSAL**	**REHEARSAL**
Course Notes	Assignments	Flashcards	Revision Notes	Mock Exam
	Combined Materials Pack (CMP)		ASET	Mock Marking
	Assignment Marking			
	Tutorials			
	Online Classroom			

The products and services are described in more detail below.

'Learning' products

Course Notes

The Course Notes will help you develop the basic knowledge and understanding of principles needed to pass the exam. They incorporate the complete Core Reading and include full explanation of all the syllabus objectives, with worked examples and questions (including some past exam questions) to test your understanding.

Each chapter includes:

- the relevant syllabus objectives
- a chapter summary
- a page of important formulae or definitions (where appropriate)
- practice questions with full solutions.

Paper B Online Resources (PBOR)

The Paper B Online Resources (PBOR) will help you prepare for the computer-based paper. Delivered through a virtual learning environment (VLE), you will have access to worked examples and practice questions. PBOR will also include the Y Assignments, which are two exam-style assessments.

'Learning & revision' products

X Assignments

The Series X Assignments are written assessments that cover the material in each part of the course in turn. They can be used to both develop and test your understanding of the material.

Combined Materials Pack (CMP)

The Combined Materials Pack (CMP) comprises the Course Notes, PBOR and the Series X Assignments.

The CMP is available in **eBook** format for viewing on a range of electronic devices. eBooks can be ordered separately or as an addition to paper products. Visit **www.ActEd.co.uk** for full details about the eBooks that are available, compatibility with different devices, software requirements and printing restrictions.

X / Y Assignment Marking

We are happy to mark your attempts at the X and/or Y assignments. Marking is not included with the Assignments or the CMP and you need to order both Series X and Series Y Marking separately. You should submit your script as an attachment to an email, in the format detailed in your assignment instructions. You will be able to download your marker's feedback via a secure link on the internet.

Don't underestimate the benefits of doing and submitting assignments:

- Question practice during this phase of your study gives an early focus on the end goal of answering exam-style questions.
- You're incentivised to keep up with your study plan and get a regular, realistic assessment of your progress.
- Objective, personalised feedback from a high quality marker will highlight areas on which to work, and help with exam technique.

In a recent study, we found that students who attempt more than half the assignments have significantly higher pass rates.

There are two different types of marking product: Series Marking and Marking Vouchers.

Series Marking

Series Marking applies to a specified subject, session and student. If you purchase Series Marking, you will **not** be able to defer the marking to a future exam sitting or transfer it to a different subject or student.

We typically provide full solutions with the Series Assignments. However, if you order Series Marking at the same time as you order the Series Assignments, you can choose whether or not to receive a copy of the solutions in advance. If you choose not to receive them with the study material, you will be able to download the solutions via a secure link on the internet when your marked script is returned (or following the final deadline date if you do not submit a script).

If you are having your attempts at the assignments marked by ActEd, you should submit your scripts regularly throughout the session, in accordance with the schedule of recommended dates set out in information provided with the assignments. This will help you to pace your study throughout the session and leave an adequate amount of time for revision and question practice.

The recommended submission dates are realistic targets for the majority of students. Your scripts will be returned more quickly if you submit them well before the final deadline dates.

Any script submitted *after* the relevant final deadline date will not be marked. It is your responsibility to ensure that we receive scripts in good time.

Marking Vouchers

Marking Vouchers give the holder the right to submit a script for marking at any time, irrespective of the individual assignment deadlines, study session, subject or person.

Marking Vouchers can be used for any assignment. They are valid for four years from the date of purchase and can be refunded at any time up to the expiry date.

Although you may submit your script with a Marking Voucher at any time, you will need to adhere to the explicit Marking Voucher deadline dates to ensure that your script is returned before the date of the exam. The deadline dates are provided with the assignments.

Tutorials

Our tutorials are specifically designed to develop the knowledge that you will acquire from the course material into the higher-level understanding that is needed to pass the exam.

We run a range of different tutorials including face-to-face tutorials at various locations, and Live Online tutorials. Full details are set out in our *Tuition Bulletin*, which is available on our website at **www.ActEd.co.uk**.

Regular and Block Tutorials

In preparation for these tutorials, we expect you to have read the relevant part(s) of the Course Notes before attending the tutorial so that the group can spend time on exam questions and discussion to develop understanding rather than basic bookwork.

You can choose ***one*** of the following types of tutorial:

- **Regular Tutorials** spread over the session
- **a Block Tutorial** held two to eight weeks before the exam.

The tutorials outlined above will focus on and develop the skills required for the written Paper A examination. Students wishing for some additional tutor support working through exam-style questions for Paper B may wish to attend a Preparation Day. These will be available Live Online or face-to-face, where students will need to provide their own device capable of running Excel or R as required.

Online Classroom

The Online Classroom acts as either a valuable add-on or a great alternative to a face-to-face or Live Online tutorial, focussing on the written Paper A examination.

At the heart of the Online Classroom in each subject is a comprehensive, easily-searched collection of tutorial units. These are a mix of:

- teaching units, helping you to really get to grips with the course material, and
- guided questions, enabling you to learn the most efficient ways to answer questions and avoid common exam pitfalls.

The best way to discover the Online Classroom is to see it in action. You can watch a sample of the Online Classroom tutorial units on our website at **www.ActEd.co.uk**.

'Revision' products

Flashcards

For most subjects, there is ***a lot of material*** to revise. Finding a way to fit revision into your routine as painlessly as possible has got to be a good strategy. Flashcards are a relatively inexpensive option that can provide a massive boost. They can also provide a variation in activities during a study day, and so help you to maintain concentration and effectiveness.

Flashcards are a set of A6-sized cards that cover the key points of the subject that most students want to commit to memory. Each flashcard has questions on one side and the answers on the reverse. We recommend that you use the cards actively and test yourself as you go.

Flashcards are available in **eBook** format for viewing on a range of electronic devices. eBooks can be ordered separately or as an addition to paper products. Visit **www.ActEd.co.uk** for full details about the eBooks that are available, compatibility with different devices, software requirements and printing restrictions.

The following questions and comments might help you to decide if flashcards are suitable for you:

- Do you have a regular train or bus journey?

  Flashcards are ideal for regular bursts of revision on the move.

- Do you want to fit more study into your routine?

  Flashcards are a good option for 'dead time', eg using flashcards on your phone or sticking them on the wall in your study.

- Do you find yourself cramming for exams (even if that's not your original plan)?

  Flashcards are an extremely efficient way to do your pre-exam memorising.

If you are retaking a subject, then you might consider using flashcards if you didn't use them on a previous attempt.

'Revision & rehearsal' products

Revision Notes

Our Revision Notes have been designed with input from students to help you revise efficiently. They are suitable for first-time sitters who have worked through the ActEd Course Notes or for retakers (who should find them much more useful and challenging than simply reading through the course again).

The Revision Notes are a set of A5 booklets – perfect for revising on the train or tube to work. Each booklet covers one main theme or a set of related topics from the course and includes:

- Core Reading with a set of integrated short questions to develop your bookwork knowledge
- relevant past exam questions with concise solutions from the last ten years
- other useful revision aids.

ActEd Solutions with Exam Technique (ASET)

The ActEd Solutions with Exam Technique (ASET) contains our solutions to eight past exam papers, plus comment and explanation. In particular, it highlights how questions might have been analysed and interpreted so as to produce a good solution with a wide range of relevant points. This will be valuable in approaching questions in subsequent examinations.

‘Rehearsal’ products

Mock Exam

The Mock Exam consists of two papers. There is a 100-mark mock exam for the written Paper A examination and a separate mock exam for the computer-based Paper B exam. These provide a realistic test of your exam readiness.

Mock Marking

We are happy to mark your attempts at the mock exams. The same general principles apply as for the Assignment Marking. In particular:

- Mock Exam Marking applies to a specified subject, session and student. In this subject it covers the marking of both papers.
- Marking Vouchers can be used for each mock exam paper. (Note that you will need two marking vouchers in order to have the two mock papers marked.)

Recall that:

- marking is not included with the products themselves and you need to order it separately
- you should submit your script via email in the format detailed in the mock exam instructions
- you will be able to download the feedback on your marked script via a secure link on the internet.

1.4 Skills

Technical skills

The Core Reading and exam papers for these subjects tend to be very technical. The exams themselves have many calculation and manipulation questions. The emphasis in the exam will therefore be on *understanding* the mathematical techniques and applying them to various, frequently unfamiliar, situations. It is important to have a feel for what the numerical answer should be by having a deep understanding of the material and by doing reasonableness checks.

As a high level of pure mathematics and statistics is generally required for the Core Principles subjects, it is important that your mathematical skills are extremely good. If you are a little rusty you may wish to consider purchasing additional material to help you get up to speed. The course 'Pure Maths and Statistics for Actuarial Studies' is available from ActEd and it covers the mathematical techniques that are required for the Core Principles subjects, some of which are beyond A-Level (or Higher) standard. You do not need to work through the whole course in order – you can just refer to it when you need help on a particular topic. An initial assessment to test your mathematical skills and further details regarding the course can be found on our website at www.ActEd.co.uk.

Study skills

Overall study plan

We suggest that you develop a realistic study plan, building in time for relaxation and allowing some time for contingencies. Be aware of busy times at work, when you may not be able to take as much study leave as you would like. Once you have set your plan, be determined to stick to it. You don't have to be too prescriptive at this stage about what precisely you do on each study day. The main thing is to be clear that you will cover all the important activities in an appropriate manner and leave plenty of time for revision and question practice.

Aim to manage your study so as to allow plenty of time for the concepts you meet in these courses to 'bed down' in your mind. Most successful students will probably aim to complete the courses at least a month before the exam, thereby leaving a sufficient amount of time for revision. By finishing the courses as quickly as possible, you will have a much clearer view of the big picture. It will also allow you to structure your revision so that you can concentrate on the important and difficult areas.

You can also try looking at our discussion forum on the internet, which can be accessed at **www.ActEd.co.uk/forums** (or use the link from our home page at **www.ActEd.co.uk**). There are some good suggestions from students on how to study.

Study sessions

Only do activities that will increase your chance of passing. Try to avoid including activities for the sake of it and don't spend time reviewing material that you already understand. You will only improve your chances of passing the exam by getting on top of the material that you currently find difficult.

Ideally, each study session should have a specific purpose and be based on a specific task, *eg 'Finish reading Chapter 3 and attempt Practice Questions 1.4, 1.7 and 1.12 '*, as opposed to a specific amount of time, *eg 'Three hours studying the material in Chapter 3'*.

Try to study somewhere quiet and free from distractions (*eg* a library or a desk at home dedicated to study). Find out when you operate at your peak, and endeavour to study at those times of the day. This might be between 8*am* and 10*am* or could be in the evening. Take short breaks during your study to remain focused – it's definitely time for a short break if you find that your brain is tired and that your concentration has started to drift from the information in front of you.

Order of study

We suggest that you work through each of the chapters in turn. To get the maximum benefit from each chapter you should proceed in the following order:

1. Read the Syllabus Objectives. These are set out in the box at the start of each chapter.
2. Read the Chapter Summary at the end of each chapter. This will give you a useful overview of the material that you are about to study and help you to appreciate the context of the ideas that you meet.
3. Study the Course Notes in detail, annotating them and possibly making your own notes. Try the self-assessment questions as you come to them. As you study, pay particular attention to the listing of the Syllabus Objectives and to the Core Reading.
4. Read the Chapter Summary again carefully. If there are any ideas that you can't remember covering in the Course Notes, read the relevant section of the notes again to refresh your memory.
5. Attempt (at least some of) the Practice Questions that appear at the end of the chapter.
6. Where relevant, work through the relevant Paper B Online Resources for the chapter(s). You will need to have a good understanding of the relevant section of the paper-based course before you attempt the corresponding section of PBOR.

It's a fact that people are more likely to remember something if they review it several times. So, do look over the chapters you have studied so far from time to time. It is useful to re-read the Chapter Summaries or to try the Practice Questions again a few days after reading the chapter itself. It's a good idea to annotate the questions with details of when you attempted each one. This makes it easier to ensure that you try all of the questions as part of your revision without repeating any that you got right first time.

Once you've read the relevant part of the notes and tried a selection of questions from the Practice Questions (and attended a tutorial, if appropriate), you should attempt the corresponding assignment. If you submit your assignment for marking, spend some time looking through it carefully when it is returned. It can seem a bit depressing to analyse the errors you made, but you will increase your chances of passing the exam by learning from your mistakes. The markers will try their best to provide practical comments to help you to improve.

To be really prepared for the exam, you should not only know and understand the Core Reading but also be aware of what the examiners will expect. Your revision programme should include plenty of question practice so that you are aware of the typical style, content and marking structure of exam questions. You should attempt as many past exam questions as you can.

Active study

Here are some techniques that may help you to study actively.

1. Don't believe everything you read. Good students tend to question everything that they read. They will ask 'why, how, what for, when?' when confronted with a new concept, and they will apply their own judgement. This contrasts with those who unquestioningly believe what they are told, learn it thoroughly, and reproduce it (unquestioningly?) in response to exam questions.

2. Another useful technique as you read the Course Notes is to think of possible questions that the examiners could ask. This will help you to understand the examiners' point of view and should mean that there are fewer nasty surprises in the exam room. Use the Syllabus to help you make up questions.

3. Annotate your notes with your own ideas and questions. This will make you study more actively and will help when you come to review and revise the material. Do not simply copy out the notes without thinking about the issues.

4. Attempt the questions in the notes as you work through the course. Write down your answer before you refer to the solution.

5. Attempt other questions and assignments on a similar basis, *ie* write down your answer before looking at the solution provided. Attempting the assignments under exam conditions has some particular benefits:

   - It forces you to think and act in a way that is similar to how you will behave in the exam.
   - When you have your assignments marked it is *much* more useful if the marker's comments can show you how to improve your performance under exam conditions than your performance when you have access to the notes and are under no time pressure.
   - The knowledge that you are going to do an assignment under exam conditions and then submit it (however good or bad) for marking can act as a powerful incentive to make you study each part as well as possible.
   - It is also quicker than trying to write perfect answers.

6. Sit a mock exam four to six weeks before the real exam to identify your weaknesses and work to improve them. You could use a mock exam written by ActEd or a past exam paper.

You can find further information on how to study in the profession's Student Handbook, which you can download from their website at:

www.actuaries.org.uk/studying

Revision and exam skills

Revision skills

You will have sat many exams before and will have mastered the exam and revision techniques that suit you. However, it is important to note that due to the high volume of work involved in the Core Principles subjects, it is not possible to leave all your revision to the last minute. Students who prepare well in advance have a better chance of passing their exams on the first sitting.

Unprepared students find that they are under time pressure in the exam. Therefore it is important to find ways of maximising your score in the shortest possible time. Part of your preparation should be to practise a large number of exam-style questions under timed exam conditions as soon as possible. This will:

- help you to develop the necessary understanding of the techniques required
- highlight the key topics, which crop up regularly in many different contexts and questions
- help you to practise the specific skills that you will need to pass the exam.

There are many sources of exam-style questions. You can use past exam papers, the Practice Questions at the end of each chapter (which include many past exam questions), assignments, mock exams, the Revision Notes and ASET.

Exam question skill levels

Exam questions are not designed to be of similar difficulty. The Institute and Faculty of Actuaries specifies different skill levels that questions may be set with reference to.

Questions may be set at any skill level:

- Knowledge – demonstration of a detailed knowledge and understanding of the topic
- Application – demonstration of an ability to apply the principles underlying the topic within a given context
- Higher Order – demonstration of an ability to perform deeper analysis and assessment of situations, including forming judgements, taking into account different points of view, comparing and contrasting situations, suggesting possible solutions and actions, and making recommendations.

Command verbs

The Institute and Faculty of Actuaries use command verbs (such as ‘Define’, ‘Discuss’ and ‘Explain’) to help students to identify what the question requires. The profession has produced a document, ‘Command verbs used in the Associate and Fellowship written examinations’, to help students to understand what each command verb is asking them to do.

It also gives the following advice:

- The use of a specific command verb within a syllabus objective does not indicate that this is the only form of question which can be asked on the topic covered by that objective.
- The Examiners may ask a question on any syllabus topic using any of the agreed command verbs, as are defined in the document.

You can find the relevant document on the profession's website at:

https://www.actuaries.org.uk/studying/prepare-your-exams

1.5 The examination

What to take to the exam

IMPORTANT NOTE: The following information was correct at the time of printing, however it is important to keep up-to-date with any changes. See the profession's website for the latest guidance.

For the written exams the examination room will be equipped with:

- the question paper
- an answer booklet
- rough paper
- a copy of the Yellow Tables.

Remember to take with you:

- black pens
- a permitted scientific calculator – please refer to **www.actuaries.org.uk** for the latest advice.

Please also refer to the profession's website and your examination instructions for details about what you will need for the computer-based Paper B exam.

Past exam papers

You can download some past exam papers and Examiners' Reports from the profession's website at **www.actuaries.org.uk**. However, please be aware that these exam papers are for the pre-2019 syllabus and not all questions will be relevant.

1.6 Queries and feedback

Questions and queries

From time to time you may come across something in the study material that is unclear to you. The easiest way to solve such problems is often through discussion with friends, colleagues and peers – they will probably have had similar experiences whilst studying. If there's no-one at work to talk to then use our discussion forum at **www.ActEd.co.uk/forums** (or use the link from our home page at **www.ActEd.co.uk**).

Our online forum is dedicated to actuarial students so that you can get help from fellow students on any aspect of your studies from technical issues to study advice. You could also use it to get ideas for revision or for further reading around the subject that you are studying. ActEd tutors will visit the site from time to time to ensure that you are not being led astray and we also post other frequently asked questions from students on the forum as they arise.

If you are still stuck, then you can send queries by email to the relevant subject email address (see Section 2.6), but we recommend that you try the forum first. We will endeavour to contact you as soon as possible after receiving your query but you should be aware that it may take some time to reply to queries, particularly when tutors are away from the office running tutorials. At the busiest teaching times of year, it may take us more than a week to get back to you.

If you have many queries on the course material, you should raise them at a tutorial or book a personal tuition session with an ActEd tutor. Information about personal tuition is set out in our current brochure. Please email **ActEd@bpp.com** for more details.

Feedback

If you find an error in the course, please check the corrections page of our website (**www.ActEd.co.uk/paper_corrections.html**) to see if the correction has already been dealt with. Otherwise please send details via email to the relevant subject email address (see Section 2.6).

Each year our tutors work hard to improve the quality of the study material and to ensure that the courses are as clear as possible and free from errors. We are always happy to receive feedback from students, particularly details concerning any errors, contradictions or unclear statements in the courses. If you have any comments on this course please email them to the relevant subject email address (see Section 2.6).

Our tutors also work with the profession to suggest developments and improvements to the Syllabus and Core Reading. If you have any comments or concerns about the Syllabus or Core Reading, these can be passed on via ActEd. Alternatively, you can send them directly to the Institute and Faculty of Actuaries' Examination Team by email to **education.services@actuaries.org.uk**.

2.1 Subject CS2 – background

History

The Actuarial Statistics subjects (Subjects CS1 and CS2) are new subjects in the Institute and Faculty of Actuaries 2019 Curriculum.

Subject CS2 is *Risk Modelling and Survival Analysis*.

Predecessors

The topics covered in the Actuarial Statistics subjects (Subjects CS1 and CS2) cover content previously in Subjects CT3, CT4, CT6 and a small amount from Subject ST9:

- Subject CS1 contains material from Subjects CT3 and CT6.
- Subject CS2 contains material from Subjects CT4, CT6 and ST9.

Exemptions

You will need to have passed or been granted an exemption from Subjects CT4 and CT6 to be eligible for a pass in Subject CS2 during the transfer process.

Links to other subjects

- This subject assumes that the student is competent with the material covered in CS1 – Actuarial Statistics – and the required knowledge for that subject.
- CM1 – Actuarial Mathematics and CM2 – Financial Engineering and Loss Reserving apply the material in this subject to actuarial and financial modelling.
- Topics in this subject are further built upon in SP1 – Health and Care Principles, SP7 – General Insurance Reserving and Capital Modelling Principles, SP8 – General Insurance Pricing Principles and SP9 – Enterprise Risk Management Principles.

2.2 Subject CS2 – Syllabus and Core Reading

Syllabus

The Syllabus for Subject CS2 is given here. To the right of each objective are the chapter numbers in which the objective is covered in the ActEd course.

Aim

The aim of Subject CS2 is to provide a grounding in mathematical and statistical modelling techniques that are of particular relevance to actuarial work, including stochastic processes and survival models and their application.

Competences

On successful completion of this subject, a student will be able to:

1. describe and use statistical distributions for risk modelling
2. describe and apply the main concepts underlying the analysis of time series models
3. describe and apply Markov chains and processes
4. describe and apply techniques of survival analysis
5. describe and apply basic principles of machine learning.

Syllabus topics

1. Random variables and distributions for risk modelling (20%)
2. Time series (20%)
3. Stochastic processes (25%)
4. Survival models (25%)
5. Machine learning (10%)

The weightings are indicative of the approximate balance of the assessment of this subject between the main syllabus topics, averaged over a number of examination sessions.

The weightings also have a correspondence with the amount of learning material underlying each syllabus topic. However, this will also reflect aspects such as:

- the relative complexity of each topic, and hence the amount of explanation and support required for it
- the need to provide thorough foundation understanding on which to build the other objectives
- the extent of prior knowledge which is expected
- the degree to which each topic area is more knowledge or application based.

Detailed syllabus objectives

1 Random variables and distributions for risk modelling (20%)

1.1 Loss distributions, with and without risk sharing (Chapters 15 and 18)

1.1.1 Describe the properties of the statistical distributions which are suitable for modelling individual and aggregate losses.

1.1.2 Explain the concepts of excesses (deductibles), and retention limits.

1.1.3 Describe the operation of simple forms of proportional and excess of loss reinsurance.

1.1.4 Derive the distribution and corresponding moments of the claim amounts paid by the insurer and the reinsurer in the presence of excesses (deductibles) and reinsurance.

1.1.5 Estimate the parameters of a failure time or loss distribution when the data is complete, or when it is incomplete, using maximum likelihood and the method of moments.

1.1.6 Fit a statistical distribution to a dataset and calculate appropriate goodness of fit measures.

1.2 Compound distributions and their applications in risk modelling (Chapters 19 and 20)

1.2.1 Construct models appropriate for short term insurance contracts in terms of the numbers of claims and the amounts of individual claims.

1.2.2 Describe the major simplifying assumptions underlying the models in 1.2.1.

1.2.3 Define a compound Poisson distribution and show that the sum of independent random variables each having a compound Poisson distribution also has a compound Poisson distribution.

1.2.4 Derive the mean, variance and coefficient of skewness for compound binomial, compound Poisson and compound negative binomial random variables.

1.2.5 Repeat 1.2.4 for both the insurer and the reinsurer after the operation of simple forms of proportional and excess of loss reinsurance.

1.3 Introduction to copulas (Chapter 17)

1.3.1 Describe how a copula can be characterised as a multivariate distribution function which is a function of the marginal distribution functions of its variates, and explain how this allows the marginal distributions to be investigated separately from the dependency between them.

1.3.2 Explain the meaning of the terms dependence or concordance, upper and lower tail dependence; and state in general terms how tail dependence can be used to help select a copula suitable for modelling particular types of risk.

1.3.3 Describe the form and characteristics of the Gaussian copula and the Archimedean family of copulas.

1.4 Introduction to extreme value theory (Chapter 16)

1.4.1 Recognise extreme value distributions, suitable for modelling the distribution of severity of loss and their relationships

1.4.2 Calculate various measures of tail weight and interpret the results to compare the tail weights.

2 Time series (20%)

2.1 Concepts underlying time series models (Chapters 13 and 14)

2.1.1 Explain the concept and general properties of stationary, $I(0)$, and integrated, $I(1)$, univariate time series.

2.1.2 Explain the concept of a stationary random series.

2.1.3 Explain the concept of a filter applied to a stationary random series.

2.1.4 Know the notation for backwards shift operator, backwards difference operator, and the concept of roots of the characteristic equation of time series.

2.1.5 Explain the concepts and basic properties of autoregressive (AR), moving average (MA), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) time series.

2.1.6 Explain the concept and properties of discrete random walks and random walks with normally distributed increments, both with and without drift.

2.1.7 Explain the basic concept of a multivariate autoregressive model.

2.1.8 Explain the concept of cointegrated time series.

2.1.9 Show that certain univariate time series models have the Markov property and describe how to rearrange a univariate time series model as a multivariate Markov model.

2.2 Applications of time series models (Chapters 13 and 14)

2.2.1 Outline the processes of identification, estimation and diagnosis of a time series, the criteria for choosing between models and the diagnostic tests that might be applied to the residuals of a time series after estimation.

2.2.2 Describe briefly other non-stationary, non-linear time series models.

2.2.3 Describe simple applications of a time series model, including random walk, autoregressive and cointegrated models as applied to security prices and other economic variables.

2.2.4 Develop deterministic forecasts from time series data, using simple extrapolation and moving average models, applying smoothing techniques and seasonal adjustment when appropriate.

3 Stochastic processes (25%)

3.1 Describe and classify stochastic processes. (Chapter 1)

3.1.1 Define in general terms a stochastic process and in particular a counting process.

3.1.2 Classify a stochastic process according to whether it:

- operates in continuous or discrete time
- has a continuous or a discrete state space
- is a mixed type

and give examples of each type of process.

3.1.3 Describe possible applications of mixed processes.

3.1.4 Explain what is meant by the Markov property in the context of a stochastic process and in terms of filtrations.

3.2 Define and apply a Markov chain. (Chapter 2)

3.2.1 State the essential features of a Markov chain model.

3.2.2 State the Chapman-Kolmogorov equations that represent a Markov chain.

3.2.3 Calculate the stationary distribution for a Markov chain in simple cases.

3.2.4 Describe a system of frequency based experience rating in terms of a Markov chain and describe other simple applications.

3.2.5 Describe a time-inhomogeneous Markov chain model and describe simple applications.

3.2.6 Demonstrate how Markov chains can be used as a tool for modelling and how they can be simulated.

3.3 Define and apply a Markov process. (Chapters 4 and 5)

3.3.1 State the essential features of a Markov process model.

3.3.2 Define a Poisson process, derive the distribution of the number of events in a given time interval, derive the distribution of inter-event times, and apply these results.

3.3.3 Derive the Kolmogorov equations for a Markov process with time independent and time/age dependent transition intensities.

3.3.4 Solve the Kolmogorov equations in simple cases.

3.3.5 Describe simple survival models, sickness models and marriage models in terms of Markov processes and describe other simple applications.

3.3.6 State the Kolmogorov equations for a model where the transition intensities depend not only on age/time, but also on the duration of stay in one or more states.

3.3.7 Describe sickness and marriage models in terms of duration dependent Markov processes and describe other simple applications.

3.3.8 Demonstrate how Markov jump processes can be used as a tool for modelling and how they can be simulated.

4 Survival models (25%)

4.1 Explain the concept of survival models.

4.1.1 Describe the model of lifetime or failure time from age x as a random variable. (Chapter 6)

4.1.2 State the consistency condition between the random variable representing lifetimes from different ages. (Chapter 6)

4.1.3 Define the distribution and density functions of the random future lifetime, the survival function, the force of mortality or hazard rate, and derive relationships between them. (Chapter 6)

4.1.4 Define the actuarial symbols ${}_tp_x$ and ${}_tq_x$ and derive integral formulae for them. (Chapter 6)

4.1.5 State the Gompertz and Makeham laws of mortality. (Chapter 6)

4.1.6 Define the curtate future lifetime from age x and state its probability function. (Chapter 6)

4.1.7 Define the symbols e_x and $\overset{\circ}{e}_x$ and derive an approximate relation between them. Define the expected value and variance of the complete and curtate future lifetimes and derive expressions for them. (Chapter 6)

4.1.8 Describe the two-state model of a single decrement and compare its assumptions with those of the random lifetime model. (Chapter 3)

4.2 Describe estimation procedures for lifetime distributions.

4.2.1 Describe the various ways in which lifetime data might be censored. (Chapter 7)

4.2.2 Describe the estimation of the empirical survival function in the absence of censoring, and what problems are introduced by censoring. (Chapter 7)

4.2.3 Describe the Kaplan-Meier (or product limit) estimator of the survival function in the presence of censoring, compute it from typical data and estimate its variance. (Chapter 7)

4.2.4 Describe the Nelson-Aalen estimator of the cumulative hazard rate in the presence of censoring, compute it from typical data and estimate its variance. (Chapter 7)

4.2.5 Describe models for proportional hazards, and how these models can be used to estimate the impact of covariates on the hazard. (Chapter 8)

4.2.6 Describe the Cox model for proportional hazards, derive the partial likelihood estimate in the absence of ties, and state the asymptotic distribution of the partial likelihood estimator. (Chapter 8)

4.3 Derive maximum likelihood estimators for transition intensities. (Chapters 3 and 4)

4.3.1 Describe an observational plan in respect of a finite number of individuals observed during a finite period of time, and define the resulting statistics, including the waiting times.

4.3.2 Derive the likelihood function for constant transition intensities in a Markov model of transfers between states given the statistics in 4.3.1.

4.3.3 Derive maximum likelihood estimators for the transition intensities in 4.3.2 and state their asymptotic joint distribution.

4.3.4 State the Poisson approximation to the estimator in 4.3.3 in the case of a single decrement.

4.4 Estimate transition intensities dependent on age (exact or census). (Chapter 9)

4.4.1 Explain the importance of dividing the data into homogeneous classes, including subdivision by age and sex.

4.4.2 Describe the principle of correspondence and explain its fundamental importance in the estimation procedure.

4.4.3 Specify the data needed for the exact calculation of a central exposed to risk (waiting time) depending on age and sex.

4.4.4 Calculate a central exposed to risk given the data in 4.4.3.

4.4.5 Explain how to obtain estimates of transition probabilities, including in the single decrement model the actuarial estimate based on the simple adjustment to the central exposed to risk.

4.4.6 Explain the assumptions underlying the census approximation of waiting times.

4.4.7 Explain the concept of the rate interval.

4.4.8 Develop census formulae given age at birthday where the age may be classified as next, last, or nearest relative to the birthday as appropriate, and the deaths and census data may use different definitions of age.

4.4.9 Specify the age to which estimates of transition intensities or probabilities in 4.4.8 apply.

4.5 Graduation and graduation tests (Chapters 10 and 11)

4.5.1 Describe and apply statistical tests of the comparison of crude estimates with a standard mortality table testing for:

- the overall fit
- the presence of consistent bias
- the presence of individual ages where the fit is poor
- the consistency of the 'shape' of the crude estimates and the standard table.

For each test describe:

- the formulation of the hypothesis
- the test statistic
- the distribution of the test statistic using approximations where appropriate
- the application of the test statistic.

4.5.2 Describe the reasons for graduating crude estimates of transition intensities or probabilities, and state the desirable properties of a set of graduated estimates.

4.5.3 Describe a test for smoothness of a set of graduated estimates.

4.5.4 Describe the process of graduation by the following methods, and state the advantages and disadvantages of each:

- parametric formula
- standard table
- spline functions

(The student will not be required to carry out a graduation.)

4.5.5 Describe how the tests in 4.5.1 should be amended to compare crude and graduated sets of estimates.

4.5.6 Describe how the tests in 4.5.1 should be amended to allow for the presence of duplicate policies.

4.5.7 Carry out a comparison of a set of crude estimates and a standard table, or of a set of crude estimates and a set of graduated estimates.

4.6 Mortality projection (Chapter 12)

4.6.1 Describe the approaches to the forecasting of future mortality rates based on extrapolation, explanation and expectation, and their advantages and disadvantages.

4.6.2 Describe the Lee-Carter, age-period-cohort, and p-spline regression models for forecasting mortality.

4.6.3 Use an appropriate computer package to apply the models in 4.6.2 to a suitable mortality dataset.

4.6.4 List the main sources of error in mortality forecasts.

5 Machine learning (10%)

5.1 Explain and apply elementary principles of machine learning. (Chapter 21)

5.1.1 Explain the main branches of machine learning and describe examples of the types of problems typically addressed by machine learning.

5.1.2 Explain and apply high-level concepts relevant to learning from data.

5.1.3 Describe and give examples of key supervised and unsupervised machine learning techniques, explaining the difference between regression and classification and between generative and discriminative models.

5.1.4 Explain in detail and use appropriate software to apply machine learning techniques (*eg* penalised regression and decision trees) to simple problems.

5.1.5 Demonstrate an understanding of the perspectives of statisticians, data scientists, and other quantitative researchers from non-actuarial backgrounds.

Core Reading

The Subject CS2 Course Notes include the Core Reading in full, integrated throughout the course.

Accreditation

The Institute and Faculty of Actuaries would like to thank the numerous people who have helped in the development of the material contained in this Core Reading.

Further reading

The exam will be based on the relevant Syllabus and Core Reading and the ActEd course material will be the main source of tuition for students.

2.3 Subject CS2 – the course structure

There are five parts to the Subject CS2 course. The parts cover related topics and have broadly equal marks in the paper-based exam. The parts are broken down into chapters.

The following table shows how the parts, the chapters and the syllabus items relate to each other. The end columns show how the chapters relate to the days of the regular tutorials. We have also given you a broad indication of the length of each chapter. This table should help you plan your progress across the study session.

Part	*Chapter*	*Title*	*No of pages*	*Syllabus objectives*	*5 full days*
1	1	Stochastic processes	35	3.1	1
	2	Markov chains	70	3.2	
	3	The two-state Markov model and the Poisson model	41	4.1.8, 4.3	
	4	Time-homogeneous Markov jump processes	71	3.3.1-3.3.5, 4.3.1-4.3.3	
2	5	Time-inhomogeneous Markov jump processes	56	3.3.1, 3.3.3-3.3.8	2
	6	Survival models	36	4.1.1-4.1.7	
	7	Estimating the lifetime distribution	59	4.2.1-4.2.4	
3	8	Proportional hazards models	42	4.2.5-4.2.6	3
	9	Exposed to risk	32	4.4	
	10	Graduation and statistical tests	62	4.5.1-4.5.3, 4.5.5, 4.5.7	
	11	Methods of graduation	31	4.5.4-4.5.7	
	12	Mortality projection	54	4.6	
4	13	Time series 1	71	2.1.1-2.1.2, 2.1.4-2.1.6, 2.1.9, 2.2.3	4
	14	Time series 2	61	2.1.3, 2.1.7-2.1.8, 2.2	
	15	Loss distributions	45	1.1.1, 1.1.5-1.1.6	
	16	Extreme value theory	45	1.4	

5	17	Copulas	57	1.3	5
	18	Reinsurance	45	1.1.2-1.1.5	
	19	Risk models 1	38	1.2.1-1.2.4	
	20	Risk models 2	43	1.2.1-1.2.2, 1.2.5	
	21	Machine learning	78	5.1	

2.4 Subject CS2 – summary of ActEd products

The following products are available for Subject CS2:

- Course Notes
- PBOR (including the Y Assignments)
- X Assignments – five assignments:
  - X1, X2, X3: 80-mark tests (you are allowed 2¾ hours to complete these)
  - X4, X5: 100-mark tests (you are allowed 3¼ hours to complete these)
- Series X Marking
- Series Y Marking
- Online Classroom – over 150 tutorial units
- Flashcards
- Revision Notes
- ASET – four years' exam papers, *ie* eight papers, covering the period April 2014 to September 2017
- Mock Exam
- Mock Exam Marking
- Marking Vouchers.

We will endeavour to release as much material as possible but unfortunately some revision products may not be available until the September 2019 or even April 2020 exam sessions. Please check the ActEd website or email ActEd@bpp.com for more information.

The following tutorials are typically available for Subject CS2:

- Regular Tutorials (five days)
- Block Tutorials (five days)
- a Preparation Day for the computer-based exam.

Full details are set out in our *Tuition Bulletin*, which is available on our website at **www.ActEd.co.uk**.

2.5 Subject CS2 – skills and assessment

Technical skills

The *Actuarial Statistics* subjects (Subjects CS1 and CS2) are very mathematical and have relatively few questions requiring wordy answers.

Exam skills

Exam question skill levels

In the CS subjects, the approximate split of assessment across the three skill types is:

- Knowledge – 20%
- Application – 65%
- Higher Order skills – 15%.

Assessment

Assessment consists of a combination of a 3¼-hour written examination and a 1¾-hour computer-based practical examination.

2.6 Subject CS2 – frequently asked questions

Q: ***What knowledge of earlier subjects should I have?***

A: Knowledge of Subject CS1, Actuarial Statistics, is assumed.

Q: ***What level of mathematics is required?***

A: Good mathematical skills are essential for Subject CS2. Calculus and algebra (including matrices) are used extensively in this course.

If your maths is a little rusty you may wish to consider purchasing additional material to help you get up to speed. The course 'Pure Maths and Statistics for Actuarial Studies' is available from ActEd and it covers the mathematical techniques that are required for the Core Principles subjects, some of which are beyond A-Level (or Higher) standard. You do not need to work through the whole course in order – you can just refer to it when you need help on a particular topic. An initial assessment to test your mathematical skills and further details regarding the course can be found on our website.

Q: ***What should I do if I discover an error in the course?***

A: If you find an error in the course, please check our website at:

www.ActEd.co.uk/paper_corrections.html

to see if the correction has already been dealt with. Otherwise please send details via email to **CS2@bpp.com**.

Q: ***Who should I send feedback to?***

A: We are always happy to receive feedback from students, particularly details concerning any errors, contradictions or unclear statements in the courses.

If you have any comments on this course in general, please email **CS2@bpp.com**.

If you have any comments or concerns about the Syllabus or Core Reading, these can be passed on to the profession via ActEd. Alternatively, you can send them directly to the Institute and Faculty of Actuaries' Examination Team by email to **education.services@actuaries.org.uk**.

1

Stochastic processes

Syllabus objectives

3.1 Describe and classify stochastic processes.

3.1.1 Define in general terms a stochastic process and in particular a counting process.

3.1.2 Classify a stochastic process according to whether it:

(a) operates in continuous or discrete time

(b) has a continuous or a discrete state space

(c) is a mixed type

and give examples of each type of process.

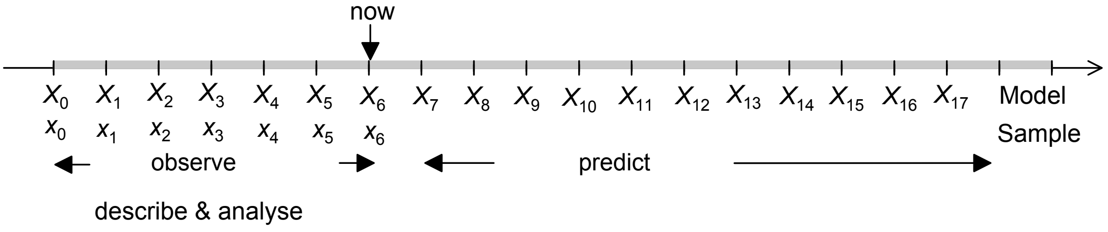
3.1.3 Describe possible applications of mixed processes.

3.1.4 Explain what is meant by the Markov property in the context of a stochastic process and in terms of filtrations.

0 Introduction

In this chapter we define the concept of a *stochastic process*. A stochastic process is a family or set of ordered random variables. The order is indicated by indexing each random variable in the family by a subscript. Usually the ordering is a result of the random variables being observed over time, so X_t is a random variable that models the value of the stochastic process at time t. The random variables in the set can be dependent on one another reflecting the nature of the process being modelled.

As in all statistical modelling we will collect a sample of data from the process being modelled. So for example, X_t might model the price of a stock at time t, and we have observations of the stock price for the last 6 trading days. We can use these data to describe the process and to analyse the nature of its past behaviour over time. The data may also be used to estimate the parameters of our stochastic process model. We could then use the estimated stochastic process model to predict the future behaviour of the stock price. It is the dependence between the random variables in the set that allows us to make predictions by extrapolating past patterns into the future.


now
time
X_0 X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10} X_{11} X_{12} X_{13} X_{14} X_{15} X_{16} X_{17}
Model
x_0 x_1 x_2 x_3 x_4 x_5 x_6
Sample
observe
predict
describe & analyse

We start with some definitions and then give examples of different types of processes. The final parts of the second section discuss some of the properties that a stochastic process may possess, *ie stationarity*, *independent increments* and the *Markov property*.

In Section 3 we look at several examples of stochastic processes, and use these to illustrate the definitions and properties we have given in the previous section. Some of these examples will be discussed in more detail in later chapters.

1 Types of stochastic processes

We begin with the definition of a stochastic process.

Stochastic process

A *stochastic process* is a model for a time-dependent random phenomenon. So, just as a single random variable describes a static random phenomenon, a stochastic process is a collection of random variables X_t, one for each time t in some set J.

The process is denoted $\{X_t : t \in J\}$.

The set of values that the random variables X_t are capable of taking is called the *state space* of the process, S.

The values in the set J are called the *time set* or *time domain* of the process. This terminology should be familiar – if $y = f(x)$, the values taken by x are the domain (or support) of the function and the values taken by y are the range of the function.

The set of random variables may be dependent. So in order to describe their statistical properties we will need to know their joint distribution. If the random variables in the set were independent (as is the case for example with statistical models of the sampling process) it would be sufficient to know the (marginal) distribution of each random variable in the set.

The random variables in the set need not be identically distributed. However, we will discover that processes that have identically distributed random variables are particularly important in the study of stochastic processes.

The state space of the stochastic process will include all of the values that can be taken by any of the random variables in the set. However, for particular random variables in the set, some of these values may have a zero probability of occurring.

For example, we might model the closing value of the FTSE100 index by a stochastic process $\{X_t\}$. The random variable X_t models the value at the end of day *t*.

Question

Explain whether the state space and time set for this process are discrete or continuous.

Solution

The state space is technically discrete as share prices are measured in pence. However, as there are very many distinct values that are close together, it is often easier to use a continuous random variable to model the share price, which results in a continuous state space.

If time is measured in days, then the values of the process are recorded at times 1, 2, 3, …. So the time set is discrete.

A possible model might say that the value of X_t depends on the values at the end of the two previous trading days X_{t-1} and X_{t-2}. If this dependence does not change with t, then we could use the model to predict future vales of X_t. These predictions may not be exact even if the model is good.

The first choice that one faces when selecting a stochastic process to model a real life situation is that of the nature (discrete or continuous) of the time set J and of the state space S.

While the process being modelled could have its value recorded continuously or at very frequent discrete times, we may choose to model only the values at discrete or less frequent discrete time points. This may be because we are only able to record measurements at these times. This could be because of physical limitations on the measurement process or because the measurement process is very expensive and we cannot afford more frequent measurements. Aside from these considerations we may be content just to model the process at these time points because, for example, predictions using this frequency will be perfectly adequate for our needs. No purpose is served by using a more elaborate model than is necessary.

There is no requirement that the labels used in the set J should be actual calendar times, merely that they should put the random variables in order.

In statistical modelling it is common to approximate (the state space of) discrete random variables by continuous random variables when the number of discrete values becomes large enough. So, for example, we often approximate a discrete binomial random variable by a continuous normal random variable when the binomial random variable has more than 20 or 30 discrete values. We know that the continuous model is not an exact representation of what is being modelled but it is adequate for our purposes and is easier to use.

1.1 Discrete state space with discrete time changes

Here is an example of a process with a discrete state space and discrete time changes.

A motor insurance company reviews the status of its customers yearly. Three levels of discount are possible $(0, 25\%, 40\%)$ depending on the accident record of the driver. In this case the appropriate state space is $S = \{0, 25, 40\}$ and the time set is $J = \{0, 1, 2, \ldots\}$ where each interval represents a year. This problem is studied in Chapter 2.

The time set often starts at 0 (whether continuous or discrete). Time 0 is taken to be the start, so that after one unit of time (day, minute *etc*) we have $t = 1$.

In principle the company could record the discount status of each policyholder on a continuous basis, but discount levels are usually only changed on the annual renewal date of the policy. So it makes sense just to model and record these values. A model with more frequent recording will be more complicated (and expensive) yet is unlikely to be more useful in managing a portfolio of motor policies.

The time set is discrete. The state space contains three discrete values.

1.2 Discrete state space with continuous time changes

A life insurance company classifies its policyholders as Healthy, Sick or Dead. Hence the state space $S=\{H,S,D\}$. As for the time set, it is natural to take $J=[0,\infty)$ as illness or death can occur at any time. On the other hand, it may be sufficient to count time in units of days, thus using $J=\{0,1,2,...\}$. This problem is studied in some detail in Chapters 4 and 5.

The description suggests that the time set could be continuous or discrete with very frequent recording. In using the model it is important that we are able to answer questions like 'What is the probability that a life that is healthy at time s is sick at time t?' This suggests that a model with a continuous time domain will be more useful.

Here the state space contains three discrete values.

1.3 Continuous state space

Claims of unpredictable amounts reach an insurance company at unpredictable times; the company needs to forecast the *cumulative claims* over $[0,t]$ in order to assess the risk that it might not be able to meet its liabilities. It is standard practice to use $[0,\infty)$ both for S and J in this problem. However, other choices are possible: claims come in units of a penny and do not really form a continuum. Similarly the intra-day arrival time of a claim is of little significance, so that $\{0,1,2,...\}$ is a possible choice for J and/or S.

The intra-day arrival time is the time at which a claim arrives on a particular day.

The choice of a discrete or continuous time set is influenced by the availability of data, *eg* are cumulative claims figures recorded on a daily basis, or are figures only available at the end of each quarter? The choice is also influenced by the purpose of the modelling, *eg* are predictions of the cumulative claims at the end of each quarter sufficient or are predictions needed more frequently?

Claim amounts may be recorded in pence or to the nearest pound and so in principle the state space is discrete, but it contains a very large number of non-negative values and so a continuous approximation is perfectly adequate.

An important class of models having a continuous state space and a discrete time set is *time series*. Many economic and financial stochastic processes fall into this class, *eg* daily prices at the close of trading for a company's shares. Time series are studied in Chapters 13 and 14.

1.4 Displaying observed data

When we take observations on a process, we obtain a sample of each of the random variables in the set making up the stochastic process.

In displaying the data we want to convey information about three features. These are:

- the size of the values, *ie* to give an idea of the means of the random variables in the set
- the volatility of the values, *ie* to give an idea of the variances of the random variables in the set
- the relationships between the values of the random variables, *ie* to give an idea of the covariances between the random variables in the set.

We do this by plotting X_t against t. Even when the time set is discrete we can join up the plotted points with straight lines. There are no observations on these lines, but they can help to show the 'shape' of the time series. You may have seen plots like this in newspapers, *eg* showing the price of a stock each day or the interest rate at the end of each quarter. Plots like this can be used even when the state space of the process is discrete.

Question

You are thinking of moving to live in Edinburgh. As part of your research into what it's like to live in Edinburgh you have collected the following sets of data:

- the maximum daily temperature each day since 1 January 2015
- whether or not it rained for each day since 1 January 2015
- the number of cyclists injured in road accidents since 1 January 2015.

(i) For each data set choose an appropriate state space and a time set. In each case state whether the state space and the time set are discrete or continuous, and give the units of measurement, *eg* dollars, weeks.

(ii) Imagine that you have data for each process. Draw sketches to display these samples of data.

Solution

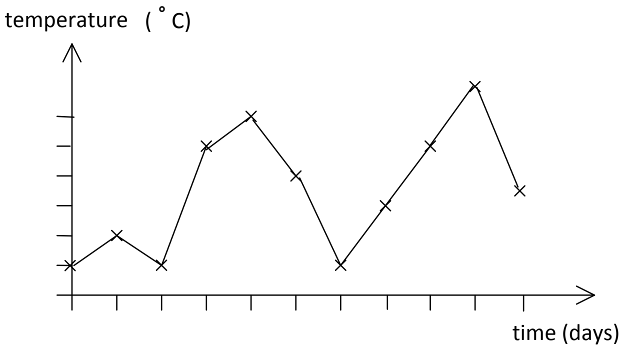
Maximum daily temperature

(i) ***State space and time set***

We would use a discrete time set, say the non-negative integers $t = 0,1,2\ldots$, where t represents the number of days since 1 January 2015.

In practice we might only quote temperature to the nearest degree, so if the number of possible values was small we could use a discrete state space. Alternatively, we might prefer to use a continuous state space if for example we were recording temperatures to the nearest 0.1 of a degree and we thought values might range from $-30^\circ C$ to $+40^\circ C$.

(ii) ***Sketch***



Daily rainfall

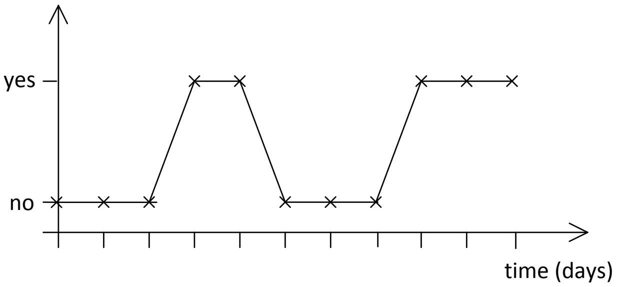
(i) ***State space and time set***

We would use a discrete time set, say the non-negative integers $t = 0,1,2\ldots$, where t represents the number of days since 1 January 2015.

The state space would be discrete consisting of two values: (yes: it rained) and (no: it didn't rain).

(ii) ***Sketch***

We can display the data in a similar way to the temperature data, but there will be more 'flat' sections in the plot.



Injured cyclists

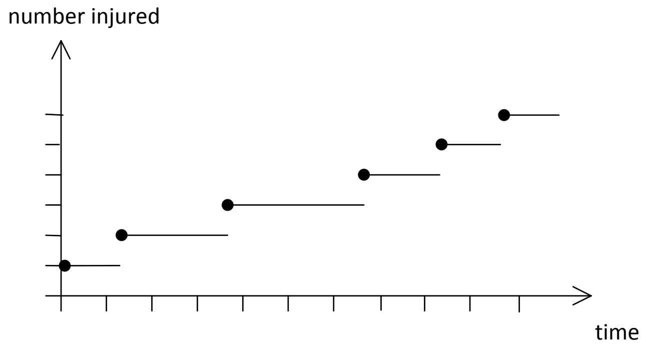
(i) ***State space and time set***

In principle there will be a value for this random variable at every point in time. So we would use a continuous time set, say the non-negative numbers $t \geq 0$, where t represents time since 1 January 2015.

The state space would be discrete – it would be the set of non-negative integers. Only if we thought that the number of possible values was very large would it be appropriate to approximate this discrete state space by a continuous one.

(ii) ***Sketch***

The display would be similar to the previous ones, except that here there is no need to use lines to link observations to show the shape. Because the time set is continuous and the process jumps when each event occurs, we use 'blobs' to show which value should be used at the points of discontinuity.



1.5 Processes of mixed type

Just because a stochastic process operates in continuous time does not mean that it cannot also change value at predetermined discrete instants; such processes are said to be of mixed type. As an example, consider a pension scheme in which members have the option to retire on any birthday between ages 60 and 65. The number of people electing to take retirement at each year of age between 60 and 65 cannot be predicted exactly, nor can the time and number of deaths among active members. Hence the number of contributors to the pension scheme can be modelled as a stochastic process of mixed type with state space $S = \{1,2,3,\ldots\}$ and time interval $J = [0,\infty)$. Decrements of random amounts will occur at fixed dates due to retirement as well as at random dates due to death.

So this process is a combination (or mixture) of two processes:

- a stochastic process modelling the number of deaths, which has a discrete state space and a continuous time set
- a stochastic process modelling the number of retirements, which has a discrete state space and a discrete time set.

We model the total number of decrements and observe the initial number of members less the total number of decrements in $(0,t)$. This is a mixed process.

Question

You run a business that sells and provides service for a range of expensive sports cars. Each car sells for between £40,000 and £50,000 (cash only) and you sell about 10 to 20 each year. The 'life blood' of the business is the regular servicing and maintenance of the cars you have sold previously.

Describe the characteristics of a stochastic process that might be a suitable model for the balance on your company's bank account.

Solution

There are two processes at work:

- the day to day transactions of the business.

  These will tend to produce a 'smoothly' changing bank balance as there are lots of transactions and the majority are for relatively small amounts. We have a continuous state space and a continuous time domain.

- the very infrequent transactions which result from car sales.

  These will produce big jumps in the bank balance. The state space will consist of a limited number of discrete values and the time domain will be continuous.

The combination of the two processes suggests that a mixed process would be appropriate.

As a rule, one can say that continuous time and continuous state space stochastic processes, although conceptually more difficult than discrete ones, are also ultimately more flexible (in the same way as it is easier to calculate an integral than to sum an infinite series).

It is important to be able to conceptualise the nature of the state space of any process which is to be analysed, and to establish whether it is most usefully modelled using a discrete, a continuous, or a mixed time domain. Usually the choice of state space will be clear from the nature of the process being studied (as, for example, with the Healthy-Sick-Dead model), but whether a continuous or discrete time set is used will often depend on the specific aspects of the process that are of interest, and upon practical issues like the time points for which data are available.

1.6 Counting processes

A counting process is a stochastic process, X, in discrete or continuous time, whose state space S is the collection of natural numbers $\{0,1,2,...\}$, with the property that $X(t)$ is a non-decreasing function of t.

2 Defining a stochastic process

2.1 Sample paths

Having selected a time set and a state space, it remains to define the process $\{X_t : t \in J\}$ itself. This amounts to specifying the joint distribution of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ for *all* $t_1, t_2, \ldots, t_n$ in J and *all* integers n. This appears to be a formidable task; in practice this is almost invariably done indirectly, through some simple intermediary process (see Section 2.3 and Section 3).

Consider a simple model of the price of a stock measured in pence. Each trading day $t = 0, 1, 2, \ldots$ the price increases by 1 pence or decreases by 1 pence with probabilities p and $1-p$ respectively. The changes each day are independent. Let the price at time t be denoted X_t and assume $X_0 = 100$, so that the initial price (time 0 in our model) is £1.

This completely determines a stochastic process, even though we haven't *explicitly* given all the joint distributions. The time set (measured in days) is $J = \{0,1,2,\ldots\}$. The state space (measured in pence) is the set of non-negative integers $\{0, 1, 2, \ldots\}$.

In fact, what we've done is to specify the process in terms of its *increments* at each time. These changes $Z_t = X_t - X_{t-1}$ form another stochastic process – the intermediary process, which we referred to above.

The stochastic process of these increments is:

$$Z_t = \begin{cases} +1 & \text{with probability } p \\ -1 & \text{with probability } 1-p \end{cases}$$

Since we have assumed that these random variables are independent, $\{Z_t\}_{t \in J}$ is a set of independent and identically distributed (IID) random variables. However, the X_t's themselves are not independent. The value of X_t is the previous value plus a random change of ± 1. The value therefore depends very much on the previous value and so they are not independent. For example $P(X_{10} = 110) > 0$ but $P(X_{10} = 110 \mid X_1 = 99) = 0$.

The X_t's cannot be identically distributed either since, for example, the possible values that X_1 can take are just 99 or 101, whereas the possible values of X_2 are $98, 100$ and 102; corresponding to two days of price falls, a one day fall and a one day rise and two days of rises.

The unconditional variance of X_t increases as t increases since we are less certain about where the share price might be. (By unconditional here we mean that we are not conditioning on previous values, other than X_0.)

The process $\{Z_t\}_{t \in J}$ is an example of a white noise process.

White noise

White noise is a stochastic process that consists of a set of independent and identically distributed random variables. The random variables can be either discrete or continuous and the time set can be either discrete or continuous.

Let's consider the stock price model a little more. Here we have defined a stochastic process, $\{X_t\}_{t\in J}$ in terms of its increments $\{Z_t\}_{t\in J}$ such that:

$$X_t = X_{t-1} + Z_t$$

where Z_t is a given white noise process.

Knowing the probabilities p and $1-p$ allows us to calculate the joint probability distributions of $X_{t_1},\ldots,X_{t_n}$ for all n and all $t_1,\ldots,t_n$. So we have a complete specification of X_t.

Question

Calculate $P(X_2 = 102, X_5 = 103 \mid X_0 = 100)$ for the stock price model discussed above.

Solution

We can do this by considering all the different ways of starting from 100 at time 0, arriving at 102 at time 2, and finishing at 103 at time 5. In order for this to occur, the price must increase on the first two days, which happens with probability p^2. Independently, it must then increase on another two days and decrease on one day, not necessarily in that order. The decrease in price can occur at times 3, 4 or 5, giving three different possibilities. Each of these has probability $p^2(1-p)$. So:

$$P(X_2 = 102,\ X_5 = 103 \mid X_0 = 100) = p^2 \times 3p^2(1-p) = 3p^4(1-p)$$

Other joint probabilities can be calculated in the same way, so all the joint distributions can be determined and the stochastic process X_t is completely specified.

Question

For the stock price model described above:

(i) calculate:

(a) $P(X_2 = 100, X_4 = 103 \mid X_0 = 100)$

(b) $P(X_2 = 100, X_4 = 102 \mid X_0 = 100)$.

(ii) write down the joint distribution of X_2, X_4 given $X_0 = 100$.

Solution

(i) ***Probabilities***

(a) $P\left(X_2=100, X_4=103 \mid X_0=100\right)=2p\left(1-p\right)\times 0=0$

(There are two ways of going from $X_0=100$ to $X_2=100$, each having a probability $p(1-p)$. It is impossible to go from $X_2=100$ to $X_4=103$ because there are not enough steps to produce an increment of 3.)

(b) $P\left(X_2=100, X_4=102 \mid X_0=100\right)=2p(1-p)p^2=2(1-p)p^3$.

(ii) ***Joint distribution***

To work out the joint distribution of X_2 and X_4 we would calculate each possibility as in (i) above. To be transparent, we've included below a table of the probabilities for the various paths from time $t=0$ through to $t=4$, given $X_0=100$.

Path	Probability	Path	Probability
100 101 102 103 104	p^4	100 99 100 101 102	$p^3(1-p)$
100 101 102 103 102	$p^3(1-p)$	100 99 100 101 100	$p^2(1-p)^2$
100 101 102 101 102	$p^3(1-p)$	100 99 100 99 100	$p^2(1-p)^2$
100 101 102 101 100	$p^2(1-p)^2$	100 99 100 99 98	$p(1-p)^3$
100 101 100 101 102	$p^3(1-p)$	100 99 98 99 100	$p^2(1-p)^2$
100 101 100 101 100	$p^2(1-p)^2$	100 99 98 99 98	$p(1-p)^3$
100 101 100 99 100	$p^2(1-p)^2$	100 99 98 97 98	$p(1-p)^3$
100 101 100 99 98	$p(1-p)^3$	100 99 98 97 96	$(1-p)^4$

From this table we can derive the joint distribution of X_2 and X_4:

$P(X_2=98, X_4=96)=(1-p)^4$

$P(X_2=98, X_4=98)=2p(1-p)^3$

$P(X_2=98, X_4=100)=p^2(1-p)^2$

$P(X_2=100, X_4=98)=2p(1-p)^3$

$P(X_2=100, X_4=100)=4p^2(1-p)^2$

$P(X_2=100, X_4=102)=2p^3(1-p)$

$P(X_2=102, X_4=100)=p^2(1-p)^2$

$P(X_2=102, X_4=102)=2p^3(1-p)$

$P(X_2=102, X_4=104)=p^4$

We can think of a random event as an experiment with outcomes of varying probability. The outcome of a given experiment would be a realisation of that random variable. A *sample path* is then just the sequence of outcomes of a particular set of experiments.

For example, suppose we toss a coin at times $\{0,1,2,3,4\}$. The outcome of each toss (*ie* experiment) will be a head *H* or tail *T*, so that the expression $HHTHT$ denotes an example sample path. The set of all sample paths will be represented by the set of all sequences of *H* and *T* of length five.

Sample paths

A joint realisation of the random variables X_t for all t in J is called a *sample path* of the process; this is a function from J to S.

To say that a sample path is a function from J to S means that with each time, *ie* with each member of J, we associate the outcome of the experiment carried out at that time, which is a member of S.

The properties of the sample paths of the process must match those observed in real life (at least in a statistical sense). If this is the case, the model is regarded as successful and can be used for prediction purposes. It is essential that at least the broad features of the real life problem be reproduced by the model; the most important of these are discussed in the next subsections.

Suppose we toss a *biased* coin 1,000 times and on each toss the coin only has a one in four chance of landing tails say (a very biased coin!). A naive model of the series of tosses with equal probabilities for heads and tails would lead to sample paths that tended to have similar numbers of heads and tails. The real life experiments, however, would differ substantially. This discrepancy between observed sample paths and predicted paths would highlight the weakness of the model.

2.2 Stationarity

Stationarity is defined as follows.

Strict stationarity

A stochastic process is said to be *stationary*, or *strictly stationary*, if the joint distributions of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ and $X_{k+t_1}, X_{k+t_2}, \ldots, X_{k+t_n}$ are identical for all $t_1, t_2, \ldots, t_n$ and $k+t_1, k+t_2, \ldots, k+t_n$ in *J* and all integers *n*. This means that the statistical properties of the process remain unchanged as time elapses.

Here 'statistical properties' refers to probabilities, expected values, variances, and so on. A stationary process will be statistically 'the same' over the time period 5 to 10 and the time period from 120 to 125, for example.

In particular, the distribution of X_t will be identical to the distribution of X_{t+k} for all t and k such that t and $t+k$ are in J. Moreover, this in turn implies that expectations $E(X_t)$ and variances $\text{var}(X_t)$ must be constant over time. The failure of any one of these conditions to hold could be used to show a process is not stationary. Showing that they all hold may be difficult though.

Question

Consider a process defined by the equation:

$$X_t = X_{t-1} + Z_t \,, \quad t = 1,2,3,\ldots$$

where, for $t = 0,1,2,\ldots$:

$$Z_t = \begin{cases} +1 & \text{with probability } p \\ -1 & \text{with probability } 1-p \end{cases}$$

and $X_0 = 0$.

Calculate $P(X_{10} = 10)$ and $P(X_2 = 10)$. Hence comment on the stationarity of the process.

Solution

$P\left(X_{10} = 10\right) = p^{10}$ but $P\left(X_2 = 10\right) = 0$. Since X_2 and X_{10} do not have the same distribution, the process is not stationary.

The process described in the question above is known as a *simple random walk*. The stock price model described earlier is technically not a random walk since it does not start at 0. For the stock price model, $X_0 = 100$.

Recall the example of Section 1.2 with the three states Healthy, Sick and Dead. One would certainly not use a strictly stationary process in this situation, as the probability of being alive in 10 years' time should depend on the age of the individual.

For example, the probability of a life aged 50 surviving to age 60 is not likely to be the same as the probability of a life aged 75 surviving to age 85.

Strict stationarity is a stringent requirement which may be difficult to test fully in real life. For this reason another condition, known as *weak stationarity* is also in use. This requires that the mean of the process $m(t) = E(X_t)$ is constant and that the covariance of the process:

$$\text{cov}(X_s, X_t) = E\left[\left(X_s - m(s)\right)\left(X_t - m(t)\right)\right]$$

depends only on the time difference $t-s$.

The time difference $t-s$ is referred to as the *lag*. Recall also that the covariance can be written:

$$\operatorname{cov}(X_s, X_t) = E(X_s X_t) - E(X_s)E(X_t)$$

If a process is strictly stationary then it is also weakly stationary. However, a weakly stationary process is not necessarily strictly stationary.

Weak stationarity considers only the first two moments of the joint distribution of the set of random variables X_t.

Weak stationarity

A process is weakly stationary if:

- $E(X_t)$ is constant for all t, and
- $\operatorname{cov}(X_t, X_{t+k})$ depends only on the lag, k.

To be weakly stationary a process must pass both these tests. If it fails either of the tests then it is not weakly stationary. So when checking for stationarity, start with the easier condition (the mean), then check the covariances.

To carry out these checks we need to use the properties of the expectation and variance functions. Recall that:

$$\operatorname{var}(X_t) = \operatorname{cov}(X_t, X_t)$$

ie the variance is equal to the covariance at lag 0. So $\operatorname{var}(X_t)$ will be constant for a weakly stationary process. In addition, if W, X and Y are random variables and a, b and c are constants, then the following hold with *no assumptions* required:

- $\operatorname{cov}(Y, X) = \operatorname{cov}(X, Y)$
- $\operatorname{cov}(X, c) = 0$
- $\operatorname{cov}(aX, bY) = ab\operatorname{cov}(X, Y)$
- $\operatorname{cov}(X+Y, W) = \operatorname{cov}(X, W) + \operatorname{cov}(Y, W)$.

Question

Show that:

$$\operatorname{cov}(aX+bY, cW+d) = ac\operatorname{cov}(W, X) + cb\operatorname{cov}(W, Y)$$

by using the properties of the covariance function given above.

Solution

We have:

$$\begin{aligned}\operatorname{cov}(aX+bY,cW+d) &= \operatorname{cov}(cW,aX+bY)+\operatorname{cov}(d,aX+bY)\\ &= \operatorname{cov}(cW,aX)+\operatorname{cov}(cW,bY)+\operatorname{cov}(d,aX)+\operatorname{cov}(d,bY)\\ &= ac\operatorname{cov}(W,X)+cb\operatorname{cov}(W,Y)+0+0\\ &= ac\operatorname{cov}(W,X)+cb\operatorname{cov}(W,Y)\end{aligned}$$

2.3 Increments

Increments

An *increment* of a process is the amount by which its value changes over a period of time, *eg* $X_{t+u}-X_t$ (where $u>0$).

The increments of a process often have simpler properties than the process itself.

Example

Let S_t denote the price of one share of a specific stock. It might be considered reasonable to assume that the distribution of the return over a period of duration u, $\frac{S_{t+u}}{S_t}$, depends on u but not on t. Accordingly the log-price process $X_t=\log S_t$ would have stationary increments:

$$X_{t+u}-X_t=\log\frac{S_{t+u}}{S_t}$$

even though X_t itself is unlikely to be stationary.

Independent increments

A process X_t is said to have *independent increments* if for all *t* and every *u* > 0 the increment $X_{t+u}-X_t$ is independent of all the past of the process $\{X_s : 0\le s\le t\}$.

In the last example it is a form of the efficient market hypothesis to assume that $X_t=\log S_t$ has independent increments.

The efficient markets hypothesis is covered in Subject CM2.

The example of Section 1.3 can also be modelled by a process with stationary independent increments. Many processes are defined through their increments: see Section 3.

We have already seen that a random walk may be defined through its increments. The increment with $u=1$ is the process $\nabla X_{t+1} = X_{t+1} - X_t = Z_{t+1}$ we discussed before. These are independent of the past of the process X_t. In fact for any $u > 0$ the increment $X_{t+u} - X_t$ is independent of the past values of the process. (Here u has to be a positive integer.)

2.4 The Markov property

A major simplification occurs if the future development of a process can be predicted from its present state alone, without any reference to its past history.

Suppose that we are at time s and the value of the process at time s is x. (In symbols we have $X_s = x$).

Stated precisely the Markov property reads:

$$\boldsymbol{P\left[X_t \in A \mid X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n, X_s = x\right] = P\left[X_t \in A \mid X_s = x\right]}$$

for all times $s_1 < s_2 < \cdots < s_n < s < t$, all states $x_1, x_2, \ldots, x_n$ and x in S and all subsets A of S. This is called the *Markov property*.

The necessity to work with subsets $A \subseteq S$ (rather than just having $X_t = a \in S$) is to cover the continuous state space cases. For these the probability that X_t takes on a particular value is zero. We therefore need to work with probabilities of X_t lying in some interval of *S*, or more generally in some subset. For discrete state spaces the Markov property has the following simplification.

Markov property for a stochastic process with a discrete state space

A stochastic process with a discrete state space has the Markov property if:

$$P\left[X_t = a \mid X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n, X_s = x\right] = P\left[X_t = a \mid X_s = x\right]$$

for all times $s_1 < s_2 < \cdots < s_n < s < t$ and all states $a, x_1, \ldots, x_n$ in *S*.

It can be argued that the example of Section 1.2 can be modelled by a Markov process: if there is full recovery from the sick state to the healthy state, past sickness history should have no effect on future health prospects.

Markov result

A process with independent increments has the Markov property.

Proof

$$\begin{aligned} &P\left[X_t \in A \mid X_{s_1} = x_1, X_{s_2} = x_2, \dots, X_{s_n} = x_n, X_s = x\right] \\ &= P\left[X_t - X_s + x \in A \mid X_{s_1} = x_1, X_{s_2} = x_2, \dots, X_{s_n} = x_n, X_s = x\right] \\ &= P\left[X_t - X_s + x \in A \mid X_s = x\right] \\ &= P\left[X_t \in A \mid X_s = x\right] \end{aligned}$$

The first equality here uses the fact that we are given $X_s = x$ so that $X_t = X_t - X_s + x$. The second equality follows from the assumption that the increment $X_t - X_s$ is independent of past increments. Finally, the third equality uses the fact that $X_s = x$ again.

If we need to check whether or not a stochastic process is Markov, then:

- we can first check if it has independent increments, if yes then it is Markov.
- if it does not have independent increments, then it may still be Markov if it satisfies the Markov definition.
- sometimes it is very difficult to check a process using the Markov definition, so we may need to resort to some general reasoning arguments to try and demonstrate that the Markov definition is satisfied.

Question

Consider a discrete-time process on the integers defined as follows: $X_t = X_{t-1} + I_t$ where I_t are random variables taking the value +1 or –1 with probabilities $p_t = e^{-|X_{t-1}|}$ and $q_t = 1 - e^{-|X_{t-1}|}$ respectively. Explain whether this process has:

(a) independent increments

(b) the Markov property.

Solution

(a) The increments, $X_t - X_{t-1}$, depend on the value of X_{t-1}. So the process does not have independent increments.

(b) It is Markov, however, since our knowledge of past values additional to the current value is irrelevant.

To recap: a process with independent increments has the Markov property, but a Markov process does not necessarily have independent increments. The two properties are not equivalent.

2.5 Filtrations

The following structures underlie any stochastic process X_t:

- **a sample space Ω: each outcome ω in Ω determines a sample path $X_t(\omega)$**
- **a set of events F: this is a collection of events, by which is meant subsets of Ω, to which a probability can be attached**
- **for each time t, a smaller collection of events $F_t \subset F$: this is the set of those events whose truth or otherwise are known at time t. In other words an event A is in F_t if it depends only on X_s, $0 \le s \le t$.**

As t increases, so does F_t: $F_t \subset F_u, t \le u$. Taken collectively, the family $(F_t)_{t \ge 0}$ is known as the (natural) filtration associated with the stochastic process $X_t, t \ge 0$; it describes the information gained by observing the process or the internal history of X_t up to time t.

This is the key thing to know about the (natural) filtration – that F_t gives the history of the process up to time t. If we had a process with a discrete time set $t = 0, 1, 2, \ldots$, then we could write the history of the process up to time n as follows:

$$X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0$$

Alternatively, we could denote this set of events by F_n.

However, if we had a process with a continuous time set, we couldn't list the complete set of events up to time n, even if we used the '…' notation. (This is because there is an uncountable number of points in the set $[0, n]$.) So in this case the complete history of the process up to time n has to be represented in terms of the filtration.

The process X_t can be said to have the Markov property if:

$$P\left[X_t \le x \middle| F_s\right] = P\left[X_t \le x \middle| X_s\right]$$

for all $t \ge s \ge 0$.

When a Markov process has a discrete state space and a discrete time set it is called a *Markov chain*; Markov chains are studied in Chapter 2. When the state space is discrete but the time set is continuous, one uses the term *Markov jump process*; Markov jump processes are studied in Chapters 3, 4 and 5.

Using the preliminaries in this section we can now show by a series of examples how to define a stochastic process.

3 Examples

In each of the examples in this section we try to appreciate which of the following properties hold:

- stationarity in the weak sense
- independent increments
- the Markov property.

3.1 White noise

Consider a discrete-time stochastic process consisting of a sequence of independent random variables $X_1,...,X_n,...$.

The Markov property holds in a trivial way.

Even though the process does not have independent increments, the Markov definition is satisfied since the future development of the process is completely independent of its past.

The process is stationary if and only if all the random variables X_n have the same distribution. Such sequences of independent identically distributed (IID for short) random variables are sometimes described as a discrete-time *white noise*.

White noise processes are normally defined as having a mean of zero at all times, that is $m(t)=E[X_t]=0$ for all values of t. They may be defined in either discrete or continuous time. In a white noise process with a mean of zero, the covariance of the process, $\text{cov}(s,t)=E\left[(X_s-m(s))(X_t-m(t))\right]$, is zero for $s \neq t$. The main use of white noise processes is as a starting point to construct more elaborate processes below.

3.2 General random walk

Start with a sequence of IID random variables $Y_1,...,Y_j,...$ and define the process:

$$X_n=\sum_{j=1}^{n} Y_j$$

with initial condition $X_0=0$.

The formula for the random walk can also be written as:

$$X_n=X_{n-1}+Y_n \ , \ n=1,2,3,...$$

This is a process with stationary independent increments, and thus a discrete-time Markov process. It is known as a general random walk. The process is not even weakly stationary, as its mean and variance are both proportional to n.

The log of the closing value of the FTSE100 index could be modelled by a general random walk, the value one day being the value on the previous day plus some random adjustment.

In the special case where the steps Y_j of the walk take only the values $+1$ and -1, the process is known as a *simple random walk*.

In addition, if:

$$Y_j = \begin{cases} +1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5 \end{cases}$$

the process is known as a *simple symmetric random walk*.

3.3 Poisson process

A *Poisson process with rate* λ is a continuous-time integer-valued process N_t, $t \geq 0$, with the following properties:

(i) $N_0 = 0$

(ii) **N_t has independent increments**

(iii) **N_t has Poisson distributed stationary increments:**

$$P\left[N_t - N_s = n\right] = \frac{\left[\lambda\left(t-s\right)\right]^n e^{-\lambda(t-s)}}{n!}, \qquad s < t, \quad n = 0, 1, \ldots$$

Property (iii) can also be expressed as follows:

$$N_t - N_s \sim Poisson\left(\lambda(t-s)\right) \text{ for } 0 \leq s < t$$

This is a Markov jump process with state space $S = \{0,1,2,\ldots\}$. It is not stationary: as is the case for the random walk, both the mean and variance increase linearly with time.

The process counts the number of events (that are occurring at a rate λ per unit time) that occur between time s and time t. It is an example of a counting process. Counting processes are defined in Section 1.6.

This process is of fundamental importance when counting the cumulative number of occurrences of some event over $[0,t]$, irrespective of the nature of the event (car accident, claim to insurance company, arrival of customer at a service point). A detailed study of this process and its extensions is one of the subjects of Chapter 4.

Since the increments have a Poisson distribution they can only take the values 0,1,2,… It follows that the process must be increasing, that is $N_t \geq N_s$ for all s and t such that $t \geq s$. The expectations must therefore be increasing, and the process is not weakly stationary.

In fact, the process can only increase by one step at a time, making it a natural counting process. It is very often used to model the number of insurance claims made by time t. The rate parameter λ is the expected number of claims arriving per unit time.

We will see in Chapter 4 that there are other equivalent definitions of a Poisson process.

3.4 Compound Poisson process

Start with a Poisson process $N_t, t \geq 0$ and a sequence of IID random variables Y_j, $j \geq 1$. A *compound Poisson process* is defined by:

$$X_t = \sum_{j=1}^{N_t} Y_j \text{ , } t \geq 0 \qquad \textbf{(1.1)}$$

We additionally assume that the random variables Y_j are independent of N_t.

When $N_t = 0$ we define $X_t = 0$. The Y_j may be discrete or continuous random variables. For example, N_t could be the number of storms up to time t, and Y_j could be the number of claims arising from the j th storm (discrete), or the cost of claims from the j th storm (continuous).

This process has independent increments and thus the Markov property holds. It serves as a model for the cumulative claim amount reaching an insurance company during $[0,t]$: N_t is the total number of claims over the period and Y_j is the amount of the j th claim.

A common application consists of estimating the *probability of ruin*:

$$\psi(u) = P\left[u + ct - X_t < 0 \text{ for some } t > 0\right]$$

for a given initial capital *u*, premium rate *c*, X_t defined as in (1.1), and some fixed distribution of the claim sizes.

If we receive income from premiums at a rate of c per unit of time, then by time t we will have received an amount ct. Also, X_t models the cumulative amount of claims incurred by the company. Starting with an initial surplus of u we will therefore have a surplus of $u + ct - X_t$ at time t. The probability of ruin is therefore just the probability that at some point in the future we will be ruined, *ie* the probability that the surplus is less than 0.

3.5 Time series

A time series is a set of observations indexed in time order, *eg* the closing value of the FTSE100 share price index at the end of each week. The observations are usually equally spaced in time, in which case they can be considered to be realisations of the random variables $X_1, X_2, X_3, \ldots$. The values of $X_1, X_2, X_3, \ldots$ are related to each other, and should not be considered as a set of independent random variables (except in the trivial case of a white noise process).

By definition, a time series has a discrete time set and a continuous state space, so that $X_1, X_2, X_3, \ldots$ are continuous random variables.

Time series will be studied in detail in Chapters 13 and 14.

Chapter 1 Summary

Stochastic processes

A stochastic process is a model for a time-dependent random phenomenon, a collection of random variables $\{X_t : t \in J\}$, one for each time t in the time set J. The time set may be discrete or continuous. The set of values that the X_t are capable of taking is called the *state space S*. The state space may also be discrete or continuous.

Defining such a process amounts to specifying the joint distribution of $X_{t_1}, X_{t_2}, .., X_{t_n}$ for all $t_1, t_2, ..., t_n$ in the time set J and all integers *n*.

Sample paths

A joint realisation of the random variables X_t for all t in J is called a sample path of the process; this is a function from J to S.

Stationarity

If the statistical properties of a process do not vary over time, the process is stationary. This makes the modelling process much easier.

Mathematically, stationarity requires the joint distribution of any set of values $\{X_{t_1}, X_{t_2}, \dots, X_{t_n}\}$ to be the same as the joint distribution of $\{X_{t_1+k}, X_{t_2+k}, \dots, X_{t_n+k}\}$, *ie* when all times are shifted across by k. This is the *strict* definition.

In practice, it is only necessary to have *weak* stationarity. This requires only the first two moments not to vary over time, *ie* $E(X_t)$ and $\text{var}(X_t)$ are constant, and $\text{cov}(X_{t_1}, X_{t_2})$ depends only on the lag $t_2 - t_1$.

Independent increments

An increment of a stochastic process (that has a numerical state space) is just the change in the value between two times, *ie* $X_{t_2} - X_{t_1}$. If this is independent of the past values of the process up to and including time t_1 then the process is said to have independent increments.

Filtration

We often need to look at expectations of the future value of a process, conditional on the known past history. For example, for a discrete-time process, we might be interested in $P[X_n \mid X_1, X_2, \dots, X_{n-1}]$.

For a continuous-time process this presents a theoretical difficulty, since it is impossible to list the values at all past times. The filtration notation F is used here, and we write, for example, $P[X_t \mid F_s]$ to represent the probability distribution at a future time t, conditional on the values up to the earlier time s.

The filtration notation can be used for both discrete-time and continuous-time processes.

Markov property

If the probabilities for the future values of a process are dependent only on the latest available value, the process has the Markov property.

Mathematically, for a process with time set $\{1, 2, 3, ...\}$ and a discrete state space:

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_1 = x_1)$$
$$= P(X_n = x_n \mid X_{n-1} = x_{n-1})$$

For a continuous-time process with a discrete state space, we need to express this in the form:

$$P(X_n = x_n \mid F_s) = P(X_n = x_n \mid X_s)$$

For a continuous-time process with a continuous state space, we need to express this in the form:

$$P(X_n \in A \mid F_s) = P(X_n \in A \mid X_s)$$

White noise

White noise is a stochastic process that consists of a set of independent and identically distributed random variables. The random variables can be either discrete or continuous and the time set can be either discrete or continuous. White noise processes are stationary and have the Markov property.

Random walk

Suppose that $Y_1, Y_2, Y_3, \ldots$ is a sequence of IID random variables and suppose that:

$$X_0 = 0$$

$$X_n = Y_1 + Y_2 + \cdots + Y_n = X_{n-1} + Y_n, \quad n = 1, 2, 3, \ldots$$

Then X_n is a (general) random walk. X_n has a discrete time set. If the random variables Y_j are discrete, then X_n has a discrete state space. If the random variables Y_j are continuous, then X_n has a continuous state space.

In the special case when each Y_j can only take the values $+1$ and -1, X_n is said to be a simple random walk. In addition, if $P(Y_j = +1) = P(Y_j = -1) = 0.5$, X_n is said to be a simple symmetric random walk.

Poisson process

N_t is a Poisson process with rate λ if it is a continuous-time, integer-valued process with the following properties:

- $N_0 = 0$
- N_t has independent increments
- N_t has Poisson distributed stationary increments:

$$P(N_t - N_s = n) = \frac{e^{-\lambda(t-s)}[\lambda(t-s)]^n}{n!}, \quad 0 \le s < t, \quad n = 0, 1, 2, \ldots$$

Compound Poisson process

Suppose that:

$$S_t = Y_1 + Y_2 + \cdots + Y_{N_t}$$

where $Y_1, Y_2, Y_3, \ldots$ is a sequence of IID random variables, N_t is a Poisson process with rate λ, and the random variables Y_j are independent of N_t. Then S_t is a compound Poisson process with rate λ.

Like a Poisson process, a compound Poisson process has a continuous time set. If the random variables Y_j are discrete, then S_t has a discrete state space. If the random variables Y_j are continuous, then S_t has a continuous state space.

Time series

A time series is a set of observations indexed in time order. A time series has a discrete time set and a continuous state space.

The practice questions start on the next page so that you can keep the chapter summaries together for revision purposes.

Chapter 1 Practice Questions

1.1 For a stochastic process X_n with time set J and state space S, define the terms:

(i) stationary

(ii) weakly stationary

(iii) increment

(iv) Markov property.

1.2 A moving average (stochastic) process, X_n, has a discrete time set and a continuous state space and is defined as:

$$X_n = Z_n + \alpha_1 Z_{n-1} + \alpha_2 Z_{n-2} + \alpha_3 Z_{n-3}$$

where $\{Z_n, n \in \mathbb{Z}\}$ are independent and identically distributed $N(0,\sigma^2)$ random variables and $\alpha_1, \alpha_2, \alpha_3$ are constants.

(i) Prove that X_n is weakly stationary.

(ii) Explain whether the Markov property holds.

(iii) Deduce whether the process has independent increments.

1.3 Explain whether a random walk has the Markov property.

1.4 (i) (a) Define a Poisson process with rate λ.

(b) Define a compound Poisson process.

(ii) Identify the circumstances in which a compound Poisson process is also a Poisson process.

(iii) The cumulative amount of claims reaching an insurance company is modelled using a compound Poisson process.

(a) Explain why the compound Poisson process has the Markov property.

(b) Comment on whether this seems reasonable for the given insurance model.

(c) State whether the compound Poisson process is weakly stationary.

(d) Explain whether you expect the cumulative insurance claims to follow a weakly stationary process.

1.5 Define a simple symmetric random walk and identify its time set and state space.

1.6 The price of an ordinary share is modelled as a stochastic process X_n; $n = 0, 1, 2, 3, \dots$ with initial condition $X_0 = x_0 > 0$, where:

$$X_n = x_0 \prod_{j=1}^{n} U_j \quad n \geq 1$$

and U_n is a white noise process.

(i) Show that the process $\ln X_n$, $n \geq 0$ has independent increments.

(ii) Explain why X_n is a Markov process.

1.7 Exam style Calculate the covariance between the values $X(t)$ and $X(t+s)$ taken by a Poisson process $X(t)$ with constant rate λ at the two times t and $t+s$, where $s > 0$. [2]

1.8 Exam style (i) X_n is a stochastic process with a discrete state space and a discrete time set. Show that if non-overlapping increments of this process are independent, then the process satisfies the Markov property. [2]

(ii) Show that a white noise process in discrete time with a discrete state space does not have independent increments, but is a Markov process. [2]

[Total 4]

1.9 Exam style An insurer has initial capital of u and receives premium income continuously at the rate of c per annum. Let $S(t)$ denote the total claim amount up to time t.

(i) Describe a model that would allow the insurer to estimate its probability of ruin (*ie* the probability that its claims outgo is more than its available funds). State any assumptions that you make. [3]

(ii) Write down an expression for the probability of ruin in terms of u, c and $S(t)$. [1]

[Total 4]

1.10 Exam style (i) In the context of a stochastic process denoted by $\{X_t : t \in J\}$, define the terms:

(a) state space

(b) time set

(c) sample path. [2]

(ii) Stochastic process models can be placed in one of four categories according to whether the state space is continuous or discrete, and whether the time set is continuous or discrete. For each of the four categories:

(a) state a stochastic process model of that type

(b) give an example of a problem an actuary may wish to study using a model from that category. [4]

[Total 6]

Chapter 1 Solutions

1.1 (i) ***Stationary***

A stochastic process X_n is stationary if the joint distributions of $X_{t_1}, X_{t_2}, \dots, X_{t_m}$ and $X_{t_1+k}, X_{t_2+k}, \dots, X_{t_m+k}$ are identical for all $t_1, t_2, \dots, t_m, t_1+k, t_2+k, \dots, t_m+k \in J$ and all integers m.

(ii) ***Weakly stationary***

The process is weakly stationary if the expectations $E\left[X_t\right]$ are constant with respect to t and the covariances $\operatorname{cov}\left(X_t, X_{t+k}\right)$ depend only on the lag k.

(iii) ***Increment***

If t and $t+u$ are in J then the increment for duration u will be $X_{t+u} - X_t$.

(iv) ***Markov property***

The Markov property states that:

$$P\left(X_t \in A \mid X_{t_1} = x_1, X_{t_2} = x_2, \dots, X_{t_m} = x_m\right) = P\left(X_t \in A \mid X_{t_m} = x_m\right)$$

for all times $t_1 < t_2 < \dots < t_m < t \in J$, all states $x_1, x_2, \dots, x_m \in S$ and all subsets A of S.

1.2 (i) ***Weak stationarity***

The Z_j are independent and identically distributed, and the α_j are constants. So:

$$E(X_n) = \left(1+\alpha_1+\alpha_2+\alpha_3\right)E(Z) = \left(1+\alpha_1+\alpha_2+\alpha_3\right)\times 0 = 0$$

and:

$$\operatorname{var}(X_n) = \operatorname{var}(Z) + \alpha_1^2 \operatorname{var}(Z) + \alpha_2^2 \operatorname{var}(Z) + \alpha_3^2 \operatorname{var}(Z) = \left(1+\alpha_1^2+\alpha_2^2+\alpha_3^2\right)\sigma^2$$

which is constant.

The covariance at lag 1 is:

$$\begin{aligned}
&\operatorname{cov}\left(X_n, X_{n+1}\right) \\
&= \operatorname{cov}\left(Z_n + \alpha_1 Z_{n-1} + \alpha_2 Z_{n-2} + \alpha_3 Z_{n-3}, Z_{n+1} + \alpha_1 Z_n + \alpha_2 Z_{n-1} + \alpha_3 Z_{n-2}\right) \\
&= \alpha_1 \operatorname{var}(Z) + \alpha_1\alpha_2 \operatorname{var}(Z) + \alpha_2\alpha_3 \operatorname{var}(Z) \\
&= \left(\alpha_1 + \alpha_1\alpha_2 + \alpha_2\alpha_3\right)\sigma^2
\end{aligned}$$

The covariance at lag 2 is:

$$
\begin{aligned}
&\text{cov}\left(X_n, X_{n+2}\right)\\
&=\text{cov}\left(Z_n+\alpha_1 Z_{n-1}+\alpha_2 Z_{n-2}+\alpha_3 Z_{n-3}, Z_{n+2}+\alpha_1 Z_{n+1}+\alpha_2 Z_n+\alpha_3 Z_{n-1}\right)\\
&=\alpha_2 \text{ var}(Z)+\alpha_1\alpha_3 \text{ var}(Z)\\
&=\left(\alpha_2+\alpha_1\alpha_3\right)\sigma^2
\end{aligned}
$$

The covariance at lag 3 is:

$$
\begin{aligned}
&\text{cov}\left(X_n, X_{n+3}\right)\\
&=\text{cov}\left(Z_n+\alpha_1 Z_{n-1}+\alpha_2 Z_{n-2}+\alpha_3 Z_{n-3}, Z_{n+3}+\alpha_1 Z_{n+2}+\alpha_2 Z_{n+1}+\alpha_3 Z_n\right)\\
&=\alpha_3 \text{ var}(Z)\\
&=\alpha_3\sigma^2
\end{aligned}
$$

The covariances at lags 4, 5, 6 … are 0.

So the covariance depends only on the lag and not on the value of n. Thus the process X_n is weakly stationary.

(ii) ***Markov?***

For a Markov process, the value of X_n only depends on the most recently known value. However, X_n depends on the previous X values so it does not possess the Markov property.

(iii) ***Independent increments?***

If the increments of a process are independent, then that process must have the Markov property. Since we've said that this process is not a Markov process, it cannot have independent increments.

1.3 A random walk has independent increments, so it has the Markov property.

1.4 (i)(a) ***Poisson process***

A Poisson process N_t, $t \geq 0$, with rate λ is a continuous-time, integer-valued process such that:

- $N_0 = 0$
- N_t has independent increments
- $N_t - N_s \sim Poisson(\lambda(t-s))$ for $0 \leq s < t$.

(i)(b) ***Compound Poisson process***

Let $\{X_n\}_{n=1}^{\infty}$ be independent identically distributed random variables. A compound Poisson process with rate λ is defined for $t \geq 0$ to be:

$$S_t = X_1 + X_2 + \cdots + X_{N_t}$$

where N_t is a Poisson process and $S_t = 0$ when $N_t = 0$.

(ii) ***When a compound Poisson process is also a Poisson process***

S_t is also a Poisson process if the random variables X_j can only take the value 0 or 1.

The situation where $X_j = 1$ for all j is a special case of this.

(iii)(a) ***Markov property***

It is sufficient to show that the compound Poisson process has independent increments, since then the Markov property must hold. However, having independent increments is part of the definition of the compound Poisson process.

(iii)(b) ***Reasonableness***

This is consistent with insurance claims, since we would only expect the cumulative insurance claims by time t to depend on the most recently known value. For example, if we know the cumulative claims after day one are £1,000, and by day ten are £15,000, we wouldn't expect the older value of £1,000 to add any useful information to the more recent value of £15,000.

(iii)(c) ***Weak stationarity***

The process cannot be stationary since, for example, $E(S_t)$ changes with t.

(iii)(d) ***Is cumulative claim amount weakly stationary?***

We wouldn't expect $E(S_t)$ to be constant since the cumulative claims generally increases with time. This would be a constant only in the trivial case where the individual claim amounts are £0, which is rather uninteresting.

In order to show that a process is not stationary, it is sufficient to show that any one of the conditions fails to hold.

1.5 A simple symmetric random walk is defined by the equation:

$$X_n = \sum_{j=1}^{n} Y_j$$

where the random variables Y_j are independent and identically distributed with common probability distribution:

$$P\left(Y_j = +1\right) = \frac{1}{2} \quad \text{and} \quad P\left(Y_j = -1\right) = \frac{1}{2}$$

The word 'symmetric' is important as it denotes a particular process that is equally likely to 'step' upwards or downwards in its walk.

In addition, the process starts at 0, *ie* $X_0 = 0$.

The simple symmetric random walk has a discrete state space consisting of the values, $\{\ldots,-2,-1,0,+1,+2,\ldots\}$ and a discrete time set consisting of the values $\{0,1,2,\ldots\}$.

1.6 (i) ***Independent increments***

By definition:

$$\ln X_n = \ln x_0 + \sum_{j=1}^{n} \ln U_j = \ln x_0 + \sum_{j=1}^{n} Z_j$$

where $Z_j = \ln U_j$ is a white noise process, *ie* are a set of independent and identically distributed random variables. Then:

$$\ln X_n - \ln X_{n-1} = \ln U_n = Z_n$$

Because $Z_n, n = 0,1,\ldots$ are independent, $\ln X_n$ has independent increments.

(ii) ***Markov process***

$\ln X_n$ has independent increments

$\Rightarrow$ $\ln X_n$ is a Markov process

$\Rightarrow$ $X_n = \exp\left(\ln X_n\right)$ is a Markov process.

because exponentiation merely rescales the state space of the process.

1.7 *The key to many results for continuous-time stochastic processes is to realise that the random variables representing behaviour in non-overlapping time periods are independent. Here the non-overlapping time periods are* $(0,t)$ *and* $(t,t+s)$. *So…*

$$\begin{aligned}\operatorname{cov}\left(X(t),X(t+s)\right)&=\operatorname{cov}\left(X(t),X(t)+\left(X(t+s)-X(t)\right)\right)\\&=\operatorname{cov}\left(X(t),X(t)\right)+\operatorname{cov}\left(X(t),X(t+s)-X(t)\right)\\&=\operatorname{var}\left(X(t)\right)+0\\&=\lambda t\end{aligned}$$

since $X(t)\sim Poisson(\lambda t)$. [2]

1.8 (i) ***Proof***

We have:

$$\begin{aligned}&P\left[X_n=a\,|\,X_{n-m}=x,X_{n-m-1}=x_{n-m-1},X_{n-m-2}=x_{n-m-2},X_{n-m-3}=x_{n-m-3},\ldots\right]\\&\quad=P\left[X_n-X_{n-m}+x=a\,|\,X_{n-m}=x,X_{n-m-1}=x_{n-m-1},X_{n-m-2}=x_{n-m-2},\ldots\right]\end{aligned}$$

for all times $m>0$ and all states $a,x,x_{n-m-1},x_{n-m-2},\ldots$ in the state space, S. [1]

$$\begin{aligned}&P\left[X_n-X_{n-m}+x=a\,|\,X_{n-m}=x,X_{n-m-1}=x_{n-m-1},X_{n-m-2}=x_{n-m-2},\ldots\right]\\&\qquad=P\left[X_n-X_{n-m}+x=a\,|\,X_{n-m}=x\right]\\&\qquad=P\left[X_n=a\,|\,X_{n-m}=x\right]\end{aligned}$$

if non-overlapping increments are independent. So X_n has the Markov property. [1]

(ii) ***White noise process***

For a discrete time, discrete state white noise process $\{Z_n:n=1,2,3,\ldots\}$, where Z_n are independent and identically distributed random variables with mean μ and variance σ^2, we have:

$$\operatorname{cov}\left(Z_n-Z_{n-1},Z_{n-1}-Z_{n-2}\right)=\operatorname{cov}\left(-Z_{n-1},Z_{n-1}\right)=-\sigma^2$$

So non-overlapping increments are not independent. [1]

However:

$$P\left[Z_n=z_n|\,Z_{n-m}=z_{n-m},Z_{n-m-1}=z_{n-m-1},Z_{n-m-2}=z_{n-m-2},\ldots\right]=P\left[Z_n=z_n\right]$$

and:

$$P\left[Z_n=z_n|\,Z_{n-m}=z_{n-m}\right]=P\left[Z_n=z_n\right]$$

because the random variables are independent. So the process satisfies the Markov property. [1]

1.9 (i) ***Model***

Let $N(t)$ denote the number of claims received by the insurer up to time t. $N(t)$ can be modelled as a Poisson process. [1]

Let X_j denote the amount of the jth claim. Then the cumulative claim amount up to time t is given by:

$$S(t) = X_1 + X_2 + \cdots X_{N(t)}$$

If we assume that the random variables X_j are independent and identically distributed, and they are independent of $N(t)$, then $S(t)$ is a compound Poisson process. [2]

(ii) ***Probability of ruin***

The probability of ruin for the insurer is the probability that, for some time t, its claims outgo up to time t is greater than its initial capital plus premium income up to time t. In symbols, this is:

$$P\left(S(t) > u + ct \text{ for some } t > 0\right)$$ [1]

1.10 *This is Subject CT4, September 2005, Question A2.*

(i)(a) ***State space***

The state space of the stochastic process $\{X_t : t \in J\}$ is the set of values that the random variables X_t can take. The state space can be discrete or continuous. [½]

(i)(b) ***Time set***

The time set for this stochastic process is J, which contains all points at which the value of the process can be observed. The time set can be discrete or continuous. [½]

(i)(c) ***Sample path***

A sample path is a joint realisation of the random variables X_t for all $t \in J$. [1]

(ii)(a) ***Examples of stochastic processes***

Discrete state space, discrete time set

Examples include Markov chains, simple random walks and discrete-time white noise processes that have discrete state spaces. [½]

Discrete state space, continuous time set

Examples include Markov jump processes (of which the Poisson process is a special case) and counting processes. [½]

Continuous state space, discrete time set

Examples include general random walks and time series. [½]

Continuous state space, continuous time set

Examples include Brownian motion, diffusion processes and compound Poisson processes where the state space is continuous. [½]

Brownian motion and diffusion processes are covered in Subject CM2.

(ii)(b) ***Examples of problems an actuary may wish to study***

Discrete state space, discrete time set

An example of this is a no claims discount system. The random variable X_t represents the discount level given to a policyholder in year t, $t = 1, 2, \ldots$. [½]

Discrete state space, continuous time set

An example of this is the health, sickness, death model, which can be used to value sickness benefits. The random variable X_t takes one of the values healthy, sick or dead for each $t \geq 0$. [½]

Continuous state space, discrete time set

An example of this is a company's share price at the end of each trading day. Another example is the annual UK inflation rate. [½]

Continuous state space, continuous time set

An example of this is the cumulative claim amount incurred on a portfolio of policies up to time t. Another example is a company's share price at time t, where t denotes time since trading began. [½]

2

Markov chains

Syllabus objectives

3.2 Define and apply a Markov chain.

3.2.1 State the essential features of a Markov chain model.

3.2.2 State the Chapman-Kolmogorov equations that represent a Markov chain.

3.2.3 Calculate the stationary distribution for a Markov chain in simple cases.

3.2.4 Describe a system of frequency-based experience rating in terms of a Markov chain and describe other simple applications.

3.2.5 Describe a time-inhomogeneous Markov chain model and describe simple applications.

3.2.6 Demonstrate how Markov chains can be used as a tool for modelling and how they can be simulated.

0 Introduction

Recall from Chapter 1 that the Markov property means that the future value of a process is independent of the past history and only depends on the current value. Any process satisfying the Markov property is a Markov process. The term *Markov chain* refers to Markov processes in discrete time and with a discrete state space.

1 An example of a Markov chain

As an example, consider the *no claims discount* (*NCD*) model run by motor insurance companies. A company might offer discounts of say 0%, 30% and 60% of the full premium.

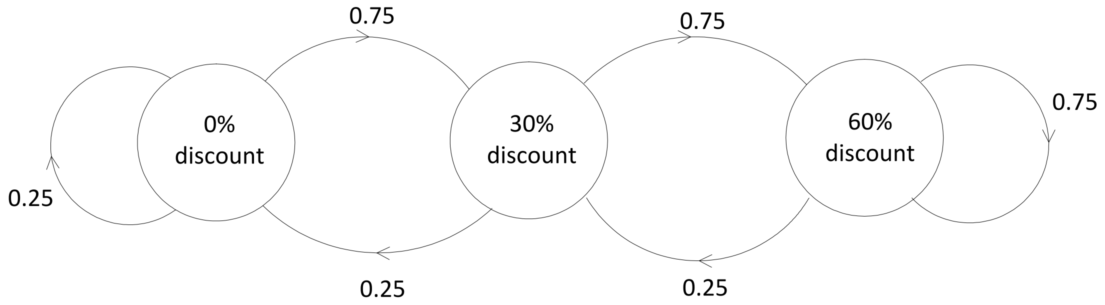
A policyholder's status is determined by the following rules:

- All new policyholders start at the 0% level.
- If no claim is made during the current year, then the policyholder moves up one discount level or remains at the 60% level.
- If one or more claims are made, the policyholder moves down one level, or remains at 0% discount.

This can be modelled using the state space $S=\{0\%,30\%,60\%\}$. When the policy is renewed each year, the policyholder moves to another level, or remains at the same level, with various probabilities depending on the chance of making a claim. Assume the chance of claiming is independent of the current level and that $P(\text{no claim})=\frac{3}{4}$.

This is an example of a discrete-time process that satisfies the Markov property (because the future only depends on the current level, and not on the past history). Hence this is a Markov chain.

One way of representing a Markov chain is by its *transition graph*.



The states are represented by the circles, and each arrow represents a possible *transition.* Staying in a state for one time period is also a possible transition, so we need an arrow to show this. Next to each arrow is written the corresponding *transition probability.* The arrows at either end correspond to policyholders who remain in the 0% or 60% states. Policyholders in the 30% discount state will always move to another state because of the rules of the discount scheme.

The transition probability between two states in unit time is therefore given explicitly. These can be written in the form of a *transition matrix*:

$$P=\begin{pmatrix} \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

The $(i,j)th$ entry in the matrix, *ie* $i\,th$ row and $j\,th$ column, gives the probability of moving in *one step* from state i to state j. In order for this notation to make sense, we need the states to be labelled to correspond with the matrix entries. In the example described above, the state '0%' corresponds to $i=1$, the state '30%' corresponds to $i=2$, and the state '60%' corresponds to $i=3$.

The entries in each row sum to 1. This can be interpreted as saying that at the end of each year, some transition into another state must occur or the process stays where it is. The sum of the probabilities of the mutually exclusive and exhaustive events that can happen when the process is in state i is 1.

It turns out that to calculate transition probabilities over 2 steps, we use the matrix P^2, and for three steps we use P^3, and so on. This greatly simplifies the analysis of such situations. This is the basic content of the *Chapman-Kolmogorov* equations, which we discuss in Section 2. The one-step transition probabilities will generally be given, and we will have to construct general transition probabilities by applying the Chapman-Kolmogorov equations.

There is an added complication however. In the stochastic process we've just looked at, the transition matrix does not depend on the current time. Such a process is said to be *time-homogeneous.* In general, however, we need to consider the possibility that even the one-step transition matrices can vary with time. It takes a while to get used to working with the matrix notation, but the theory itself is not too difficult.

In Section 3 we concentrate on the time-homogeneous case, which simplifies things. We look at the time-inhomogeneous case in Section 4.

Section 5 contains several examples, including further discussion of the NCD model introduced above. We also consider random walks on both finite and infinite state spaces.

In Section 6, we study the long-term behaviour of Markov chains. This is important. For example, in the NCD case above, we would expect after a while that the process would settle down, and that the same proportion of policyholders would be in each discount level at any one time. This does not mean that each individual stays put, but that, although each individual moves around, the process as a whole reaches an equilibrium or stationary position.

Mathematically, the problem of finding the proportion of people who are in each state in the long run can easily be tackled using the transition matrices defined above. The problem reduces to solving a set of simultaneous equations.

Not all Markov chains have a single stationary distribution. Some chains may have no stationary distribution and some chains may have more than one stationary distribution. Chains with a single stationary distribution may be such that this distribution is never reached.

We will describe three classifications of Markov chains, and use these classifications to define categories so that all those chains in the same category have the same long-run behaviour. For one category there is a unique long-term stationary distribution that will be reached after a sufficient length of time. We will describe how to find this unique stationary distribution.

2 The Chapman-Kolmogorov equations

Recall from Chapter 1 that the term *Markov chain* is reserved for discrete-time Markov processes with a finite or countable state space *S*; so a *Markov chain* is a sequence of random variables $X_0, X_1, \ldots, X_n, \ldots$ with the following property:

$$P\left[X_n = j \middle| X_0 = i_0, X_1 = i_1, \ldots, X_{m-1} = i_{m-1}, X_m = i\right] = P\left[X_n = j \mid X_m = i\right] \qquad \textbf{(2.1)}$$

for all integer times $n > m$ and states $i_0, i_1, \ldots, i_{m-1}, i, j$ in S.

This definition of the Markov property is appropriate only when the time set is discrete (as is the case for a Markov chain). In general, it need not be.

The *Markov property* (2.1) has the following interpretation: given the present state of the process $X_m = i$, the additional knowledge of the past is irrelevant for the calculation of the probability distribution of future values of the process.

Some knowledge of the past may be incorporated in X_m. It is the additional information contained in the earlier values of the process that does not provide any help in predicting the future behaviour of the process.

The conditional probabilities on the right-hand side of (2.1) are the key objects for the description of a Markov chain; we call them *transition probabilities*, and we denote them by:

$$P\left[X_n = j \middle| X_m = i\right] = p_{ij}^{(m,n)}$$

So $p_{ij}^{(m,n)}$ is the probability of being in state j at time n having been in state i at time m.

In particular, we can define the *one-step* transition probabilities:

$$P(X_{m+1} = j \mid X_m = i) = p_{ij}^{(m,m+1)}$$

These tell us in a probabilistic sense what will happen at the next step at any time m. These one-step transitions therefore describe the immediate future. The NCD transition matrix P is an example of this.

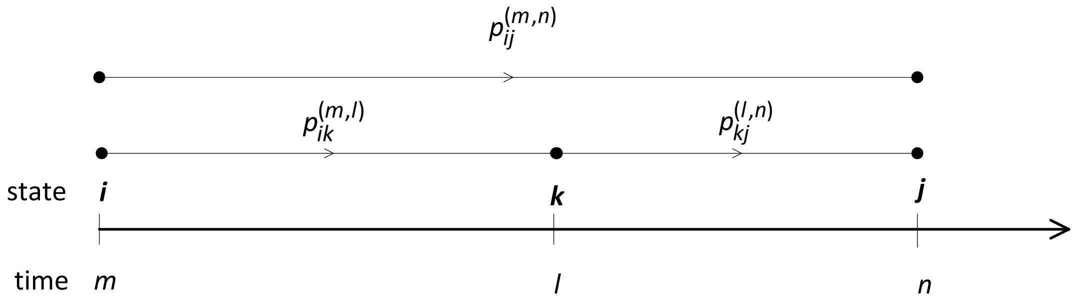
If we know all such probabilities, then intuitively we should be able to calculate any long-term transition probability, from time m to time $n > m$, by considering a sequence of such one-step transitions. This can be deduced from the following fundamental result.

The transition probabilities of a discrete-time Markov chain obey the *Chapman-Kolmogorov equations:*

$$p_{ij}^{(m,n)} = \sum_{k \in S} p_{ik}^{(m,l)} \, p_{kj}^{(l,n)}$$

for all states i, j in S and all integer times $m < l < n$.

Before giving a proof of this result we should be clear about its interpretation. We can split any path between times m and n into two parts by introducing some fixed intermediate time l. Assume at this time we are in some state k.



The associated transition probability for the whole of that path will then be the product of the transition probabilities for each part, namely $p_{ik}^{(m,l)}\ p_{kj}^{(l,n)}$. So $p_{ik}^{(m,l)}\ p_{kj}^{(l,n)}$ is the probability of going from state i at time m to state j at time n, via state k at time l. The probabilities are 'chained together', which is the reason we call these processes Markov *chains*.

If we start in state i at time m, and finish in state j at time n, then in general there will be several possibilities for this intermediate state k. To take into account all of these different paths we must therefore sum over all the mutually exclusive and exhaustive possibilities. This gives us the right-hand side of the above equation. We have used the phrase 'finish in' here rather than 'go to', because it could be that the transitions involve staying in the same state, *ie* no movement at all.

Although the above equations may appear to be rather daunting at first, it should be noted that they can be simplified vastly by considering the transition probability $p_{ij}^{(m,n)}$ as the i,jth entry of a *transition matrix* $P^{(m,n)}$. The above equations can then be written using matrix multiplication as $P^{(m,n)} = P^{(m,l)}P^{(l,n)}$. We discuss this approach in more detail in the next section. First we derive the Chapman-Kolmogorov equations mathematically.

Proof

Students should understand this proof, but they will not be expected to reproduce it in the examination. This is based on the Markov property (2.1) and on the law of total probability in its conditional form.

If $A_1, A_2, \ldots, A_k, \ldots$ form a complete set of disjoint events, *ie*:

$$\bigcup_{k=1}^{\infty} A_k = \Omega, \qquad A_k \cap A_j = \varnothing,\ k \neq j$$

then for any two events B, C :

$$P\left[B\middle|C\right] = \sum_{k=1}^{\infty} P\left[B\middle|C, A_k\right] P\left[A_k\middle|C\right]$$

Now let B be '$X_n = j$', let C be '$X_m = i$', and let A_k be '$X_l = k$'.

Thus:

$$\boldsymbol{P\left[X_n = j \middle| X_m = i\right] = \sum_{k \in S} P\left[X_n = j \middle| X_m = i, X_l = k\right] P\left[X_l = k \middle| X_m = i\right]}$$

$$\boldsymbol{= \sum_{k \in S} P\left[X_n = j \middle| X_l = k\right] P\left[X_l = k \middle| X_m = i\right]}$$

using the Markov property (note $l > m$).

This is the stated result.

The Chapman-Kolmogorov equations allow us to calculate general transition probabilities in terms of the *one-step transition probabilities* $p_{ij}^{(n,n+1)}$.

For example, if we wish to calculate the two-step transition probabilities, we can take the intermediate time $l = m + 1$ and apply the equations. Once we have the two-step transition probabilities we can use them to calculate the three-step transitions and, by iterating the procedure, the transition probabilities of any order can be found.

Hence the distribution of a Markov chain is fully determined once the following are specified:

- **the one-step transition probabilities $p_{ij}^{(n,n+1)}$**
- **the initial probability distribution $q_k = P\left[X_0 = k\right]$.**

Indeed we can deduce from these the probability of any path:

$$\boldsymbol{P\left[X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\right] = q_{i_0}\, p_{i_0 i_1}^{(0,1)}\, p_{i_1, i_2}^{(1,2)} \ldots p_{i_{n-1}, i_n}^{(n-1,n)}}$$

It is therefore convenient, where possible, to determine states in a manner that forms a Markov chain. The model in Section 5.2 illustrates this.

This is referring to the fact that a chain may be given in a form that isn't Markov. In these cases we can't apply the techniques described above to tackle the problem. However, it is sometimes possible to change the state space so that the process *is* given as a Markov chain. For example, this is the case for Model 5.2, which we will meet later in this chapter. Generally, when this can be done, it will simplify the analysis of the problem.

3 Time-homogeneous Markov chains

A simplification occurs if the one-step transition probabilities are time-independent:

$$p_{ij}^{(n,n+1)} = p_{ij} \tag{2.2}$$

In this case, we say that the Markov chain is *time-homogeneous*.

It follows easily from (2.2) that general transition probabilities depend only on time differences:

$$P\left[X_{l+m} = j \,|\, X_m = i\right] = p_{ij}^{(l)} \tag{2.3}$$

This equation defines $p_{ij}^{(l)}$ to be $p_{ij}^{(m,l+m)}$. However, the definition only makes sense if the left-hand side is independent of m.

We refer to (2.3) as the l*-step transition probability*. For time-homogeneous Markov chains, the Chapman-Kolmogorov equations read:

$$p_{ij}^{(n-m)} = \sum_{k \in S} p_{ik}^{(l-m)}\, p_{kj}^{(n-l)} \qquad \text{for } m < l < n$$

This has a very simple interpretation. The *transition matrix* P of a time-homogeneous Markov chain is a square $N \times N$ matrix where N is the number of states in S (possibly infinite), with the elements P_{ij} being the one-step transition probabilities p_{ij}:

$$P_{ij} = p_{ij}$$

The l-step transition probability $p_{ij}^{(l)}$ can be obtained by calculating the entry (i, j) of the l-th power of the matrix P:

$$p_{ij}^{(l)} = \left(P^l\right)_{ij}$$

Recall that the (i, j) th entry in a matrix A is denoted by $(A)_{ij}$ (or just A_{ij}); i refers to the row number, and j to the column. Expressions such as $(A)_{12}$ and $(A)_{31}$ represent numbers and not matrices. Similarly $(A)_{ij}$ is a number, namely the (i, j) th entry of the matrix A.

Recall also that powers are written in the same way as for ordinary numbers. For example, A^2 means AA, as writing two matrices side by side denotes matrix multiplication.

In the same way that we can think of the transition probabilities as the entries of a matrix, we can think of probability distributions $P(X_n = i)$ as the entries of a row vector ($1 \times N$ matrix, N being the number of elements in the state space, as above).

In this chapter we use the following notation:

- the random variable X_0 denotes the state occupied at time 0
- the vector $\underline{X}_0$ gives the probability of being in each state at time 0
- q_k denotes the k th entry of the vector $\underline{X}_0$, *ie* $q_k = P(X_0 = k)$.

For a time-homogeneous Markov chain, we have seen that, $P(X_1 = i) = \sum_{k \in S} q_k p_{ki}$.

The distribution of X_1 can now also be viewed as a row vector, $\underline{X}_1$, with $i th$ entry $\sum_{k \in S} q_k p_{ki}$.

The probabilities $P(X_1 = j)$, $j \in S$ can be expressed in matrix form as:

$$\underline{X}_1 = \underline{X}_0 P$$

This is a shorthand notation for the original equation with summation over indices.

The order in which we multiply numbers doesn't matter. However, the order in which we multiply matrices does matter. In fact, the expression $\underline{X}_1 = P\underline{X}_0$ doesn't even make sense, as we cannot multiply a row vector on the left by an $N \times N$ matrix (unless N happens to equal 1). If instead of a row vector we have a column vector, say the transposed vector $\underline{X}_0^T$, then this order does make sense. For example, the following equation is valid $\underline{X}_1^T = P^T \underline{X}_0^T$.

Question

For a time-inhomogeneous process, the one-step transition matrices are dependent on time and so can be labelled $P^{(n,n+1)}$, where n refers to the time.

(i) Give a matrix equation representing the distribution of the random variable X_5 in terms of the initial distribution and transition matrices.

(ii) Explain how this simplifies for a time-homogeneous chain.

Solution

(i) ***Matrix equation***

With the given notation:

$$\underline{X}_5 = \underline{X}_0 \, P^{(0,1)} P^{(1,2)} P^{(2,3)} P^{(3,4)} P^{(4,5)}$$

(ii) ***Simplification***

In the time-homogeneous case, $P^{(0,1)} = P^{(1,2)} = P^{(2,3)} = P^{(3,4)} = P^{(4,5)} = P$. So $\underline{X}_5 = \underline{X}_0 P^5$.

The normalisation condition $\sum_{j \in S} p_{ij} = 1$ holds for all i, *ie* each row of P must add up to one.

More generally:

$$\sum_{j \in S} p_{ij}^{(l)} = 1 \text{ for all } i$$

It is often revealing to draw the *transition graph* of a Markov chain: this is a diagram in which each state in S is represented as a node of the graph and an arrow is drawn from node i to node j whenever $p_{ij} > 0$, indicating that a direct transition from state i to state j is possible. The value of p_{ij} can be recorded above the arrow.

Question

Consider a Markov process with state space $S = \{0,1,2\}$ and transition matrix, P:

$$P = \begin{pmatrix} p & q & 0 \\ 0.5 & 0 & 0.5 \\ p-0.5 & 0.7 & 0.2 \end{pmatrix}$$

(i) Determine the values of p and q.

(ii) Calculate the transition probabilities $p_{ij}^{(3)}$.

(iii) Draw the transition graph for the process represented by P.

Solution

(i) ***Values of p and q***

Since each row must sum to one we have $p = 0.6$ from the third row and $q = 0.4$ from the first row.

(ii) ***Transition probabilities***

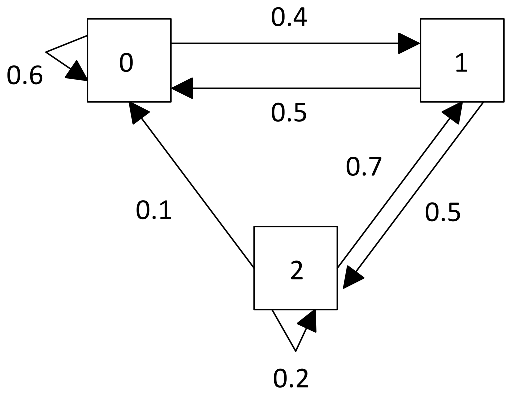
We use the fact that $p_{ij}^{(3)} = (P^3)_{ij}$:

$$P^3 = \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.5 & 0 & 0.5 \\ 0.1 & 0.7 & 0.2 \end{pmatrix}^3 = \begin{pmatrix} 0.56 & 0.24 & 0.20 \\ 0.35 & 0.55 & 0.10 \\ 0.43 & 0.18 & 0.39 \end{pmatrix} \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.5 & 0 & 0.5 \\ 0.1 & 0.7 & 0.2 \end{pmatrix} = \begin{pmatrix} 0.476 & 0.364 & 0.160 \\ 0.495 & 0.210 & 0.295 \\ 0.387 & 0.445 & 0.168 \end{pmatrix}$$

We can work out powers of matrices either way round since $P^2P = PP^2 = P^3$. In fact, for any three matrices A, B and C, we have $A(BC) = (AB)C$.

(iii) ***Transition graph***

The transition graph is:



4 Time-inhomogeneous Markov chains

For a time-inhomogeneous Markov chain, the transition probabilities cannot simply be denoted by p_{ij} because they will depend on the absolute values of time, rather than just the time difference.

For a time-inhomogeneous chain, the one-step transition probabilities are denoted by $p_{ij}^{(n,n+1)}$.

The value of 'time' can be represented by many factors, for example the time of year, age or duration.

So for a time-inhomogeneous Markov chain, the probability of going from state i at time 0 to state j at time n is not necessarily the same as going from state i at time m ($m \neq 0$) to state j at time $m+n$, even though both time intervals are of length n . For a time-inhomogeneous chain, the transition probabilities depend not only on the length of the time interval, but also on when it starts.

Question

The stochastic process X is defined as follows:

$$X_0 = 0$$

$$X_n = X_{n-1} + Y_n , \qquad n = 1,2,3,...$$

where Y_n can only take the values 0 and 1, and the corresponding probabilities are:

$$P(Y_1 = 1) = \frac{1}{4}$$

$$P(Y_{n+1} = 1) = \frac{1}{6}\left(1 + \frac{X_n}{n}\right), \; n = 1,2,3,...$$

Draw the transition diagram for X_n covering all transitions that could happen in the first three time periods, including transition probabilities.

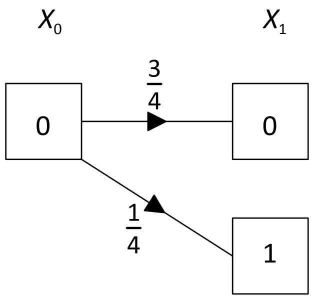
Solution

The possible values of X_1 are 0 and 1, and the corresponding probabilities are:

$$P(X_1 = 1) = P(Y_1 = 1) = \frac{1}{4}$$

$$P(X_1 = 0) = 1 - \frac{1}{4} = \frac{3}{4}$$

So the transition diagram for the first time period is:

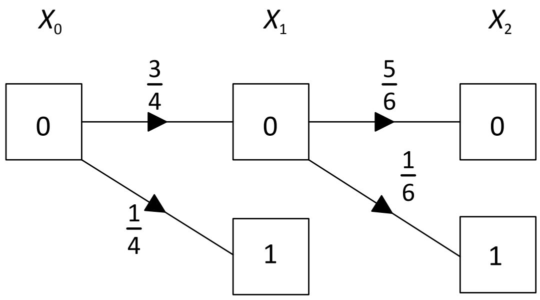


If the process is in state 0 at time 1, it may be in state 0 or state 1 at time 2. The corresponding probabilities are:

$$P(X_2=1|X_1=0)=P(Y_2=1|X_1=0)=\frac{1}{6}\left(1+\frac{0}{1}\right)=\frac{1}{6}$$

$$P(X_2=0|X_1=0)=1-\frac{1}{6}=\frac{5}{6}$$

Adding these to the diagram gives:

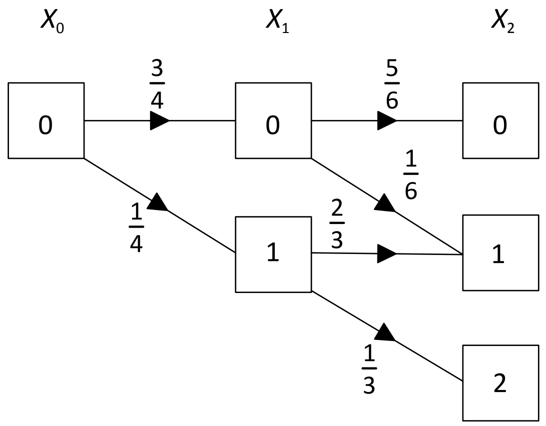


Similarly, if the process in state 1 at time 1, it may be in state 1 or state 2 at time 2. The corresponding probabilities are:

$$P(X_2=2|X_1=1)=P(Y_2=1|X_1=1)=\frac{1}{6}\left(1+\frac{1}{1}\right)=\frac{1}{3}$$

$$P(X_2=1|X_1=1)=1-\frac{1}{3}=\frac{2}{3}$$

So the transition diagram covering all the transitions that could happen in the first two time periods is:



We can consider the third time period in a similar way.

If the process is in state 0 at time 2, it may be in state 0 or state 1 at time 3. The corresponding probabilities are:

$$P(X_3 = 1 \mid X_2 = 0) = P(Y_3 = 1 \mid X_2 = 0) = \frac{1}{6}\left(1 + \frac{0}{2}\right) = \frac{1}{6}$$

$$P(X_3 = 0 \mid X_2 = 0) = 1 - \frac{1}{6} = \frac{5}{6}$$

If the process is in state 1 at time 2, it may be in state 1 or state 2 at time 3. The corresponding probabilities are:

$$P(X_3 = 2 \mid X_2 = 1) = P(Y_3 = 1 \mid X_2 = 1) = \frac{1}{6}\left(1 + \frac{1}{2}\right) = \frac{1}{4}$$

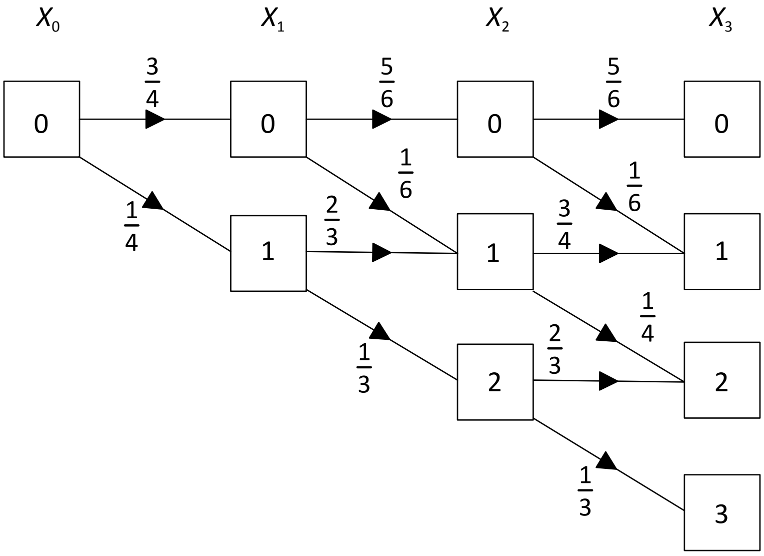
$$P(X_3 = 1 \mid X_2 = 1) = 1 - \frac{1}{4} = \frac{3}{4}$$

If the process is in state 2 at time 2, it may be in state 2 or state 3 at time 3. The corresponding probabilities are:

$$P(X_3 = 3 \mid X_2 = 2) = P(Y_3 = 1 \mid X_2 = 2) = \frac{1}{6}\left(1 + \frac{2}{2}\right) = \frac{1}{3}$$

$$P(X_3 = 2 \mid X_2 = 2) = 1 - \frac{1}{3} = \frac{2}{3}$$

Adding these to the diagram, it becomes:



Looking at the defining equation for X_n in the question immediately above, *ie* $X_n = X_{n-1} + Y_n$, we see that the process X has the Markov property. In contrast, the process Y does not have the Markov property as the probability distribution for Y_{n+1} depends on X_n, which itself depends on all of $Y_1, Y_2, \dots, Y_n$.

The defining equation for X_n is similar to that for a general random walk, which we defined in Section 3.2 of Chapter 1. X is *not* a random walk, however, as the random variables $Y_1, Y_2, Y_3, \dots$ are not identically distributed.

5 Models

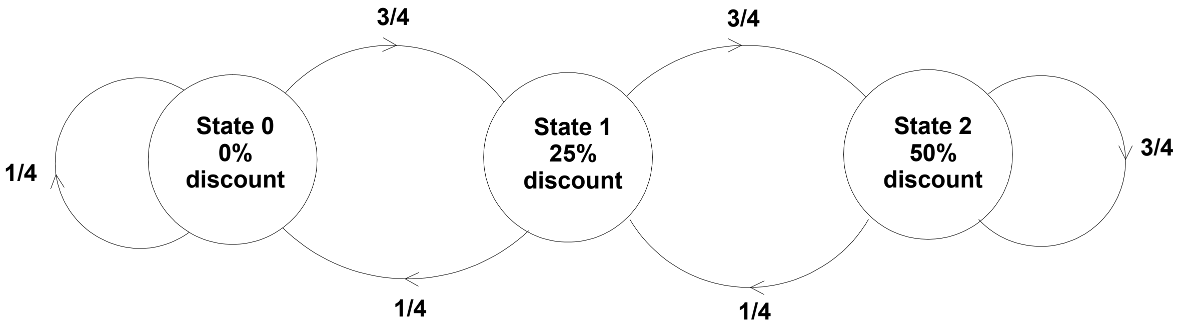
Throughout the rest of this chapter we will refer to the models in the following sections by their section number. For example, Model 5.1 will mean the model in Section 5.1.

5.1 A simple model of a No Claims Discount (NCD) policy

The No Claims Discount system (NCD) in motor insurance, whereby the premium charged depends on the driver's claim record, is a prime application of Markov chains. We present two simple models and we suggest various possible improvements.

A motor insurance company grants its customers either no discount (state 0) or 25% discount (state 1) or 50% discount (state 2). A claim-free year results in a transition to the next higher state the following year (or in the retention of the maximum discount); similarly, a year with one claim or more causes a transition to the next lower state (or the retention of the zero discount status).

Under these rules, the discount status of a policyholder forms a Markov chain with state space $S=\{0,1,2\}$; if the probability of a claim-free year is ¾, the transition graph and transition matrix are:



The transition matrix is given by:

$$P=\begin{pmatrix} \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

The probability of holding the maximum discount in year $n+3$ given that you do not qualify for any discount in year n is:

$$p_{0,2}^{(3)}=\left(P^3\right)_{1,3}=\tfrac{9}{16}$$

$p_{0,2}^{(3)}$ denotes the probability of going from state 0 to state 2 in 3 time steps. This is equal to $\left(P^3\right)_{1,3}$, the entry in the first row and the third column of the matrix P^3 (the cube of the matrix P). So we could calculate P^3 and identify the $(1,3)th$ entry. This turns out to be $\tfrac{9}{16}$. However, calculating the cube of P is time-consuming and is not actually necessary given that we only need one probability and there aren't that many possible paths.

One alternative approach is the following. Given that the policyholder does not qualify for any discount in year n, we must be starting in state 0. We want the probability of ending in state 2 after three (one-step) transitions. This can only happen if our path is:

$$0 \to 0 \to 1 \to 2 \text{ with probability } 1/4 \times 3/4 \times 3/4$$

or:

$$0 \to 1 \to 2 \to 2 \text{ with probability } 3/4 \times 3/4 \times 3/4$$

The sum of these gives the result $9/16$. So we see that the matrix representation is a notation for the way we have always calculated probabilities.

However, if there are many possible paths, the above approach can be tedious. Probably the most efficient way to proceed with a problem like this is to work as follows.

Since we know that the distribution at time n is $(1,0,0)$, we can calculate the probability distribution at time $n+1$ by post-multiplying the vector $(1,0,0)$ by the transition matrix P. This gives:

$$\left(\frac{1}{4},\frac{3}{4},0\right)$$

Then post-multiplying $\left(\frac{1}{4},\frac{3}{4},0\right)$ by the transition matrix P, we obtain the probability distribution at time $n+2$, which is:

$$\left(\frac{1}{4},\frac{3}{16},\frac{9}{16}\right)$$

Post-multiplying this by P gives us the probability distribution at time $n+3$. However, since we just need the probability of being on maximum discount in year $n+3$, we only have to multiply the vector $\left(\frac{1}{4},\frac{3}{16},\frac{9}{16}\right)$ by the last column of P. We find that the probability distribution at time $n+3$ is of the form:

$$\left(*,*,\frac{9}{16}\right)$$

So the required probability is $\frac{9}{16}$.

Question

Calculate the probability in the above model of starting with a discount level of 25% and ending up 4 years later at the same level.

Solution

Repeated post-multiplication of the vector $(0,1,0)$ by the transition matrix P gives:

$$
\begin{aligned}
(0,1,0) &\rightarrow \left(\frac{1}{4},0,\frac{3}{4}\right) \\
&\rightarrow \left(\frac{1}{16},\frac{3}{8},\frac{9}{16}\right) \\
&\rightarrow \left(\frac{7}{64},\frac{3}{16},\frac{45}{64}\right) \\
&\rightarrow \left(*,\frac{33}{128},*\right)
\end{aligned}
$$

So the required probability is $\frac{33}{128}$.

Alternatively, we can consider each possible sample path separately and sum the probabilities as follows. The only paths are:

$1\rightarrow 0\rightarrow 0\rightarrow 0\rightarrow 1$ with probability $\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{3}{4}=\frac{3}{256}$

$1\rightarrow 0\rightarrow 1\rightarrow 0\rightarrow 1$ with probability $\frac{1}{4}\times\frac{3}{4}\times\frac{1}{4}\times\frac{3}{4}=\frac{9}{256}$

$1\rightarrow 0\rightarrow 1\rightarrow 2\rightarrow 1$ with probability $\frac{1}{4}\times\frac{3}{4}\times\frac{3}{4}\times\frac{1}{4}=\frac{9}{256}$

$1\rightarrow 2\rightarrow 1\rightarrow 0\rightarrow 1$ with probability $\frac{3}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{3}{4}=\frac{9}{256}$

$1\rightarrow 2\rightarrow 1\rightarrow 2\rightarrow 1$ with probability $\frac{3}{4}\times\frac{1}{4}\times\frac{3}{4}\times\frac{1}{4}=\frac{9}{256}$

$1\rightarrow 2\rightarrow 2\rightarrow 2\rightarrow 1$ with probability $\frac{3}{4}\times\frac{3}{4}\times\frac{3}{4}\times\frac{1}{4}=\frac{27}{256}$

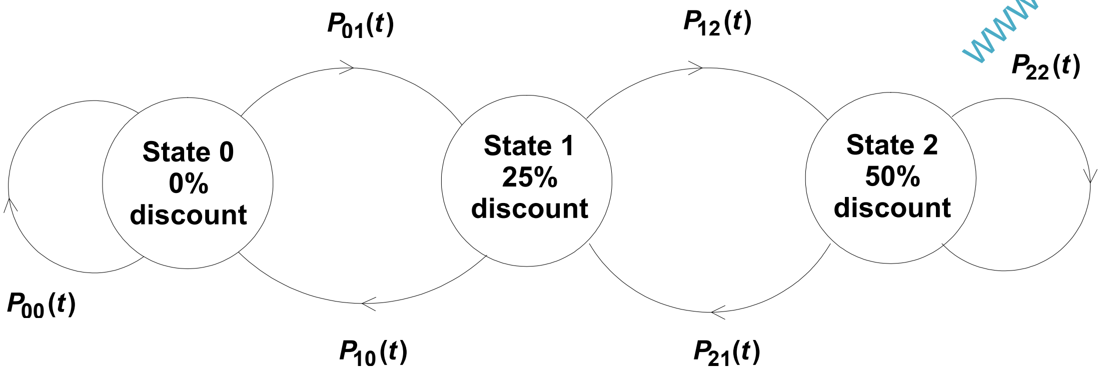
Adding up probabilities for each path we get $\frac{66}{256}=\frac{33}{128}$ as above.

Time-inhomogeneous model

For a time-inhomogeneous case of this model, the probability of an accident would be time-dependent to reflect changes in traffic conditions. This could be due to general annual trends in the density of traffic and/or propensity to claim.

The probability of an accident may also be affected by changes in weather conditions, which is another possible reason for using a time-inhomogeneous model.

The transition graph and matrix would then become:



and:

$$\boldsymbol{P(t)} = \begin{pmatrix} P_{00}(t) & P_{01}(t) & P_{02}(t) \\ P_{10}(t) & P_{11}(t) & P_{12}(t) \\ P_{20}(t) & P_{21}(t) & P_{22}(t) \end{pmatrix}$$

In fact, the transition matrix simplifies to:

$$P(t) = \begin{pmatrix} P_{00}(t) & P_{01}(t) & 0 \\ P_{10}(t) & 0 & P_{12}(t) \\ 0 & P_{21}(t) & P_{22}(t) \end{pmatrix}$$

5.2 Another model of an NCD policy

In the model in Section 5.1 there are 3 discount levels, namely 0%, 25% and 50%.

Modify the previous model as follows: there are now four levels of discount:

0 :	**no discount**
1 :	**25% discount**
2 :	**40% discount**
3 :	**60% discount**

The rules for moving up the discount scale are as before, but in the case of a claim during the current year, the discount status moves down one or two steps (if this is possible) according to whether or not the previous year was claim-free.

So the discount level next year depends on the number of claims during this year *and* last year. If there is a claim this year, then we move down one level if last year was claim-free and two levels if there was a claim last year.

Under these rules, the discount status X_n of a policyholder does *not* form a Markov chain on $S=\{0,1,2,3\}$ because:

$$P\left[X_{n+1}=0 \middle| X_n=2, X_{n-1}=1\right]=0$$

whereas:

$$P\left[X_{n+1}=0 \middle| X_n=2, X_{n-1}=3\right]>0$$

$P\left[X_{n+1}=0 \middle| X_n=2, X_{n-1}=1\right]$ is the probability of a policyholder being on 0% discount in year $n+1$ given that they were on 40% discount in year n *and* 25% discount in year $n-1$. This probability is zero since there was no claim in year $n-1$. (If there had been, the policyholder wouldn't have moved up to 40% discount.)

The Core Reading equations above show that the future value of the process depends not only on its current value, but also on the past. Earlier we commented that it is sometimes possible to transform a chain that isn't Markov into one that is. We can do this here by altering the state space.

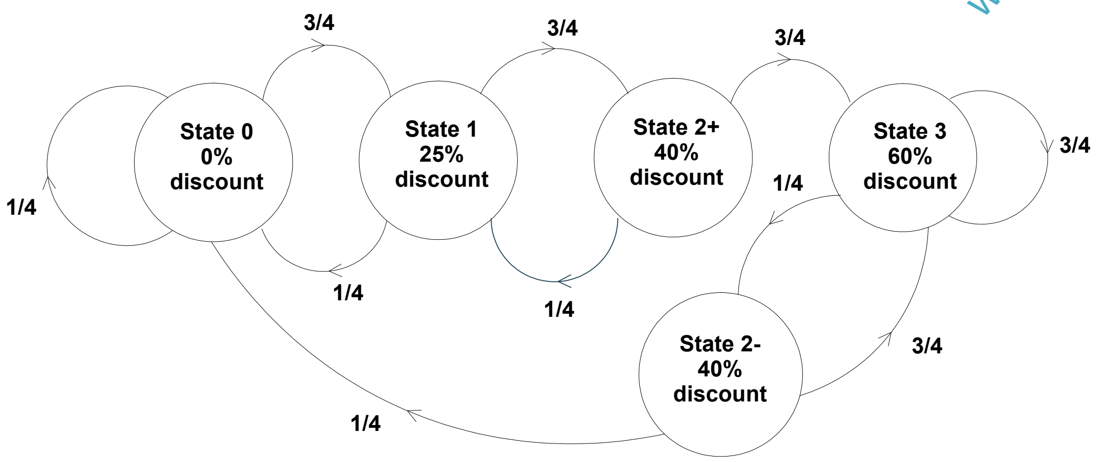
To construct a Markov chain $\{Y_n, n=0, 1, 2, \ldots\}$, one needs to incorporate some information on the previous year into the state; in fact this is necessary only for state 2, which we split as:

2^+ :	**40% discount and no claim in the previous year**
2^- :	**40% discount and claim in the previous year**

We can see why it's only state 2 that needs to be split, by reasoning as follows:

- If the policyholder is on 0% discount this year and makes a claim, the discount level next year will be 0%.
- If the policyholder is on 25% discount this year and makes a claim, the discount level next year will be 0%.
- If the policyholder is on 40% discount this year and makes a claim, the discount level next year will be either 0% or 25%, depending on whether the policyholder claimed last year.
- If the policyholder is on 60% discount this year, last year must have been claim-free. So, if a claim is made this year, the discount level next year will 40%.

Assuming as before a probability of ¾ of no claim in any given year, we have a Markov chain on the state space $S' = \{0, 1, 2^+, 2^-, 3\}$ with transition graph:



and transition matrix:

$$P = \begin{pmatrix} \tfrac{1}{4} & \tfrac{3}{4} & 0 & 0 & 0 \\ \tfrac{1}{4} & 0 & \tfrac{3}{4} & 0 & 0 \\ 0 & \tfrac{1}{4} & 0 & 0 & \tfrac{3}{4} \\ \tfrac{1}{4} & 0 & 0 & 0 & \tfrac{3}{4} \\ 0 & 0 & 0 & \tfrac{1}{4} & \tfrac{3}{4} \end{pmatrix}$$

The probability of being at the 60% discount level in year $n+3$ given that you hold 25% in year n is:

$$p_{1,3}^{(3)} = \left(P^3\right)_{2,5} = {}^{27}\!/_{64}$$

Question

A policyholder starts at 0% discount. Calculate the probability that this policyholder is on the maximum level of discount after 5 years.

Solution

The initial distribution is $(1,0,0,0,0)$. Successive post-multiplication by the transition matrix P shown above gives:

$$(1,0,0,0,0) \to \left(\frac{1}{4},\frac{3}{4},0,0,0\right) \to \left(\frac{1}{4},\frac{3}{16},\frac{9}{16},0,0\right)$$

$$\to \left(\frac{7}{64},\frac{21}{64},\frac{9}{64},0,\frac{27}{64}\right) \to \left(\frac{7}{64},\frac{15}{128},\frac{63}{256},\frac{27}{256},\frac{27}{64}\right)$$

$$\to \left(*,*,*,*,\frac{297}{512}\right)$$

So the required probability is $\dfrac{297}{512}$.

Alternatively, we could calculate the (1,5)th entry in the matrix P^5. In order to do this, we first calculate the two matrices P^2 and P^3. We have:

$$P^2 = \frac{1}{16}\begin{pmatrix} 1 & 3 & 0 & 0 & 0 \\ 1 & 0 & 3 & 0 & 0 \\ 0 & 1 & 0 & 0 & 3 \\ 1 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix}^2 = \frac{1}{16}\begin{pmatrix} 4 & 3 & 9 & 0 & 0 \\ 1 & 6 & 0 & 0 & 9 \\ 1 & 0 & 3 & 3 & 9 \\ 1 & 3 & 0 & 3 & 9 \\ 1 & 0 & 0 & 3 & 12 \end{pmatrix}$$

Also:

$$P^3 = \frac{1}{64}\begin{pmatrix} 1 & 3 & 0 & 0 & 0 \\ 1 & 0 & 3 & 0 & 0 \\ 0 & 1 & 0 & 0 & 3 \\ 1 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix}\begin{pmatrix} 4 & 3 & 9 & 0 & 0 \\ 1 & 6 & 0 & 0 & 9 \\ 1 & 0 & 3 & 3 & 9 \\ 1 & 3 & 0 & 3 & 9 \\ 1 & 0 & 0 & 3 & 12 \end{pmatrix} = \frac{1}{64}\begin{pmatrix} 7 & 21 & 9 & 0 & 27 \\ 7 & 3 & 18 & 9 & 27 \\ 4 & 6 & 0 & 9 & 45 \\ 7 & 3 & 9 & 9 & 36 \\ 4 & 3 & 0 & 12 & 45 \end{pmatrix}$$

The (1,5)th entry is of P^5 is:

$$\sum_{k=1}^{5}\left(P^3\right)_{1k}\left(P^2\right)_{k5} = \frac{1}{64\times 16}\left(7\times 0 + 21\times 9 + 9\times 9 + 0\times 9 + 27\times 12\right) = \frac{594}{1{,}024} = \frac{297}{512}$$

Time-inhomogeneous model

This basic model is amenable to numerous improvements. For instance the accident probability can be made to depend on the discount status to reflect the influence of the latter on driver care. Also the accident probability can be time-dependent (leading to a time-inhomogeneous chain) to reflect changes in traffic conditions (as described in Section 5.1).

5.3 Simple random walk on $S = \{\ldots -2, -1, 0, 1, 2, \ldots\}$

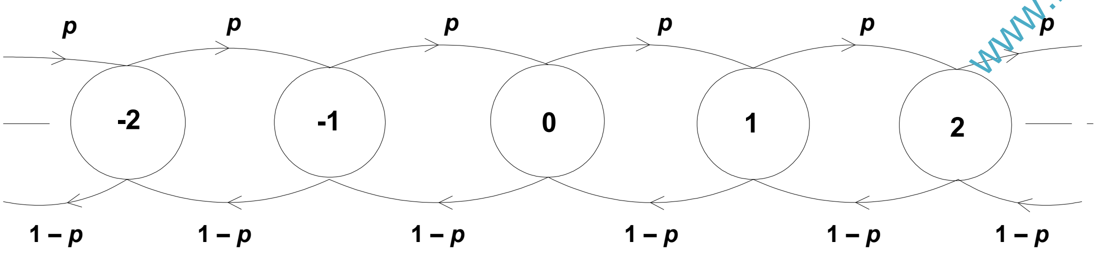
This is defined as $X_n = Y_1 + Y_2 + \ldots + Y_n$ where the random variables Y_j (the steps of the walk) are mutually independent with the common probability distribution:

$$P\left[Y_j = 1\right] = p, \qquad P\left[Y_j = -1\right] = 1 - p$$

The Markov property holds because the process has independent increments.

The transition graph and the transition matrix are infinite.

The transition diagram is:



The transition matrix is:

$$P = \begin{bmatrix} \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \ddots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 1-p & 0 & p & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & 1-p & 0 & p & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \ddots & \ddots & \ddots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & 1-p & 0 & p & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & 1-p & 0 & p & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \ddots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \ddots \end{bmatrix}$$

In order to get from i to j in n steps, the random walk must make $u = \frac{1}{2}(n + j - i)$ steps in an upward direction and $n - u$ in a downward direction. Since the distribution of the number of upward jumps in n steps is binomial with parameters n and p, the n-step transition probabilities can be calculated as:

$$p_{ij}^{(n)} = \begin{cases} \dbinom{n}{u} p^u (1-p)^{n-u} & \text{if } 0 \le n+j-i \le 2n \text{ and } n+j-i \text{ is even} \\ 0 & \text{otherwise} \end{cases}$$

Question

Prove that the n-step transition probabilities for a simple random walk on the integers are given by the formula above.

Solution

Consider going from i to j in n steps. Let the number of upward steps be u and the number of downward steps be d. Since we make n steps in total, we must have:

$$u + d = n$$

Also, since the net upward movement must equal the excess of upward steps over downward steps, we have:

$$u - d = j - i$$

Solving these simultaneous equations, we see that:

$$u = ½(n + j - i) \quad \text{and} \quad d = n - u = n - ½(n + j - i) = ½(n - j + i)$$

Since u must be a non-negative integer, it must be the case that $n + j - i$ is a non-negative *even* number.

We also know that $j - i \le n$. So:

$$0 \le n + j - i \le 2n$$

The order in which the upward and downward steps occur doesn't matter. There are $\binom{n}{u} = \binom{n}{½(n+j-i)}$ ways of choosing where the u upward steps occur in the sequence of n steps. Each upward step occurs with probability p.

Putting all this together we have:

$$p_{ij}^{(n)} = \begin{cases} \binom{n}{u} p^u (1-p)^d & \text{if } 0 \le n + j - i \le 2n \text{ and } n + j - i \text{ is even} \\ 0 & \text{otherwise} \end{cases}$$

Note that, in addition to being time-homogeneous, a simple random walk is also *space-homogeneous*:

$$\boldsymbol{p_{i\,j}^{(n)} = p_{i+r\;j+r}^{(n)}}$$

This means that only the time taken and the overall distance travelled (including minus sign if necessary) affect the transition probability. Exactly when and where they occur doesn't matter. So, for example, the probability of going from state 4 at time 4 to state -1 at time 11, is the same as the probability of going from state 8 at time 3 to state 3 at time 10. In both cases we move 5 steps to the left in a time of 7 units.

In the special case when $p = 1 - p = ½$ (and the initial state is 0), this process is a simple symmetric random walk.

5.4 Simple random walk on $\{0, 1, 2, \ldots, b\}$

This is similar to the previous model, except that *boundary conditions* have to be specified at 0 and b; these will depend on the interpretation given to the chain.

We are also relaxing the assumption that the process starts at 0.

Commonly used boundary conditions include:

Reflecting boundary: $P\left[X_{n+1}=1 \mid X_n=0\right]=1$

Absorbing boundary: $P\left[X_{n+1}=0 \mid X_n=0\right]=1$

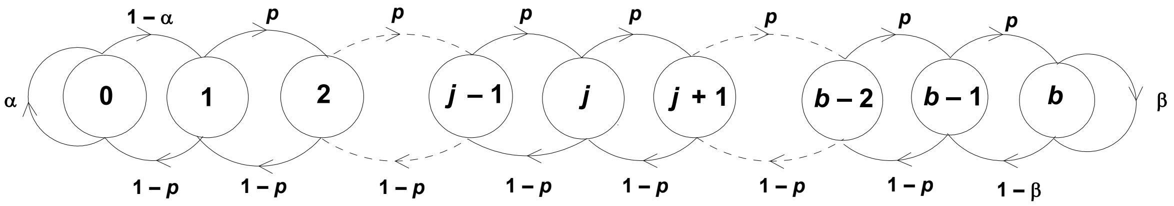
Mixed boundary: $\begin{cases} P\left[X_{n+1}=0 \mid X_n=0\right]=\alpha \\ P\left[X_{n+1}=1 \mid X_n=0\right]=1-\alpha \end{cases}$

One way of viewing a random walk is to picture a particle randomly moving from place to place.

- If the particle is absorbed by a state then the probability of it moving to another state is 0 and hence the terminology 'absorbing state.'
- Similarly, if the particle is fully reflected or bounces back out with probability 1, then the state is called a reflecting state.
- The third type of boundary condition is a mixture of the previous two. There is some chance the particle will be reflected and some that it will stay put. If there is a non-zero chance of being reflected, then this will eventually happen.

A random walk with absorbing boundary conditions at both 0 and b can be used (for example) **to describe the wealth of a gambler who will continue to gamble until either his fortune reaches a target b or his fortune hits 0 and he is ruined; in either case, reaching the boundary means staying there forever.**

In the general case, with mixed boundary conditions, the transition graph is:



and the transition matrix is:

$$P=\left[\begin{array}{ccccccccc}
0 & 1 & 2 & \cdots & \cdots & \cdots & b-2 & b-1 & b \\
\alpha & 1-\alpha & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
1-p & 0 & p & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & 1-p & 0 & p & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & 1-p & 0 & p & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \ddots & \ddots & \ddots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & 1-p & 0 & p & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & 1-p & 0 & p & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1-p & 0 & p \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1-\beta & \beta
\end{array}\right]$$

Reflecting and absorbing boundary conditions are obtained as special cases, taking α, β equal to 0 or 1.

The simple NCD model of Section 5.1 is another practical example of a bounded random walk.

Question

State the boundary conditions for the NCD model of Section 5.1.

Solution

In Model 5.1 the boundary conditions are:

$$P\left[X_{n+1}=0 \mid X_n=0\right]=\frac{1}{4}$$

$$P\left[X_{n+1}=1 \mid X_n=0\right]=\frac{3}{4}$$

$$P\left[X_{n+1}=2 \mid X_n=2\right]=\frac{3}{4}$$

$$P\left[X_{n+1}=1 \mid X_n=2\right]=\frac{1}{4}$$

5.5 A model of accident proneness

For a given driver, any period j is either accident free ($Y_j=0$) or gives rise to exactly one accident ($Y_j=1$).

The possibility of more than one accident in any time period is ignored for simplicity.

The probability of an accident in the next period is estimated using the driver's past record as follows (all variables y_j are either 0 or 1):

$$P\left[Y_{n+1}=1 \middle| Y_1=y_1, Y_2=y_2, \ldots, Y_n=y_n\right]=\frac{f\left(y_1+y_2+\ldots+y_n\right)}{g(n)}$$

where f and g are two given increasing functions satisfying $0 \leq f(m) \leq g(m)$. Of course:

$$P\left[Y_{n+1}=0 \middle| Y_1=y_1, Y_2=y_2, \ldots, Y_n=y_n\right]=1-\frac{f\left(y_1+y_2+\ldots+y_n\right)}{g(n)}$$

Question

(i) Explain why the functions f and g have to be increasing functions.

(ii) Explain why the functions f and g must satisfy the given inequalities.

(iii) Interpret this model in the case where $f(m)=m$ and $g(n)=n$.

Solution

(i) ***Why the functions are increasing***

$\dfrac{f(y_1+y_2+\dots+y_n)}{g(n)}$ is the probability of a driver having an accident in the next time period. We would expect this to be higher for a driver who has had more accidents in the last n time periods. In other words, for fixed n, we expect the probability to be higher for larger $y_1+y_2+\dots+y_n$. So f is an increasing function.

On the other hand, if two drivers have the same number of accidents, but for one driver these accidents occurred within a shorter time period, then we would expect this driver to have a higher probability of having another accident. In other words, if n is smaller for a fixed value of $y_1+y_2+\dots+y_n$ then we expect the probability to be higher. In turn this says that g must be smaller. Thus g must also be an increasing function.

(ii) ***Inequalities***

We assume that both f and g are positive. If one of these were negative, the other would have to be negative too, since their ratio is positive. In addition, they would then both have to be decreasing. We can ignore this possibility.

In order that $\dfrac{f(y_1+y_2+\dots+y_n)}{g(n)}$ is a probability we must have:

$$0 \le \frac{f(y_1+y_2+\dots+y_n)}{g(n)} \le 1$$

The above inequality is equivalent to:

$$0 \le f(y_1+y_2+\dots+y_n) \le g(n)$$

In particular, for the 'maximum' case when all the y's are equal to 1, we have $y_1+y_2+\dots+y_n=n$, and we obtain $0\le f(n)\le g(n)$ as required.

(iii) ***Special case***

In this case:

$$\frac{f(y_1+y_2+\dots+y_n)}{g(n)} = \frac{1}{n}(y_1+y_2+\dots+y_n)$$

So we're estimating the probability of a claim next year using the average number of years that had a claim in the past.

The dependence on the past record means that $Y_1, Y_2, \ldots, Y_n, \ldots$ does *not* have the Markov property (it depends on all previous values of Y_j).

Consider, however, the cumulative number of accidents suffered by the driver:

$$X_n = \sum_{j=1}^{n} Y_j$$

This is a Markov chain with state space $S = \{0, 1, 2, \ldots\}$.

It possesses the Markov property because:

$$P\left[X_{n+1} = 1 + x_n \middle| X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right]$$

$$= P\left[X_{n+1} = 1 + x_n \middle| Y_1 = x_1, Y_2 = x_2 - x_1, \ldots, Y_n = x_n - x_{n-1}\right]$$

Since $\sum_{j=1}^{n} Y_j = X_n$, the condition $Y_1 = x_1, Y_2 = x_2 - x_1, \ldots, Y_n = x_n - x_{n-1}$ is a function only of X_n and hence:

$$P\left[X_{n+1} = 1 + x_n \middle| X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right]$$

$$= P\left[X_{n+1} = 1 + x_n \middle| X_n = x_n\right] = \frac{f(x_n)}{g(n)}$$

This is independent of the values of $X_1, X_2, \ldots, X_{n-1}$. So the process $\{X_n\}$ has the Markov property.

Note that the chain is only time-homogeneous if $g(n)$ is constant.

However this is neither very realistic nor useful.

As an example, consider the following two drivers. The first is a 45-year old who has had two accidents in the last 20 years of motoring (both in the last year), and the other is an 18-year old who has had two accidents in the last year. For any time-homogeneous model as described above, with time period of one month say, the probabilities of the two drivers having an accident next month would be the same. A more meaningful model should take into account the length of time over which previous accidents have occurred. In the above example, this is the length of time a person has been driving. So we should use a time-inhomogeneous model here.

Question

For time periods of one year, let $f(x_n) = 0.5 + x_n$ and $g(n) = 1 + n$ so that:

$$P(Y_{n+1} = 1 | X_n = x_n) = \frac{0.5 + x_n}{1 + n}$$

Determine:

(i) the probability that a driver has an accident in the second year, given that they did not have an accident in the first year

(ii) the probability that a driver has an accident in the 11th year, given that they had an accident in each of the first 10 years

(iii) the (i,j) th entry in the one-step transition matrix $P^{(n,n+1)}$ of the Markov chain X_n.

Solution

(i) $P(Y_2=1\,|\,X_1=0)=\dfrac{0.5+0}{2}=0.25$

(ii) $P(Y_{11}=1\,|\,X_{10}=10)=\dfrac{0.5+10}{11}=\dfrac{10.5}{11}=0.955$

(iii) $$\left(P^{(n,n+1)}\right)_{ij}=P(X_{n+1}=j\,|\,X_n=i)=\begin{cases}\dfrac{0.5+i}{1+n} & \text{if } j=i+1\\ \dfrac{0.5+n-i}{1+n} & \text{if } j=i\\ 0 & \text{otherwise}\end{cases}$$

6 The long-term distribution of a Markov chain

6.1 The stationary probability distribution

We say that π_j, $j \in S$ is a *stationary probability distribution* for a Markov chain with transition matrix P if the following conditions hold for all j in S:

- $$\pi_j = \sum_{i \in S} \pi_i p_{ij} \qquad \textbf{(2.4)}$$
- $$\pi_j \geq 0$$
- $$\sum_{j \in S} \pi_j = 1$$

Note how (2.4) can be stated in the compact form $\pi = \pi P$ where π is viewed as a row vector.

The interpretation of (2.4) is that, *if we take π as our initial probability distribution*, that is to say $P[X_0 = i] = \pi_i$, then the distribution at time 1 is again given by π:

$$P[X_1 = j] = \sum_{i \in S} P\left[X_1 = j \middle| X_0 = i\right] P[X_0 = i] = \sum_{i \in S} p_{ij} \pi_i = \pi_j$$

The same is true at all times $n \geq 1$, so that π is an *invariant* probability distribution; in fact the chain is then a *stationary process* in the sense of Chapter 1.

So if the chain ever reaches the distribution π at some time n, *ie* $P(X_n = i) = \pi_i$ for all values of i, then (because the transition matrix sends π back to itself, *ie* $\pi = \pi P$) the distribution of X_t will be π for all subsequent times $t \geq n$. The statistical properties of the process do not change over time, so the chain is a stationary process.

In general a Markov chain need not have a stationary probability distribution, and if it exists it need not be unique. For instance no stationary probability distribution exists for Model 5.3, whereas in Model 5.4 uniqueness depends on the values of α, β. When the state space S is *finite*, the situation is simple.

Stationary distribution result (1)

A Markov chain with a finite state space has at least one stationary probability distribution.

The proof of this result is beyond the syllabus.

As an example, we will compute the stationary probability for NCD Model 5.2. The equations (2.4) read:

$$\begin{aligned} \pi_0 &= \tfrac{1}{4}\pi_0 + \tfrac{1}{4}\pi_1 + \tfrac{1}{4}\pi_{2-} \\ \pi_1 &= \tfrac{3}{4}\pi_0 + \tfrac{1}{4}\pi_{2+} \\ \pi_{2+} &= \tfrac{3}{4}\pi_1 \\ \pi_{2-} &= \tfrac{1}{4}\pi_3 \\ \pi_3 &= \tfrac{3}{4}\pi_{2+} + \tfrac{3}{4}\pi_{2-} + \tfrac{3}{4}\pi_3 \end{aligned} \qquad \textbf{(2.5)}$$

The coefficients in these equations correspond to the columns in the matrix P.

This linear system is not linearly independent since adding up all the equations results in an identity (this is a general feature of equations $\pi = \pi P$ due to the property $\sum_{j \in S} p_{ij} = 1$).

Because of this we can discard any one of the equations, say the last one.

To say that the above equations are not linearly independent means that any four of them will always rearrange to give the remaining one, which is therefore redundant. This will always be true for equations of the form $\pi = \pi P$ if P is a matrix whose rows sum to 1. As a result we may discard one of them – it doesn't matter which – and solve the remaining system. We usually discard the most complicated looking one.

Note also that by linearity, any multiple of a solution of (2.5) is again a solution; uniqueness comes only as a result of the normalisation $\sum_{j \in S} \pi_j = 1$. For this reason, it is good practice to solve for the components of π in terms of one of them (say π_1 here), which we will refer to as the working variable. The value of the working variable is determined at the last step by normalisation.

Once the value of the working variable has been established, the others can be deduced as well.

Although uniqueness comes only as a result of the normalisation $\sum_{j \in S} \pi_j = 1$, this does not mean that uniqueness has to come at all. It might be the case that even after applying the normalisation condition the solution is not unique. In addition, if the state space is not finite, then there may not be a stationary distribution at all.

We now summarise the method and apply it to the above example.

Step 1: Discard one of the equations. Here the first or the last one are obvious choices; delete the final one, say.

Step 2: Select one of the π_j's as working variable. Here $\pi_1, \pi_{2+}, \pi_{2-}$ or π_3 are reasonable choices; choose π_1.

Step 3: Rewrite remaining equations in terms of the working variable.

$$3\pi_0 - \pi_{2-} = \pi_1 \qquad \text{(a)}$$

$$3\pi_0 + \pi_{2+} = 4\pi_1 \qquad \text{(b)}$$

$$\pi_{2+} = {}^3\!/_4\,\pi_1 \qquad \text{(c)}$$

$$4\pi_{2-} - \pi_3 = 0 \qquad \text{(d)}$$

Step 4: Solve the equations in terms of the working variable.

In general we might do this by Gaussian elimination.

Gaussian elimination is a general method for solving a system of linear equations.

However, here the equations are so simple that the solution can be read off if we take them in the right order:

$$\pi_{2+} = {}^3\!/_4\,\pi_1$$

We get this directly from (c). Substituting this into (b) then gives:

$$\pi_0 = \frac{\pi_1}{3}\left(4 - {}^3\!/_4\right) = {}^{13}\!/_{12}\,\pi_1$$

Now substitute for π_0 in (a) to get:

$$\pi_{2-} = \pi_1\left(-1 + {}^{13}\!/_4\right) = {}^9\!/_4\,\pi_1$$

Finally:

$$\pi_3 = 9\pi_1$$

from (d).

Step 5: Solve for the working variable.

We have:

$$\pi = \pi_1\left({}^{13}\!/_{12}, 1, {}^3\!/_4, {}^9\!/_4, 9\right)$$

and:

$$\sum_j \pi_j = \frac{\pi_1}{12}\left(13 + 12 + 9 + 27 + 108\right) = {}^{169}\!/_{12}\,\pi_1 = 1$$

So:

$$\pi_1 = {}^{12}\!/_{169}$$

Step 6: Combine the results of the last two steps to obtain the solution.

$$\pi = \left(\frac{13}{169}, \frac{12}{169}, \frac{9}{169}, \frac{27}{169}, \frac{108}{169}\right)$$

It is good practice to use the equation discarded earlier to verify that the calculated solution is correct.

In the above example it has turned out that there is only one solution, but this won't always be the case. A sufficient condition, though not a necessary one, is given below. This requires the introduction of a further classification of Markov chains into those that are irreducible and those that are not.

The question of *uniqueness* of the stationary distribution is more delicate than existence; we shall consider only *irreducible* chains.

Irreducibility

A Markov chain is said to be irreducible if any state j can be reached from any other state i. In other words, a chain is irreducible if, given any pair of states i, j there exists an integer n with $p_{ij}^{(n)} > 0$.

This is a property that can normally be judged from the transition graph alone.

It is not necessary to include probabilities on the graph as we are only looking to see if there exists a path from i to j for any two states i and j.

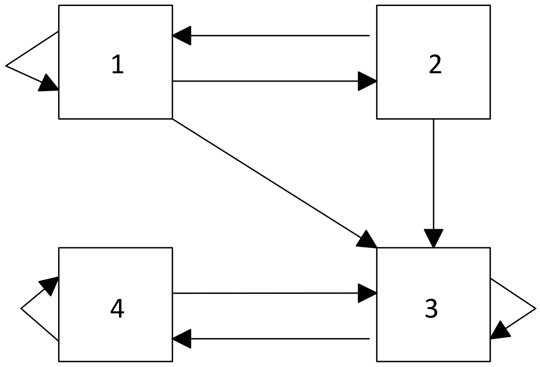
Question

Determine whether the process with the following transition matrix is irreducible:

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

Solution

Labelling the states as 1, 2, 3, and 4, the transition graph is as follows:



From the transition graph we see that this process is not irreducible. For example, we cannot get from state 4 to state 1.

The Markov chains of Models 5.1, 5.2 and 5.3 are irreducible; so is 5.4 except when either boundary is absorbing ($\alpha = 1$ or $\beta = 1$). Such absorbing states occur in many practical situations (*eg* ruin).

Question

Explain why a random walk with absorbing barriers is not irreducible.

Solution

If the absorbing boundary is at state i, then there can be no path from i to any other state. So a random walk with an absorbing boundary is not irreducible.

Irreducible Markov chains have the following property.

Stationary distribution result (2)

An irreducible Markov chain with a finite state space has a unique stationary probability distribution.

The proof of this result is beyond the syllabus.

Now consider the Markov chain with transition matrix:

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/3 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2/3 & 1/3 \\ 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

Suppose that the states are labelled 1, 2, 3, 4 and 5.

This chain is not irreducible. Looking at the transition matrix we see that:

- it is not possible to leave the subset of states $\{1,2\}$
- similarly it is not possible to leave the subset of states $\{4,5\}$
- it is not possible to leave state 3.

Each of these three subsets of states, $\{1,2\}$, $\{3\}$ and $\{4,5\}$ is effectively an irreducible Markov process on its own. Therefore each of these has a unique stationary distribution.

In order to determine the stationary distribution of the chain formed by states 1 and 2, we need to solve the matrix equation:

$$\begin{pmatrix}\pi_0 & \pi_1\end{pmatrix}\begin{pmatrix}1/2 & 1/2 \\ 1/3 & 2/3\end{pmatrix}=\begin{pmatrix}\pi_0 & \pi_1\end{pmatrix}$$

This is equivalent to the following system of equations:

$$1/2\,\pi_0 + 1/3\,\pi_1 = \pi_0$$

$$1/2\,\pi_0 + 2/3\,\pi_1 = \pi_1$$

Taking all terms to the left-hand side in both equations we get:

$$-1/2\,\pi_0 + 1/3\,\pi_1 = 0$$

$$1/2\,\pi_0 - 1/3\,\pi_1 = 0$$

These equations are equivalent (since they just have opposite signs). So either one of them can be discarded. Solving either one of them gives $\pi_1 = 3/2\,\pi_0$.

We also require $\pi_0 + \pi_1 = 1$. So we have the unique solution:

$$\pi_0 = 2/5 \quad \text{and} \quad \pi_1 = 3/5$$

We can check the solution by substituting these values back into the discarded equation.

Similarly, the stationary distribution of the chain formed by states 4 and 5 is:

$$\pi_4 = 3/5 \quad \text{and} \quad \pi_5 = 2/5$$

Combining these, we see that $\frac{1}{5}(2,3,0,0,0)$ and $\frac{1}{5}(0,0,0,3,2)$ are both stationary distributions for the 5-state process.

In addition, the stationary distribution corresponding to state 3 is:

$$(0,0,1,0,0)$$

Since these possible stationary distributions are all independent, we can have linear combinations of them.

A general stationary distribution is therefore of the form:

$$\frac{1}{5a+b+5c}(2a,3a,b,3c,2c)$$

where a, b and c are arbitrary non-negative constants. For example, setting $a=1$, $b=2$ and $c=3$ gives the stationary distribution $\frac{1}{22}(2,3,2,9,6)$.

It is common for Markov chains with infinite state spaces to have no stationary probability distribution, even if the chain is irreducible; this is the case for the simple random walk of Model 5.3.

6.2 The long-term behaviour of Markov chains

The importance of the stationary distribution comes from its connection with the long-term behaviour of a Markov chain. Under suitable conditions (to be made precise below), a Markov chain will 'settle down' to its stationary distribution after a sufficiently long period of time.

It is natural to expect the distribution of a Markov chain to tend to the invariant distribution π for large times. This is why the stationary distribution is so important: if the above convergence holds, $p_{ij}^{(n)}$ will be close to π_j for an overwhelming fraction of the time in the long run.

Certain phenomena complicate the above picture somewhat.

The period of a state

A state i is said to be *periodic* with period $d > 1$ if a return to i is possible only in a number of steps that is a multiple of d (*ie* $p_{ii}^{(n)} = 0$ unless $n = md$ for some integer m).

A state is said to be *aperiodic* if it is not periodic.

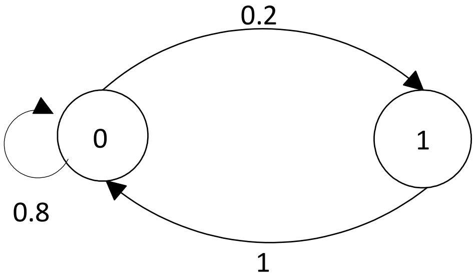
It is only for aperiodic states that $\lim\limits_{n\to\infty} p_{ii}^{(n)}$ can exist.

One can check using the transition graphs that, in Models 5.1 and 5.2, all states are aperiodic, whereas in Model 5.3, all states have period 2. Finally in Model 5.4 all states are aperiodic unless both α and β are either 0 or 1.

It is not necessarily the case that return to an aperiodic state is possible after an *arbitrary* number of steps, only that return is not constrained to be in a multiple of some number $d > 1$. So, in effect, the highest common factor of the return times for the state needs to be 1 in order for the state to be aperiodic.

Drawing a transition diagram might help us to decide whether or not a state is periodic. If a state has an arrow back to itself, that state is aperiodic because a return to that state is possible in any number of steps. However, even if there is no arrow from the state back to itself, the state may still be aperiodic.

Consider a time-homogeneous Markov chain on the state space $S = \{0,1\}$ with transition probabilities $p_{00} = 0.8$, $p_{01} = 0.2$, $p_{10} = 1$ and $p_{11} = 0$. The transition diagram is:



Since state 0 has an arrow back to itself, it is aperiodic.

A return to state 1 is not possible in 1 step, but is possible in 2, 3, 4, … steps. The highest common factor of 2, 3, 4, … is 1. So state 1 is aperiodic.

In fact, once we have decided that state 0 is aperiodic, we can say straight away that state 1 is also aperiodic by applying the following important result for irreducible chains.

Periodicity result

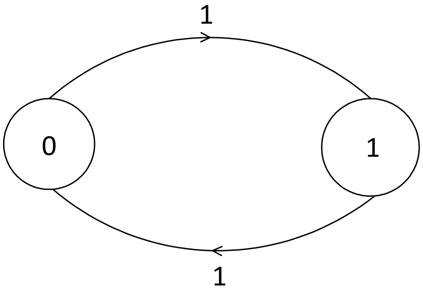
If a Markov chain is *irreducible* all its states have the same period (or *all* are aperiodic).

This greatly simplifies the problem of finding periodicities for irreducible chains as only one state need be considered. The chain in the two-state example above *is* irreducible (as every state can be reached from every other state).

Aperiodic Markov chains

A Markov chain is said to be aperiodic if all its states are aperiodic.

Now consider a time-homogeneous Markov chain on the state space $S = \{0,1\}$ with transition probabilities $p_{01} = 1$ and $p_{10} = 1$. The transition graph is:



The chain has a finite number of states. Each state can be reached from the other state so the chain is irreducible. So it must have a unique stationary distribution. It isn't difficult to see that this must be $(½, ½)$, *ie* an equal chance of being in either state. (We'll prove this in the next question.)

From the diagram we see that a return to state 0 is possible only in an even number of steps. So the period of state 0 is 2. The same is true of state 1. So the chain is not aperiodic and the process may not conform to the stationary distribution in the long term.

Although the process has a stationary distribution, if the process doesn't start off in that distribution, then it will never reach it. The idea that a process can 'start off in a distribution' might be confusing as any particular run of the process must start in a particular state. However, we can have a lack of information about which state that is. For example, we might only know that at time 0 there is a 10% chance that it's in state 0, and a 90% chance that it's in state 1. At time 1, these probabilities will be reversed, with a 90% chance of being in state 0, and only 10% chance of being in state 1. The process doesn't settle down to an equilibrium position.

An alternative way of thinking about this is to suppose we have a large number of independent copies of the Markov process running. To be concrete, let's assume that each independent process describes the state of a person, as in the case of policyholders following an NCD Markov chain. We can then picture the idea of the starting distribution in terms of numbers of people. For example, the 10%/90% split referred to above would correspond to 10% of the people starting in state 0, and 90% starting in state 1. Because of the transition probabilities, each person will change state. This will continue at each time step, so that the process never settles down to equilibrium.

Question

Show by solving the necessary matrix equation that the stationary distribution for the process in the example above is $(½, ½)$.

Solution

The transition matrix is given by $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. So for a stationary distribution we need to solve:

$$\begin{pmatrix} \pi_0 & \pi_1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \pi_0 & \pi_1 \end{pmatrix}$$

This is equivalent to $\pi_0 = \pi_1$. Together with the normalisation condition this gives $\pi_0 = \pi_1 = ½$. This is unique, as we would expect for an irreducible Markov chain on a finite state space.

Next consider a time-homogeneous Markov chain on the state space $S = \{0,1\}$ with transition probabilities $p_{00} = p_{01} = p_{10} = p_{11} = ½$. So no matter what state the process is in, the probability of remaining in that state and the probability of moving to the other state are both 0.5.

Again the process is finite and irreducible, so a unique stationary distribution exists. Moreover, this stationary distribution is again $(½, ½)$.

However, in this case a sample path starting in state 0, say, can return to state 0 after any number of steps, hence the state is aperiodic. Exactly the same is true of state 1.

Furthermore, $p_{00}^{(2)} = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}$, since to go from 0 to 0 in 2 time steps, we either stay still for two time steps, or go from 0 to 1 and back to 0 again.

In fact, since $P^2 = P$, we must have $P^n = P$ by induction. Therefore $p_{00}^{(n)} = \tfrac{1}{2}$ for all $n = 1,2,3,\ldots$. It follows that $\lim_{n\to\infty} p_{00}^{(n)} = \tfrac{1}{2}$.

In contrast to the process in the previous example, this process does settle down to the equilibrium distribution. In fact this occurs after one time step. But, in general, this would take longer.

We can now state a result on convergence to the stationary probability distribution.

Stationary distribution result (3)

Let $p_{ij}^{(n)}$ be the n-step transition probability of an irreducible aperiodic Markov chain on a finite state space. Then for every i and j:

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$$

where π is the stationary probability distribution.

Note how the above limit is independent of the starting state i. The proof is beyond the syllabus.

This result is saying that no matter what state the process is currently in, the probability of ending up in state j after a very long time, is the same as the probability of being in state j as given by the stationary distribution π. This is the same as saying that, after a very long time, the distribution is constant and equal to the stationary distribution.

Summary

- A Markov chain with a finite state space has at least one stationary distribution.
- An irreducible Markov chain with a finite state space has a unique stationary distribution.
- An irreducible, aperiodic Markov chain with a finite state space will settle down to its unique stationary distribution in the long run.

7 Modelling using Markov chains

Using the principle of economy of effort, it is common to start the modelling process by attempting to fit a simple stochastic model, such as a Markov chain, to a set of observations. If tests show that this is inadequate, a more sophisticated model can be attempted at the next stage of the modelling process.

This section assumes that the model being fitted is time-homogeneous. The situation is generally more complicated when fitting a time-inhomogeneous model.

7.1 Estimating transition probabilities

The first thing to fix when setting up a Markov model is the state space. As shown by the example in Section 5.2, the state space which first springs to mind may not be the most suitable and may need some modification before a Markov model can be fitted.

Recall that the example referred to was the NCD system where it was required to split one of the discount levels into two, depending on the previous state.

Once the state space is determined, however, the Markov model must be fitted to the data by estimating the transition probabilities p_{ij}.

Denote by $x_1, x_2, \dots x_N$ the available observations and define:

- **n_i as the number of times t $(1 \le t \le N-1)$ such that $x_t = i$;**
- **n_{ij} as the number of times t $(1 \le t \le N-1)$ such that $x_t = i$ and $x_{t+1} = j$.**

Thus n_{ij} is the observed number of transitions from state i to j, n_i the observed number of transitions from state i.

So $n_i = \sum_{j \in S} n_{ij}$.

The reason that the definition of n_i only allows t to go up to $N-1$, rather than N, is so that it equals the number of transitions out of state i, and not just the number of times the process is in state i.

Then the best estimate of p_{ij} is $\hat{p}_{ij} = \dfrac{n_{ij}}{n_i}$.

If a confidence interval is required for a transition probability, the fact that the conditional distribution of N_{ij} given N_i is *Binomial*(N_i, p_{ij}) means that a confidence interval may be obtained by standard techniques.

An approximate 95% confidence interval for p_{ij} is given by:

$$\hat{p}_{ij} \pm 1.96\sqrt{\frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{n_i}}$$

Confidence intervals are covered in detail in Subject CS1.

7.2 Assessing the fit

The next step is to ensure that the fit of the model to the data is adequate, or in other words to check that the Markov property seems to hold.

For a general Markov chain model a full verification of the Markov property would involve a great deal of work and a voluminous supply of data. In practice it is generally considered sufficient to look at triplets of successive observations.

Triplets test

Denote by n_{ijk} the number of times t $(1 \le t \le N-2)$ such that $x_t = i$, $x_{t+1} = j$ and $x_{t+2} = k$. If the Markov property holds we expect n_{ijk} to be an observation from a binomial distribution with parameters n_{ij} and p_{jk}. A simple but effective test, therefore, is the chi-square goodness-of-fit test based on the test statistic:

$$\chi^2 = \sum_i \sum_j \sum_k \frac{\left(n_{ijk} - n_{ij}\hat{p}_{jk}\right)^2}{n_{ij}\hat{p}_{jk}}$$

This is of the familiar form $\sum_{i,j,k} \frac{(O-E)^2}{E}$ where O represents the observed frequency and E represents the expected frequency. If $N_{ijk} \sim Binomial(n_{ij}, p_{jk})$, then $E(N_{ijk}) = n_{ij}p_{jk}$. When calculating the expected frequencies, n_{ij} is calculated as $\sum_k n_{ijk}$ to ensure that the observed and expected frequencies tally.

Question

A 3-state process has been observed over a period of time and the sequence of states occupied is as follows:

1,3,2,2,1,3,3,2,3,1,2,3,2,1,1,2,2,1,3,3

(i) Calculate the values of n_{ijk}, n_{ij} and n_i.

(ii) Estimate the one-step transition probabilities.

(iii) State the null and alternative hypotheses for the triplets test.

(iv) Calculate the test statistic for the triplets test (without combining triplets with small expected frequencies).

Solution

(i) ***Values***

The values of n_{ijk} are shown in the matrices below:

$$\left(n_{1jk}\right)=\begin{pmatrix}0&1&0\\0&1&1\\0&1&2\end{pmatrix}\qquad\left(n_{2jk}\right)=\begin{pmatrix}1&0&2\\2&0&0\\1&1&0\end{pmatrix}\qquad\left(n_{3jk}\right)=\begin{pmatrix}0&1&0\\1&1&1\\0&1&0\end{pmatrix}$$

The first row of the matrix $\left(n_{1jk}\right)$ contains the entries n_{111}, n_{112} and n_{113}; the second row consists of n_{121}, n_{122} and n_{123}, *etc*.

The value of n_{ij} is the ij th entry of the following matrix:

$$\left(n_{ij}\right)=\begin{pmatrix}1&2&3\\3&2&2\\1&3&2\end{pmatrix}$$

and the n_i values are the row sums of the matrix $\left(n_{ij}\right)$:

$$\left(n_i\right)=\begin{pmatrix}6\\7\\6\end{pmatrix}$$

(ii) ***One-step transition probabilities***

Using the formula $\hat{p}_{ij}=\dfrac{n_{ij}}{n_i}$, we obtain the following estimates:

$$\begin{pmatrix}\frac{1}{6}&\frac{2}{6}&\frac{3}{6}\\\frac{3}{7}&\frac{2}{7}&\frac{2}{7}\\\frac{1}{6}&\frac{3}{6}&\frac{2}{6}\end{pmatrix}=\begin{pmatrix}\frac{1}{6}&\frac{1}{3}&\frac{1}{2}\\\frac{3}{7}&\frac{2}{7}&\frac{2}{7}\\\frac{1}{6}&\frac{1}{2}&\frac{1}{3}\end{pmatrix}=\begin{pmatrix}0.167&0.333&0.5\\0.429&0.286&0.286\\0.167&0.5&0.333\end{pmatrix}$$

(iii) ***Hypotheses***

The null hypothesis is:

H_0: the process has the Markov property

The alternative hypothesis is:

H_1: the process does not have the Markov property

Under the null hypothesis, $N_{ijk}\sim Binomial(n_{ij},p_{jk})$.

(iv) *Chi-squared test*

The formula for the test statistic is:

$$\sum_{i,j,k} \frac{\left(n_{ijk} - n_{ij}\hat{p}_{jk}\right)^2}{n_{ij}\hat{p}_{jk}}$$

where n_{ij} is calculated using the formula $n_{ij} = \sum_k n_{ijk}$. This gives rise to the same values for n_{ij} as shown in the matrix in (i), except for n_{33}, as the string of observations ends with 3,3. Using this formula gives $n_{33} = 1$. The final 3,3 in the list of observations is not counted here because this is unable to give rise to an observation of the form 3,3,*k*.

We have:

ijk	Observed frequency n_{ijk}	n_{ij}	$\hat{p}_{jk}$	Expected frequency $n_{ij}\,\hat{p}_{jk}$	$\frac{\left(n_{ijk} - n_{ij}\hat{p}_{jk}\right)^2}{n_{ij}\hat{p}_{jk}}$
111	0	1	0.167	0.167	0.167
112	1	1	0.333	0.333	1.333
113	0	1	0.500	0.500	0.500
121	0	2	0.429	0.857	0.857
122	1	2	0.286	0.571	0.321
123	1	2	0.286	0.571	0.321
131	0	3	0.167	0.500	0.500
132	1	3	0.500	1.500	0.167
133	2	3	0.333	1.000	1.000
211	1	3	0.167	0.500	0.500
212	0	3	0.333	1.000	1.000
213	2	3	0.500	1.500	0.167
221	2	2	0.429	0.857	1.524
222	0	2	0.286	0.571	0.571
223	0	2	0.286	0.571	0.571
231	1	2	0.167	0.333	1.333
232	1	2	0.500	1.000	0.000
233	0	2	0.333	0.667	0.667
311	0	1	0.167	0.167	0.167

312	1	1	0.333	0.333	1.333
313	0	1	0.500	0.500	0.500
321	1	3	0.429	1.286	0.063
322	1	3	0.286	0.857	0.024
323	1	3	0.286	0.857	0.024
331	0	1	0.167	0.167	0.167
332	1	1	0.500	0.500	0.500
333	0	1	0.333	0.333	0.333

The observed value of the test statistic is the sum of the numbers in the final column, *ie*:

$$\sum_{i,j,k} \frac{\left(n_{ijk} - n_{ij}\hat{p}_{jk}\right)^2}{n_{ij}\hat{p}_{jk}} = 14.61$$

Recall that, when carrying out a chi-squared test, the expected frequencies should ideally all be 5 or more. If this is not the case, the validity of the test is questionable.

In order to complete the chi-squared test, we would need to know the number of degrees of freedom to use. The formula for the number of degrees of freedom is beyond the scope of Subject CS2. If this comes up in the exam, the number of degrees of freedom should be stated or a formula will be given.

An additional method in frequent use for assessing goodness of fit is to run some simulations of the fitted chain and to compare graphs of the resulting trajectories with a graph of the process actually observed. This method often highlights deficiencies that are missed by the chi-square test.

For example, given a sequence $y_1, y_2, \ldots, y_N$ of closing values of an exchange rate, one model which suggests itself is to let x_t be the nearest integer to $K \log y_t$ where K is a scaling constant of suitable magnitude, and to model $x_1, x_2, \ldots, x_N$ as a random walk, with transition probabilities:

$$p_{i,i+1} = \theta, \quad p_{i,i-1} = \phi, \quad p_{i,i} = 1 - \theta - \phi$$

The parameters θ and ϕ can be estimated quite satisfactorily in practice, but a visual comparison of a simulated random walk with the observed trajectory of x tends to show that the real exchange rate remains relatively constant for long periods with occasional bursts of increased volatility, whereas the Markov chain model is incapable of simulating such behaviour.

7.3 Simulation

Simulating a time-homogeneous Markov chain is fairly straightforward, as the Markov property means that the conditional distribution of X_{t+1} given the history of X up until time t is only dependent on X_t.

If the state space of X is finite, there are only a limited number of distributions, all discrete, from which the program needs to be able to sample; these can be listed individually, along with instructions telling the program which distribution to use for each step.

Consider a two-state Markov chain with transition matrix:

$$\begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix} \end{array}$$

If the process is in state 0 at time 0, then we could simulate a series of observations from this process as follows.

Row 1 of the transition matrix is the conditional distribution of X_1 given that $X_0 = 0$.

We can simulate a value for X_1 as follows:

- generate a random number u from $U(0,1)$
- set $X_1 = \begin{cases} 0 & \text{if } u \le 0.6 \\ 1 & \text{if } u > 0.6 \end{cases}$

If the simulated value is 0, we repeat the simulation above to obtain a simulated value of X_2. If the simulated value is 1, we can simulate a value for X_2 (using the probabilities in the second row of the matrix) as follows:

- generate a random number u from $U(0,1)$
- set $X_2 = \begin{cases} 0 & \text{if } u \le 0.3 \\ 1 & \text{if } u > 0.3 \end{cases}$

This process is repeated to simulate additional values of the Markov chain.

Models that assume an infinite state space usually have a simple transition structure, often based on the distribution of the increments. The random walk, which has independent increments, is one such example; another might be a process which can only make transitions of the form $x \mapsto x+1$ or $x \mapsto x-1$, with respective probabilities θ_x and $1-\theta_x$.

The second example in the paragraph above is not a random walk because the increments ($+1$ or -1) are not identically distributed. The associated probabilities depend on the current state.

In addition to commercial simulation packages, which are able to simulate Markov chains without difficulty, even standard spreadsheet software can easily cope with the practical aspects of estimating transition probabilities and performing a simulation.

In R, the package `markovchain` **can be used to create/simulate a Markov chain.**

As an example consider a Markov chain with three states: Employed (Emp), Unemployed (claiming benefit) (Unemp) and Inactive in the labour force (Inactive), measured at the end of each month. Suppose the transition matrix, with the states in the order given above, is:

$$\begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.5 & 0.4 & 0.1 \\ 0.4 & 0.0 & 0.6 \end{pmatrix}$$

To create a Markov chain in R use:

```
Employment = new("markovchain", states = c("Emp", "Unemp",
"Inactive"),
transitionMatrix = matrix(data = c(0.8, 0.1,0.1,
0.5,0.4, 0.1,
0.4,0.0,0.6),byrow = byRow, nrow = 3),
name = "Employmt")
```

Note that R will give an error message unless the markovchain package has been preloaded. This is covered in more detail in the R part of Subject CS2.

Suppose the process begins in state Emp. To see the probability distribution after 3 and 6 months, use:

```
InitialState = c(1,0,0)
After3Months = InitialState * (Employment * Employment *
Employment)
After6Months = InitialState * (Employment^6)
```

After 6 months the probabilities are:

Employed 0.687, Unemployed 0.116, Inactive 0.197

Chapter 2 Summary

Markov chains

A Markov process with a discrete time set and discrete state space is called a Markov chain.

Chapman-Kolmogorov equations

$$p_{ij}^{(m,n)} = \sum_{k\in S} p_{ik}^{(m,l)}\, p_{kj}^{(l,n)}$$

for all states i,j in S and all integer times $m<l<n$.

Time-homogeneous Markov chains

A simplification occurs if the one-step transition probabilities are time independent:

$$p_{ij}^{(n,n+1)} = p_{ij}$$

Then the Chapman-Kolmogorov equations are:

$$p_{ij}^{(n-m)} = \sum_{k\in S} p_{ik}^{(l-m)}\, p_{kj}^{(n-l)}$$

The transition matrix

The transition matrix P is a square $N\times N$ matrix, where N is the number of states in S. The entry in the i *th* row and j *th* column is p_{ij}.

In the time-homogeneous case, the l-step transition probability $p_{ij}^{(l)}$ (*ie* the probability of moving from state i to state j in exactly l steps) can be obtained by calculating the (i,j) *th* entry of the matrix P^l.

Random walks

Random walks are important examples of Markov chains. The increments of a random walk are IID. In other words, the values move up or down by completely random amounts at each step.

A simple random walk has step-sizes of ± 1, *ie*:

$$P(X_n = x+1 \mid X_{n-1} = x) = p\,,\quad P(X_n = x-1 \mid X_{n-1} = x) = 1-p (=q)$$

In a simple symmetric random walk, $p = q = ½$.

Stationary distributions

These probabilities must satisfy the vector equation $\pi = \pi P$ with $\pi_i \geq 0$ and $\sum \pi_i = 1$.

A solution to these equations is called a stationary distribution.

Irreducible chains

A Markov chain is said to be irreducible if every state can be reached from every other state.

Periodicity

A state i is said to be periodic with period $d > 1$ if a return to state i is possible only in a number of steps that is a multiple of d.

A state is said to be aperiodic if it is not periodic.

A Markov chain is said to be aperiodic if all its states are aperiodic.

If a Markov chain is irreducible, all its states have the same period or are all aperiodic.

Long-term behaviour of Markov chains

A Markov chain with a finite state space has at least one stationary distribution.

An irreducible Markov chain with a finite state space has a unique stationary distribution.

An irreducible, aperiodic Markov chain with a finite state space will settle down to its unique stationary distribution in the long run.

Estimating transition probabilities

The transition probability p_{ij} is estimated by:

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i} = \frac{\text{number of transitions from state } i \text{ to state } j}{\text{number of transitions out of state } i}$$

Testing the Markov assumption

We can test the Markov assumption using a triplets test. The formula for the test statistic is:

$$\sum_{i,j,k} \frac{\left(n_{ijk} - n_{ij}\hat{p}_{jk}\right)^2}{n_{ij}\hat{p}_{jk}}$$

Chapter 2 Practice Questions

2.1 A simple no claims discount system for motor insurance has four levels of discount – 0%, 20%, 40% and 60%. A new policyholder starts on 0% discount. At the end of each policy year, policyholders will change levels according to the following rules:

- At the end of a claim-free year, a policyholder moves up one level, or remains on the maximum discount.
- At the end of a year in which exactly one claim was made, a policyholder drops back one level, or remains at 0%.
- At the end of a year in which more than one claim was made, a policyholder drops back to zero discount.

For a particular driver in any year, the probability of a claim-free year is 0.9, the probability of exactly one claim is 0.075, and the probability of more than one claim is 0.025.

Mike took out a policy for the first time on 1 January 2015, and by 1 January 2018 he had made only one claim, on 3 May 2017. Calculate the probability that he is on 20% discount in 2020.

2.2 A Markov chain is determined by the transition matrix:

$$P=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \end{bmatrix}$$

Determine the period of each of the states in this chain.

2.3 A Markov chain $\{X_n\}_{n=0}^{\infty}$ has a discrete state space S. The initial probability distribution is given by $P\left[X_0=i\right]=q_i$. The one-step transition probabilities are denoted by $P\left[X_{m+1}=i_{m+1}\,|\,X_m=i_m\right]=p_{i_m i_{m+1}}^{(m,m+1)}$.

(i) State the Markov property for such a process.

(ii) Write down expressions for the following in terms of p's and q's.

(a) $P\left[X_0=i_0,X_1=i_1,\ldots,X_n=i_n\right]$

(b) $P\left[X_4=i\right]$

2.4 A new actuarial student is planning to sit one exam each session. He expects that his performance in any exam will only be dependent on whether he passed or failed the last exam he sat. If he passes a given exam, the probability of passing the next will be α, regardless of the nature of the exam. If he fails an exam, the probability of passing the next will be β.

(i) Obtain an expression for the probability that:

(a) the first exam he fails is the seventh, given that he passes the first

(b) he passes the fifth exam, given that he fails the first three.

(ii) Explain the results above in terms of a Markov chain, specifying the state space and transition matrix. (For the purposes of this model, assume that we are only interested in predicting passing or failing, not in the number of exams passed so far.)

2.5 The stochastic process $\{X_t\}$ is defined by the relationship $X_t = Z_t + Z_{t-1}$, where $\{Z_t\}$ is a sequence of independent random variables with probability function:

$$Z_t = \begin{cases} 1 & \text{with probability } p \\ 1{,}000 & \text{with probability } q \end{cases}$$

where $p+q=1$ and $q<p$.

(i) Obtain expressions in terms of p and q for each of the following probabilities:

(a) $P(X_5 = 1{,}001)$

(b) $P(X_5 = 1{,}001 \mid X_4 = 1{,}001)$

(c) $P(X_5 = 1{,}001 \mid X_4 = 1{,}001, X_3 = 1{,}001)$.

(ii) State, with reasons, whether $\{X_t\}$ has the Markov property.

2.6 Determine all the stationary distributions for a Markov chain with transition matrix:

$$P = \begin{pmatrix} 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{5} & \frac{4}{5} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ 0 & \frac{4}{5} & \frac{1}{5} & 0 & 0 \\ \frac{1}{2} & \frac{3}{10} & \frac{1}{5} & 0 & 0 \end{pmatrix}$$

2.7 Exam style

At the end of each year an independent organisation ranks the performance of the unit trusts invested in a particular sector, and classifies them into four quartiles (with quartile 1 relating to the best performance). Past experience has shown that, at the end of each year, a fund will either remain in the same quartile or will move to a neighbouring quartile.

In fact, there is a probability $1-2\alpha$ that a fund will remain in the same quartile and, where upward or downward movements are both possible, these are equally likely. However, it has been found that a fund that has remained in the top or bottom quartile for two consecutive years has a probability of $1-\beta$ ($\beta<\alpha$) of remaining in the same quartile the following year.

(i) Construct a Markov chain with six states to model this situation, defining the states in your model and drawing a transition diagram. [3]

(ii) Write down the transition matrix for your model. [2]

(iii) Explain whether this Markov chain is irreducible and/or periodic. [2]

(iv) Show that, if a stationary distribution exists with a quarter of the funds in each quartile, then $\beta=\dfrac{\alpha(1-2\alpha)}{1-\alpha}$. [4]

(v) Last year 20% of funds in the second quartile moved up to the top quartile. Assuming the fund rankings have reached a stationary state, estimate the probability that a fund that has been in the top quartile for the last two years will remain in the top quartile for a third consecutive year. [2]

[Total 13]

2.8 Exam style

A simple NCD system has four levels of discount – 0%, 20%, 40% and 60%. A new policyholder starts on 0% discount. At the end of each policy year, policyholders will change levels according to the following rules:

- At the end of a claim-free year, a policyholder moves up one level, or remains on the maximum discount.
- At the end of a year in which exactly one claim was made, a policyholder drops back one level, or remains at 0%.
- At the end of a year in which more than one claim was made, a policyholder drops back to zero discount.

For a particular policyholder in any year, the probability of a claim-free year is ${}^{7}\!/_{10}$, the probability of exactly one claim is ${}^{1}\!/_{5}$ and the probability of more than one claim is ${}^{1}\!/_{10}$.

(i) Write down the transition matrix for this time-homogeneous Markov chain. [2]

(ii) Calculate the 2-step transition probabilities from state i to state j, $p_{ij}^{(2)}$. [3]

(iii) If the policyholder starts with no discount, calculate the probability that this policyholder is at the maximum discount level 5 years later. [5]

(iv) If a large number of people having the same claim probabilities take out policies at the same time, calculate the proportion would you expect to be in each discount category in the long run. [5]

[Total 15]

2.9 Consider the following two Markov chains:

Exam style

- Chain I is defined on the state space $\{1,2,3,4\}$ and has transition matrix:

$$\begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & \tfrac{1}{2} & 0 & \tfrac{1}{2} \\ \tfrac{1}{2} & 0 & \tfrac{1}{2} & 0 \\ 0 & \tfrac{1}{2} & 0 & \tfrac{1}{2} \\ \tfrac{1}{2} & 0 & \tfrac{1}{2} & 0 \end{pmatrix} \end{array}$$

- Chain II is defined on the state space $\{1,2,3,4,5\}$ and has transition matrix:

$$\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & \tfrac{1}{2} & 0 & 0 & \tfrac{1}{2} \\ \tfrac{1}{2} & 0 & \tfrac{1}{2} & 0 & 0 \\ 0 & \tfrac{1}{2} & 0 & \tfrac{1}{2} & 0 \\ 0 & 0 & \tfrac{1}{2} & 0 & \tfrac{1}{2} \\ \tfrac{1}{2} & 0 & 0 & \tfrac{1}{2} & 0 \end{pmatrix} \end{array}$$

Let X_t denote the state occupied at time t. For each chain:

(i) Draw a transition diagram, including on your diagram the probability of each possible transition. [2]

(ii) Calculate:

(a) $P\left(X_2 = 1 \mid X_0 = 1\right)$

(b) $P\left(X_4 = 1 \mid X_0 = 1\right)$ [4]

(iii) Explain whether the chain is irreducible and/or aperiodic. [3]

(iv) Explain whether or not the process will converge to a stationary distribution given that it is in State 1 at time 0. If it does converge, determine the stationary distribution. [3]

[Total 12]

2.10 Exam style

An author is about to start writing a book that will contain 20 chapters. The author plans to write a new chapter each week. However, when he reviews his work at the end of each week, there is a probability of 0.25 (which is independent of the current state of the book) that he will not be happy with one of the chapters he has written. In this case, he will spend the following week rewriting that particular chapter instead of embarking on a new one. He may decide to rewrite any one chapter, including a new one he has just finished or one that he has previously rewritten.

Let X_k denote the number of chapters that the author is happy with at the end of week k, and define $X_0 = 0$.

(i) Explain why X_k can be modelled as a Markov chain. [2]

(ii) Calculate the probability that the author will complete the book in exactly 25 weeks. [2]

(iii) Calculate the expected number of weeks it will take the author to complete the book. [3]

[Total 7]

2.11 Exam style

The credit-worthiness of debt issued by companies is assessed at the end of each year by a credit rating agency. The ratings are A (the most credit-worthy), B and D (debt defaulted). Historic evidence supports the view that the credit rating of a debt can be modelled as a Markov chain with one-year transition matrix:

$$P = \begin{pmatrix} 0.92 & 0.05 & 0.03 \\ 0.05 & 0.85 & 0.1 \\ 0 & 0 & 1 \end{pmatrix}$$

(i) Determine the probability that a company currently rated A will never be rated B in the future. [2]

(ii) (a) Calculate the second-order transition probabilities of the Markov chain.

(b) Hence calculate the expected number of defaults within the next two years from a group of 100 companies, all initially rated A. [2]

The manager of a portfolio investing in company debt follows a ‘downgrade trigger’ strategy. Under this strategy, any debt in a company whose rating has fallen to B at the end of a year is sold and replaced with debt in an A-rated company.

(iii) Calculate the expected number of defaults for this investment manager over the next two years, given that the portfolio initially consists of 100 A-rated bonds. [2]

(iv) Comment on the suggestion that the downgrade trigger strategy will improve the return on the portfolio. [2]

[Total 8]

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 2 Solutions

2.1 Mike starts with 0% discount in 2015. He makes no claims in 2015 or 2016, and so has a 40% discount in 2017. He makes exactly one claim that year so he falls back to 20% discount for 2018. So we are looking for the probability of being at the 20% discount level in 2020, given that in 2018 the discount was also 20%.

The transition matrix is:

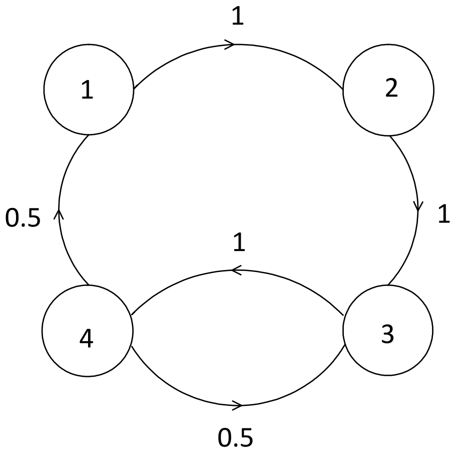
$$P = \begin{pmatrix} 0.1 & 0.9 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0 \\ 0.025 & 0.075 & 0 & 0.9 \\ 0.025 & 0 & 0.075 & 0.9 \end{pmatrix}$$

We want the (20%,20%) entry (or (2,2)th entry) in P^2.

Multiplying the 2nd row of P by the 2nd column gives:

$$0.1 \times 0.9 + 0.9 \times 0.075 = 0.1575$$

2.2 The transition graph for this Markov chain is shown below:



We can see from this that the chain is irreducible, so all the states have the same period.

We only need to find the period of one of the states. The chain can return to state 1 having started in state 1 after 4, 6, 8, 10, … moves. The highest common factor of these numbers is 2, so the period of all the states in the chain is 2.

2.3 (i) ***Markov property***

The Markov property means a lack of dependence on the past of the process:

$$P\left[X_n = j \middle| X_0 = i_0, X_1 = i_1, \ldots, X_{m-1} = i_{m-1}, X_m = i\right] = P\left[X_n = j \middle| X_m = i\right]$$

for all integer times $n > m$ and states $i_0, i_1, \ldots, i_{m-1}, i, j$ in S.

(ii) ***Probabilities***

(a) $P\left[X_0=i_0, X_1=i_1, \ldots, X_n=i_n\right]=q_{i_0} p_{i_0 i_1}^{(0,1)} p_{i_1 i_2}^{(1,2)} \ldots p_{i_{n-1} i_n}^{(n-1,n)}$

This is the probability of the process taking a unique given path.

(b) $P\left[X_4=i\right]=\sum_{i_0 \in S} \sum_{i_1 \in S} \sum_{i_2 \in S} \sum_{i_3 \in S} q_{i_0} p_{i_0 i_1}^{(0,1)} p_{i_1 i_2}^{(1,2)} p_{i_2 i_3}^{(2,3)} p_{i_3 i}^{(3,4)}$

Here we need to sum over all the possible starting points and then over all paths from these starting points to end up in state i at time 4.

2.4 (i) ***Probabilities***

(a) $\alpha^5(1-\alpha)$

(b) $(1-\beta)\beta+\beta\alpha=\beta(\alpha-\beta+1)$

(ii) ***Explanation***

Because the student's performance depends only upon whether he passed or failed the last exam, we can think of the problem as a Markov chain on the state space $\{F, P\}$ representing 'failed the last exam' and 'passed the last exam' respectively. The transition matrix is:

$$\begin{pmatrix} 1-\beta & \beta \\ 1-\alpha & \alpha \end{pmatrix}$$

In (i)(a) we are considering the probability of the unique path:

$$P \rightarrow P \rightarrow P \rightarrow P \rightarrow P \rightarrow P \rightarrow F$$

given that we start in P.

In (i)(b) we are considering the probabilities of the paths:

$$F \rightarrow F \rightarrow P \quad \text{and} \quad F \rightarrow P \rightarrow P$$

where the F at the start of these sequences represents the event that he fails the third exam.

Alternatively, we could view this as the transition probability $p_{FP}^{(2)}$, which is the FP entry (ie (1,2)th entry) in the matrix :

$$\begin{pmatrix} 1-\beta & \beta \\ 1-\alpha & \alpha \end{pmatrix}^2=\begin{pmatrix} (1-\beta)^2+\beta(1-\alpha) & (1-\beta)\beta+\beta\alpha \\ (1-\alpha)(1-\beta)+(1-\alpha)\alpha & (1-\alpha)\beta+\alpha^2 \end{pmatrix}$$

2.5 (i) ***Probabilities***

(a)

$$P(X_5=1{,}001)=P(Z_5+Z_4=1{,}001)$$
$$=P(Z_5=1{,}000,Z_4=1)+P(Z_5=1,Z_4=1{,}000)$$
$$=qp+pq=2pq$$

(b) Using the result in (i)(a) to evaluate the denominator:

$$P(X_5=1{,}001\,|\,X_4=1{,}001)=\frac{P(X_5=1{,}001,X_4=1{,}001)}{P(X_4=1{,}001)}$$
$$=\frac{P(Z_5+Z_4=1{,}001,Z_4+Z_3=1{,}001)}{2pq}$$
$$=\frac{P(Z_5=1{,}000,Z_4=1,Z_3=1{,}000)}{2pq}$$
$$+\frac{P(Z_5=1,Z_4=1{,}000,Z_3=1)}{2pq}$$
$$=\frac{q^2p+p^2q}{2pq}$$
$$=\frac{pq(q+p)}{2pq}$$
$$=\tfrac{1}{2} \qquad \text{because } p+q=1$$

(c) Using the expression for the numerator in (i)(b) to evaluate the denominator:

$$P(X_5=1{,}001\,|\,X_4=1{,}001,X_3=1{,}001)$$
$$=\frac{P(X_5=1{,}001,X_4=1{,}001,X_3=1{,}001)}{P(X_4=1{,}001,X_3=1{,}001)}$$
$$=\frac{P(Z_5+Z_4=1{,}001,Z_4+Z_3=1{,}001,Z_3+Z_2=1{,}001)}{pq\left(p+q\right)}$$
$$=\frac{P(Z_5=Z_3=1{,}000,Z_4=Z_2=1)}{pq\left(p+q\right)}$$
$$+\frac{P(Z_5=Z_3=1,Z_4=Z_2=1{,}000)}{pq\left(p+q\right)}$$
$$=\frac{2p^2q^2}{pq\left(p+q\right)}$$
$$=\frac{2pq}{\left(p+q\right)}$$
$$=2pq \qquad \text{because } p+q=1$$

(ii) ***Markov?***

If $\{X_t\}$ had the Markov property, the probabilities in (i)(b) and (i)(c) would be the same. Since they are not, it doesn't. (Note that $2pq < ½$ when $q < p$.)

2.6 We are solving:

$$\begin{pmatrix}\pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5\end{pmatrix}\begin{pmatrix} 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{5} & \frac{4}{5} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ 0 & \frac{4}{5} & \frac{1}{5} & 0 & 0 \\ \frac{1}{2} & \frac{3}{10} & \frac{1}{5} & 0 & 0 \end{pmatrix} = \begin{pmatrix}\pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5\end{pmatrix}$$

This gives five equations:

$$\tfrac{1}{2}\pi_5 = \pi_1$$

$$\tfrac{1}{5}\pi_2 + \tfrac{4}{5}\pi_4 + \tfrac{3}{10}\pi_5 = \pi_2$$

$$\tfrac{4}{5}\pi_2 + \tfrac{1}{5}\pi_4 + \tfrac{1}{5}\pi_5 = \pi_3$$

$$\tfrac{1}{2}\pi_1 + \tfrac{1}{3}\pi_3 = \pi_4$$

$$\tfrac{1}{2}\pi_1 + \tfrac{2}{3}\pi_3 = \pi_5$$

Rearranging we obtain:

$$-\pi_1 + \tfrac{1}{2}\pi_5 = 0$$

$$-\tfrac{4}{5}\pi_2 + \tfrac{4}{5}\pi_4 + \tfrac{3}{10}\pi_5 = 0$$

$$\tfrac{4}{5}\pi_2 - \pi_3 + \tfrac{1}{5}\pi_4 + \tfrac{1}{5}\pi_5 = 0$$

$$\tfrac{1}{2}\pi_1 + \tfrac{1}{3}\pi_3 - \pi_4 = 0$$

$$\tfrac{1}{2}\pi_1 + \tfrac{2}{3}\pi_3 - \pi_5 = 0$$

We will ignore the third equation since one equation is always redundant. So we are trying to solve:

$$-\pi_1 + \tfrac{1}{2}\pi_5 = 0$$

$$-\tfrac{4}{5}\pi_2 + \tfrac{4}{5}\pi_4 + \tfrac{3}{10}\pi_5 = 0$$

$$\tfrac{1}{2}\pi_1 + \tfrac{1}{3}\pi_3 - \pi_4 = 0$$

$$\tfrac{1}{2}\pi_1 + \tfrac{2}{3}\pi_3 - \pi_5 = 0$$

We will choose π_1 as the working variable.

From the first equation we have $\pi_5 = 2\pi_1$.

Substituting this in the fourth equation gives $\pi_3 = \frac{9}{4}\pi_1$.

Using the third we can then obtain $\pi_4 = \frac{1}{2}\pi_1 + \frac{1}{3} \times \frac{9}{4}\pi_1 = \frac{5}{4}\pi_1$.

Finally from the second equation we see that:

$$\pi_2 = \frac{5}{4}\left(\frac{4}{5} \times \frac{5}{4}\pi_1 + \frac{3}{10} \times 2\pi_1\right) = 2\pi_1$$

Thus our solution in terms of π_1 is $\left(1, 2, \frac{9}{4}, \frac{5}{4}, 2\right)\pi_1$.

Now apply the condition of summing to 1 to get:

$$\pi_1 = \frac{1}{\left(1+2+\frac{9}{4}+\frac{5}{4}+2\right)} = \frac{2}{17}$$

and therefore the stationary distribution is $\left(\frac{2}{17}, \frac{4}{17}, \frac{9}{34}, \frac{5}{34}, \frac{4}{17}\right)$.

This chain has a finite state space and is irreducible, so it has a unique stationary distribution.

2.7 (i) ***Markov chain***

We need to subdivide the top and bottom quartiles in order to satisfy the Markov property. This results in the following 6 states:

State 11: Funds in the 1st quartile this year and last year

State 1: Funds in the 1st quartile this year but not last year

State 2: Funds in the 2nd quartile this year

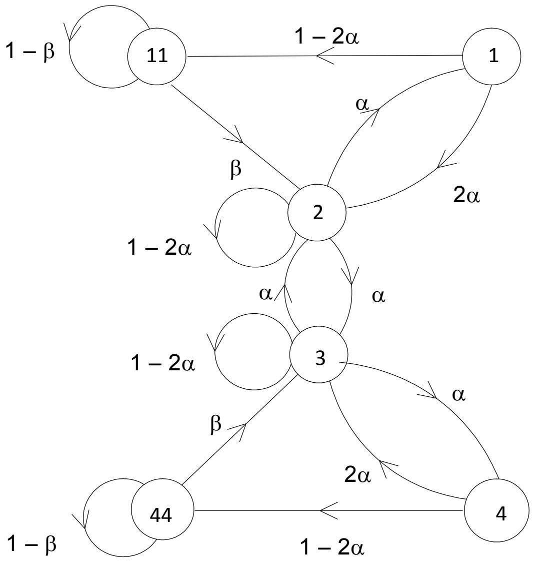
State 3: Funds in the 3rd quartile this year

State 4: Funds in the 4th quartile this year but not last year

State 44: Funds in the 4th quartile this year and last year [1]

The labels for the states need not match the ones given here.

The transition diagram then looks like this:



[2]

(ii) ***Transition matrix***

The transition matrix is:

$$
\begin{array}{cc}
 & \begin{array}{cccccc} 11 & 1 & 2 & 3 & 4 & 44 \end{array} \\
\begin{array}{c} 11 \\ 1 \\ P = \; 2 \\ 3 \\ 4 \\ 44 \end{array} &
\left[\begin{array}{cccccc}
1-\beta & & \beta & & & \\
1-2\alpha & & 2\alpha & & & \\
 & \alpha & 1-2\alpha & \alpha & & \\
 & & \alpha & 1-2\alpha & \alpha & \\
 & & & 2\alpha & & 1-2\alpha \\
 & & & \beta & & 1-\beta
\end{array} \right]
\end{array}
$$

[2]

(iii) ***Irreducible and periodic?***

This chain is irreducible since it is possible to move from each state to any other, *eg* by following the route $\cdots \to 11 \to 2 \to 3 \to 4 \to 44 \to 3 \to 2 \to 1 \to 11 \to \cdots$. [1]

A periodic chain is one in which a state can only be revisited at multiples of some fixed number $d > 1$. State 11 is aperiodic as it can be revisited after any number of steps. Also, since this chain is irreducible, all the states have the same periodicity. So the chain is aperiodic. [1]

(iv) ***Proof***

If a stationary distribution exists with a quarter of the funds in each quartile, then the stationary probabilities π_i must satisfy:

$$\pi_{11} + \pi_1 = \pi_2 = \pi_3 = \pi_4 + \pi_{44} = \tfrac{1}{4}$$ [1]

The stationary probabilities also satisfy the matrix equation $\pi = \pi P$.

The first column of this matrix equation tells us that:

$$(1-\beta)\pi_{11}+(1-2\alpha)\pi_1=\pi_{11} \quad \Rightarrow -\beta\pi_{11}+(1-2\alpha)\pi_1=0 \quad \Rightarrow \pi_{11}=\frac{(1-2\alpha)}{\beta}\pi_1 \qquad [1]$$

But we want $\pi_{11}+\pi_1=\frac{1}{4}$. So:

$$\frac{(1-2\alpha)}{\beta}\pi_1+\pi_1=\tfrac{1}{4} \quad \textit{ie} \quad \left(1+\frac{1-2\alpha}{\beta}\right)\pi_1=\tfrac{1}{4} \qquad [1]$$

The second column of this matrix equation tells us that:

$$\alpha\pi_2=\pi_1 \quad \textit{ie} \quad \alpha\times\tfrac{1}{4}=\pi_1$$

Combining these two equations gives:

$$\left(1+\frac{1-2\alpha}{\beta}\right)\alpha\times\tfrac{1}{4}=\tfrac{1}{4} \quad \Rightarrow \left(1+\frac{1-2\alpha}{\beta}\right)\alpha=1 \quad \Rightarrow \beta=\frac{\alpha(1-2\alpha)}{1-\alpha} \qquad [1]$$

(v) ***Estimated probability***

The probability of a fund in the second quartile moving up to the top quartile is α. So we estimate $\hat{\alpha}=0.2$. Hence the probability of the fund remaining in the top quartile for a third consecutive year is estimated to be:

$$1-\hat{\beta}=1-\frac{\hat{\alpha}(1-2\hat{\alpha})}{1-\hat{\alpha}}=1-\frac{0.2\times0.6}{0.8}=0.85 \qquad [2]$$

2.8 (i) ***Transition matrix***

The one-step transition matrix is:

$$P=\frac{1}{10}\begin{pmatrix}3&7&0&0\\3&0&7&0\\1&2&0&7\\1&0&2&7\end{pmatrix}=\begin{pmatrix}0.3&0.7&0&0\\0.3&0&0.7&0\\0.1&0.2&0&0.7\\0.1&0&0.2&0.7\end{pmatrix} \qquad [2]$$

(ii) ***Two-step transition probabilities***

We use the fact that $p_{ij}^{(2)}=\left(P^2\right)_{ij}$.

$$P^2=\frac{1}{100}\begin{pmatrix}3&7&0&0\\3&0&7&0\\1&2&0&7\\1&0&2&7\end{pmatrix}\begin{pmatrix}3&7&0&0\\3&0&7&0\\1&2&0&7\\1&0&2&7\end{pmatrix}=\frac{1}{100}\begin{pmatrix}30&21&49&0\\16&35&0&49\\16&7&28&49\\12&11&14&63\end{pmatrix} \qquad [3]$$

(iii) ***Probability of being at maximum discount in 5 years***

We shall represent the states 0%, 20%, 40% and 60% by 0,1,2 and 3 respectively. In order to calculate $p_{0,3}^{(5)}$ we can use $\left(P^5\right)_{0,3}=\sum_{k=0}^{3}\left(P^2\right)_{0,k}\left(P^3\right)_{k,3}$. So we can first calculate the fourth column of P^3:

$$P^3=\frac{1}{1{,}000}\begin{pmatrix}30&21&49&0\\16&35&0&49\\16&7&28&49\\12&11&14&63\end{pmatrix}\begin{pmatrix}3&7&0&0\\3&0&7&0\\1&2&0&7\\1&0&2&7\end{pmatrix}=\frac{1}{1{,}000}\begin{pmatrix}*&*&*&343\\ *&*&*&343\\ *&*&*&539\\ *&*&*&539\end{pmatrix}$$ [2]

Now we have:

$$\left(P^5\right)_{0,3}=\sum_{k=0}^{3}\left(P^2\right)_{0,k}\left(P^3\right)_{k,3}=\frac{1}{100{,}000}\left(30\times 343+21\times 343+49\times 539\right)$$

$$=\frac{43{,}904}{100{,}000}=0.43904$$ [3]

(iv) ***Long-term proportions on each discount level***

This is equivalent to finding the stationary distribution, *ie* solving the matrix equation:

$$\left(\pi_0\ \pi_1\ \pi_2\ \pi_3\right)\begin{pmatrix}3&7&0&0\\3&0&7&0\\1&2&0&7\\1&0&2&7\end{pmatrix}=10\left(\pi_0\ \pi_1\ \pi_2\ \pi_3\right)$$

This matrix equation is equivalent to the simultaneous equations:

$$\begin{aligned}3\pi_0+3\pi_1+\pi_2+\pi_3&=10\pi_0\\7\pi_0+2\pi_2&=10\pi_1\\7\pi_1+2\pi_3&=10\pi_2\\7\pi_2+7\pi_3&=10\pi_3\end{aligned}$$ [1]

We can ignore one of the equations, say the first. Rearranging we get:

$$\begin{aligned}7\pi_0-10\pi_1+2\pi_2&=0\\7\pi_1-10\pi_2+2\pi_3&=0\\7\pi_2-3\pi_3&=0\end{aligned}$$ [1]

Use π_3 (say) as the working variable. From the third equation we have $\pi_2 = \frac{3}{7}\pi_3$. Substituting in the second:

$$\pi_1 = \left(\frac{10}{7} \times \frac{3}{7} - \frac{2}{7}\right)\pi_3 = \frac{16}{49}\pi_3 \qquad [1]$$

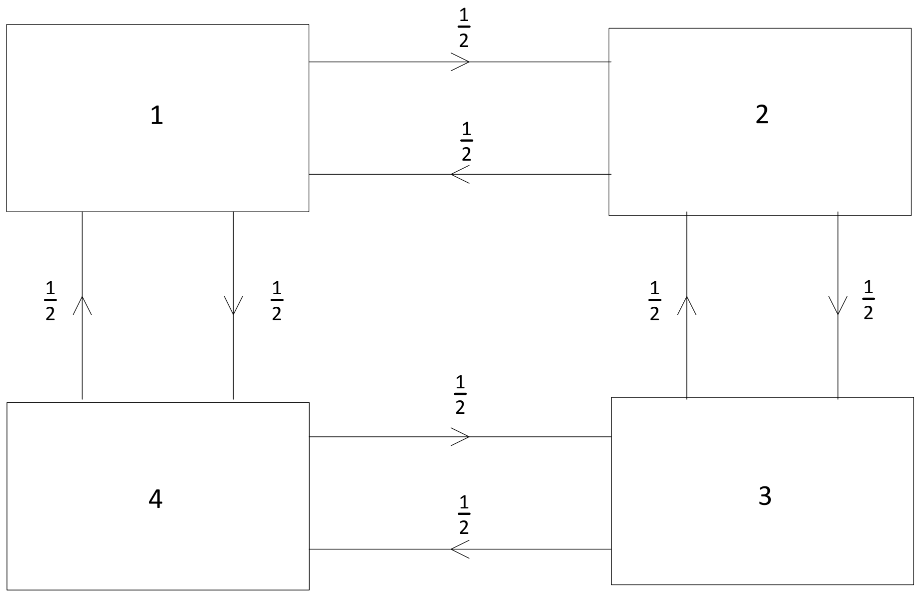
Finally from the first equation $\pi_0 = \left(\frac{10}{7} \times \frac{16}{49} - \frac{2}{7} \times \frac{3}{7}\right)\pi_3 = \frac{118}{343}\pi_3$. [1]

So we have the stationary distribution $(118,112,147,343)\frac{\pi_3}{343}$. Since the probabilities must sum to 1, the stationary distribution is:

$$\frac{1}{720}(118,112,147,343) = (0.1639, 0.1556, 0.2042, 0.4764) \qquad [1]$$

2.9 *Chain I*

(i) *Transition diagram*



[1]

(ii) *Probabilities*

The initial distribution is $(1,0,0,0)$. Repeated postmultiplication of this vector by the transition matrix for Chain I gives:

$$(1,0,0,0) \to \left(0,\tfrac{1}{2},0,\tfrac{1}{2}\right) \to \left(\tfrac{1}{2},0,\tfrac{1}{2},0\right) \to \left(0,\tfrac{1}{2},0,\tfrac{1}{2}\right) \to \left(\tfrac{1}{2},0,\tfrac{1}{2},0\right) \to \cdots$$

So:

(a) $P(X_2 = 1 \mid X_0 = 1) = \frac{1}{2}$ [1]

(b) $P(X_4 = 1 | X_0 = 1) = \frac{1}{2}$ [1]

Alternatively, because we are only asked about particular probabilities, we could evaluate all the possible paths corresponding to each event and add their probabilities.

$P(X_2 = 1 | X_0 = 1)$

Time	0	1	2	Probability
Path	1	2	1	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
Path	1	4	1	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

$P(X_4 = 1 | X_0 = 1)$

Time	0	1	2	3	4	Probability
Path	1	2	1	2	1	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$
Path	1	2	3	2	1	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$
Path	1	4	1	4	1	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$
Path	1	4	3	4	1	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$
Path	1	2	1	4	1	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$
Path	1	4	1	2	1	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$
Path	1	2	3	4	1	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$
Path	1	4	3	2	1	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

(iii) ***Irreducible and/or aperiodic?***

Chain I is irreducible since every state can be reached from every other state. [½]

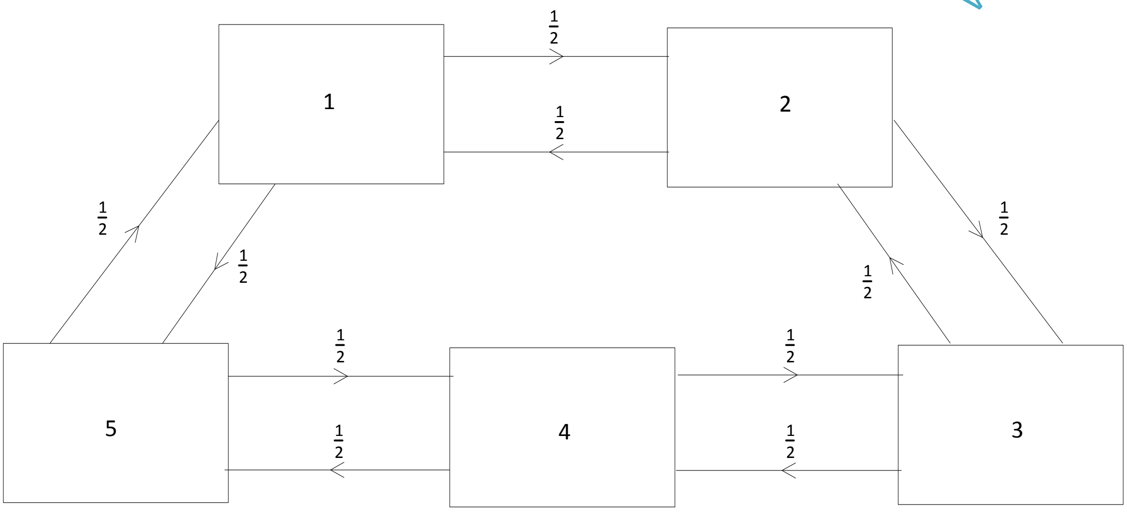
Because the chain is irreducible every state will have the same period. It is possible to return to state 1 in 2, 4, 6, 8 … moves. State 1 has a period of 2 and so every state has a period of 2. The chain is not aperiodic. [1]

(iv) ***Will the process converge to a stationary distribution?***

The process has a finite number of states and is irreducible, so it has a unique stationary distribution, but this process will not converge to its stationary distribution. In the solution to part (ii), we saw that the distribution will alternate between $\left(\frac{1}{2}, 0, \frac{1}{2}, 0\right)$ and $\left(0, \frac{1}{2}, 0, \frac{1}{2}\right)$. [1]

Chain II

(i) ***Transition diagram***



[1]

(ii) ***Probabilities***

The initial distribution is $(1,0,0,0,0)$. Repeated postmultiplication of this vector by the transition matrix for Chain II gives:

$$(1,0,0,0,0) \to \left(0,\tfrac{1}{2},0,0,\tfrac{1}{2}\right) \to \left(\tfrac{1}{2},0,\tfrac{1}{4},\tfrac{1}{4},0\right) \to \left(0,\tfrac{3}{8},\tfrac{1}{8},\tfrac{1}{8},\tfrac{3}{8}\right) \to \left(\tfrac{3}{8},\ldots,\ldots,\ldots,\ldots\right)$$

So:

(a) $P(X_2 = 1 \mid X_0 = 1) = \frac{1}{2}$ [1]

(b) $P(X_4 = 1 \mid X_0 = 1) = \frac{3}{8}$ [1]

Alternatively, because we are only asked about particular probabilities, we could evaluate all the possible paths corresponding to each event and add their probabilities.

$P(X_2 = 1 \mid X_0 = 1)$

Time	0	1	2	Probability
Path	1	2	1	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
Path	1	5	1	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

$P(X_4 = 1 | X_0 = 1)$

Time	0	1	2	3	4	Probability
Path	1	2	1	2	1	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$
Path	1	2	3	2	1	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$
Path	1	5	1	5	1	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$
Path	1	2	1	5	1	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$
Path	1	5	1	2	1	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$
Path	1	5	4	5	1	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$

(iii) ***Irreducible and/or aperiodic?***

Chain II is irreducible since every state can be reached from every other state. [½]

Because the chain is irreducible every state will have the same period. It is possible to return to state 1 in 2, 4, 5, 6, 7, 8 … moves. State 1 has a period of 1 (it is aperiodic) and so every state is aperiodic. The chain is aperiodic. [1]

(iv) ***Will the process converge to a stationary distribution?***

Yes. The chain has a finite number of states, is irreducible and is aperiodic. So there will be a unique stationary distribution that the process will conform to in the long term. [1]

By symmetry, this stationary distribution is $\left(\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5}\right)$. [1]

2.10 (i) ***Markov chain***

The process has the Markov property since the probability of moving on to the next chapter does not depend on the number of chapters currently written (so it is not dependent on the past history of the process). [1]

In fact, we have:

$$X_k = \begin{cases} X_{k-1}+1 & \text{with probability } 0.75 \\ X_{k-1} & \text{with probability } 0.25 \end{cases}$$

for $X_{k-1} \neq 20$ *and* $P(X_k = 20 | X_{k-1} = 20) = 1$.

X_t has a discrete state space, namely $\{0,1,2,...,20\}$, and a discrete time set since the value of the process is recorded at the end of each week. So the process is a Markov chain. [1]

(ii) ***Probability***

To calculate the probability that the book is finished in exactly 25 weeks, we need the probability that the last chapter is completed in the 25th week and, in the first 24 weeks there were 5 chapters rewritten. So the probability is:

$$\binom{24}{5}0.25^5 \times 0.75^{19} \times 0.75 = 0.13163 \qquad [2]$$

(iii) ***Expected number of weeks until completion***

Let m_k be the expected time until the book is finished, given that there are currently k chapters completed. Then, for $k = 0,1,...,19$:

$$m_k = 1 + 0.75m_{k+1} + 0.25m_k \qquad [1]$$

That is, in one week's time, there is a 75% chance of having $k+1$ completed chapters and a 25% chance of still having k completed chapters.

Rearranging this equation, we get:

$$0.75m_k = 1 + 0.75m_{k+1}$$

or:

$$m_k = \frac{1}{0.75} + m_{k+1} \qquad [½]$$

Since $m_{20} = 0$, we have:

$$m_{19} = \frac{1}{0.75}$$

$$m_{18} = \frac{1}{0.75} + \frac{1}{0.75} = \frac{2}{0.75}$$

$$m_{17} = \frac{1}{0.75} + \frac{2}{0.75} = \frac{3}{0.75}$$

and so on. In general, we have:

$$m_k = \frac{20-k}{0.75} \qquad [1]$$

So the expected time until the book is completed is:

$$m_0 = \frac{20}{0.75} = 26.67 \text{ weeks} \qquad [½]$$

Alternatively, let N denote the number of weeks it takes to complete the book. The possible values of N are 20, 21, 22, … and:

$$P(N=20)=0.75^{20}$$

$$P(N=21)=\binom{20}{1}0.25\times 0.75^{20}=\binom{20}{19}0.25\times 0.75^{20}$$

$$P(N=22)=\binom{21}{2}0.25^2\times 0.75^{20}=\binom{21}{19}0.25^2\times 0.75^{20}$$

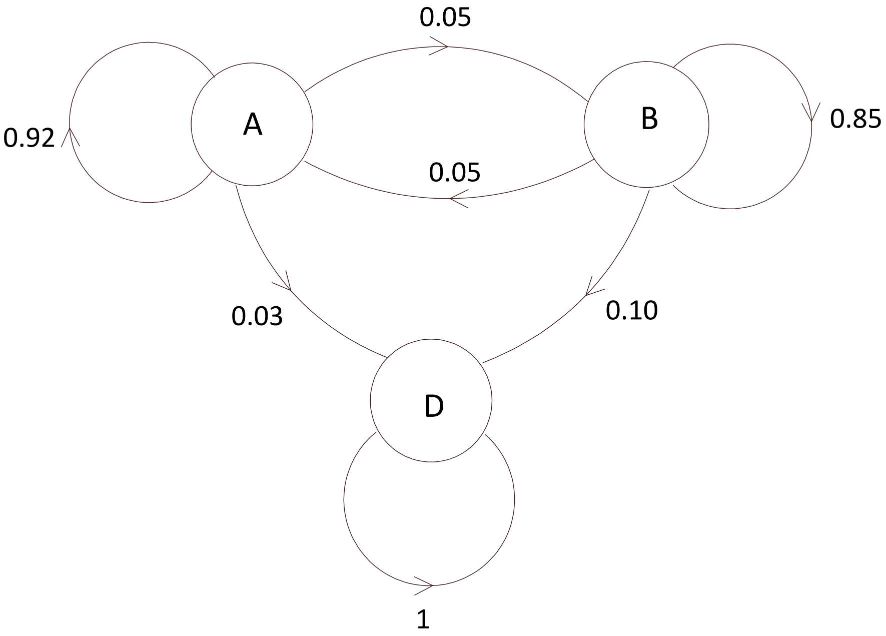
and so on. So N has a Type 1 negative binomial distribution with $k=20$ and $p=0.75$. Hence:

$$E(N)=\frac{k}{p}=\frac{20}{0.75}=26.67 \text{ weeks}$$

2.11 *This question is Subject CT4, September 2006, Question A4.*

(i) ***Probability of never being rated B in the future***

We have the following transition diagram:



A company that is never rated B in the future will:

(a) remain in State A for some period of time, and

(b) then move to State D and remain there.

So we can sum over all future times at which the single transition from State A to State D can take place. This gives us the following expression:

$$0.03+0.92\times 0.03+(0.92)^2\times 0.03+(0.92)^3\times 0.03+\cdots$$ [1]

This is an infinite geometric progression, whose sum is:

$$\frac{0.03}{1-0.92}=0.375$$ [1]

So the probability that a company is never rated B in the future is 0.375.

(ii)(a) ***Second-order transition probabilities***

The second-order transition probabilities are given by:

$$P^2=\begin{pmatrix}0.92&0.05&0.03\\0.05&0.85&0.1\\0&0&1\end{pmatrix}\begin{pmatrix}0.92&0.05&0.03\\0.05&0.85&0.1\\0&0&1\end{pmatrix}=\begin{pmatrix}0.8489&0.0885&0.0626\\0.0885&0.7250&0.1865\\0&0&1\end{pmatrix}$$ [1]

(ii)(b) ***Expected number of defaults***

The probability that a company rated A at time zero is in State D at time 2 is 0.0626. So the expected number of companies in this state out of 100 is 6.26. [1]

(iii) ***Expected number of defaults***

For this manager we use the original matrix P . After one year, the expected number of companies in each state will be:

$$\begin{pmatrix}100&0&0\end{pmatrix}\begin{pmatrix}0.92&0.05&0.03\\0.05&0.85&0.1\\0&0&1\end{pmatrix}=\begin{pmatrix}92&5&3\end{pmatrix}$$ [½]

If the five state B's are moved to State A and the process repeated, we have:

$$\begin{pmatrix}97&0&3\end{pmatrix}\begin{pmatrix}0.92&0.05&0.03\\0.05&0.85&0.1\\0&0&1\end{pmatrix}=\begin{pmatrix}89.24&4.85&5.91\end{pmatrix}$$ [1]

So the expected number of defaults by the end of the second year under this arrangement is 5.91. [½]

(iv) ***Comment***

The downgrade trigger strategy will reduce the expected number of defaults, as we have seen. However, the return on the portfolio will also be a function of the yields on the debt. Companies rated B are likely to have bonds with a higher yield (because of the higher risk), so excluding these may in fact reduce the yield on the portfolio. [1]

Also, the actual number of defaults may not match the expected number. The return depends on the actual progress of the portfolio, rather than the expected outcome. [½]

There will also be a cost incurred when buying and selling bonds. [½]

3

The two-state Markov model and the Poisson model

Syllabus objectives

4.1 Explain the concept of survival models.

4.1.8 Describe the two-state model of a single decrement and compare its assumptions with those of the random lifetime model. (This model will be discussed in detail in Chapter 6.)

4.3 Derive maximum likelihood estimators for transition intensities.

4.3.1 Describe an observational plan in respect of a finite number of individuals observed during a finite period of time, and define the resulting statistics, including the waiting times.

4.3.2 Derive the likelihood function for constant transition intensities in a Markov model of transfers between states given the statistics in 4.3.1.

4.3.3 Derive maximum likelihood estimators for the transition intensities in 4.3.2 and state their asymptotic joint distribution.

4.3.4 State the Poisson approximation to the estimator in 4.3.3 in the case of a single decrement.

0 Introduction

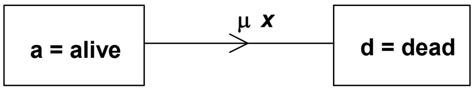
In this chapter we consider a formulation of the problem in which we analyse the random process by which a life passes from one state (alive) to another (dead). The results are consistent with those that we obtain when we model a person's future lifetime as a continuous random variable. We will discuss this alternative model in Chapter 6.

The model discussed in this chapter is an example of a Markov jump process. These processes are discussed further in Chapters 4 and 5, where we consider models with multiple states.

This chapter is based on the paper 'An Actuarial Survey of Statistical Models for Decrement and Transition Data' by A S Macdonald, BAJ 2 (1996), by kind permission of the editor of BAJ.

1 The two-state Markov model

The two-state model is illustrated in the figure below. There is an alive state and a dead state, with transitions in one direction only.



We define a transition probability ${}_tq_x$ where:

$${}_tq_x = P\left[\text{person in the dead state at age } x+t \,\middle|\, \text{in the alive state at age } x\right]$$

and an occupancy or survival probability ${}_tp_x$ where:

$${}_tp_x = P\left[\text{person in the alive state at age } x+t \,\middle|\, \text{in the alive state at age } x\right]$$

The probability that a life alive at a given age will be dead at any subsequent age is governed by the age-dependent transition intensity μ_{x+t} $(t \geq 0)$ in a way made precise by Assumption 2 below.

Transition intensities are also sometimes called *forces of transition* or *transition rates*.

1.1 Assumptions underlying the model

There are three assumptions underlying the simple two-state model.

Assumption 1

The probabilities that a life at any given age will be found in either state at any subsequent age depend only on the ages involved and on the state currently occupied. This is the *Markov* assumption.

So, past events do not affect the probability of a future event.

In particular, the past history of an individual – for example, current state of health, spells of sickness, occupation – is excluded from the model. If we knew these factors, we could:

(a) treat each combination of factors as a separate model; in other words, *stratify* the problem; or

(b) specify a model which took them into account; in other words, treat the problem as one of *regression*.

We will consider approach (b) in Chapter 8, where we look at proportional hazards models.

Assumption 2

For a short time interval of length *dt*:

$$_{dt}q_{x+t} = \mu_{x+t}\,dt + o(dt) \qquad (t \geq 0)$$

In other words, the probability of dying in a very short time interval of length dt is equal to the transition intensity multiplied by the time interval, plus a small correction term. This is equivalent to assuming that $_{dt}q_{x+t} \approx \mu_{x+t}\,dt$.

Remember that a function $g(t)$ is said to be '$o(t)$' if $\lim\limits_{t\to 0}\dfrac{g(t)}{t} = 0$, in other words if $g(t)$ tends to zero 'faster' than t itself. Where we are not concerned about the precise form of $g(t)$, we can use the term $o(t)$ in an equation to denote any function that is $o(t)$.

Assumption 2 can also be stated as follows:

$$\mu_{x+t} = \lim_{dt\to 0}\frac{_{dt}q_{x+t}}{dt}$$

For the purpose of inference, we restrict our attention to ages between x and $x+1$, and introduce a further assumption.

Assumption 3

μ_{x+t} is a constant μ for $0 \leq t < 1$.

Our investigation will consist of many observations of small segments of lifetimes, *ie* single years of age. Assumption 3 simplifies the model by treating the transition intensity as a constant for all individuals aged x last birthday. This does not mean that we believe that the transition intensity will increase by a discrete step when an individual reaches age $x+1$, although this is a consequence of the assumption.

1.2 Comparison with other models

It is important to emphasise that this two-state model is not the same as the model based on the future lifetime random variable T_x, which is discussed in Chapter 6; we start with different assumptions. The model in Chapter 6 is formulated in terms of a random variable T representing future lifetime. The model in this chapter is in terms of a transition intensity between states. It is easy to impose some mild conditions under which the models are equivalent, but when we consider more than one decrement these two formulations lead in different directions.

We will consider models with more than one decrement in Chapters 4 and 5, including a simple multiple-state model with three states: healthy, sick and dead. In that particular model, a life in the healthy state can move to the sick state or the dead state. Similarly, a life in the sick state can move to the healthy state or the dead state. For now, though, we will concentrate on the simple two-state model to derive some important results, many of which can be generalised to multiple-state models.

2 Survival probabilities

Since we have specified the model in terms of a transition intensity, we must see how to compute transition probabilities.

Consider the survival probability ${}_{t+dt}p_x$, and condition on the state occupied at age $x+t$.

Here we are thinking about the probability of surviving from age x to $x+t$ and onwards from age $x+t$ to $x+t+dt$.

By the Markov assumption (Assumption 1), nothing else affects the probabilities of death or survival after age $x+t$.

$$\begin{aligned}
{}_{t+dt}p_x &= {}_tp_x \times P[\text{Alive at } x+t+dt \mid \text{Alive at } x+t] \\
&\qquad + {}_tq_x \times P[\text{Alive at } x+t+dt \mid \text{Dead at } x+t] \\
&= ({}_tp_x \times {}_{dt}p_{x+t}) + ({}_tq_x \times 0) \\
&= {}_tp_x \times \left(1-\mu_{x+t}\,dt + o(dt)\right)
\end{aligned}$$

The last equality follows from Assumption 2.

Therefore:

$$\begin{aligned}
\frac{\partial}{\partial t}\,{}_tp_x &= \lim_{dt\to 0+} \frac{{}_{t+dt}p_x - {}_tp_x}{dt} \\
&= -{}_tp_x\mu_{x+t} + \lim_{dt\to 0+}\frac{o(dt)}{dt} \\
&= -{}_tp_x\mu_{x+t} \qquad\qquad (3.1)
\end{aligned}$$

So:

$${}_tp_x = \exp\left(-\int_0^t \mu_{x+s}\,ds\right)$$

Question

Show that the solution of the differential equation $\frac{\partial}{\partial t}\,{}_tp_x = -{}_tp_x\,\mu_{x+t}$ is:

$${}_tp_x = \exp\left(-\int_0^t \mu_{x+s}\,ds\right)$$

A reminder of two techniques that can be used to solve first-order differential equations is given in the appendix in Section 8 of this chapter.

Solution

Separating the variables gives:

$$\frac{\frac{\partial}{\partial t}\,{}_t p_x}{{}_t p_x} = -\mu_{x+t}$$

or equivalently:

$$\frac{\partial}{\partial t}\ln {}_t p_x = -\mu_{x+t}$$

Changing the variable from t to s (since we want to use t as one of the limits of the integral), this becomes:

$$\frac{\partial}{\partial s}\ln {}_s p_x = -\mu_{x+s}$$

Integrating both sides of this equation with respect to s between the limits $s=0$ and $s=t$ gives:

$$\int_0^t \frac{\partial}{\partial s}\ln {}_s p_x \, ds = \left[\ln {}_s p_x\right]_0^t = -\int_0^t \mu_{x+s}\, ds$$

So:

$$\ln {}_t p_x - \ln {}_0 p_x = -\int_0^t \mu_{x+s}\, ds$$

Now ${}_0 p_x = 1$, since this is the probability that a life aged x survives for at least 0 years, and $\ln {}_0 p_x = 0$.

Taking exponentials then gives:

$${}_t p_x = \exp\left(-\int_0^t \mu_{x+s}\, ds\right)$$

as required.

This is an extremely important result. It is restated below and it is also given on page 32 of the *Tables*.

Relationship between survival probabilities and force of mortality

For any $x \geq 0$ and $t \geq 0$:

$${}_t p_x = \exp\left(-\int_0^t \mu_{x+s}\, ds\right)$$

Assumption 3 has not been used so far. Under the assumption that the force of mortality is constant between exact age x and exact age $x+t$, this result simplifies to give:

$${}_tp_x = e^{-\mu t}$$

where μ represents the constant force.

We will obtain the same result in Chapter 6 when we formulate a survival model in terms of T, the lifetime distribution.

The important point is that it has been derived here strictly from the assumptions of the two-state model, and that the method is easily extended to models with more states. In the Markov framework, (3.1) is an example of the *Kolmogorov forward (differential) equations*. These are discussed in detail in Chapters 4 and 5.

3 Statistics

3.1 Definitions

Next we define our observations.

We suppose that we observe a total of N lives during some finite period of observation, between the ages of x and $x+1$.

We could suppose that lives were observed, or not, as a result of some random mechanism (not depending on any parameter of interest), but here we suppose that data are analysed retrospectively, so we regard N as a non-random quantity. We need not assume that we observe the N lives simultaneously, nor need we assume that we observe each life for the complete year of age. We do assume that all N lives are identical and statistically independent.

In reality no two lives are truly identical. Here we are using the word 'identical' to refer to the fact that all the lives follow the same stochastic model of living and dying. So the lives will all have the same value of μ, but they won't all die at the same time.

For $i=1,\ldots,N$ define:

- **$x+a_i$ to be the age at which observation of the ith life starts**
- **$x+b_i$ to be the age at which observation of the ith life must cease if the life survives to that age.**

$x+b_i$ will be either $x+1$, or the age of the ith life when the investigation ends, whichever is smaller.

For simplicity we consider Type I censoring.

This means that the value of b_i is known, when the period of observation starts at $x+a_i$. So b_i is a fixed number, and not a random variable. If we plan to observe a life from 52.25 until 52.75, then $b_i = 0.75$, but of course not all lives will survive to the end of the planned period of observation. To complete our model we will need another random variable that measures whether it was death before 52.75 or survival to 52.75, that ended the period of observation.

The approach can be extended to more realistic forms of censoring.

In other words, we could modify the derivation to allow for lives leaving the investigation at random times through decrements other than death.

In this case b_i would be a random variable. If we plan to observe a life from 52.25 until death or retirement, whichever event occurs first, then b_i is a random variable (and to complete our model we will need another random variable that measures whether it was death or retirement that ended the period of observation).

Types of censoring are considered in more detail in Chapter 7.

Define a random variable D_i as follows:

$$D_i = \begin{cases} 1 & \text{if the } i\text{th life is observed to die} \\ 0 & \text{if the } i\text{th life is not observed to die} \end{cases}$$

D_i is an example of an indicator random variable; it indicates the occurrence of death.

In the above definition, we are talking about whether or not the i th life dies during the planned observation period from age $x+a_i$ to age $x+b_i$. D_i is the extra random variable that completes our model.

The expected value of D_i is:

$$E\left[D_i\right]=0\times P\left[D_i=0\right]+1\times P\left[D_i=1\right]=P\left[D_i=1\right]={}_{b_i-a_i}q_{x+a_i}$$

ie $E\left[D_i\right]$ is just the probability of a death being observed.

Define a random variable T_i as follows:

$$x+T_i = \text{the age at which observation of the } i\text{th life ends}$$

Notice that D_i and T_i are not independent, since:

$$D_i=0 \Leftrightarrow T_i=b_i$$

ie if no death has been observed, the life must have survived to $x+b_i$.

$$D_i=1 \Leftrightarrow a_i<T_i<b_i$$

ie an observed death must have occurred between $x+a_i$ and $x+b_i$.

It will often be useful to work with the time spent under observation, so define:

$$V_i=T_i-a_i$$

V_i is called the *waiting time*. It has a mixed distribution, with a probability mass at the point b_i-a_i.

A mixed distribution has a discrete and a continuous part.

For example, suppose that observation of life i begins at exact age 82 years and 3 months and observation will continue until the earlier of the life's 83rd birthday or death. In this case, $a_i=0.25$, $b_i=1$ and V_i is a random variable taking values between 0 and 0.75. V_i has a mixed distribution with a probability mass at 0.75.

3.2 Joint density function

The pair (D_i,V_i) comprise a *statistic*, meaning that the outcome of our observation is a sample (d_i,v_i) drawn from the distribution of (D_i,V_i).

Let $f_i(d_i,v_i)$ be the joint distribution of (D_i,V_i).

It is easily written down by considering the two cases $D_i=0$ and $D_i=1$.

If $D_i=0$, no death has been observed and the life is known to have survived for the period of length b_i-a_i from exact age $x+a_i$ to exact age $x+b_i$.

If $D_i=1$, the life is known to have survived for the period v_i $(0<v_i<b_i-a_i)$ from exact age $x+a_i$ to exact age $x+a_i+v_i$ before dying at exact age $x+a_i+v_i$.

Therefore, $f_i(d_i,v_i)$ has a distribution that is specified by the following expression, which is a combination of a probability mass (corresponding to $d_i=0$) and a probability density (corresponding to $d_i=1$):

$$f_i(d_i,v_i)=\begin{cases} {}_{b_i-a_i}p_{x+a_i} & (d_i=0) \\ {}_{v_i}p_{x+a_i}\cdot\mu_{x+a_i+v_i} & (d_i=1) \end{cases}$$

$$=\begin{cases} \exp\left(-\int_0^{b_i-a_i}\mu_{x+a_i+t}\,dt\right) & (d_i=0) \\ \exp\left(-\int_0^{v_i}\mu_{x+a_i+t}\,dt\right)\mu_{x+a_i+v_i} & (d_i=1) \end{cases}$$

$$=\exp\left(-\int_0^{v_i}\mu_{x+a_i+t}\,dt\right)\mu_{x+a_i+v_i}^{d_i}$$

Now assume that μ_{x+t} is a constant μ for $0\le t<1$ (this is the first time we have needed Assumption 3) and so $f_i(d_i,v_i)$ takes on the simple form:

$$f_i(d_i,v_i)=e^{-\mu v_i}\,\mu^{d_i}$$

We can then write down an expression for the joint probability function, provided that we can assume that the lifetimes of all the lives involved are statistically independent.

The joint probability function of all the (D_i,V_i), by independence, is proportional to:

$$\prod_{i=1}^{N}e^{-\mu v_i}\,\mu^{d_i}=e^{-\mu(v_1+\ldots+v_N)}\,\mu^{d_1+\ldots+d_N}=e^{-\mu v}\,\mu^{d}$$

where $d=\sum_{i=1}^{N}d_i$ and $v=\sum_{i=1}^{N}v_i$.

In other words, define random variables D and V to be the total number of deaths and the total waiting time, respectively, and the joint probability function of all the (D_i,V_i) can be simply expressed in terms of D and V.

For a known transition intensity, we can calculate the likelihood of any combination of deaths and waiting time. However, in practice the value of the transition intensity is unknown. We use statistical inference to calculate the value of the transition intensity that is most plausible given the observed data, *ie* the maximum likelihood estimate of μ. This is the subject of the next section.

Question

Suppose that observation of life i begins at exact age 82 years and 3 months and observation will continue until the earlier of the life's 83rd birthday or death. Assuming that $\mu = 0.1$, determine:

(i) the probability function of D_i

(ii) $E[D_i]$

(iii) the probability density/mass function of V_i

(iv) $E[V_i]$.

Solution

We will need to use the result from Section 2 that:

$${}_t p_x = \exp\left(-\int_0^t \mu_{x+s}\, ds\right) = e^{-\mu t} = e^{-0.1t}$$

(i) ***Probability function***

The random variable D_i can only take the value 0 or 1. Its probability function is:

$$P[D_i = 0] = {}_{0.75}p_{82.25} = e^{-0.1\times 0.75} = 0.9277$$

$$P[D_i = 1] = 1 - 0.9277 = 0.0723$$

(ii) ***Expected value of D_i***

The expected value of D_i is:

$$E[D_i] = 0\times 0.9277 + 1\times 0.0723 = 0.0723$$

(iii) ***Probability density/mass function***

The probability density/mass function of V_i is:

$$f(v_i) = \begin{cases} {}_{v_i}p_{82.25}\,\mu_{82.25+v_i} & \text{if } v_i < 0.75 \\ {}_{0.75}p_{82.25} & \text{if } v_i = 0.75 \end{cases}$$

$$= \begin{cases} 0.1e^{-0.1v_i} & \text{if } v_i < 0.75 \\ 0.9277 & \text{if } v_i = 0.75 \end{cases}$$

(iv) ***Expected value of V_i***

The expected value of V_i is:

$$E[V_i] = \int_0^{0.75} t \times 0.1e^{-0.1t}\,dt + 0.9277 \times 0.75$$

Integrating by parts:

$$\int_0^{0.75} t \times 0.1e^{-0.1t}\,dt = \left[-t\,e^{-0.1t}\right]_0^{0.75} + \int_0^{0.75} e^{-0.1t}dt$$

$$= -0.75\,e^{-0.075} - 10\left[e^{-0.1t}\right]_0^{0.75}$$

$$= -0.6958 - 10 \times \left[e^{-0.075} - 1\right]$$

$$= -0.6958 + 0.7226 = 0.0268$$

So:

$$E[V_i] = 0.0268 + 0.9277 \times 0.75 = 0.7226$$

As expected, the answer to (iv) is just less than 0.75.

4 The maximum likelihood estimator

Maximum likelihood estimation is covered in Subject CS1. A brief reminder of the process is given below.

Maximum likelihood estimation

The steps involved in maximum likelihood estimation are as follows:

- write down the likelihood function L – this is the probability/PDF of obtaining the values we have observed
- take logs and simplify the resulting expression
- differentiate the log-likelihood with respect to each parameter to be estimated – this will involve partial differentiation if there is more than one parameter to be estimated
- set the derivatives equal to 0 and solve the equations simultaneously
- check that the resulting values are maxima. This will usually involve differentiating a second time. Strictly speaking, when there are two or more parameters to estimate, checking for maxima involves examination of the Hessian matrix, which is beyond the scope of Subject CS2.

4.1 Maximising the likelihood function

We have already seen that the joint probability function of all the (D_i, V_i) is:

$$\prod_{i=1}^{N} e^{-\mu v_i} \mu^{d_i} = e^{-\mu(v_1+\ldots+v_N)} \mu^{d_1+\ldots+d_N} = e^{-\mu v} \mu^d$$

where $d = \sum_{i=1}^{N} d_i$ and $v = \sum_{i=1}^{N} v_i$.

This probability function immediately furnishes the likelihood for μ :

$$\boldsymbol{L(\mu; d, v) = e^{-\mu v} \mu^d}$$

which yields the maximum likelihood estimate (MLE) for μ .

Maximum likelihood estimate of μ under the two-state Markov model

$$\boldsymbol{\hat{\mu} = d / v}$$

Recall that we are now assuming that the force of mortality is constant over the year of age $(x, x+1)$.

Question

Show that $\hat{\mu} = \frac{d}{v}$.

Solution

Taking the log of the likelihood given above, we obtain:

$$\log L = -\mu v + d \log \mu$$

Differentiating with respect to μ:

$$\frac{\partial}{\partial \mu} \log L = -v + \frac{d}{\mu}$$

This is equal to 0 when $\mu = \frac{d}{v}$. So there is a stationary point when $\mu = \frac{d}{v}$.

Differentiating again:

$$\frac{\partial^2}{\partial \mu^2} \log L = -\frac{d}{\mu^2}$$

This second derivative is negative when $\mu = \frac{d}{v}$. (In fact, it's always negative.) So we have a maximum.

Hence:

$$\hat{\mu} = \frac{d}{v}$$

It is reassuring that the mathematical approach produces a result that is intuitive, *ie* that the maximum likelihood estimate of the hazard rate is the number of observed deaths divided by the total time for which lives were exposed to the hazard.

The measurement of the total time for which lives are exposed to the hazard is one of the fundamental techniques covered by this course. It enables accurate assessment of risks, from the probability of a policyholder dying to the probability of a claim under a motor insurance policy. The technical term is 'exposed to risk'. We will study it in more detail in Section 6 of this chapter and also in Chapter 10.

4.2 Properties of the maximum likelihood estimator

The *estimate* $\hat{\mu}$, being a function of the sample values *d* and *v*, can itself be regarded as a sample value drawn from the distribution of the corresponding *estimator*.

Maximum likelihood estimator of μ under the two-state Markov model

$$\tilde{\mu} = D / V$$

As usual we are using capital letters to denote random variables, and lower case letters to denote sample values.

So, the estimator $\tilde{\mu}$ is a random variable and the estimate $\hat{\mu}$ is the observed value of that random variable.

It is important in applications to be able to estimate the moments of the estimator $\tilde{\mu}$, for example to compare the experience with that of a standard table. At least, we need to estimate $E[\tilde{\mu}]$ and $\text{var}[\tilde{\mu}]$.

In order to derive the properties of the estimator $\tilde{\mu}$ we will use two results that link the random variables D and V.

The following *exact* results are obtained:

$$E[D_i - \mu V_i] = 0 \tag{3.2}$$

$$\text{var}[D_i - \mu V_i] = E[D_i] \tag{3.3}$$

Note that the first of these can also be written as $E[D_i] = \mu.E[V_i]$.

In the case that the $\{a_i\}$ and $\{b_i\}$ are known constants, this follows from integrating/summing the probability function of (D_i, V_i) over all possible events to obtain:

$$\int_0^{b_i - a_i} e^{-\mu v_i} \mu \, dv_i + e^{-\mu(b_i - a_i)} = 1 \tag{*}$$

and then differentiating with respect to μ, once to obtain the mean and twice to obtain the variance.

We will show how to use this to prove Result (3.2) above in a moment, but first we need to derive formula for $E[D_i]$ and $E[V_i]$. (The derivation of Result (3.3) is covered in the questions at the end of this chapter.)

We have already seen that:

$$E[D_i] = 0 \times P(D_i = 0) + 1 \times P(D_i = 1) = P(D_i = 1) = {}_{b_i - a_i}q_{x+a_i}$$

ie it is the probability that life i dies between exact age $x + a_i$ and exact age $x + b_i$. This can also be expressed in integral form as:

$$\int_0^{b_i - a_i} {}_{v_i}p_{x+a_i} \mu_{x+a_i+v_i} \, dv_i$$

The integrand can be considered to be the 'probability' that the life dies at exact age $x+a_i+v_i$. The probability of dying between exact age $x+a_i$ and exact age $x+b_i$ is obtained by integrating this expression over the relevant range of values of v_i, *ie* from $v_i=0$ to $v_i=b_i-a_i$.

Now, under the assumption that the force of mortality is constant over the year of age $(x,x+1)$:

$$_{v_i}p_{x+a_i}=e^{-\mu v_i}$$

and hence:

$$E[D_i]=\int_0^{b_i-a_i}\mu e^{-\mu v_i}\,dv_i$$

We have also already seen that V_i is a mixed random variable. It can take any value between 0 and b_i-a_i, and has a point mass at b_i-a_i. Its probability/mass function is:

$$f(v_i)=\begin{cases} {}_{v_i}p_{x+a_i}\,\mu_{x+a_i+v_i} & \text{if } v_i<b_i-a_i \\ {}_{b_i-a_i}p_{x+a_i} & \text{if } v_i=b_i-a_i \end{cases}$$

Under the assumption that the force of mortality is constant over the year of age $(x,x+1)$, this is:

$$f(v_i)=\begin{cases} \mu e^{-\mu v_i} & \text{if } v_i<b_i-a_i \\ e^{-\mu(b_i-a_i)} & \text{if } v_i=b_i-a_i \end{cases}$$

So:

$$E[V_i]=\int_0^{b_i-a_i}v_i\mu e^{-\mu v_i}\,dv_i+(b_i-a_i)e^{-\mu(b_i-a_i)}$$

We are now able to prove result (3.2).

Proof of (3.2)

We start from result (*), which is restated below:

$$\int_0^{b_i-a_i}e^{-\mu v_i}\,\mu\,dv_i+e^{-\mu(b_i-a_i)}=1 \qquad (*)$$

Differentiating (*) with respect to μ gives:

$$\int_0^{b_i-a_i}\left(e^{-\mu v_i}-\mu v_i e^{-\mu v_i}\right)dv_i-(b_i-a_i)e^{-\mu(b_i-a_i)}=0$$

(Because the limits of the integral don't depend on μ, this just involves differentiating the expressions inside the integral with respect to μ.)

The equation immediately above can be rewritten as follows:

$$\int_0^{b_i-a_i} e^{-\mu v_i}\,dv_i = \int_0^{b_i-a_i} \mu v_i e^{-\mu v_i}\,dv_i + (b_i-a_i)e^{-\mu(b_i-a_i)} = E[V_i]$$

Multiplying through by μ then gives:

$$\mu \int_0^{b_i-a_i} e^{-\mu v_i}\,dv_i = \mu E[V_i]$$

The expression on the left-hand side is equal to $E[D_i]$. So $E[D_i]-\mu E[V_i]=0$ as required.

4.3 Asymptotic distribution of $\tilde{\mu}$

To find the asymptotic distribution of $\tilde{\mu}$, consider:

$$\frac{1}{N}(D-\mu V)=\frac{1}{N}\sum_{i=1}^{N}(D_i-\mu V_i)$$

We know that $E\left[D_i-\mu V_i\right]=0$ and that $\text{var}\left[D_i-\mu V_i\right]=E\left[D_i\right]$.

So, by the Central Limit Theorem:

$$\frac{1}{N}(D-\mu V) \sim Normal\left(0,\frac{E[D]}{N^2}\right)$$

Now, since $\tilde{\mu}=\frac{D}{V}$, it follows that:

$$\tilde{\mu}-\mu=\frac{D}{V}-\mu=\frac{D-\mu V}{V}=\frac{N}{V}\left(\frac{D-\mu V}{N}\right)$$

Then note that (not rigorously):

$$\lim_{N\to\infty}(\tilde{\mu}-\mu)=\lim_{N\to\infty}\frac{N}{V}\left(\frac{D}{N}-\frac{\mu V}{N}\right)$$

By the law of large numbers, $V/N \to E(V_i)$.

Technically, this refers to 'convergence in probability'. So asymptotically:

$$E(\tilde{\mu}-\mu)=\frac{1}{E(V_i)}E\left(\frac{D-\mu V}{N}\right)=0$$

Also:

$$\text{var}(\tilde{\mu}-\mu) = \text{var}\left[\left(\frac{D-\mu V}{N}\right)\times\frac{1}{E\left(V_i\right)}\right]$$

$$= \frac{E\left(D\right)}{N^2\left[E\left(V_i\right)\right]^2}$$

$$= \frac{E[D]}{\left[E\left(V_1+V_2+\cdots+V_N\right)\right]^2}$$

because $E\left(V_1\right)=E\left(V_2\right)=\cdots=E\left(V_N\right)$.

Now since $V=\sum_{i=1}^{N}V_i$, we have:

$$\text{var}(\tilde{\mu}-\mu)=\frac{E(D)}{[E(V)]^2}$$

So:

$$(\tilde{\mu}-\mu)\sim N\left(0,\frac{E\left(D\right)}{\left[E\left(V\right)\right]^2}\right)$$

But we know that $E(D-\mu V)=0$. So:

$$E(D)=\mu E(V)$$

and:

$$\mu=\frac{E(D)}{E(V)}$$

Hence:

$$\text{var}(\tilde{\mu})=\text{var}(\tilde{\mu}-\mu)=\frac{E(D)}{[E(V)]^2}=\frac{\mu}{E(V)}$$

We now have the following asymptotic result.

Asymptotic distribution of $\tilde{\mu}$

Asymptotically:

$$\tilde{\mu}\sim \textbf{\textit{Normal}}\left(\mu,\frac{\mu}{\mathbf{E[V]}}\right)$$

We can use this result to calculate probabilities and confidence intervals.

Question

A scientist identifies 1,282 newborn wildebeest and observes them during their first year of life on the savannah. The scientist wishes to calculate the constant transition intensity over this period covering all types of death, including natural causes and ending up as a tasty snack for passing carnivores.

If the true transition intensity is 0.18, calculate the probability that the scientist observes a mortality rate in excess of 0.2.

Solution

The expected number of deaths is:

$$E(D) = 1{,}282\,q_0 = 1{,}282(1 - p_0) = 1{,}282(1 - e^{-0.18})$$

Then, using the result $E(D) = \mu E(V)$, we have:

$$E(V) = \frac{E(D)}{\mu} = \frac{1{,}282(1 - e^{-0.18})}{0.18} = 1{,}173.24$$

Alternatively, we could calculate the expected waiting time for the i th animal as follows:

$$\begin{aligned}
E\left[V_i\right] &= \int_0^1 0.18\,t\,e^{-0.18t}\,dt + e^{-0.18} \\
&= \left[-t\,e^{-0.18t}\right]_0^1 + \int_0^1 e^{-0.18t}\,dt + e^{-0.18} \\
&= -\cancel{e^{-0.18}} + \left[-\frac{1}{0.18}e^{-0.18t}\right]_0^1 + \cancel{e^{-0.18}} \\
&= \frac{1}{0.18}\left(1 - e^{-0.18}\right) \\
&= 0.915165
\end{aligned}$$

So the total expected waiting time is:

$$E[V] = 1{,}282 \times E[V_i] = 1{,}173.24$$

as before.

Asymptotically:

$$\tilde{\mu} \sim N\left(\mu, \frac{\mu}{E[V]}\right) \equiv N\left(0.18, \frac{0.18}{1{,}173.24}\right) \equiv N\left(0.18,\ 0.01239^2\right)$$

Hence the required probability is:

$$P(\tilde{\mu} > 0.2) \approx 1 - \Phi\left(\frac{0.2 - 0.18}{0.01239}\right) = 1 - \Phi(1.6147) = 1 - 0.9468 = 0.0532$$

This probability is approximate since the sample size is not that large.

5 Alternative method of obtaining the asymptotic distribution

In this section we describe another way of obtaining the asymptotic distribution of $\tilde{\mu}$, the maximum likelihood estimator of μ.

We have already seen that $\hat{\mu}=\dfrac{d}{v}$, $\tilde{\mu}=\dfrac{D}{V}$ and $\dfrac{d^2\log L}{d\mu^2}=-\dfrac{d}{\mu^2}$.

Now instead of deriving results for the expectation and variance of $D_i-\mu V_i$ as in Section 4.2, we can use the asymptotic properties of maximum likelihood estimators.

These estimators are asymptotically normal and unbiased. So $E\left(\tilde{\mu}\right)=\mu$. It just remains for us to find an expression for $\text{var}\left(\tilde{\mu}\right)$. This is given by the Cramér-Rao lower bound:

$$\text{var}\left(\tilde{\mu}\right)=\frac{-1}{E\left(\dfrac{d^2\log L}{d\mu^2}\right)}=\frac{-1}{E\left(\dfrac{-D}{\mu^2}\right)}=\frac{\mu^2}{E(D)}$$

We are using D rather than d in the line above since we are thinking about the variance of the estimator of μ.

So, asymptotically:

$$\tilde{\mu}\sim N\left(\mu,\frac{\mu^2}{E(D)}\right)$$

This is consistent with the result in Section 4.3 since $E(D)=\mu E(V)$.

In practice, we will not know the exact variance, so we need to estimate it. This can be done by:

- replacing μ by $\hat{\mu}$, its estimated value, and
- replacing $E(D)$ by d, the observed number of deaths.

This gives:

$$\text{var}\left(\tilde{\mu}\right)\approx\frac{\hat{\mu}^2}{d}$$

Also, since $\hat{\mu}=\dfrac{d}{v}$, we have:

$$\frac{\hat{\mu}^2}{d}=\frac{\hat{\mu}}{d}\times\frac{d}{v}=\frac{\hat{\mu}}{v}$$

So:

$$\text{var}\left(\tilde{\mu}\right) \approx \frac{\hat{\mu}}{v}$$

In actuarial terminology, the observed waiting time at age *x*, which we have denoted *v*, is often called the *central exposed to risk* and is denoted E_x^c .

We use this notation in the next section on the Poisson model and consider this concept in more detail later in this course. In the meantime you should be prepared to use either term.

6 The Poisson model

6.1 The Poisson distribution

The Poisson distribution is a discrete probability distribution in which the random variable can only take non-negative integer values.

A random variable X is said to have a Poisson distribution with mean $\lambda\,(\lambda > 0)$ if the probability function of X is:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad \text{for } x = 0,1,2,\ldots$$

Remember that $E(X) = \lambda$ and $\text{var}(X) = \lambda$.

The Poisson distribution is used to model the number of times 'rare' events occur during a given period of time, *eg* the number of particles emitted by a radioactive source in a minute. Such analogies suggest the Poisson distribution could be used as a model for the number of deaths among a group of lives, given the time spent exposed to risk.

6.2 The Poisson model of mortality

If we assume that we observe N individuals as before, for a total of E_x^c person-years, and that the force of mortality is a constant μ , then a Poisson model is given by the assumption that D has a Poisson distribution with parameter μE_x^c. That is:

$$P(D = d) = \frac{e^{-\mu E_x^c}(\mu E_x^c)^d}{d\,!}$$

As before we are assuming that the lives are independent and identical in terms of their mortality.

Under the observational plan described above, the Poisson model is not an exact model, since it allows a non-zero probability of more than N deaths, but it is often a very good approximation.

The probability of more than N deaths is usually negligible.

Question

A large computer company always maintains a workforce of exactly 5,000 young workers, immediately replacing any worker who leaves.

Use the Poisson model to calculate the probability that there will be fewer than 3 deaths during any 6-month period, assuming that all workers experience a constant force of mortality of 0.0008 per annum.

Solution

We have a constant population of 5,000 individuals with a constant force of mortality of 0.0008 *pa*. If we assume that deaths are independent, the Poisson model applies and the number of deaths during any 6-month period has a Poisson distribution with mean:

$$0.0008\times 5{,}000\times\frac{6}{12}=2$$

So:

$$P(\text{No deaths})=e^{-2}=0.1353$$

$$P(\text{Exactly 1 death})=2e^{-2}=0.2707$$

$$P(\text{Exactly 2 deaths})=\frac{2^2}{2!}e^{-2}=0.2707$$

and hence the probability of fewer than 3 deaths is:

$$0.1353+0.2707+0.2707=0.6767$$

6.3 Estimating the underlying force of mortality

We would like to use our knowledge about the number of deaths observed and the total exposed to risk (waiting time) to estimate the unknown true force of mortality.

The Poisson likelihood leads to the following estimator of (constant) μ .

Maximum likelihood estimator of μ under the Poisson model

$$\tilde{\mu}=\frac{\boldsymbol{D}}{\boldsymbol{E_x^c}}$$

Question

Derive the above formula for the maximum likelihood estimator of μ .

Solution

The likelihood of observing d deaths if the true value of the hazard rate is μ is:

$$L(\mu)=\frac{(\mu E_x^c)^d e^{-\mu E_x^c}}{d!}$$

This can be maximised by maximising its log:

$$\log L(\mu) = d(\log \mu + \log E_x^c) - \mu E_x^c - \log d!$$

Differentiating with respect to μ :

$$\frac{\partial}{\partial \mu} \log L(\mu) = \frac{d}{\mu} - E_x^c$$

This is zero when:

$$\mu = \frac{d}{E_x^c}$$

This is a maximum since $\frac{\partial^2}{\partial \mu^2} \log L(\mu) = -\frac{d}{\mu^2} < 0$.

So $\hat{\mu} = \frac{d}{E_x^c}$ is the maximum likelihood estimate of μ . It is the realised value of the maximum likelihood estimator $\tilde{\mu} = \frac{D}{E_x^c}$.

The estimator $\tilde{\mu}$ has the following properties:

(i) $E[\tilde{\mu}] = \mu$

So $\tilde{\mu}$ is an unbiased estimator of μ .

(ii) $\text{var}[\tilde{\mu}] = \frac{\mu}{E_x^c}$

In practice, we will substitute $\hat{\mu}$ for μ to estimate these from the data.

Question

Prove these results for $E[\tilde{\mu}]$ and $\text{var}[\tilde{\mu}]$.

Solution

If $D \sim Poisson(\mu E_x^c)$, where E_x^c is a fixed quantity, then $E[D] = \text{var}[D] = \mu E_x^c$ and hence:

$$E[\tilde{\mu}] = E\left[\frac{D}{E_x^c}\right] = \frac{E[D]}{E_x^c} = \frac{\mu E_x^c}{E_x^c} = \mu$$

$$\text{var}[\tilde{\mu}] = \text{var}\left[\frac{D}{E_x^c}\right] = \frac{\text{var}[D]}{(E_x^c)^2} = \frac{\mu E_x^c}{(E_x^c)^2} = \frac{\mu}{E_x^c}$$

Since maximum likelihood estimators are asymptotically normally distributed, we have the following result:

Asymptotic distribution of μ

When E_x^c is large, the distribution of the estimator $\tilde{\mu}$ is:

$$\tilde{\mu} \sim \text{Normal}\left(\mu, \frac{\mu}{E_x^c}\right)$$

These properties show that this is a sensible estimator to use. Its mean value equals the true value of μ and it varies as little as possible from the true value. The normal approximation allows us to calculate approximate probabilities and confidence intervals for μ.

Question

In a mortality investigation covering a 5-year period, where the force of mortality can be assumed to be constant, there were 46 deaths and the population remained approximately constant at 7,500.

Calculate an approximate 95% confidence interval for the force of mortality.

Solution

The maximum likelihood estimate of the force of mortality is:

$$\hat{\mu} = \frac{d}{E_x^c} = \frac{46}{7{,}500 \times 5} = 0.001227$$

An approximate 95% confidence interval for μ is given by:

$$\hat{\mu} \pm 1.96\sqrt{\text{var}(\tilde{\mu})} = \hat{\mu} \pm 1.96\sqrt{\frac{\mu}{E_x^c}}$$

Using the given sample values and replacing μ by $\hat{\mu}$ in order to estimate the variance gives the following confidence interval:

$$0.001227 \pm 1.96\sqrt{\frac{0.001227}{7,500 \times 5}} = 0.001227 \pm 0.000354 = (0.00087, 0.00158)$$

6.4 Links to the two-state Markov model

Under the two-state model, $E[\tilde{\mu}] = \mu$ and $\text{var}[\tilde{\mu}] = \dfrac{\mu}{E[V]}$, but the true values of μ and $E[V]$ are unknown and must be estimated from the data as $\hat{\mu}$ and E_x^c respectively. So although the estimators are different, we obtain the same numerical estimates of the parameter and of the first two moments of the estimator, in either case.

6.5 Estimating death probabilities

Once we have calculated $\hat{\mu}$, the estimated (constant) force of mortality that applies over the age range x to $x+1$, we can use this to estimate q_x (the probability of dying over that year of age) as follows:

$$\hat{q}_x = 1 - e^{-\hat{\mu}}$$

7 Comment on application

Having estimated piecewise constant intensities over single years of age, we can use these (if required) to estimate the function μ_x as a smooth function of age (the process of smoothing is called graduation). For this purpose we usually assume that $\hat{\mu}$ estimates $\mu_{x+½}$.

The process of graduation is covered in Chapters 10 and 11.

Both the two-state Markov model and the Poisson model assume that the force of mortality is constant over the year of age x to $x+1$, *ie* it is constant over the interval during which a life is aged x last birthday. In fact, the force of mortality is likely to vary over this year of age; what we are really estimating is the average force of mortality between the ages of x and $x+1$. We assume that this represents the force of mortality at the midway point of the year of age x to $x+1$, which is age $x+½$.

In other situations, we may be considering lives aged x next birthday or aged x nearest birthday. We will consider this situation in more detail in Part 3 of the course.

We can calculate any required probabilities from:

$$ {}_t p_x = \exp\left(-\int_0^t \mu_{x+s}\, ds \right) $$

using numerical methods if necessary.

8 Appendix – solving first-order differential equations

In this appendix, we give a reminder of two methods that can be used to solve first-order differential equations. These are the separation method and the integrating factor method.

The separation method

The separation method can be used to solve equations of the form:

$$\frac{dy}{dx} = g(x)\,h(y)$$

where $g(x)$ is a function of x and $h(y)$ is a function of y. The variables are separated by rewriting the equation as:

$$\frac{dy}{h(y)} = g(x)\,dx$$

Each side is then integrated to obtain the solution. If we are given an initial condition or a boundary condition, this can be used to determine the value of the constant of integration.

Question

Solve the differential equation $\dfrac{dy}{dx} = (x+1)y$ for $y > 0$, subject to the initial condition $y(0) = 2$.

Solution

Separating the variables we obtain:

$$\frac{dy}{y} = (x+1)\,dx$$

Then integrating both sides gives:

$$\ln y = ½x^2 + x + C$$

where C denotes a constant of integration. Taking exponentials, this becomes:

$$y = e^{½x^2 + x + C} = Ae^{½x^2 + x}$$

where $A = e^C$. Finally, using the initial condition:

$$y(0) = 2 \Rightarrow Ae^0 = 2 \Rightarrow A = 2 \Rightarrow y(x) = 2\,e^{½x^2 + x}$$

The integrating factor method

The integrating factor method can be used to solve equations of the form:

$$\frac{dy}{dx}+P(x)\,y=Q(x) \qquad (*)$$

where $P(x)$ and $Q(x)$ are both functions of x. In the context of this course, y will usually denote some probability.

The first step is to calculate the integrating factor (IF):

$$IF=e^{\int P(x)dx}$$

Then multiply each term in (*) by the integrating factor:

$$\frac{dy}{dx}e^{\int P(x)dx}+P(x)\,e^{\int P(x)dx}\,y=Q(x)\,e^{\int P(x)dx} \qquad (**)$$

Now integrate both sides of (**) with respect to x. The left-hand side will be:

$$y\,e^{\int P(x)dx}=y\times IF$$

(We can check this by applying the product rule for differentiation to the product $y\times IF$.) Finally, we divide through by IF to obtain an expression for y.

Question

Solve the differential equation $x\frac{dy}{dx}=2x-(x+1)y$ for $x>0$, subject to the condition $y(1)=0$.

Solution

We first write the differential equation in the form $\frac{dy}{dx}+P(x)\,y=Q(x)$:

$$\frac{dy}{dx}+\left(\frac{x+1}{x}\right)y=2$$

The integrating factor is given by:

$$\exp\left[\int\left(\tfrac{x+1}{x}\right)dx\right]=\exp\left[\int\left(1+\tfrac{1}{x}\right)dx\right]=\exp\left[x+\ln x\right]=e^{x}\,e^{\ln x}=x\,e^{x}$$

We don't have to bother about the constant of integration at this stage. The constants will cancel out when we multiply every term by the integrating factor.

Multiplying both sides of the differential equation by the integrating factor gives:

$$\frac{dy}{dx}xe^x + \left(1+\tfrac{1}{x}\right)xe^x y = 2\,x\,e^x$$

Integrating the left-hand side with respect to x gives:

$$y \times IF = y\,x\,e^x$$

Integrating the right-hand side with respect to x (using integration by parts), we obtain:

$$\int 2xe^x dx = 2xe^x - \int 2e^x\,dx = 2xe^x - 2e^x + C$$

Equating these gives:

$$yxe^x = 2xe^x - 2e^x + C$$

$$\Rightarrow y = 2 - \frac{2}{x} + \frac{C}{xe^x}$$

Finally, from the condition $y(1)=0$, we have:

$$0 = 2 - 2 + \frac{C}{e} \Rightarrow C = 0$$

So the required solution is:

$$y(x) = 2 - \frac{2}{x}$$

Chapter 3 Summary

Two-state Markov model

We can model mortality as a Markov process with two states (alive and dead) and a transition rate (or transition intensity) μ_x.

Assumptions

1. The probabilities that a life at any given age will be found in either state at any subsequent age depend only on the ages involved and on the state currently occupied. This is the Markov assumption.

2. For a short time interval of length h :

$$_hq_{x+t} = h\mu_{x+t} + o(h)$$

3. μ_{x+t} is a constant μ for $0 \le t < 1$.

Survival probabilities

From this model we can derive the following formula for the survival probability:

$$_tp_x = \exp\left(-\int_0^t \mu_{x+s}\,ds\right)$$

Waiting times

The waiting time for a life is the time spent under observation. The observed waiting time is often called the central exposed to risk.

We can use the observed total waiting time and the observed number of deaths to estimate the underlying transition intensity. The estimation can be done using the method of maximum likelihood. To proceed, we have to consider the 'probability' of getting the results we have observed from our mortality investigation.

Joint distribution of an observed sample

$$f_i(d_i, v_i) = e^{-\mu v_i}\,\mu^{d_i}$$

Joint distribution of all observed samples (likelihood function)

$$L(\mu) = e^{-\mu v}\,\mu^d$$

where $d = \sum_{i=1}^{N} d_i$ and $v = \sum_{i=1}^{N} v_i$.

Maximum likelihood estimator (two-state model)

The maximum likelihood estimate of μ is $\hat{\mu}=\dfrac{d}{v}$ and the maximum likelihood estimator is $\tilde{\mu}=\dfrac{D}{V}$. Asymptotically, $\tilde{\mu}\sim N\left(\mu,\dfrac{\mu}{E[V]}\right)$ or, equivalently, $\tilde{\mu}\sim N\left(\mu,\dfrac{\mu^2}{E[D]}\right)$.

We assume that the estimated transition intensity $\hat{\mu}$ estimates $\mu_{x+½}$.

Poisson model

Under the Poisson model, we assume that the force of mortality is constant between integer ages and the number of deaths has a Poisson distribution with mean μE_x^c:

$$D \sim Poisson\ (\mu E_x^c) \quad \text{and} \quad P[D=d]=\frac{e^{-\mu E_x^c}(\mu E_x^c)^d}{d\,!}$$

Maximum likelihood estimator (Poisson model)

The maximum likelihood estimate of μ is $\hat{\mu}=\dfrac{d}{E_x^c}$ and the maximum likelihood estimator is $\tilde{\mu}=\dfrac{D}{E_x^c}$. Asymptotically, $\tilde{\mu}\sim N\left(\mu,\dfrac{\mu}{E_x^c}\right)$.

This model is an approximation to the two-state model and provides the same numerical estimate of μ.

Chapter 3 Practice Questions

3.1 Show that, under the assumptions of the two-state Markov model for mortality:

$$\text{var}[D_i - \mu V_i] = E[D_i]$$

3.2 A survival model for an elderly population has two states A and D, representing alive and dead. The force of mortality at age *t* years $\mu(t)$, *ie* the transition rate from A to D, is given by:

$$\mu(t) = 0.0001 \times (1.10)^t$$

Calculate the probability that a 60-year-old will survive to age 80.

3.3 A certain species of small mammal is subject to a constant force of mortality of 0.2 *pa* over the first year of life. Calculate the expected number of deaths in the first year of life from a population of 1,000 new births.

3.4 In a mortality investigation, females aged 65 last birthday were observed. The following data values were recorded:

Total waiting time = 916 years

Observed number of deaths = 10

Calculate an approximate 95% confidence interval for the force of mortality of females aged 65 last birthday, assuming that the force is constant over this year of age.

3.5 Exam style An investigation took place into the mortality of residents of a care home. The investigation began on 1 January 2017 and ended on 1 January 2018. The table below gives the data collected in this investigation for 8 lives.

Date of birth	Date of entry into observation	Date of exit from observation	Whether or not exit was due to death (1) or other reason (0)
1 April 1946	1 January 2017	1 January 2018	0
1 October 1946	1 January 2017	1 January 2018	0
1 November 1946	1 March 2017	1 September 2017	1
1 January 1947	1 March 2017	1 June 2017	1
1 January 1947	1 June 2017	1 September 2017	0
1 March 1947	1 September 2017	1 January 2018	0
1 June 1947	1 January 2017	1 January 2018	0
1 October 1947	1 June 2017	1 January 2018	0

The force of mortality, μ, between exact ages 70 and 71 is assumed to be constant.

(i) (a) Estimate the constant force of mortality, μ, using a two-state model and the data for the 8 lives in the table.

(b) Hence or otherwise estimate q_{70}. [7]

(ii) Show that the maximum likelihood estimate of the constant force, μ, using a Poisson model of mortality is the same as the estimate using the two-state model. [5]

(iii) Outline the differences between the two-state model and the Poisson model when used to estimate transition rates. [2]

[Total 14]

Chapter 3 Solutions

3.1 We can start with the usual formula for variances:

$$\text{var}[D_i - \mu V_i] = E[(D_i - \mu V_i)^2] - \{E[D_i - \mu V_i]\}^2$$

Since $E[D_i - \mu V_i] = 0$, this is just:

$$\text{var}[D_i - \mu V_i] = E[(D_i - \mu V_i)^2]$$

We can evaluate this using the definition of the expectation, by considering that death will either occur at some time v_i in the interval $(0, b_i - a_i)$, in which case $D_i - \mu V_i = 1 - \mu v_i$, or the life will survive to the end of this interval, in which case $D_i - \mu V_i = 0 - \mu(b_i - a_i)$. So we get:

$$\text{var}[D_i - \mu V_i] = \int_0^{b_i - a_i} (1 - \mu v_i)^2 \mu e^{-\mu v_i}\, dv_i + [\mu(b_i - a_i)]^2 e^{-\mu(b_i - a_i)}$$

If we expand the integrand, we get:

$$\text{var}[D_i - \mu V_i] = \int_0^{b_i - a_i} \mu e^{-\mu v_i} dv_i$$

$$-2\mu \int_0^{b_i - a_i} v_i\, \mu e^{-\mu v_i} dv_i + \mu^2 \int_0^{b_i - a_i} v_i^2 \mu e^{-\mu v_i} dv_i + \mu^2 (b_i - a_i)^2 e^{-\mu(b_i - a_i)}$$

The first of these integrals is just $E[D_i]$, as shown in this chapter. So we have:

$$\text{var}[D_i - \mu V_i] = E[D_i]$$

$$+\mu^2 \left\{ -2\int_0^{b_i - a_i} v_i\, e^{-\mu v_i} dv_i + \int_0^{b_i - a_i} v_i^2 \mu e^{-\mu v_i} dv_i + (b_i - a_i)^2 e^{-\mu(b_i - a_i)} \right\}$$

So we just need to show that the three terms in the curly brackets sum to zero.

From equation (*) in Section 4.2, we know that:

$$\int_0^{b_i - a_i} \mu e^{-\mu v_i} dv_i + e^{-\mu(b_i - a_i)} = 1$$

If we differentiate this with respect to μ, we get:

$$\int_0^{b_i - a_i} (e^{-\mu v_i} - v_i\, \mu e^{-\mu v_i}) dv_i - (b_i - a_i) e^{-\mu(b_i - a_i)} = 0$$

Differentiating this again with respect to μ, we get:

$$\int_0^{b_i - a_i} (-2 v_i\, e^{-\mu v_i} + v_i^2 \mu e^{-\mu v_i}) dv_i + (b_i - a_i)^2 e^{-\mu(b_i - a_i)} = 0$$

Since the LHS of this equation is the same expression as in the curly brackets, we have established the result:

$$\text{var}[D_i - \mu V_i] = E[D_i]$$

3.2 The survival probability is given by:

$$ {}_{20}p_{60} = \exp\left(-\int_{60}^{80} 0.0001(1.1)^t \, dt \right) = \exp\left(-0.0001 \left[\frac{1.1^t}{\ln 1.1} \right]_{60}^{80} \right) = 0.16046$$

3.3 The probability that a new-born mammal survives for one year is:

$$p_0 = e^{-0.2} = 0.81873$$

So the probability of death within the first year is:

$$q_0 = 1 - p_0 = 1 - 0.81873 = 0.18127$$

and the expected number of deaths from an initial population of 1,000 new births is:

$$1{,}000\, q_0 = 1{,}000 \times 0.18127 = 181.27$$

3.4 An approximate 95% confidence interval for μ is:

$$\hat{\mu} \pm 1.96\sqrt{\text{var}(\tilde{\mu})}$$

We have:

$$\hat{\mu} = \frac{d}{v} = \frac{10}{916}$$

and, using the result in Section 5, $\text{var}(\tilde{\mu})$ is estimated by:

$$\frac{\hat{\mu}}{v} = \frac{d}{v^2} = \frac{10}{916^2}$$

So an approximate 95% confidence interval for μ is:

$$\frac{10}{916} \pm 1.96\sqrt{\frac{10}{916^2}} = 0.01092 \pm 0.00677 = (0.00415, 0.01768)$$

3.5 (i)(a) ***Estimate of μ***

We need to find the central exposed to risk for age 70 for each of the lives. This is the period that we observed each life during the study (*ie* between 1 January 2017 and 1 January 2018) when the lives were in the age range $(70, 71)$.

The table below summarises the calculations. We start by writing down the dates of each life's 70th and 71st birthdays. If the date of entry is after the life's 70th birthday or the date of exit is before the life's 71st birthday, we adjust the start/end date accordingly. The exposed to risk is then calculated by subtracting the start date from the end date.

Life	Start date	End date	Exposed to risk
1	~~1 April 2016~~ (age 70) 1 January 2017 (entry)	1 April 2017 (age 71)	3 months
2	~~1 October 2016~~ (age 70) 1 January 2017 (entry)	1 October 2017 (age 71)	9 months
3	~~1 November 2016~~ (age 70) 1 March 2017 (entry)	~~1 November 2017~~ (age 71) 1 September 2017 (exit)	6 months
4	~~1 January 2017~~ (age 70) 1 March 2017 (entry)	~~1 January 2018~~ (age 71) 1 June 2017 (exit)	3 months
5	~~1 January 2017~~ (age 70) 1 June 2017 (entry)	~~1 January 2018~~ (age 71) 1 September 2017 (exit)	3 months
6	~~1 March 2017~~ (age 70) 1 September 2017 (entry)	~~1 March 2018~~ (age 71) 1 January 2018 (exit)	4 months
7	1 June 2017 (age 70)	~~1 June 2018~~ (age 71) 1 January 2018 (exit)	7 months
8	1 October 2017 (age 70)	~~1 October 2018~~ (age 71) 1 January 2018 (exit)	3 months

In this type of question, be very careful counting the number of months. It is very easy to miscalculate these by one month.

So: $E_{70}^{c} = 3+9+6+3+3+4+7+3 = 38$ months [3]

From the final column of the table given in the question (the reason for exit), we see that Life 3 and Life 4 both died in the age range $(70,71)$ during the period of investigation. So $d = 2$ and:

$$\hat{\mu} = \frac{2}{38/12} = 0.63158$$ [2]

(i)(b) ***Estimate of q_{70}***

Since we are told that μ is the *constant* force of mortality over the year of age $(70,71)$, we can estimate q_{70} as:

$$\hat{q}_{70}=1-e^{-\hat{\mu}}=1-e^{-0.63158}=0.46825 \qquad [2]$$

(ii) ***Maximum likelihood estimate of μ using the Poisson model***

Under the Poisson model the observed number of deaths will follow a Poisson distribution with a parameter $\frac{38}{12}\mu$.

Since we observed 2 deaths, the likelihood function is:

$$L=\frac{\left(\frac{38}{12}\mu\right)^2\exp\left(-\frac{38}{12}\mu\right)}{2!}=C\mu^2\exp\left(-\frac{38}{12}\mu\right) \qquad [1]$$

where C is a constant.

The log-likelihood function is:

$$\log L=\log C+2\log\mu-\frac{38}{12}\mu \qquad [1]$$

Differentiating gives:

$$\frac{d\log L}{d\mu}=\frac{2}{\mu}-\frac{38}{12} \qquad [1]$$

This is equal to 0 when $\mu=\frac{2}{38/12}=0.63158$. [1]

We also have:

$$\frac{d^2\log L}{d\mu^2}=-\frac{2}{\mu^2}<0 \qquad [½]$$

So this solution maximises the log-likelihood and hence the maximum likelihood estimate of μ is 0.63158. [½]

This is the same as the estimate we obtained in part (i)(a) based on the two-state model.

(iii) *Differences between the two-state and the Poisson model*

The Poisson model can be considered to be an approximation to the two-state model. [½]

The Poisson model is not exact since it allows a non-zero probability of more than N deaths, where N is the total number of lives involved in the investigation. However, since this probability is usually negligible, the Poisson model often provides a good approximation. [½]

The estimation of the transition rates in the two-state model involves the measurement of two random variables – the observed number of decrements and the exposed to risk that gave rise to these decrements. [½]

The Poisson model assumes that the exposed to risk remains constant and estimation of the transition rates in the model only involves the measurement of the observed number of decrements. [½]

4

Time-homogeneous Markov jump processes

Syllabus objectives

3.3 Define and apply a Markov process.

3.3.1 State the essential features of a Markov process model.

3.3.2 Define a Poisson process, derive the distribution of the number of events in a given time interval, derive the distribution of inter-event times, and apply these results.

3.3.3 Derive the Kolmogorov equations for a Markov process with time independent and time/age dependent transition intensities.

3.3.4 Solve the Kolmogorov equations in simple cases.

3.3.5 Describe simple survival models, sickness models and marriage models in terms of Markov processes and describe other simple applications.

4.3 Derive maximum likelihood estimators for transition intensities.

4.3.1 Describe an observational plan in respect of a finite number of individuals observed during a finite period of time, and define the resulting statistics, including the waiting times.

4.3.2 Derive the likelihood function for constant transition intensities in a Markov model of transfers between states given the statistics in 4.3.1.

4.3.3 Derive maximum likelihood estimators for the transition intensities in 4.3.2 and state their asymptotic joint distribution.

0 Introduction

In Chapter 3 we considered a simple two-state Markov model. In this chapter we will show how the model and the results can be extended to any number of states.

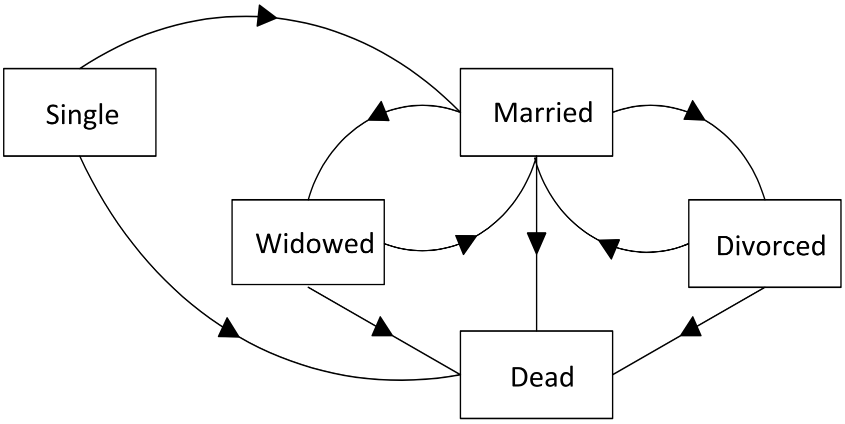
One important aspect of the simple two-state model is that transition is possible in one direction only, from *alive* to *dead*. In practice, we may wish to study a model in which transition between states is possible in both directions. This opens up the possibility of a life entering a particular state more than once.

An example of this is a model in which the states relate to marital status. Such a model might comprise five states – *single*, *married*, *divorced*, *widowed* and *dead*.

Question

Draw this model, showing clearly the possible transitions between the five states.

Solution



Some of the problems that we will consider in this chapter are:

- how to calculate the probability of a life remaining in a particular state for a period of length t given there is more than one possible way of exiting that state
- how to use real-life observations to estimate the transition rates
- how to calculate the probability of a particular decrement occurring based on our estimates of the transition rates.

The results that we derive here form the basic building blocks of several actuarial techniques. Some financial applications of this theory are discussed in Subject CM1, where we use transition probabilities to calculate expected present values.

Much of the theory is analogous to that for Markov chains. The Chapman-Kolmogorov equations can be written in the same format for example. However, in discrete time we have the central notion of the one-step transition probabilities. In the continuous case there is no longer the same fundamental significance to the unit time interval, as we can consider time intervals of arbitrarily small length h. As is often the case with continuous variables, the natural thing to do is to consider limits as $h \to 0$. This leads to a reformulation of the Chapman-Kolmogorov equations as *differential equations*. Much of our time will be spent constructing and interpreting such differential equations, along with their *integral equation* analogues.

These differential and integral equations can be solved to give results for the transition probabilities in terms of the transition rates. All the versions of the equations will have the same solution for a particular model. For some models, one of the equations may be more straightforward to solve than the others. Exam questions sometimes guide us towards a particular equation, rather than us having to choose one for ourselves.

In this chapter we consider only time-homogeneous Markov jump process. These are processes in which the transition rates do not vary over time, so the transition probabilities $P\left(X_t = j \mid X_s = i\right)$ depend only on the length of the time interval, $t - s$.

1 Notation

Different authors tend to use different notation for the same quantities, and the Markov model is an example of this. Actuaries often use notations derived from the standardised International Actuarial Notation, in which the 'transition rate' is the force of mortality μ_x, and the corresponding probabilities are the life table probabilities ${}_tp_x$ and ${}_tq_x$. Moreover, the index x is generally understood to indicate age (*eg* age now, or age at policy inception) and the index t indicates duration since age x. Probabilists and statisticians tend to use different notations.

We met the notation ${}_tp_x$ and ${}_tq_x$ in Chapter 3. Recall that ${}_tp_x$ denotes the probability that a life aged x survives for at least another t years, and ${}_tq_x = 1 - {}_tp_x$ is the probability that a life aged x dies within the next t years.

The non-homogeneous (*ie* time-inhomogeneous) **Markov model offers particularly rich, and potentially confusing, opportunities to invent different notations for the same quantities. To try to limit any such confusion, we make the following remarks.**

1. **We have written $p_{ij}(s,t)$ to mean the probability of the process being in state j at time t, conditional on being in state i at time $s \le t$.**

   The traditional actuarial notation would reserve the symbol t for duration since time s, in which case the above probability would be expressed $p_{ij}(s,s+t)$. Just as likely, the life table symbol ${}_tp_s$ would be adapted, so that $p_{ij}(s,s+t)$ would be written as ${}_tp_s^{ij}$. Other variants, such as ${}_tp_{ij}(s)$, may be encountered.

   For time-homogeneous processes, it is just the length of the time interval that is important, not when it starts. So $p_{ij}(0,t) = p_{ij}(s,s+t)$ for all s, and we will use the notation $p_{ij}(t)$ to denote this probability.

2. **We have written $\mu_{ij}(s)$ to mean the transition rate from state i to state j at time s. Following the actuarial tradition, the time (or age) may be indicated by a subscript, so that the same rate may be written μ_s^{ij}.**

   For time-homogeneous processes, the transition rates are constant and we will denote these by μ_{ij}. You may also see the notation σ_{ij} instead of μ_{ij} used to denote the transition rate from state i to state j. In particular, the formulae given on page 38 of the *Tables* use the σ_{ij} notation.

While a standard international actuarial notation was adopted for the life table and its derived quantities, the same is not true for the richer models needed to represent insurance contracts that depend on more than just being alive or dead. The actuarial reader must always be prepared to assimilate the notation that each particular author decides to use.

So the notation used in exam questions (and other questions) may not be the same as the notation used in this chapter. You should try to be flexible and accept whatever notation is given to you in a question. You should also try to stick to the notation given in a question when writing your answer to that question.

2 The Poisson process

The Poisson process forms the simplest example of a Markov jump process in continuous time. In studying the Poisson process we shall encounter features which are of general applicability in this chapter.

2.1 Definition

The standard time-homogeneous Poisson process is a counting process in continuous time, $\{N_t, t \geq 0\}$, where N_t records the number of occurrences of some type of event within the time interval from 0 to t. The events of interest occur singly and may occur at any time.

In fact, we have already given a definition of a Poisson process with parameter λ in Chapter 1. Recall that it is a continuous-time process, starting at 0, with stationary independent increments, and, over a time period of length t, these increments follow a Poisson distribution with parameter λt. An alternative definition is given below, and we will show that they are equivalent.

The probability that an event occurs during the short time interval from time t to time $t+h$ is approximately equal to λh for small h; the parameter λ is called the *rate* of the Poisson process.

The Poisson process is very commonly used to model the occurrence of unpredictable incidents, such as car accidents or arrival of claims at an office.

The above definition should be made more precise if it is to be used for calculations. Formally, an integer-valued process $\{N_t, t \geq 0\}$, with filtration $\{F_t, t \geq 0\}$, is a Poisson process if:

$$P\left[N_{t+h} - N_t = 1 \mid F_t\right] = \lambda h + o(h)$$

$$P\left[N_{t+h} - N_t = 0 \mid F_t\right] = 1 - \lambda h + o(h) \qquad (4.1)$$

$$P\left[N_{t+h} - N_t \neq 0,1 \mid F_t\right] = o(h)$$

where the statement that $f(h) = o(h)$ as $h \to 0$ means $\lim\limits_{h \to 0} \dfrac{f(h)}{h} = 0$.

As may be seen from the definition, the increment $N_{t+h} - N_t$ of the Poisson process is independent of past values of the process and has a distribution which does not depend on t. It therefore follows that the Poisson process is a process with stationary, independent increments and, in addition, satisfies the Markov property.

It is far from obvious that the process defined above coincides with the Poisson process characterised in Chapter 1 as having independent, stationary, Poisson-distributed increments. That is one of the properties that we shall prove.

Distribution of increments

N_t is a Poisson random variable with mean λt. More generally, $N_{t+s} - N_s$ is a Poisson random variable with mean λt, independent of anything that has occurred before time s.

Proof

Define $p_j(t) = P(N_t = j)$, the probability that there have been exactly j events by time t. The proof will be complete if we can verify that, for each $j \geq 0$,

$$p_j(t) = \frac{e^{-\lambda t}(\lambda t)^j}{j!} \qquad (4.2)$$

We will do this by setting up a differential equation and a boundary condition that $p_j(t)$ must satisfy. It will then be possible to check that the given expression does satisfy this condition.

For any $j > 0$, and for small positive h:

$$\begin{aligned} p_j(t+h) &= P(N_{t+h} = j) \\ &= P(N_t = j \text{ and } N_{t+h} = N_t) + P(N_t = j-1 \text{ and } N_{t+h} = N_t + 1) + o(h) \\ &= p_j(t)(1-\lambda h) + p_{j-1}(t)\lambda h + o(h) \end{aligned}$$

Rearranging this equation, and letting $h \to 0$, we obtain, again for $j > 0$:

$$\frac{dp_j(t)}{dt} = -\lambda p_j(t) + \lambda p_{j-1}(t) \qquad (4.3)$$

with initial condition $p_j(0) = 0$.

The same analysis yields, in the case $j = 0$:

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) \qquad (4.4)$$

with $p_0(0) = 1$. It is now straightforward to verify that the suggested solution (4.2) satisfies both the differential equations (4.3) and (4.4) as well as the initial conditions.

Question

Verify that the function $p_j(t) = \dfrac{e^{-\lambda t}(\lambda t)^j}{j!}$ satisfies the equations given above.

Solution

If $j=0$ then:

$$p_0(t)=e^{-\lambda t}$$

and: $$p_0'(t)=-\lambda e^{-\lambda t}=-\lambda p_0(t)$$

as required.

Otherwise:

$$p_j(t)=\frac{e^{-\lambda t}(\lambda t)^j}{j!}$$

and: $$p_j'(t)=-\lambda\frac{e^{-\lambda t}(\lambda t)^j}{j!}+j\lambda(\lambda t)^{j-1}\frac{e^{-\lambda t}}{j!}=-\lambda p_j(t)+\lambda p_{j-1}(t)$$

Also:

$$p_j(0)=\frac{e^{-\lambda 0}(\lambda 0)^j}{j!}=0 \text{ if } j>0 \text{ and } p_0(0)=1$$

Hence the boundary condition is also satisfied.

In view of the fact that the increments of N are stationary and are independent of the past, this result may be generalised to a statement that $N_{t+s}-N_s$ is a Poisson random variable with mean λt, independent of anything that has occurred before time s.

A Poisson process could be used to model motor insurance claims. The events in this case could be occurrences of claims events (*eg* accidents, fires, thefts *etc*) or claims reported to the insurer. The parameter λ represents the average rate of occurrence of claims (*eg* 50 per day). The assumption that, in a sufficiently short time interval, there can be at most one claim is satisfied because we are working in continuous time. If there is a motorway pile-up, we can say that claims occurred at times 3:00, 3:01, 3:02 *etc*.

2.2 Sums of independent Poisson processes

Suppose that claims are made to two insurance companies, A and B. The numbers of claims made to each are independent and follow Poisson processes with parameters λ_A (claims per day) and λ_B respectively. Then the combined number of claims $(A+B)_t$ is a Poisson process with parameter $\lambda_A+\lambda_B$. This can be verified by checking the three defining properties of a Poisson process that are given in Chapter 1.

Firstly, as both processes start at 0, trivially so does their sum. It remains to show that the increments are independent and stationary, and that the parameter for the combined process is $\lambda_A+\lambda_B$.

Since the processes are independent of one another, it follows that their increments are independent of one another. These increments are Poisson with parameters $\lambda_A(t-s)$ and $\lambda_B(t-s)$. Their sum is therefore Poisson, with parameter $(\lambda_A+\lambda_B)(t-s)$ (as shown in the following question). They are therefore also stationary and independent. So we do have a Poisson process with parameter $\lambda_A+\lambda_B$.

Question

Let $X \sim Poisson(\lambda)$ and $Y \sim Poisson(\mu)$ be independent random variables. Prove that $X+Y \sim Poisson(\lambda+\mu)$.

Solution

Consider the moment generating function of $X+Y$. We have:

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)M_Y(t) \quad \text{by independence} \\ &= e^{\lambda(e^t-1)}e^{\mu(e^t-1)} \\ &= e^{(\lambda+\mu)(e^t-1)} \end{aligned}$$

Since this is the same as the MGF of $Poisson(\lambda+\mu)$, we can apply the uniqueness property of MGFs to give the required result.

Alternatively, if we use the convolution approach, we have:

$$\begin{aligned} P(X+Y=k) &= \sum_{i=0}^{k} P(X=i, Y=k-i) \\ &= \sum_{i=0}^{k} P(X=i)P(Y=k-i) \qquad \text{by independence} \\ &= \sum_{i=0}^{k} \frac{e^{-\lambda}\lambda^i}{i!}\frac{e^{-\mu}\mu^{k-i}}{(k-i)!} \\ &= \frac{e^{-(\lambda+\mu)}}{k!}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}\lambda^i\mu^{k-i} \\ &= \frac{e^{-(\lambda+\mu)}}{k!}(\lambda+\mu)^k \qquad \text{by the binomial expansion} \end{aligned}$$

This is the probability function for the $Poisson(\lambda+\mu)$ distribution.

We have shown that if we have two independent Poisson processes with parameters λ and μ, then the sum of the processes is Poisson with parameter $\lambda+\mu$. This conforms to intuition. For example, suppose that the arrivals of two different types of insurance claim follow a Poisson process, one at the rate of 5 per day the other at the rate of 6 per day. We would expect that the total arrivals follows a Poisson process with a rate of 11 per day. This is true as long as the processes are independent.

2.3 Thinning of Poisson processes

It is also useful to know that a Poisson process behaves in an intuitive way when considering the problem of *thinning* or *sampling*. Again, consider insurance claims arriving such that they follow a Poisson process with rate 10 per day. Then if one in every 10 claims is of a certain type, *eg* those over £10,000, the arrival of these will be Poisson with rate 1 per day. This assumes that such claims occur randomly within the arrivals of all claims. So every claim that arrives is over £10,000 with probability 0.1, independently of anything else. Here we have 'thinned' the Poisson process.

Question

An insurance company has two types of policy, A and B. Claims arriving under A follow a Poisson process with a rate of 5 per day. Claims arrive independently under B and follow a Poisson process with a rate of 3 per day. A randomly selected claim from A has a probability of $\frac{1}{5}$ of being over £10,000 while a randomly selected claim from B has probability $\frac{2}{3}$ of being over £10,000.

Calculate the number of claims over £10,000 that are expected per day.

Solution

We need to calculate the Poisson parameter for claims over £10,000. This is the sum of the parameters for claims over £10,000 from each of A and B.

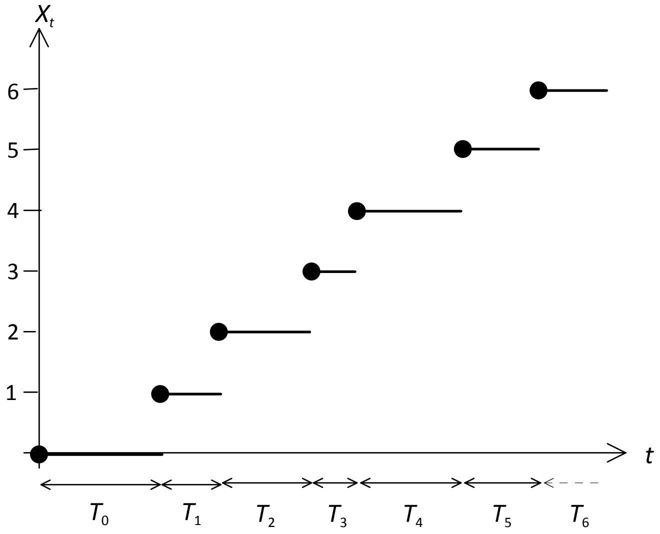
By the 'thinning rule', claims under A that are over £10,000 arrive as a Poisson process with rate $\frac{1}{5}\times 5=1$ per day.

Similarly, for B, the rate is $\frac{2}{3}\times 3=2$ per day.

So the expected number of claims over £10,000 is 3 per day.

2.4 Inter-event times

Since the Poisson process N_t changes only by unit upward jumps, its sample paths are fully characterised by the times at which the jumps take place. Denote by $T_0, T_1, T_2, ...$ the successive *inter-event times* (or holding times), a sequence of random variables.



Note that we choose (by convention) the sample paths of X_t to be right-continuous so that $X_{T_0} = 1, X_{T_0+T_1} = 2$, … .

So:

- $N_t = 0$ for values of t in the interval $[0, T_0)$
- $N_t = 1$ for values of t in the interval $[T_0, T_0 + T_1)$
- $N_t = 2$ for values of t in the interval $[T_0 + T_1, T_0 + T_1 + T_2)$

and so on. Because we have chosen the sample paths to be right-continuous, N_t is constant over intervals of the form $[a, b)$. If we had chosen the sample paths to be left-continuous, N_t would have been constant over intervals of the form $(a, b]$.

Distribution of holding time random variables

$T_0, T_1, T_2, ...$ is a sequence of independent exponential random variables, each with parameter λ.

Proof

$P(T_0 > t)$ is the probability that no events occur between time 0 and time t, which is also equal to $P(N_t = 0) = p_0(t) = e^{-\lambda t}$.

Now the distribution function of T_0 is $F(t) = P(T_0 \le t) = 1 - e^{-\lambda t}$, $t > 0$, implying that T_0 is exponentially distributed.

Consider now the conditional distribution of T_1 given the value of T_0.

$$\begin{aligned} P\left[T_1 > t \middle| T_0 = s\right] &= P\left[N_{t+s} = 1 \middle| T_0 = s\right] \\ &= P\left[N_{t+s} - N_s = 0 \middle| T_0 = s\right] \\ &= P\left[N_{t+s} - N_s = 0\right] \\ &= p_0(t) \\ &= e^{-\lambda t} \end{aligned}$$

where the third equality reflects the independence of the increment $N_{t+s} - N_s$ from the past of the process (up to and including time s).

The above calculation proves two results at once: T_1 is independent of T_0 and has the same exponential distribution. The calculation can be repeated for $T_2, T_3, \ldots$.

In summary, all of the inter-event times are independent and are exponentially distributed with parameter λ. We will see shortly that for a time-homogeneous Markov jump process, the holding time in any given state is exponentially distributed.

Question

Claims from a certain group of policies follow a Poisson process with a rate of 5 per day and claims can be logged 24 hours a day. Calculate:

(i) the probability that there will be fewer than 2 claims reported on a given day

(ii) the probability that at least one claim will be reported during the next hour

(iii) the expected time before a claim comes in, given that there haven't been any claims for over a week.

Solution

(i) ***Fewer than 2 claims in a day***

Let X denote the number of claims reported in a day. Then $X \sim Poisson(5)$ and:

$$P(X \le 1) = \frac{e^{-5}5^0}{0!} + \frac{e^{-5}5^1}{1!} = 6e^{-5} = 0.0404$$

(ii) ***At least one claim in the next hour***

Let Y denote the number of claims reported in an hour. Then $Y \sim Poisson\left(5/24\right)$ and:

$$P(X \ge 1) = 1 - P(X = 0) = 1 - e^{-5/24} = 0.1881$$

Alternatively, we could define T to be the waiting time in hours until the next reported claim. Then $T \sim Exp(5/24)$ and:

$$P(T \leq 1) = F_T(1) = 1 - e^{-5/24} = 0.1881$$

(iii) ***Expected time until the next claim***

The waiting time has the memoryless property, so the time before another claim comes in is independent of the time since the last one. The expected time is therefore the expected value of the exponential distribution, which in this case is 0.2 days.

In real life, the assumptions of a uniform rate and independence may not be valid. If there haven't been any claims reported for a week this may be because of a 'blockage' in the system (*eg* an IT malfunction) and there may well be a 'catch-up' effect the next day.

The exponential distribution of the holding times gives us a third definition of the Poisson process. We summarise these definitions below.

Summary of definitions of a Poisson process

Let $\{N_t\}_{t\geq 0}$ be an increasing, integer-valued process starting at 0 (and continuous from the right). Let $\lambda > 0$. Then $\{N_t\}_{t\geq 0}$ is a Poisson process if any of the following three equivalent conditions hold:

(1) $\{N_t\}_{t\geq 0}$ has stationary, independent increments and for each t, N_t has a Poisson distribution with parameter λt.

(2) $\{N_t\}_{t\geq 0}$ is a Markov jump process with independent increments and transition probabilities over a short time period h given by:

$$P\left[N_{t+h} - N_t = 1 \mid F_t\right] = \lambda h + o(h)$$

$$P\left[N_{t+h} - N_t = 0 \mid F_t\right] = 1 - \lambda h + o(h)$$

$$P\left[N_{t+h} - N_t \neq 0,1 \mid F_t\right] = o(h)$$

(3) The holding times, $T_0, T_1, \ldots$ of $\{N_t\}_{t\geq 0}$ are independent exponential random variables with parameter λ and $N_{T_0+T_1+\cdots+T_{n-1}} = n$.

There is also a fourth definition, which is given below. This is a restatement of (2) using the terminology of general Markov jump processes, which we will meet shortly. For completeness we will include it here, although you will have to wait for the definition of a general transition rate μ_{ij}.

(4) $\{N_t\}_{t\geq 0}$ is a Markov jump process with independent increments and transition rates given by:

$$\mu_{ij} = \begin{cases} -\lambda & \text{if } j = i \\ \lambda & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

3 Features of time-homogeneous Markov jump processes

We start with the definition of a Markov jump process.

Markov jump process

A continuous-time Markov process X_t, $t \geq 0$ with a discrete (*ie* finite or countable) state space S is called a *Markov jump process*.

3.1 The Chapman-Kolmogorov equations

In this chapter consideration will be given to the *time-homogeneous* case, where probabilities $P(X_t = j | X_s = i)$ depend only on the length of the time interval, $t - s$.

The transition probabilities of the Markov jump process:

$$p_{ij}(t) = P\left(X_t = j | X_0 = i\right)$$

obey the Chapman-Kolmogorov equations:

$$p_{ij}(t+s) = \sum_{k \in S} p_{ik}(s)\, p_{kj}(t) \qquad \text{for all } s, t > 0 \qquad (4.5)$$

The derivation of the Chapman-Kolmogorov equations in continuous time is identical to the derivation in discrete time. See Chapter 2.

3.2 The transition matrix

Denoting by $P(t)$ the matrix with entries $p_{ij}(t)$, known as the *transition matrix*, Equation (4.5) reads:

$$P(t+s) = P(s)P(t) \quad \text{for all } s, t > 0$$

If we know the transition matrix $P(t)$ and the initial probability distribution $q_i = P(X_0 = i)$, we can find general probabilities involving the process X_t by using the Markov property.

For instance, when $0 < t_1 < t_2 < \ldots < t_n$:

$$P\left[X_0 = i, X_{t_1} = j_1, X_{t_2} = j_2, \ldots, X_{t_n} = j_n\right] = q_i\, p_{ij_1}(t_1)\, p_{j_1 j_2}(t_2 - t_1) \ldots p_{j_{n-1} j_n}(t_n - t_{n-1})$$

Adding over all states i gives:

$$P\left[X_{t_1} = j_1, X_{t_2} = j_2, \ldots, X_{t_n} = j_n\right] = \sum_{i \in S} q_i\, p_{ij_1}(t_1)\, p_{j_1 j_2}(t_2 - t_1) \ldots p_{j_{n-1} j_n}(t_n - t_{n-1})$$

The above results are similar to the results given in Section 2 of Chapter 2.

3.3 Transition rates

For Markov chains we have the fundamental notion of the one-step transition probabilities. This is because Markov chains operate in discrete time. Together with the starting distribution $P\left[X_0 = i\right] = q_i$, these fully determine the distribution of the chain. When we come to deal with Markov jump processes, however, we may consider transitions over arbitrarily small times, so that time steps of one unit are no longer of the same fundamental importance.

For a continuous-time process, we consider transition probabilities over a very short time interval of time h. Dividing by h expresses this as a probability of transition in unit time. Taking limits as h tends to 0 leads to the concept of a *transition rate*. We have seen this before in the two-state Markov model of Chapter 3. Recall that transition rates are also sometimes referred to as *transition intensities* or *forces of transition*.

These transition rates are the fundamental concept in continuous time; they are analogous to the one-step transition probabilities in the discrete case. Unlike probabilities, these transition rates can take values greater than 1 (as frequently happens with annual recovery rates). For example, if, on average, you spend half an hour in a particular state before leaving, then the transition rate out will be 2 per hour.

In order to differentiate the transition probabilities and avoid technical problems with the mathematics, we will make the following assumption.

We will *assume* that the functions $p_{ij}(t)$ are continuously differentiable. This is a large assumption to make; indeed, the full theory of Markov jump processes permits transition probabilities that do not satisfy this requirement. Such processes are called *irregular*. They are of little use in practical modelling, however, and the loss involved in restricting our attention to regular Markov processes is not significant for the purposes of this course.

Noting that:

$$p_{ij}(0) = \delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \tag{4.6}$$

the assumption of differentiability implies the existence of the following quantities:

$$\mu_{ij} = \frac{d}{dt} p_{ij}(t)\Big|_{t=0} = \lim_{h \to 0} \frac{p_{ij}(h) - \delta_{ij}}{h}$$

μ_{ij} is the force of transition from state i to state j. Transition rates in time-homogeneous processes do not vary over time. The function δ_{ij} in the expression above is known as the Kronecker delta.

Question

Explain Equation (4.6).

Solution

$p_{ij}(0)$ is the probability of simultaneously being in state i and state j at time 0. This is 1 if $i=j$ but 0 otherwise.

Equivalently, the following relations hold as $h \to 0$ ($h > 0$):

$$p_{ij}(h) = \begin{cases} h\,\mu_{ij} + o(h) & \text{if } i \neq j \\ 1 + h\mu_{ii} + o(h) & \text{if } i = j \end{cases} \tag{4.7}$$

The interpretation of the first line of (4.7) is simply that the probability of a transition from i to j during any short time interval $[s, s+h]$ is proportional to h; hence the name *transition rate* or *transition intensity* given to μ_{ij}.

So the first line of (4.7) says that if i and j are different states, then the probability of going from state i to state j in a short time interval of length h is:

$$\left(h \times \text{the force of transition from state } i \text{ to state } j\right) + o(h)$$

This is similar to Assumption 2 for the two-state Markov model in Chapter 3, which states that:

$${}_hq_{x+t} = h\mu_{x+t} + o(h) \quad \text{for small } h$$

We also assume that the probability of more than one transition in a short time interval of length h is $o(h)$.

Note finally that as a result of (4.7) $\mu_{ij} \geq 0$ for $i \neq j$, but $\mu_{ii} \leq 0$. In fact differentiating the identity $\sum_{j \in S} p_{ij}(t) = 1$ with respect to t at $t = 0$ yields:

$$\mu_{ii} = -\sum_{j \neq i} \mu_{ij}$$

Alternatively, we could argue as follows:

$$\mu_{ii} = \lim_{h\to 0} \frac{p_{ii}(h) - 1}{h} = \lim_{h\to 0} \frac{1 - \sum_{j\neq i} p_{ij}(h) - 1}{h} = -\sum_{j\neq i} \lim_{h\to 0} \frac{p_{ij}(h)}{h} = -\sum_{j\neq i} \mu_{ij}$$

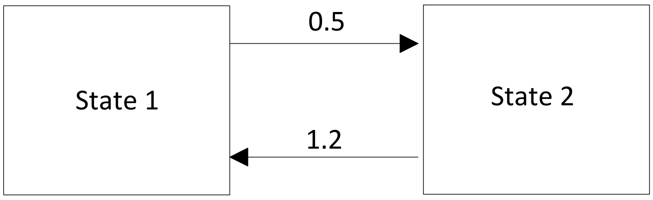
Generator matrix

The generator matrix A of a Markov jump process is the matrix of transition rates. In other words, the i,jth entry of A is μ_{ij}.

Hence each row of the matrix A has zero sum.

The relationship $\mu_{ii} = -\sum_{j \neq i} \mu_{ij}$ is often used as a working definition of μ_{ii}. The transition rate μ_{ii} is then defined as minus the sum of the transition rates out of state i.

As an example, consider the following two-state Markov jump process with transition rates as shown below:

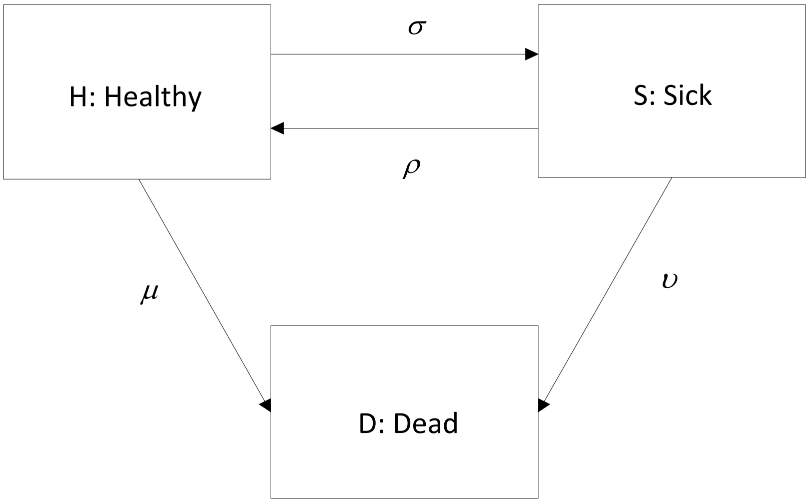


Taking the states in the order 1, 2, the generator matrix is:

$$\begin{pmatrix} -0.5 & 0.5 \\ 1.2 & -1.2 \end{pmatrix}$$

3.4 The time-homogeneous health-sickness-death model

Consider the following health-sickness-death (HSD) model with constant transition rates.



The transition rate from sick to dead is denoted by the Greek letter υ (pronounced nu).

A life may be in the healthy state or the sick state on a number of separate occasions before making the one-way transition to the dead state. Alternatively, a life may pass from the healthy state to the dead state without ever having been in the sick state.

Question

Give expressions for μ_{SH}, μ_{HH} and μ_{DD} using the notation of the HSD model shown above.

Solution

Using this notation, the rates are:

$$\mu_{SH} = \rho, \quad \mu_{HH} = -(\sigma + \mu), \quad \mu_{DD} = 0$$

The generator matrix for the HSD model is:

$$A = \begin{pmatrix} -\sigma - \mu & \sigma & \mu \\ \rho & -\rho - v & v \\ 0 & 0 & 0 \end{pmatrix}$$

Here the order of the states has been taken to be H, S, then D (as usual).

The rows of this matrix sum to 0, which is consistent with our earlier equation. A common mistake is to think that the rate from dead to dead is 1. It isn't. The transition probability is 1, but the transition rate is the derivative, and hence the constant 1 differentiates to 0.

Another way to think of the last row of this matrix is as follows. We can't go from the dead state to the healthy state, so the force of transition from dead to healthy is 0. Similarly, the force of transition from the dead state to the sick state is 0. Each row of the generator matrix must sum to 0, so the DD entry must also be 0.

4 Kolmogorov's forward differential equations

Transition rates are of fundamental importance in that they characterise fully the distribution of Markov jump processes. In order to see this, substitute $t = h$ and $s = t$ in (4.5):

$$\boldsymbol{p_{ij}(t+h) = \sum_{k \in S} p_{ik}(t)\, p_{kj}(h) = p_{ij}(t) + h \sum_{k \in S} p_{ik}(t)\, \mu_{kj} + o(h)}$$

The second equality follows from the relationship:

$$p_{kj}(h) = \begin{cases} h\mu_{kj} + o(h) & \text{if } j \neq k \\ 1 + h\mu_{kk} + o(h) & \text{if } j = k \end{cases}$$

This leads to the differential equation:

$$\boldsymbol{\frac{d}{dt} p_{ij}(t) = \sum_{k \in S} p_{ik}(t)\, \mu_{kj} \quad \text{for all } i, j} \qquad \textbf{(4.8)}$$

Either lower case $p_{ij}(t)$ or upper case $P_{ij}(t)$ may be used to denote these transition probabilities.

Question

Derive this differential equation.

Solution

From the first Core Reading equation in this section, we have:

$$p_{ij}(t+h) = \sum_{k \in S} p_{ik}(t)\, p_{kj}(h) = p_{ij}(t) + h \sum_{k \in S} p_{ik}(t)\, \mu_{kj} + o(h)$$

Rearranging this we have:

$$\frac{p_{ij}(t+h) - p_{ij}(t)}{h} = \sum_{k \in S} p_{ik}(t)\, \mu_{kj} + \frac{o(h)}{h}$$

Taking the limit as $h \to 0$ gives the desired result:

$$\frac{d}{dt} p_{ij}(t) = \sum_{k \in S} p_{ik}(t)\, \mu_{kj}$$

since $\lim_{h \to 0} \frac{o(h)}{h} = 0$.

These differential equations are known as *Kolmogorov's forward equations.*

Kolmogorov's forward differential equations (time-homogeneous case)

These can be written in compact (*ie* matrix) **form as:**

$$\frac{d}{dt}\boldsymbol{P(t)} = \boldsymbol{P(t)\,A}$$

where $\boldsymbol{A}$ is the matrix with entries μ_{kj} .

Recall that $\boldsymbol{A}$ is often called the *generator matrix* of the Markov jump process.

Equipped with this general equation, we can write down specific equations for a given Markov jump process and solve them in simple cases.

Example

For the HSD model given in Section 3.4, the forward differential equation for $p_{HH}(t)$ can be obtained by using the general forward equation as a template. This gives:

$$\frac{d}{dt}p_{HH}(t) = p_{HH}(t)\mu_{HH} + p_{HS}(t)\mu_{SH} + p_{HD}(t)\mu_{DH}$$

Now substituting in for the transition rates, we have:

$$\frac{d}{dt}p_{HH}(t) = -p_{HH}(t)(\sigma+\mu) + p_{HS}(t)\rho$$

It is important to be able to write down any forward equation, such as the last one, fairly quickly. This means not relying on the general template every time, but instead recognising that these equations follow a pattern. When writing down the equation for *H* to *H* above, we are including a term for each possible path. To get from *H* at the outset to *H* at the end, we must be in either at *H* or *S* in the 'middle'. The two types of paths to include are therefore $H \to H \to H$ or $H \to S \to H$. So we can consider the RHS as follows:

- start with the probability of going from *H* to *H* over an interval of length t (*ie* $p_{HH}(t)$) and multiply this by the force of transition that keeps us in *H* at time t (*ie* $-(\sigma+\mu)$)
- then add on the probability of going from *H* to *S* over an interval of length t (*ie* $p_{HS}(t)$) and multiply this by the force of transition that takes us from *S* to *H* at time t (*ie* ρ).

Question

Write down Kolmogorov's forward differential equation for the transition probability $p_{HS}(t)$.

Solution

Kolmogorov's forward differential equation is:

$$\begin{aligned}\frac{d}{dt}p_{HS}(t) &= p_{HH}(t)\mu_{HS} + p_{HS}(t)\mu_{SS} + p_{HD}(t)\mu_{DS} \\ &= p_{HH}(t)\sigma - p_{HS}(t)(\rho+\nu)\end{aligned}$$

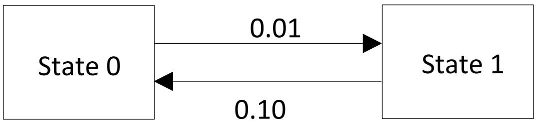
As we see from the solution above, the equation for $p_{HS}(t)$ involves $p_{HH}(t)$ as well. This will also be unknown initially so this equation cannot be solved in its own right. The forward equations for all transitions $i \to j$ will often need to be constructed and solved as a set of *simultaneous* differential equations. Generally, writing down such sets of equations is straightforward, but solving them is much more difficult.

We need to be able to solve such equations in simple cases. We are usually able to use one of two methods: separation of variables or the integrating factor method. We gave a brief review of these in an appendix to Chapter 3.

Example

A Markov jump process has two states, labelled state 0 and state 1, with forces of transition $\mu_{01} = 0.01$ and $\mu_{10} = 0.10$.

The transition diagram for this process is as follows:



Kolmogorov's forward differential equation for $p_{01}(t)$ is:

$$\begin{aligned}\frac{d}{dt}p_{01}(t) &= p_{00}(t)\mu_{01} + p_{01}(t)\mu_{11} \\ &= 0.01p_{00}(t) - 0.10p_{01}(t)\end{aligned}$$

since μ_{11} is minus the (total) force of transition out of state 1.

Now, since there are only two states, we have:

$$p_{00}(t) = 1 - p_{01}(t)$$

So the differential equation can be written as:

$$\frac{d}{dt}p_{01}(t) = 0.01\left[1 - p_{01}(t)\right] - 0.10p_{01}(t) = 0.01 - 0.11p_{01}(t)$$

This equation can be solved using the integrating factor method, by first rewriting it in the form:

$$\frac{d}{dt}p_{01}(t)+0.11p_{01}(t)=0.01$$

The integrating factor in this case is $e^{0.11t}$. Multiplying every term in the previous equation by the integrating factor gives:

$$e^{0.11t}\frac{d}{dt}p_{01}(t)+0.11e^{0.11t}p_{01}(t)=0.01e^{0.11t}$$

Then integrating both sides with respect to t, we get:

$$\begin{aligned}e^{0.11t}p_{01}(t)&=\int 0.01e^{0.11t}\,dt\\&=\frac{1}{11}e^{0.11t}+C\end{aligned}$$

where C is a constant of integration. We can calculate the value of C using the initial condition $p_{01}(0)=0$. This gives:

$$0=\frac{1}{11}+C$$

So $C=-\frac{1}{11}$. Hence:

$$e^{0.11t}p_{01}(t)=\frac{1}{11}\left(e^{0.11t}-1\right)$$

and dividing through by $e^{0.11t}$ gives:

$$p_{01}(t)=\frac{1}{11}\left(1-e^{-0.11t}\right)$$

If the state space S is finite, (4.8) gives for each fixed i a *finite* linear system of differential equations (in fact the index i enters only through the initial condition (4.6)). Accordingly, for given transition rates μ_{ij}, Equation (4.8) has a unique solution compatible with (4.6). For this reason Markov models are normally formulated simply by specifying their transition rates μ_{ij}.

5 Kolmogorov's backward differential equations

Substituting $s = h$ in (4.5) and proceeding as before, we obtain a different set of equations, known as Kolmogorov's backward equations.

Kolmogorov's backward differential equations (time-homogeneous case)

These can be written in matrix form as:

$$\frac{d}{dt}\boldsymbol{P}(\boldsymbol{t}) = \boldsymbol{AP}(\boldsymbol{t})$$

Question

Derive this differential equation.

Solution

Substituting $s = h$ into (4.5) gives:

$$p_{ij}(t+h) = \sum_{k \in S} p_{ik}(h)\, p_{kj}(t)$$

Now, since:

$$p_{ik}(h) = h\,\mu_{ik} + o(h) \quad \text{for } k \neq i$$

and: $$p_{ii}(h) = 1 - \sum_{k \neq i} p_{ik}(h) = 1 - h\sum_{k \neq i} \mu_{ik} + o(h) = 1 + h\mu_{ii} + o(h)$$

we have:

$$\begin{aligned} p_{ij}(t+h) &= \sum_{k \neq i} h\mu_{ik}\, p_{kj}(t) + \left(1 + h\mu_{ii}\right) p_{ij}(t) + o(h) \\ &= p_{ij}(t) + h\sum_{k \in S} \mu_{ik}\, p_{kj}(t) + o(h) \end{aligned}$$

If we then take the $p_{ij}(t)$ term to the left-hand side, divide by h and then take the limit $h \to 0$ we obtain the differential equation:

$$\frac{d}{dt} p_{ij}(t) = \sum_{k \in S} \mu_{ik}\, p_{kj}(t) \quad \text{for all } i, j$$

or, equivalently:

$$\frac{d}{dt} P(t) = AP(t)$$

Example

For the time-homogeneous HSD model in Section 3.4, Kolmogorov's backward differential equation for $p_{HH}(t)$ can be obtained using the general backward equation as a template. This gives:

$$\frac{d}{dt}p_{HH}(t)=\mu_{HH}\,p_{HH}(t)+\mu_{HS}\,p_{SH}(t)+\mu_{HD}\,p_{DH}(t)$$

Now substituting in the transition rates, we have:

$$\frac{d}{dt}p_{HH}(t)=-(\sigma+\mu)\,p_{HH}(t)+\sigma\,p_{SH}(t)$$

Once again, it is important to be able to write these equations down without resorting to the general equation. We can think about the RHS of the equation above in the following way:

- start with the force of transition that keeps us in state H at the start (*ie* $-(\sigma+\mu)$) and multiply this by the probability of going from *H* to *H* over an interval of length t (*ie* $p_{HH}(t)$)
- then add on the force of transition that takes us from *H* to S at the start (*ie* σ) multiplied by the probability of going from *S* to *H* over an interval of length t (*ie* $p_{SH}(t)$).

Question

Write down the backward equation for the transition probability $p_{HS}(t)$.

Solution

The backward differential equation is:

$$\begin{aligned}\frac{d}{dt}p_{HS}(t)&=\mu_{HH}p_{HS}(t)+\mu_{HS}p_{SS}(t)+\mu_{HD}p_{DS}(t)\\&=-(\sigma+\mu)p_{HS}(t)+\sigma p_{SS}(t)\end{aligned}$$

Under 'normal' circumstances, the forward and the backward systems are equivalent; this is so in particular when the transition rates are bounded:

$$\sup_{i,j}|\mu_{ij}|<\infty$$

Here 'sup' stands for 'supremum'. Technically, this is the least upper bound of a set. With finite sets, this is the largest value in the set, and you could write 'max' instead of 'sup'. For example, the supremum of the set $\{0,1,2\}$ is 2, the same as the maximum value. The supremum of the set $(0,2)$ is also 2, since 2 is the smallest number that is greater than or equal to all the numbers in the set $(0,2)$.

However, when this condition fails, the backward system is of more fundamental importance.

The forward equations are more useful in numerical work for actuarial applications because we usually have an initial condition such as knowing that a policyholder is healthy when a policy is sold, thus we want equations that we can solve forwards in time from that starting point.

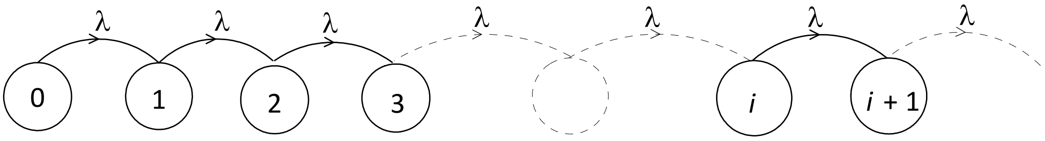
6 The Poisson process revisited

We have already mentioned (at the end of Section 2) that the Poisson process can be formulated as a Markov jump process. We now revisit this idea.

Consider the Markov jump process with state space $S = \{0, 1, 2, ...\}$ and transition rates:

$$\mu_{ij} = \begin{cases} -\lambda & \text{if } j = i \\ \lambda & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

The diagram representation is:



Recall that, in a Poisson process, events occur one at a time, and it is impossible to move to a lower-numbered state. So, when it leaves state i, the process must enter state $i+1$.

The generator matrix A in Kolmogorov's equations is:

$$A = \begin{pmatrix} -\lambda & \lambda & & & \mathbf{0} \\ & -\lambda & \lambda & & \\ & & -\lambda & \lambda & \\ & & & \ddots & \\ \mathbf{0} & & & & \ddots \end{pmatrix}$$

This leads to the forward equations:

$$\begin{cases} p'_{i0}(t) &= -\lambda p_{i0}(t) \\ p'_{ij}(t) &= \lambda p_{ij-1}(t) - \lambda p_{ij}(t), \; j > 0 \end{cases}$$

essentially identical to (4.3) and (4.4).

It is interesting also to consider the backward equations:

$$p'_{ij}(t) = -\lambda p_{ij}(t) + \lambda p_{i+1,j}(t)$$

which of course have the same solution as the forward equations despite looking dissimilar.

7 Holding times and occupancy probabilities

We shall see that time-homogeneous processes are an extension of the Poisson process. Rather than increasing in unit steps there can be transitions from each state to any other one. However, the way in which this occurs and the timing of the transitions has a lot in common with the Poisson process.

The exponential character of the holding times of the Poisson process is no accident. The *memoryless property:*

$$P[T > t+u \mid T > t] = P[T > u]$$

which characterises exponentially distributed random variables, is a necessary requirement for the holding times of time-homogeneous Markov processes.

Consider the first holding time $T_0 = \inf\{t : X_t \neq X_0\}$.

The infimum, inf, is the greatest lower bound of a set. The first holding time is therefore the length of time before the process first changes state.

Distribution of the first holding time

The first holding time of a time-homogeneous Markov jump process with transition rates μ_{ij} is exponentially distributed with parameter:

$$\lambda_i = -\mu_{ii} = \sum_{j \neq i} \mu_{ij}$$

In other words:

$$P[T_0 > t \mid X_0 = i] = e^{-\lambda_i t}$$

The proof of this result is beyond the syllabus.

Here we are defining λ_i to be the total force of transition out of state i.

We will also use the notation $p_{\overline{ii}}(t)$ to denote the probability of remaining in a state i throughout a period of length t, so $p_{\overline{ii}}(t) = e^{-\lambda_i t}$ also. Unlike the Poisson process, the first holding time depends on the initial state i. However, when given i, the holding time is still exponentially distributed with parameter λ_i.

Question

State the expected value of the first holding time for a time-homogeneous Markov jump process that starts in state i.

Solution

The first holding time in i has an exponential distribution with parameter λ_i. The average time is therefore $\dfrac{1}{\lambda_i}$.

In the Poisson process the timing of the jumps is everything.

This is because the value of the process goes up by 1 at a time.

However, in the more general setting of a Markov jump process, we must also characterise the state to which the process jumps; this is remarkably simple: the jump takes place from $X_0 = i$ to $X_{T_0} = j$ with probability *proportional to the transition rate* μ_{ij} and, moreover, the destination of the jump X_{T_0} is *independent* of the holding time T_0. In order to see this consider for $j \neq i$:

$$\begin{aligned}
P\left[X_{t+h} = j, t < T_0 \leq t+h \mid X_0 = i\right] &= P\left[X_{t+h} = j, T_0 > t \mid X_0 = i\right] \\
&= P\left[X_{t+h} = j \mid X_0 = i, T_0 > t\right] P\left[T_0 > t \mid X_0 = i\right] \\
&= P\left[X_{t+h} = j \mid X_s = i,\ 0 \leq s \leq t\right] e^{-\lambda_i t} \\
&= p_{ij}(h)\, e^{-\lambda_i t}
\end{aligned}$$

Now, divide by h and let $h \to 0$: the joint probability distribution/density of X_{T_0} and T_0 is, conditionally on $X_0 = i$, equal to:

$$\mu_{ij} e^{-\lambda_i t}$$

So it is the product of the density of the holding time $\lambda_i e^{-\lambda_i t}$ and of $\dfrac{\mu_{ij}}{\lambda_i}$.

This proves two results at once: the probability that the jump out of i is to state j is:

$$P\left[X_{T_0} = j \mid X_0 = i\right] = \frac{\mu_{ij}}{\lambda_i} \qquad (j \neq i)$$

and moreover X_{T_0} is independent of T_0.

These results are important and they are worth restating.

Probability that the process goes into state *j* when it leaves state *i*

Given that a time-homogeneous Markov jump process is currently in state i, the probability that it moves into state j when it leaves state i is given by:

$$\frac{\mu_{ij}}{\lambda_i} = \frac{\text{the force of transition from state } i \text{ to state } j}{\text{the total force of transition out of state } i}$$

Also, given a jump has occurred, the time at which it took place does not affect the probability of the jump being to a particular state.

As a result of the Markov property, the pattern is identical for successive jumps: after some state j is entered, the process stays there for an exponentially distributed time with parameter λ_j. It then jumps to state k with probability $\frac{\mu_{jk}}{\lambda_j}$.

This is a key result for time-homogeneous Markov jump processes, so we will restate it.

Distribution of holding time random variables and occupancy probabilities

For a time-homogeneous Markov jump process, let W_i denote the holding time (or waiting time) in state i. Then:

$$W_i \sim Exp(\lambda_i)$$

where λ_i is the total force of transition out of state i.

So the probability of staying in state i for at least t time units (*ie* the occupancy probability for state i) is:

$$P(W_i > t) = p_{\overline{ii}}(t) = e^{-\lambda_i t}$$

Note that the mean holding time of state j is $\frac{1}{\lambda_j}$; this is an important thing to remember when assigning numerical values to the transition rates.

So if, for example, the transition rate is 12 per hour, the mean holding time is $1/12$ hour, *ie* 5 minutes.

8 Expected time to reach state *k* starting from state *i*

Let m_i denote the expected time for the process to reach state k given that it is currently in state i. Then m_i can be calculated using the recursive formula:

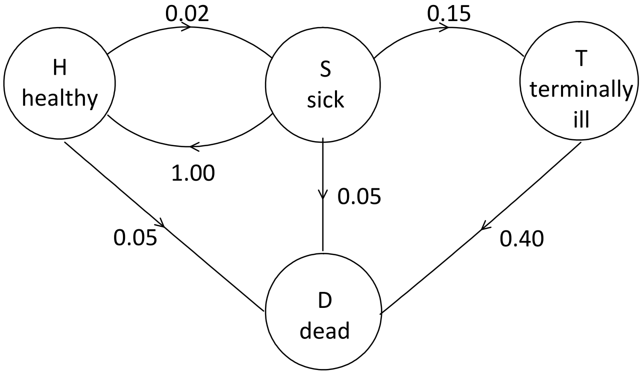
$$m_i = \frac{1}{\lambda_i} + \sum_{j \neq i,k} \frac{\mu_{ij}}{\lambda_i} m_j$$

This formula is given on page 38 of the *Tables*. Note that the *Tables* use the notation σ_{ij} instead of μ_{ij} to denote the force of transition from state i to state j.

The first term on the RHS is the expected holding time in state i. When the process leaves state i, the probability that it goes into state j is $\frac{\mu_{ij}}{\lambda_i}$, as we have just seen in Section 7. We then multiply this probability by the expected time to reach state k starting from state j, namely m_j, and sum over all possible values of j.

Question

Consider the following Health, Sickness, Death model with the addition of an extra 'Terminally ill' state, T. The rates given are per year.



(i) Calculate the expected holding time in state S.

(ii) Calculate the probability that a sick life goes into state D when it leaves the sick state.

(iii) Calculate the expected future lifetime of a healthy life.

Solution

(i) ***Expected holding time in state S***

The total rate out is 1.2 so the expected holding time is $\frac{1}{1.2} = \frac{5}{6}$ years.

(ii) ***Probability that a sick life goes into state D when it leaves the sick state***

This is the proportion of the total rate out of S that goes to D, *ie*:

$$\frac{0.05}{1.2}=\frac{1}{24}$$

(iii) ***Expected future lifetime of a healthy life***

Let m_i denote the expected future lifetime of a life in state i. We have:

$$\lambda_H = 0.02 + 0.05 = 0.07$$

So:

$$m_H = \frac{1}{0.07}+\frac{0.02}{0.07}m_S+\frac{0.05}{0.07}m_D=\frac{100}{7}+\frac{2}{7}m_S$$

since $m_D = 0$. Also:

$$\lambda_S = 1.00 + 0.15 + 0.05 = 1.20$$

So:

$$m_S = \frac{1}{1.20}+\frac{1.00}{1.20}m_H+\frac{0.15}{1.20}m_T$$

But:

$$m_T = \frac{1}{0.40}=2.5$$

So:

$$m_S = \frac{1}{1.20}+\frac{1.00}{1.20}m_H+\frac{0.15}{1.20}\times 2.5=\frac{5}{6}m_H+\frac{55}{48}$$

and:

$$m_H = \frac{100}{7}+\frac{2}{7}m_S=\frac{100}{7}+\frac{2}{7}\left(\frac{5}{6}m_H+\frac{55}{48}\right)$$

This rearranges to give:

$$\left(1-\frac{5}{21}\right)m_H=\frac{100}{7}+\frac{2}{7}\times\frac{55}{48}=\frac{2,455}{168}$$

ie: $m_H = 19.18$

So the expected future lifetime of a healthy life is 19.18 years.

9 The jump chain

If a Markov jump process is examined only at the times of its transitions, the resulting process, denoted $\{\hat{X}_n : n = 0,1,\ldots\}$, where $\hat{X}_0$ is the initial state, and for $n \geq 1$:

$$\hat{X}_n = X_{T_0+T_1+\cdots+T_{n-1}}$$

is called the *jump chain* associated with X.

The jump chain is also sometimes called the embedded chain. It is the sequence of states that the process is observed to take. The time spent in each state is ignored.

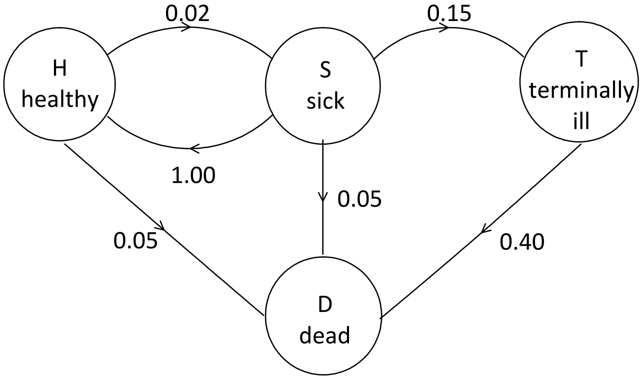
The foregoing analysis shows that $\hat{X}_n$ is independent of $T_0+T_1+\cdots+T_{n-1}$, *ie* the time of the nth transition, and is also independent of anything that happened prior to the $(n-1)$th transition: in fact, the distribution of $\hat{X}_n$ depends only on $\hat{X}_{n-1}$. In other words, the jump chain possesses the Markov property and is a Markov chain in its own right.

The only way in which the jump chain differs from a standard Markov chain is when the jump process $\{X_t, t \geq 0\}$ encounters an absorbing state. From that time on it makes no further transitions, implying that time stops for the jump chain. In order to deal with the jump chain entirely within the framework of Markov chains it is permissible to treat the absorbing state in the same way as for a Markov chain, so that transitions continue to occur but the chain remains in the same state after the transition.

Questions dealing solely with the sequence of states visited by the Markov jump process, such as 'What is the probability that it visits state i_0 before it reaches the absorbing state?' or 'Is state j visited infinitely often?', can be answered equally well with reference to the jump chain, since the two processes take identical paths through the state space. The theory of Markov chains can therefore be employed to arrive at solutions to such questions. Questions dealing with the time taken to visit a state, however, are likely to have very different answers in the two cases and are only accessible using the theory of Markov jump processes.

Question

Consider the following Health, Sickness, Death model with the addition of an extra 'Terminally ill' state, T. The rates given are per year.



Calculate the probability that a life in the sick state dies without ever recovering to the healthy state.

Solution

Here we can use the jump chain since the times are irrelevant. The life must either go straight to the dead state at the next jump, or to state T. If the life goes to state T, then it definitely dies without recovering (as it cannot then re-enter state H or state S). The probability is therefore:

$$\frac{0.05+0.15}{1.00+0.05+0.15}=\frac{0.20}{1.20}=\frac{1}{6}$$

Alternatively, we could calculate the required probability as:

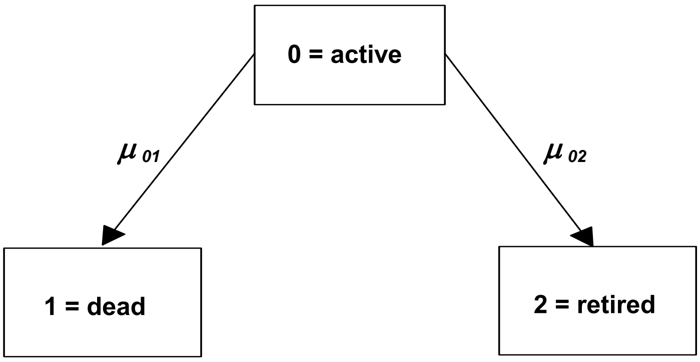
$$1-P(\text{life enters state H when it leaves state S})$$

ie:

$$1-\frac{1.00}{1.20}=\frac{1}{6}$$

10 Solutions of Kolmogorov equation in elementary cases

In certain elementary cases, the solutions of the Kolmogorov equation can simply be written down, and the two-state model is often an intuitive guide. For example, consider the two-decrement model, in which the transition intensities are constant.



Note that the term *active* is usually applied to individuals in employment, in order to differentiate them from individuals who are healthy but who have retired.

We have:

$$p_{01}(x,x+t)=\frac{\mu_{01}}{\mu_{01}+\mu_{02}}\left[1-e^{-(\mu_{01}+\mu_{02})t}\right]$$

$$p_{02}(x,x+t)=\frac{\mu_{02}}{\mu_{01}+\mu_{02}}\left[1-e^{-(\mu_{01}+\mu_{02})t}\right]$$

Here the Core Reading is using the notation $p_{ij}(x,x+t)$ to denote the probability that a life is in state j at age $x+t$, given that it was in state i at age x. We could equally well have used the notation $p_{ij}(t)$ (since the transition probabilities depend only on the length of the time interval, t).

Question

Write down the Kolmogorov forward differential equations for $p_{01}(x,x+t)$ and $p_{02}(x,x+t)$. Hence derive the two equations above.

Solution

The Kolmogorov forward equations for this two-decrement model are:

$$\frac{\partial}{\partial t}p_{01}(x,x+t)=p_{00}(x,x+t)\mu_{01}$$

and:

$$\frac{\partial}{\partial t}p_{02}(x,x+t)=p_{00}(x,x+t)\mu_{02}$$

Since it is impossible to leave the active state and subsequently return to it:

$$p_{00}(x,x+t)=p_{\overline{00}}(x,x+t)$$

So:

$$p_{00}(x,x+t)=e^{-(\mu_{01}+\mu_{02})t}$$

So the Kolmogorov equations can be written as:

$$\frac{\partial}{\partial t}p_{01}(x,x+t)=\mu_{01}\,e^{-(\mu_{01}+\mu_{02})t}$$

and:

$$\frac{\partial}{\partial t}p_{02}(x,x+t)=\mu_{02}\,e^{-(\mu_{01}+\mu_{02})t}$$

Integrating the first of these equations with respect to t gives:

$$p_{01}(x,x+t)=-\frac{\mu_{01}}{\mu_{01}+\mu_{02}}e^{-(\mu_{01}+\mu_{02})t}+C$$

where C is a constant of integration.

Since $p_{01}(x,x)=0$, it follows that:

$$C=\frac{\mu_{01}}{\mu_{01}+\mu_{02}}$$

So:

$$p_{01}(x,x+t)=\frac{\mu_{01}}{\mu_{01}+\mu_{02}}\left(1-e^{-(\mu_{01}+\mu_{02})t}\right)$$

Similarly, integrating the second equation and using the initial condition $p_{02}(x,x)=0$, we obtain:

$$p_{02}(x,x+t)=\frac{\mu_{02}}{\mu_{01}+\mu_{02}}\left(1-e^{-(\mu_{01}+\mu_{02})t}\right)$$

These probability formulae are easily interpreted – the term in brackets is the probability of having left the active state, and the fraction gives the conditional probability of each decrement having occurred, given that one of them has occurred.

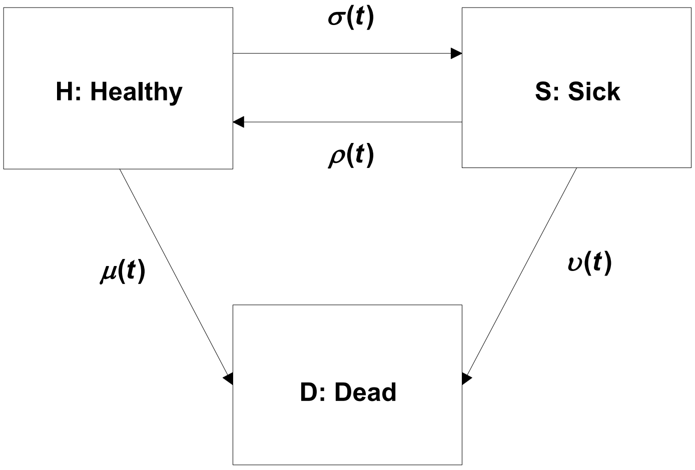
$p_{\overline{00}}(x,x+t)=e^{-(\mu_{01}+\mu_{02})t}$ is the probability that an active life aged x stays in the active state (state 0) up to age $x+t$.

However, in practice we do not always work with such simple models, or with constant transition intensities, and it is not possible to rely on solving the equations explicitly. Fortunately this does not matter; the Kolmogorov equations are quite simple to solve using ordinary numerical techniques.

In the computer-based part of Subject CS2, we will see how R can be used to calculate transition probabilities.

11 The maximum likelihood estimator in the general model

As we saw earlier in this chapter, the two-state model can be extended to any number of states, with arbitrary transitions between them, including increments and repeated transitions. Consider again the illness-death model, which has three states: healthy (H), sick (S) and dead (D):



The observations in respect of a single life are now:

(a) the times between successive transitions; and

(b) the numbers of transitions of each type.

If the transition intensities are constant, each spell of length *t* in the healthy or sick states contributes a factor of the form $e^{-(\mu+\sigma)t}$ or $e^{-(\upsilon+\rho)t}$ respectively to the likelihood, so it suffices to record the total waiting time spent in each state.

We saw in Section 7 that the probability of staying in state i for at least another t time units is $e^{-\lambda_i t}$, where λ_i denotes the total force of transition out of state i.

Question

Given that the chance of becoming sick or dying increases with age, comment on the appropriateness of the assumption that transition intensities are constant.

Solution

Although the chance of becoming sick or dying does usually increase with age, we are usually observing a large number of lives simultaneously over a narrow age interval, *ie* between ages x and $x+1$. Provided we confine our study to such intervals, it may be appropriate to assume that transition intensities over these intervals are constant.

It is harder to justify the assumption that the transition rate for recovery is constant. In practice, this will vary significantly with the duration of sickness.

The Core Reading now defines some notation. This is not standard notation, and it is quite cumbersome. You should be able to deal with whatever notation is used in a given situation. We will introduce some more general notation at the end of Section 11.1.

Define:

V_i = **Waiting time of the i th life in the healthy state**

W_i = **Waiting time of the i th life in the sick state**

S_i = **Number of transitions healthy → sick by the ith life**

R_i = **Number of transitions sick → healthy by the ith life**

D_i = **Number of transitions healthy → dead by the ith life**

U_i = **Number of transitions sick → dead by the ith life**

We also need to define totals $V = \sum_{i=1}^{N} V_i$ **(and so on).**

11.1 Maximum likelihood estimators

Using lower case symbols for the observed samples as usual, it is easily shown that the likelihood for the four parameters, $\mu, \upsilon, \sigma, \rho$**, given the data is proportional to:**

$$L(\mu,\upsilon,\sigma,\rho) = e^{-(\mu+\sigma)v} e^{-(\upsilon+\rho)w} \mu^d \upsilon^u \sigma^s \rho^r$$

This result is obtained using a similar method to that for the two-state model, as set out in Chapter 3. The likelihood function $L(\mu,\upsilon,\sigma,\rho)$ for the i th life reflects:

- the probability of the life remaining in the healthy state for total time v_i and in the sick state for time w_i, giving the factors $e^{-(\mu+\sigma)v_i}$ and $e^{-(\upsilon+\rho)w_i}$ respectively
- the probability of the life making the relevant number of transitions between states, giving the factors $\mu^{d_i}, \upsilon^{u_i}, \sigma^{s_i}$ and ρ^{r_i}.

The likelihood factorises into functions of each parameter of the form $e^{-\mu v}\mu^d$ **:**

ie
$$L(\mu,\upsilon,\sigma,\rho) = e^{-(\mu+\sigma)v} e^{-(\upsilon+\rho)w} \mu^d \upsilon^u \sigma^s \rho^r$$
$$= \left(e^{-\mu v}\mu^d\right) \times \left(e^{-\sigma v}\sigma^s\right) \times \left(e^{-\upsilon w}\upsilon^u\right) \times \left(e^{-\rho w}\rho^r\right)$$

So the log-likelihood is:

$$\log L = -(\mu+\sigma)v - (\upsilon+\rho)w + d\log\mu + u\log\upsilon + s\log\sigma + r\log\rho$$

Differentiating this with respect to each of the four parameters gives:

$$\frac{\partial \log L}{\partial \mu} = -v + \frac{d}{\mu} \qquad \frac{\partial \log L}{\partial \upsilon} = -w + \frac{u}{\upsilon}$$

$$\frac{\partial \log L}{\partial \sigma} = -v + \frac{s}{\sigma} \qquad \frac{\partial \log L}{\partial \rho} = -w + \frac{r}{\rho}$$

Setting each of these derivatives equal to 0 and solving the resulting equations, we see that:

$$\mu = \frac{d}{v} \qquad \upsilon = \frac{u}{w} \qquad \sigma = \frac{s}{v} \qquad \rho = \frac{r}{w}$$

When there is more than one parameter to be estimated, the second order condition to check for maxima is that the Hessian matrix is negative definite, or equivalently, the eigenvalues of the Hessian matrix are all negative. The Hessian matrix is the matrix of second derivatives. So in this case we consider the matrix:

$$\begin{pmatrix} \frac{\partial^2 \ln L}{\partial \mu^2} & \frac{\partial^2 \ln L}{\partial \mu \partial \upsilon} & \frac{\partial^2 \ln L}{\partial \mu \partial \sigma} & \frac{\partial^2 \ln L}{\partial \mu \partial \rho} \\ \frac{\partial^2 \ln L}{\partial \upsilon \partial \mu} & \frac{\partial^2 \ln L}{\partial \upsilon^2} & \frac{\partial^2 \ln L}{\partial \upsilon \partial \sigma} & \frac{\partial^2 \ln L}{\partial \upsilon \partial \rho} \\ \frac{\partial^2 \ln L}{\partial \sigma \partial \mu} & \frac{\partial^2 \ln L}{\partial \sigma \partial \upsilon} & \frac{\partial^2 \ln L}{\partial \sigma^2} & \frac{\partial^2 \ln L}{\partial \sigma \partial \rho} \\ \frac{\partial^2 \ln L}{\partial \rho \partial \mu} & \frac{\partial^2 \ln L}{\partial \rho \partial \upsilon} & \frac{\partial^2 \ln L}{\partial \rho \partial \sigma} & \frac{\partial^2 \ln L}{\partial \rho^2} \end{pmatrix} = \begin{pmatrix} -\frac{d}{\mu^2} & 0 & 0 & 0 \\ 0 & -\frac{u}{\upsilon^2} & 0 & 0 \\ 0 & 0 & -\frac{s}{\sigma^2} & 0 \\ 0 & 0 & 0 & -\frac{r}{\rho^2} \end{pmatrix}$$

Since this is a negative definite matrix, the maximum likelihood estimates of $\mu, \upsilon, \sigma, \rho$ are:

$$\hat{\mu} = \frac{d}{v} \qquad \hat{\upsilon} = \frac{u}{w} \qquad \hat{\sigma} = \frac{s}{v} \qquad \hat{\rho} = \frac{r}{w}$$

However, checking the Hessian matrix is beyond the scope of the Subject CS2 syllabus.

The maximum likelihood estimators are:

$$\tilde{\mu} = \frac{\boldsymbol{D}}{\boldsymbol{V}}, \quad \tilde{\upsilon} = \frac{\boldsymbol{U}}{\boldsymbol{W}}, \quad \tilde{\sigma} = \frac{\boldsymbol{S}}{\boldsymbol{V}}, \quad \tilde{\rho} = \frac{\boldsymbol{R}}{\boldsymbol{W}}$$

What we have just seen is a special case of a general result.

Estimating transition rates in a time-homogeneous Markov jump process

The maximum likelihood estimate of the transition rate μ_{ij} is:

$$\hat{\mu}_{ij} = \frac{n_{ij}}{t_i}$$

where n_{ij} is the number of transitions from state i to state j, and t_i is the total waiting time (or total holding time) in state i.

Example

During a large study into rates of sickness:

- 2,710 healthy lives fell sick and 2,490 sick lives recovered
- 70 healthy lives and 120 sick lives died.

For the whole group, the periods of health and sickness totalled 41,200 and 6,700 years, respectively.

So we have:

$$t_H = 41{,}200 \qquad t_S = 6{,}700$$

$$n_{HS} = 2{,}710 \qquad n_{SH} = 2{,}490$$

$$n_{HD} = 70 \qquad n_{SD} = 120$$

and the maximum likelihood estimates of the transition rates are:

$$\hat{\sigma} = \frac{n_{HS}}{t_H} = \frac{2{,}710}{41{,}200} = 0.0658 \qquad \hat{\rho} = \frac{n_{SH}}{t_S} = \frac{2{,}490}{6{,}700} = 0.3716$$

$$\hat{\mu} = \frac{n_{HD}}{t_H} = \frac{70}{41{,}200} = 0.0017 \qquad \hat{\upsilon} = \frac{n_{SD}}{t_S} = \frac{120}{6{,}700} = 0.0179$$

11.2 Properties of the estimators

The asymptotic properties of these estimators follow from results similar to equations (3.2) and (3.3) in Section 4.2 of Chapter 3, and the fact that the random variables $(D_i - \mu V_i)$, $(U_i - \upsilon W_i), (S_i - \sigma V_i), (R_i - \rho W_i)$ are uncorrelated, that is:

$$E\left[(D_i - \mu V_i)(U_i - \upsilon W_i)\right] = 0 \;\; etc$$

Recall that:

V_i = Waiting time of the i th life in the healthy state

W_i = Waiting time of the i th life in the sick state

S_i = Number of transitions healthy $\rightarrow$ sick by the i th life

R_i = Number of transitions sick $\rightarrow$ healthy by the i th life

D_i = Number of transitions healthy $\rightarrow$ dead by the i th life

U_i = Number of transitions sick $\rightarrow$ dead by the i th life

The estimators are not independent: D_i and U_i are both 0 or 1, but $D_i U_i \neq 1$, while (assuming that the i th life starts in the able state) $S_i = R_i$ or $R_i + 1$.

Question

Explain in words why:

(i) D_i and U_i are both 0 or 1, but $D_i U_i \neq 1$

(ii) assuming that the i th life starts in the healthy state, $S_i = R_i$ or $R_i + 1$.

Solution

(i) A life must be in one of two states at the point of death. The life may be in the healthy state ($D_i = 1$) *or* it may be in the sick state ($U_i = 1$). It cannot be in both states, so $D_i U_i \neq 1$. (In fact, $D_i U_i$ always equals zero.)

(ii) Suppose that a life starts in the healthy state. If it is in the healthy state at the point of death, then it must have made the same number of transitions from healthy to sick as from sick to healthy (*ie* $S_i = R_i$). If it is in the sick state at the point of death, then it must have made one more transition from healthy to sick than it did from sick to healthy, in which case $S_i = R_i + 1$.

The estimators are, however, asymptotically independent: the same argument as in the two-state model shows that:

- **the vector $(\tilde{\mu}, \tilde{\upsilon}, \tilde{\sigma}, \tilde{\rho})$ has an asymptotic multivariate normal distribution;**
- **each component has a marginal asymptotic distribution of the same form as before:**

$$\tilde{\mu} \sim \textbf{Normal}\left(\mu, \frac{\mu}{\mathbf{E[V]}}\right) \textbf{ etc}$$

- **asymptotically, the components are uncorrelated and so independent (being normal).**

Recall that defining several random variables simultaneously on a sample space gives rise to a multivariate distribution.

Question

State the marginal asymptotic distributions of $\tilde{\upsilon}$, $\tilde{\sigma}$ and $\tilde{\rho}$.

Solution

$$\tilde{\upsilon} \sim N\left(\upsilon, \frac{\upsilon}{E[W]}\right) \qquad \tilde{\sigma} \sim N\left(\sigma, \frac{\sigma}{E[V]}\right) \qquad \tilde{\rho} \sim N\left(\rho, \frac{\rho}{E[W]}\right)$$

Recall that maximum likelihood estimators have the following properties:

- they are asymptotically normally distributed
- they are asymptotically unbiased (*ie* if $\tilde{\theta}$ is an estimator of θ, then $E(\tilde{\theta}) = \theta$)
- asymptotically, the variance of a maximum likelihood estimator is equal to the Cramér-Rao lower bound (CRLB). The formula for the CRLB is given on page 23 of the *Tables*.

So the maximum likelihood estimators $\tilde{\mu}_{ij}$ of the transition rates μ_{ij} all have the above properties. These results can be used to construct confidence intervals for the transition intensities or as the basis for hypothesis tests.

11.3 Calculating the total waiting time

The calculation of the estimates $\hat{\mu}$, *etc*, requires the total waiting time to be computed. This can be done exactly in some circumstances, but, if the exposure data are in census form, the simple census formulae in Chapter 9 provide estimates. Multiple state models are, therefore, especially well suited to the data available in many actuarial investigations.

In order to calculate total waiting time exactly, we would need to know the exact timing of each transition. This may not be possible in practice if full information is not available. Alternatively, it may be possible to perform the calculations but the process may be too time-consuming for it to be worthwhile.

A simpler approach to data collection is the census approach, in which a series of observations ('snapshots') of a population is recorded, usually at regular intervals. Data in census form do not allow us to calculate waiting times exactly, but some simplifying assumptions allow us to calculate estimates quite accurately.

For example, we may observe 100 nonagenarians on 1 January 2018 and find that only 84 of these individuals are still alive at 1 January 2019. We could estimate the total waiting time by assuming that deaths occurred half way through the year on average. The accuracy of our estimated transition intensities would depend on the suitability of our assumption.

Actuarial investigations typically use the census approach to data collection, *eg* population studies from national censuses or the analysis of the experience of a pension scheme or a life insurance company as part of the regular valuation process. We will look at this area in more detail in Chapter 9.

A range of packages have been written in R to implement multiple state models. Several of these are described, with illustrative code, in Beyersmann, J., Schumacher, M. and Allignol, A. *Competing Risks and Multistate Models with R* (London, Springer, 2012).

Chapter 4 Summary

Poisson processes

Let $\{N_t\}_{t\geq 0}$ be an increasing, integer-valued process starting at 0 (and continuous from the right). Let $\lambda > 0$. Then $\{N_t\}_{t\geq 0}$ is a Poisson process with rate λ if any of the following four equivalent conditions hold:

(1) $\{N_t\}_{t\geq 0}$ has stationary, independent increments and for each t, N_t has a Poisson distribution with parameter λt.

(2) $\{N_t\}_{t\geq 0}$ is a Markov jump process with independent increments and transition probabilities over a short time period of length h given by:

$$P\left[N_{t+h} - N_t = 1 \mid F_t\right] = \lambda h + o(h)$$

$$P\left[N_{t+h} - N_t = 0 \mid F_t\right] = 1 - \lambda h + o(h)$$

$$P\left[N_{t+h} - N_t \neq 0,1 \mid F_t\right] = o(h)$$

(3) The holding times, $T_0, T_1, \ldots$ of $\{N_t\}_{t\geq 0}$ are independent exponential random variables with parameter λ and $N_{T_0+T_1+\cdots+T_{n-1}} = n$.

(4) $\{N_t\}_{t\geq 0}$ is a Markov jump process with independent increments and transition rates given by:

$$\mu_{ij} = \begin{cases} -\lambda & \text{if } j = i \\ \lambda & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

Sums of Poisson processes

If we have two independent Poisson processes with parameters λ and μ, then the sum of the two processes is another Poisson process with parameter $\lambda + \mu$.

Thinning of a Poisson process

When the events in a Poisson process are of different types, each type occurring at random with a certain probability, the events of a particular type form a *thinned* process. The thinned process is also a Poisson process, with rate equal to the original rate multiplied by the probability for the type of event.

Inter-event times in a Poisson process

Suppose that $T_0, T_1, T_2, \ldots$ are the successive inter-event times or holding times in a Poisson process with parameter λ. Then $T_0, T_1, T_2, \ldots$ are independent $Exp(\lambda)$ random variables.

Markov jump processes

A Markov jump process is a stochastic process with a continuous time set and discrete state space that satisfies the Markov property.

Time-homogeneous Markov jump processes

A Markov jump process is said to be time-homogeneous if the transition probabilities $P\left(X_t = j \mid X_s = i\right)$ depend only on the length of the time interval, $t-s$. Then:

$$p_{ij}(t) = P\left(X_t = j \mid X_0 = i\right)$$

Chapman-Kolmogorov equations

$$p_{ij}(s+t) = \sum_k p_{ik}(s) p_{kj}(t) \qquad \text{for all } s,t > 0$$

Transition rates

The transition rates (or transition intensities or forces of transition) for a time-homogeneous Markov jump process are given by:

$$\mu_{ij} = \frac{d}{dt} p_{ij}(t)\bigg|_{t=0} = \lim_{h \to 0} \frac{p_{ij}(h) - p_{ij}(0)}{h}$$

This is equivalent to:

$$p_{ij}(h) = \begin{cases} h\mu_{ij} + o(h) & if\ i \neq j \\ 1 + h\mu_{ii} + o(h) & if\ i = j \end{cases}$$

for small values of h.

Generator matrix

The generator matrix is the matrix of transition rates. It is usually denoted by A. Each row of the generator matrix sums to zero since $\mu_{ii} = -\sum_{j \neq i} \mu_{ij}$.

Backward and forward differential equations (time-homogeneous case)

Forward:
$$\frac{d}{dt}p_{ij}(t)=\sum_{k\in S}p_{ik}(t)\mu_{kj}$$

$$\frac{d}{dt}P(t)=P(t)A \quad \text{(matrix form)}$$

Backward:
$$\frac{d}{dt}p_{ij}(t)=\sum_{k\in S}\mu_{ik}p_{kj}(t)$$

$$\frac{d}{dt}P(t)=AP(t) \quad \text{(matrix form)}$$

Holding time random variables

For a time-homogeneous Markov jump process, let T_i denote the holding time in state i. Then $T_i \sim Exp(\lambda_i)$, where λ_i is the total force of transition out of state i. The expected holding time in state i is $\frac{1}{\lambda_i}$.

Occupancy probabilities

The probability of remaining in state i for at least t time units is:

$$P\left(T_i > t\right)=p_{\overline{ii}}(t)=e^{-\lambda_i t}$$

Probability that the process goes into state *j* when it leaves state *i*

Given that a time-homogeneous Markov jump process is currently in state i, the probability that it moves into state j when it leaves state i is $\frac{\mu_{ij}}{\lambda_i}$.

Expected time to reach a given state

To calculate the expected time to reach a given state, state k, starting from state i, we can apply the following formula recursively:

$$m_i=\frac{1}{\lambda_i}+\sum_{j\neq i,k}\frac{\mu_{ij}}{\lambda_i}m_j$$

This formula is given on page 38 of the *Tables*.

Jump chains

The jump chain (or embedded chain) of a Markov jump process is the sequence of states that the process enters. The time spent in each state is ignored. The jump chain is a Markov chain in its own right.

Estimating transition rates

The maximum likelihood estimate of the transition rate μ_{ij}, $i \neq j$, is:

$$\hat{\mu}_{ij} = \frac{n_{ij}}{t_i}$$

where n_{ij} is the number of transitions from state i to state j, and t_i is the total waiting time (or total holding time) in state i.

The maximum likelihood estimate of the transition rate μ_{ii} is $\hat{\mu}_{ii} = -\sum_{j \neq i} \hat{\mu}_{ij}$.

The maximum likelihood estimator of μ_{ij} has the following properties:

- it is asymptotically normally distributed
- it is asymptotically unbiased
- asymptotically, its variance is given by the Cramér-Rao lower bound (CRLB). The formula for the CRLB is given on page 23 of the *Tables*.

Chapter 4 Practice Questions

4.1 For a Poisson process with rate λ:

(i) state the distribution of the inter-arrival time random variable, T

(ii) give an expression for the probability that exactly one event will occur during a finite time interval of length t.

4.2 Claims on a portfolio of policies occur according to a Poisson process with a mean rate of 5 claims per day. Claim amounts are 10, 20 or 30. 20% of claims are of amount 10, 70% are of amount 20 and 10% are of amount 30.

(i) Calculate the expected waiting time until the first claim of amount 30.

(ii) Calculate the probability that there are at least 10 claims during the first 2 days, given that there were exactly 6 claims during the first day.

(iii) Calculate the probability that there are at least 2 claims of amount 20 during the first day and at least 3 claims of amount 20 during the first 2 days.

(iv) Calculate the conditional variance of the number of claims during the first day, given that there are 2 claims of amount 10 during the first day.

4.3 $\{X_t\}$ is a Markov jump process with state space $S=\{0,1,2,\ldots\}$ and $X_0=0$. The transition rates are given by:

$$\mu_{ij}=\begin{cases}\lambda & \text{if } j=i+1\\ -\lambda & \text{if } j=i\\ 0 & \text{otherwise}\end{cases}$$

(i) Write down the transition probabilities $p_{ij}(t)$.

(ii) Define the term *holding time*.

(iii) Find the distribution of the first holding time T_0.

(iv) State the value of X_{T_0}.

(v) Given that the increments are stationary and independent, state the distributions of $T_0, T_1, T_2, \ldots$. Justify your answer.

4.4 A particular machine is in constant use. Regardless of how long it has been since the last repair, it tends to break down once a day (*ie* once every 24 hours of operation) and on average it takes the repairman 6 hours to fix.

You are modelling the machine's status as a time-homogeneous Markov jump process $\{X(t): t \geq 0\}$ with two states: 'being repaired' denoted by 0, and 'working' denoted by 1.

Let $P_{i,j}(t)$ denote the probability that the process is in state j at time t given that it was in state i at time 0 and suppose that t is measured in days.

(i) State the two main assumptions that you make in applying the model and discuss briefly how you could test that each of them holds.

(ii) Draw the transition graph for the process, showing the numerical values of the transition rates.

(iii) State Kolmogorov's backward and forward differential equations for the probability $P_{0,0}(t)$.

(iv) Solve the forward differential equation in (iii) to show that:

$$P_{0,0}(t)=\frac{1}{5}+\frac{4}{5}e^{-5t}$$

4.5 Exam style

Claims on an insurer's travel insurance policies arriving in the claims department (state A) wait for an average of two days before being logged and classified by a claims administrator as requiring:

- investigation by a loss adjuster (state L),
- more details from the insured (state I),
- no further information is required and the claim should be settled immediately (state S).

Only one new claim in ten is classified as requiring investigation by a loss adjuster, and five in ten require further information from the insured.

If needed, investigation by a loss adjuster takes an average of 10 days, after which 30% of cases require further information from the insured and 70% are sent for immediate settlement.

Collecting further information from the insured takes an average of 5 days to complete, and immediate settlement takes an average of 2 days before the settlement process is complete (state C).

It is suggested that a time-homogeneous Markov process with states A, L, I, S and C could be used to model the progress of claims through the settlement system with the ultimate aim of reducing the average time to settle a claim.

(i) Calculate the generator matrix, $\{\mu_{ij}; i,j=A,L,I,S,C\}$, of such a model. [2]

(ii) Calculate the proportion of claims that eventually require more details from the insured. [2]

(iii) Derive a forward differential equation for the probability that a claim is yet to be logged and classified by a claims administrator at time t. Solve this equation to obtain an expression for the probability. [4]

[Total 8]

4.6

Exam style

An n-state, time-homogeneous Markov jump process with transition probability matrix $P(t)$ over a period of length t, is said to have a stationary distribution, $\underline{\pi} = (\pi_1, \ldots, \pi_n)$, if:

(1) $\underline{\pi} P(t) = \underline{\pi}$

(2) $0 \le \pi_i \le 1$ for each $i = 1, 2, \ldots, n$

(3) $\sum_{i=1}^{n} \pi_i = 1$

(i) Explain why condition (1) is equivalent to the condition $\underline{\pi} A = \underline{0}$ where A is the generator matrix and $\underline{0}$ is an n-dimensional vector whose entries are all 0. [1]

In a particular company the salary scale has only two different levels. On average, an employee spends 2 years at level 1 before moving on to the higher level, or leaving the company. An employee at the maximum level spends an average of 5 years before leaving. Nobody is demoted, promotion can occur at any time, and mortality can be ignored.

Upon leaving level 1, the probability that an employee moves to level 2 is 50%.

(ii) Explain how you could model this as a Markov process, commenting on any assumptions that you make. [2]

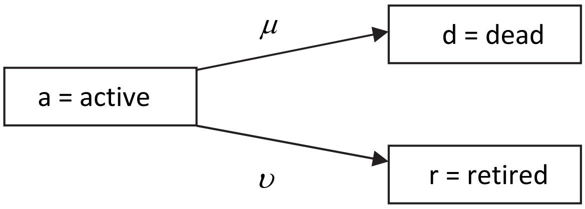
(iii) Derive the generator matrix of the Markov jump process. [2]

(iv) The company currently has 1,000 employees. The proportions at levels 1 and 2 are 60% and 40% respectively. Use a forward differential equation to determine the distribution of these employees in five years' time. You should assuming that nobody joins the company in the future. [6]

[Total 11]

4.7 Exam style

(i) The following multiple state model has been suggested as a representation of deaths and retirements between the ages of 59 and 60. There are no other decrements and the forces of decrement μ and υ are constant. Let ${}_tp_x^{ij}$ denote the probability that a life is in state j at age $x+t$ given that it was in state i at age x.



(a) State the assumptions underlying the above model.

(b) Show that ${}_tp_x^{aa} = e^{-(\mu+\upsilon)t}$ for $59 \le x \le x+t \le 60$.

(c) Suppose that you make the following observations in respect of n identical and statistically independent lives:

v = time spent in the active state

d = number of deaths

r = number of retirements

Assuming that lives are only observed to the earlier of death or retirement, show that the likelihood for μ and υ given these observations is:

$$L(\mu,\upsilon) = e^{-(\mu+\upsilon)v}\mu^d \upsilon^r$$

(d) Give formulae (without proof) for:

– the maximum likelihood estimator of the parameter υ

– the asymptotic expected value of the estimator

– an estimated standard error of the estimator. [16]

(ii) Suppose that you learn that retirements can only take place on a birthday, so that r is the number of retirements at exact age 60. In addition to v, d and r you also observe:

$m =$ number of lives attaining exact age 60, where $m \le n$. Suppose that any life attaining exact age 60 will retire with probability k, where $0 < k < 1$.

(a) State the likelihood for μ and k, given v, d, r and m.

(b) Give a formula (without proof) for the maximum likelihood estimate of the parameter *k*. [4]

[Total 20]

4.8 Exam style

Vehicles in a certain country are required to be assessed every year for road-worthiness. At one vehicle assessment centre, drivers wait for an average of 15 minutes before the road-worthiness assessment of their vehicle commences. The assessment takes on average 20 minutes to complete. Following the assessment, 80% of vehicles are passed as road-worthy allowing the driver to drive home. A further 15% of vehicles are categorised as a 'minor fail'; these vehicles require on average 30 minutes of repair work before the driver is allowed to drive home. The remaining 5% of vehicles are categorised as a 'significant fail'; these vehicles require on average three hours of repair work before the driver can go home.

A continuous-time Markov model is to be used to model the operation of the vehicle assessment centre, with states W (waiting for assessment), A (assessment taking place), M (minor repair taking place), S (significant repair taking place) and H (travelling home).

(i) Identify the distribution of the time spent in each state. [1]

(ii) Write down the generator matrix for this process. [2]

(iii) (a) Use Kolmogorov's forward equations to write down differential equations satisfied by $p_{WM}\left(t\right)$ and by $p_{WA}\left(t\right)$.

(b) Verify that $p_{WA}(t)=4e^{-t/20}-4e^{-t/15}$ for $t\geq 0$, where t is measured in minutes.

(c) Derive an expression for $p_{WM}\left(t\right)$ for $t\geq 0$. [7]

(iv) Let T_i be the expected length of time (in minutes) until the vehicle can be driven home given that the assessment process is currently in state i.

(a) Explain why $T_W = 15+T_A$.

(b) Derive corresponding equations for T_A, T_M and T_S.

(c) Calculate T_W. [4]

[Total 14]

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 4 Solutions

4.1 (i) ***Distribution of the inter-arrival time***

The distribution of the inter-arrival time random variable is $Exp(\lambda)$.

(ii) ***Probability of exactly one event***

The distribution of the number of occurrences in a time period of length t is $Poisson(\lambda t)$. So the probability of exactly one event is $\lambda t e^{-\lambda t}$.

4.2 (i) ***Expected waiting time until the first claim of amount 30***

Claims of amount 30 occur according to a Poisson process with a mean of $0.1 \times 5 = 0.5$ per day. So the waiting time until the first claim of amount 30 has an $Exp(0.5)$ distribution and the expected waiting time is $\frac{1}{0.5} = 2$ days.

(ii) ***Probability that there are at least 10 claims during the first 2 days, given that there were exactly 6 claims during the first day***

Let $N(t)$ denote the number of claims during the interval $\left[0,t\right]$. Then:

$$\begin{aligned} P\left(N(2) \ge 10 \middle| N(1) = 6\right) &= P\left(N(2) - N(1) \ge 4 \middle| N(1) - N(0) = 6\right) \\ &= P\left(N(2) - N(1) \ge 4\right) \end{aligned}$$

since $N\left(0\right) = 0$ and the numbers of claims in non-overlapping time intervals are independent.

Now $N(2) - N(1) \sim Poisson(5)$, so:

$$\begin{aligned} P\left(N(2) \ge 10 \middle| N(1) = 6\right) &= P\left(Poisson(5) \ge 4\right) \\ &= 1 - P\left(Poisson(5) \le 3\right) \\ &= 1 - e^{-5}\left(1 + \frac{5^1}{1!} + \frac{5^2}{2!} + \frac{5^3}{3!}\right) \\ &= 0.73497 \end{aligned}$$

(iii) ***Probability that there are at least 2 claims of amount 20 during the first day and at least 3 claims of amount 20 during the first 2 days***

Let $N_{20}(t)$ denote the number of claims of amount 20 in the interval $[0,t]$. We want:

$$P\left(N_{20}(1) \ge 2, N_{20}(2) \ge 3\right)$$

If we have 3 or more claims during the first day, then the second condition is automatically satisfied. If we have exactly 2 claims on the first day, then we need at least 1 claim on the second day. So the required probability is:

$$P\left(N_{20}(1)\geq 3\right)+P\left(N_{20}(1)=2, N_{20}(2)-N_{20}(1)\geq 1\right)$$

Now $N_{20}(1)$ and $N_{20}(2)-N_{20}(1)$ are both Poisson with mean $0.7\times 5=3.5$. Also, $N_{20}(1)$ and $N_{20}(2)-N_{20}(1)$ are independent. So:

$$P\left(N_{20}(1)\geq 3\right)=1-P\left(N_{20}(1)\leq 2\right)=1-e^{-3.5}\left(1+\frac{3.5^1}{1!}+\frac{3.5^2}{2!}\right)=0.679153$$

and:

$$\begin{aligned}P\left(N_{20}(1)=2, N_{20}(2)-N_{20}(1)\geq 1\right)&=P\left(N_{20}(1)=2\right)P\left(N_{20}(2)-N_{20}(1)\geq 1\right)\\&=P\left(N_{20}(1)=2\right)\left[1-P\left(N_{20}(2)-N_{20}(1)=0\right)\right]\\&=\frac{e^{-3.5}\,3.5^2}{2!}\left[1-e^{-3.5}\right]\\&=0.179374\end{aligned}$$

The required probability is:

$$P\left(N_{20}(1)\geq 3\right)+P\left(N_{20}(1)=2, N_{20}(2)-N_{20}(1)\geq 1\right)=0.679153+0.179374=0.85853$$

(iv) ***Conditional variance***

Let $N_j(t)$, $j=10,20,30$, denote the number of claims of amount j in the interval $[0,t]$. Then:

$$N(1)=N_{10}(1)+N_{20}(1)+N_{30}(1)$$

and:

$$\begin{aligned}\text{var}\left[N(1)\,\middle|\,N_{10}(1)=2\right]&=\text{var}\left[N_{10}(1)+N_{20}(1)+N_{30}(1)\,\middle|\,N_{10}(1)=2\right]\\&=\text{var}\left[2+N_{20}(1)+N_{30}(1)\right]\\&=\text{var}\left[N_{20}(1)+N_{30}(1)\right]\\&=\text{var}\left[N_{20}(1)\right]+\text{var}\left[N_{30}(1)\right]\end{aligned}$$

by independence.

Now, since $N_{20}(1)\sim \textit{Poisson}(3.5)$ and $N_{30}(1)\sim \textit{Poisson}(0.5)$:

$$\text{var}\left[N(1)\,\middle|\,N_{10}(1)=2\right]=3.5+0.5=4$$

4.3 The process defined is a Poisson process with parameter λ.

(i) ***Transition probabilities***

Since the distribution of the increments is $Poisson(\lambda t)$, these are:

$$p_{ij}(t) = \frac{e^{-\lambda t}(\lambda t)^{j-i}}{(j-i)!} \quad \text{for } j \geq i$$

(ii) ***Holding time***

The holding times are inter-event times. In other words, the time spent in a particular state between transitions. For the process given, the i th holding time T_{i-1} is the time spent in state $i-1$ before the transition to state i.

(iii) ***Distribution of the first holding time***

We have:

$$P\left[T_0 > t \mid X_0 = 0\right] = P\left[X_t = 0 \mid X_0 = 0\right] = P_{00}(t) = e^{-\lambda t}$$

and it follows that T_0 has an $Exp(\lambda)$ distribution.

(iv) ***Value of X_{T_0}***

$X_{T_0} = 1$ since we choose the sample paths to be right-continuous. So at time T_0 it has just jumped to 1.

(v) ***Distribution of i th holding time***

Consider T_i:

$$P\left[T_i > t \mid X_0 = 0, \sum_{j=0}^{i-1} T_j = s\right] = P\left[X_{t+s} - X_s = 0 \mid X_0 = 0, \sum_{j=0}^{i-1} T_j = s\right]$$

$$= P\left[X_t - X_0 = 0\right] = P_{00}(t) = e^{-\lambda t}$$

The second equality is due to the increments being independent and stationary. Hence T_i also has an exponential distribution with parameter λ.

4.4 (i) ***Assumptions***

We are assuming that the process is Markov and that the transition rates are constant.

The Markov property of the underlying jump chain can be tested using a chi-squared test based on triplets of successive observations.

A chi-squared test can also be used to test whether the waiting times are exponentially distributed with constant parameter.

(ii) ***Transition graph***

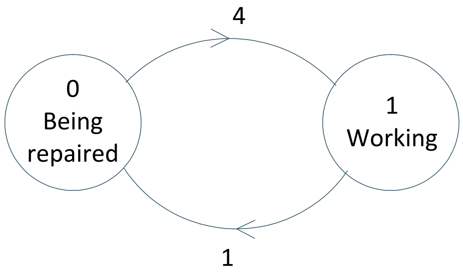
The expected holding time in the broken down state (State 0) is 0.25 days and the expected holding time in the working state (State 1) is 1 day. Since the holding time in State i is $Exp(\lambda_i)$, we have:

$$E(T_0)=\frac{1}{\lambda_0}=0.25 \Rightarrow \lambda_0=4$$

and:

$$E(T_1)=\frac{1}{\lambda_1}=1 \Rightarrow \lambda_1=1$$

The transition graph is as follows:



(iii) ***Kolmogorov's differential equations***

The backward differential equation is:

$$\frac{d}{dt}P_{0,0}(t)=-4P_{0,0}(t)+4P_{1,0}(t)$$

and the forward differential equation is:

$$\frac{d}{dt}P_{0,0}(t)=P_{0,0}(t)\times(-4)+P_{0,1}(t)\times 1=-4P_{0,0}(t)+P_{0,1}(t)$$

(iv) ***Proof***

Since $P_{0,1}(t)=1-P_{0,0}(t)$, we have the differential equation:

$$\frac{d}{dt}P_{0,0}(t)=1-5P_{0,0}(t)$$

together with the boundary condition $P_{0,0}(0)=1$.

We can solve this equation and boundary condition using an integrating factor of e^{5t}:

$$\frac{d}{dt}P_{0,0}(t)e^{5t}+5P_{0,0}(t)e^{5t}=e^{5t}$$

$$\Rightarrow\frac{d}{dt}\left(P_{0,0}(t)e^{5t}\right)=e^{5t}$$

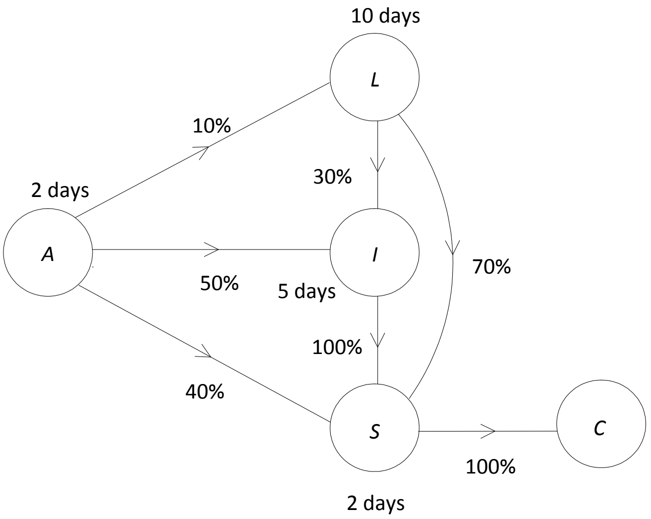
$$\Rightarrow P_{0,0}(t)e^{5t}=\frac{1}{5}e^{5t}+K$$

$$\Rightarrow P_{0,0}(t)=\frac{1}{5}+Ke^{-5t}$$

Applying the boundary condition $P_{0,0}(0)=1\Rightarrow K=\frac{4}{5}$. So we have the required result.

4.5 (i) ***Generator matrix***

The information given in the question about the occupancy times in each state and the transition probabilities in the Markov jump chain can be summarised as:



Note that this is not a proper transition diagram, as a transition diagram must show the forces of transition.

The total force out of each state is equal to the reciprocal of the expected holding time. The percentages indicate how the total force is divided between destination states.

The generator matrix is:

$$\begin{array}{c} \begin{array}{ccccc} A & L & I & S & C \end{array} \\ \begin{bmatrix} -0.5 & 0.5\times 0.1 & 0.5\times 0.5 & 0.5\times 0.4 & 0 \\ 0 & -0.1 & 0.1\times 0.3 & 0.1\times 0.7 & 0 \\ 0 & 0 & -0.2 & 0.2\times 1 & 0 \\ 0 & 0 & 0 & -0.5 & 0.5\times 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$= \begin{bmatrix} -0.5 & 0.05 & 0.25 & 0.2 & 0 \\ 0 & -0.1 & 0.03 & 0.07 & 0 \\ 0 & 0 & -0.2 & 0.2 & 0 \\ 0 & 0 & 0 & -0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

[2]

(ii) ***Proportion of claims that require further details from the insured***

We can list all the paths that correspond to the event of visiting state I if the process starts in state A.

These are $A \to L \to I \to S \to C$ and $A \to I \to S \to C$. [1]

The probabilities of these paths are $0.10\times 0.30\times 1\times 1 = 0.03$ and $0.50\times 1\times 1 = 0.50$. The total probability is 0.53. [1]

Alternatively we can use a more general approach. This has the advantage of working in more complicated situations where the path counting approach becomes very cumbersome.

Let $p_i = P\left[\text{never visit state } I \,\middle|\, \text{currently in state } i\right]$*, then using the Markov jump chain transition matrix:*

$$\begin{bmatrix} 0 & 0.10 & 0.50 & 0.40 & 0 \\ 0 & 0 & 0.30 & 0.70 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

we can write:

$$p_A = 0.1p_L + 0.5p_I + 0.4p_S = 0.07 + 0.40 = 0.47$$

$$p_L = 0.3p_I + 0.7p_S = 0.7$$

$$p_I = 0$$

$$p_S = 1$$

$$p_C = 1$$

The required probability is $1-p_A=1-0.47=0.53$.

(iii) ***Probability that a claim is yet to be logged and classified by a claims administrator at time t***

The Chapman-Kolmogorov equation is:

$$P_{AA}(t+h)=P_{AA}(t)P_{AA}(h)$$

This assumes that the process satisfies the Markov property. [½]

Then the law of total probability allows us to write:

$$P_{AA}(h)+P_{AL}(h)+P_{AI}(h)+P_{AS}(h)+P_{AC}(h)=1 \qquad [½]$$

The definition of the transition rates gives:

$$P_{AL}(h)=0.05h+o(h)$$

$$P_{AI}(h)=0.25h+o(h)$$

$$P_{AS}(h)=0.20h+o(h) \qquad [½]$$

Also, $P_{AC}(h)=o(h)$ because it involves more than one transition. [½]

Substituting we obtain:

$$P_{AA}(h)=1-0.05h-0.25h-0.20h+o(h)$$

$$\Rightarrow P_{AA}(t+h)=P_{AA}(t)\big(1-0.05h-0.25h-0.20h+o(h)\big)$$

$$\Rightarrow \frac{P_{AA}(t+h)-P_{AA}(t)}{h}=-0.50P_{AA}(t)+\frac{o(h)}{h}$$

$$\Rightarrow \frac{d}{dt}P_{AA}(t)=-0.50P_{AA}(t) \qquad [1]$$

Separating the variables:

$$\frac{\frac{d}{dt}P_{AA}(t)}{P_{AA}(t)}=\frac{d}{dt}\ln P_{AA}(t)=-0.50$$

Then integrating with respect to t:

$$\ln P_{AA}(t)=-0.50t+C$$

where C is a constant of integration. Using the initial condition $P_{AA}(0)=1$, we see that $C=0$ and hence:

$$P_{AA}\left(t\right)=e^{-0.50t} \qquad [1]$$

4.6 (i) ***Equivalent first condition for stationary***

Since $A = P'(0)$ and the vector $\underline{\pi}$ is constant, differentiating the equation $\underline{\pi}P(t) = \underline{\pi}$ with respect to t gives and setting $t = 0$ gives:

$$\underline{\pi}A = \underline{0}$$ [1]

(ii) ***Modelling as a Markov process***

This is a 3-state Markov jump process. The states are (1) level 1; (2) level 2; (3) left the company. [½]

We have made the Markov assumption, *ie* that the probability of jumping to any particular state depends only on knowing the current state that is occupied. [1]

We have assumed that transition rates between states are constant over time. [½]

(iii) ***Generator matrix***

The average waiting time in each state, i is exponentially distributed with mean $\frac{1}{\lambda_i}$. The mean times in states 1 and 2 are 2 and 5 years respectively. The values of the exponential parameters are:

$$\lambda_1 = \frac{1}{2} \quad \lambda_2 = \frac{1}{5}$$ [½]

The transition matrix of the jump chain, p_{ij} is:

$$\begin{array}{ccc} \textit{level } 1 & \textit{level } 2 & \textit{left} \\ \end{array}$$
$$\begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$ [½]

The off-diagonal elements of the generator matrix of transition rates, μ_{ij} are given by:

$$\mu_{ij} = \lambda_i p_{ij}$$

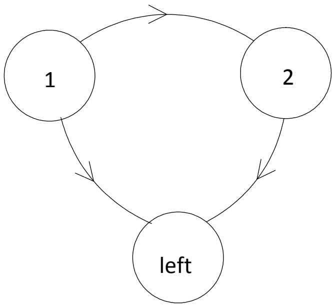
The diagonal elements are chosen to make each row of the matrix sum to 0.

The generator matrix (matrix of transition rates) is:

$$\begin{array}{ccc} \textit{level } 1 & \textit{level } 2 & \textit{left} \\ \end{array}$$
$$\begin{bmatrix} -0.50 & 0.25 & 0.25 \\ 0 & -0.20 & 0.20 \\ 0 & 0 & 0 \end{bmatrix}$$ [1]

(iv) ***Distribution of employees in five years***

The model is:



We are given the initial state as $\begin{bmatrix}600 & 400 & 0\end{bmatrix}$. *We can use the five-year transition probabilities to estimate the numbers in each state in five years' time.*

The number in state 1 will be:

$$600P_{11}(5)$$

The number in state 2 will be:

$$600P_{12}(5)+400P_{22}(5)$$

and the number of employees who have left the company can be obtained by deducting the numbers in states 1 and 2 from 1,000.

The occupancy probabilities for states 1 and 2 are given by:

$$P_{11}(t)=P_{\overline{11}}(t)=e^{-0.5t}$$ [½]

$$P_{22}(t)=P_{\overline{22}}(t)=e^{-0.2t}$$ [½]

Using the generator matrix, we can write the forward differential equation for $P_{12}(t)$:

$$\frac{d}{dt}P_{12}(t)=0.25P_{11}(t)-0.2P_{12}(t)$$

$$\Rightarrow \quad \frac{d}{dt}P_{12}(t)+0.2P_{12}(t)=0.25e^{-0.5t}$$ [1]

This can be solved using the integrating factor method. The integrating factor is $e^{0.2t}$. Multiplying through by the integrating factor gives:

$$e^{0.2t}\frac{d}{dt}P_{12}(t)+0.2e^{0.2t}P_{12}(t)=0.25e^{-0.5t}e^{0.2t}=0.25e^{-0.3t}$$ [½]

Integrating both sides:

$$e^{0.2t}P_{12}(t) = -\frac{5}{6}e^{-0.3t} + C$$ [½]

The boundary condition is $P_{12}(0) = 0$. So:

$$0 = -\frac{5}{6} + C \Rightarrow C = \frac{5}{6}$$ [½]

Simplifying then gives:

$$P_{12}(t) = \frac{5}{6}\left(e^{-0.2t} - e^{-0.5t}\right)$$ [½]

So the number of employees on level 1 in 5 years' time is:

$$600P_{11}(5) = 600e^{-2.5} = 49$$ [½]

and the number of employees on level 2 in 5 years' time is:

$$600P_{12}(5) + 400P_{22}(5) = 600 \times \frac{5}{6}\left(e^{-1} - e^{-2.5}\right) + 400e^{-1} = 290$$ [1]

The number of lives who have left the company is:

$$1{,}000 - 49 - 290 = 661$$ [½]

4.7 *In this question you have to be very careful not to mix up v (v for victor) and υ (the Greek letter 'nu').*

(i)(a) ***Assumptions***

We are assuming that the transition probabilities depend only upon the individual's current state. They do not depend upon the previous transitions for the individual, so we are assuming that the Markov property holds. [½]

We are also assuming that the probability of two or more transitions in a short time interval of length h is $o(h)$... [½]

... and for small values of h and $59 \le x \le x+t \le 60$:

$${}_hp_x^{ad} = h\mu + o(h)$$

and: $${}_hp_x^{ar} = h\upsilon + o(h)$$ [1]

(i)(b) ***Proof***

A life who remains active for $t+h$ years must first remain active for t years, then remain active for a further h years (where h represents a short time interval). Expressed in terms of probabilities, this is:

$${}_{t+h}p_x^{aa} = {}_t p_x^{aa} \times {}_h p_{x+t}^{aa}$$ [1]

This follows from the Markov property, *ie* that the probabilities in different time periods are independent of each other. [1]

During a short time period $(t, t+h)$, an active life must remain active, die or retire. So:

$${}_h p_{x+t}^{aa} + {}_h p_{x+t}^{ad} + {}_h p_{x+t}^{ar} = 1$$ [1]

Using the formulae given in (i)(a), this becomes:

$${}_h p_{x+t}^{aa} + h\mu + h\upsilon + o(h) = 1$$

So:

$${}_{t+h}p_x^{aa} = {}_t p_x^{aa} \times [1 - h(\mu+\upsilon) + o(h)]$$ [½]

Rearranging and letting $h \to 0$ gives:

$$\frac{\partial}{\partial t}\,{}_t p_x^{aa} = -(\mu+\upsilon)\,{}_t p_x^{aa}$$

since $\lim\limits_{h\to 0} \dfrac{o(h)}{h} = 0$. [1]

Rearranging this, we see that:

$$\frac{\frac{\partial}{\partial t}\,{}_t p_x^{aa}}{{}_t p_x^{aa}} = \frac{\partial}{\partial t}\log {}_t p_x^{aa} = -(\mu+\upsilon)$$ [½]

Integrating with respect to t with limits of 0 and s:

$$\left[\log {}_t p_x^{aa}\right]_0^s = -(\mu+\upsilon)s$$ [½]

Since ${}_0 p_x^{aa} = 1$, it follows that:

$${}_s p_x^{aa} = e^{-(\mu+\upsilon)s}$$ [½]

for $59 \le x \le x+s \le 60$.

(i)(c) ***Likelihood***

Here are two possible approaches to this part.

During the year, individual i will either survive to the end, die or retire. Using the result in (i)(b) and writing t_i for this individual's waiting time in the active state, the likelihood corresponding to each of these is:

Survival: $e^{-(\mu+\upsilon)t_i}$

Death: $e^{-(\mu+\upsilon)t_i} \times \mu$

Retirement: $e^{-(\mu+\upsilon)t_i} \times \upsilon$

Since the experiences of the individuals are assumed to be independent, the overall likelihood for all the lives will be:

$$L(\mu,\upsilon) = \prod_{survivors} e^{-(\mu+\upsilon)t_i} \times \prod_{deaths} \mu e^{-(\mu+\upsilon)t_i} \times \prod_{retirements} \upsilon e^{-(\mu+\upsilon)t_i} \qquad [3]$$

This can be simplified to give:

$$L(\mu,\upsilon) = \prod_{all\,lives} e^{-(\mu+\upsilon)t_i} \times \prod_{deaths} \mu \times \prod_{retirements} \upsilon$$

$$= e^{-(\mu+\upsilon)\Sigma t_i} \times \mu^d \times \upsilon^r = e^{-(\mu+\upsilon)v} \times \mu^d \times \upsilon^r \qquad [2]$$

Alternatively, we can write down the probability density/mass function for life i as a single function:

$$f_i(d_i,r_i,v_i) = \begin{cases} {}_{v_i}p_x & (d_i=0, r_i=0) \\ {}_{v_i}p_x\,\mu & (d_i=1) \\ {}_{v_i}p_x\,\upsilon & (r_i=1) \end{cases}$$

Here d_i and r_i represent the number of deaths and retirements experienced by this individual during the year (which will be 0 or 1).

We can then express these three 'combinations' in a single formula as:

$$f_i(d_i,r_i,v_i) = {}_{v_i}p_x \times \mu^{d_i} \times \upsilon^{r_i} = \exp[-(\mu+\upsilon)\,v_i] \times \mu^{d_i} \times \upsilon^{r_i}$$

So the joint likelihood for the whole group will be:

$$L = \prod_{i=1}^{N} \exp[-(\mu+\upsilon)\,v_i] \times \mu^{d_i} \times \upsilon^{r_i} = \exp[-(\mu+\upsilon)v]\,\mu^d \upsilon^r$$

(i)(d) ***Formulae***

The MLE of υ is $\tilde{\upsilon} = \dfrac{R}{V}$. [1]

Asymptotically, this has moments:

Mean: $E(\tilde{\upsilon}) = \upsilon$ [1]

Estimated standard error: $ese(\tilde{\upsilon}) = \sqrt{\dfrac{\upsilon}{v}}$ [1]

Recall that the standard error of an estimator is the square root of its variance.

(ii)(a) ***Likelihood***

The likelihood function is now found by combining the likelihood of observing d deaths during the year with the likelihood of observing r retirements out of the m lives who survived to age 60. This second part is a binomial probability, and we get:

$$e^{-\mu v}\mu^{d} \times \binom{m}{r} k^{r}(1-k)^{m-r}$$ [3]

(ii)(b) ***Maximum likelihood estimate of k***

Since we have m lives at age 60 and r are observed to retire, the maximum likelihood estimate of k is the binomial proportion, $\hat{k} = \dfrac{r}{m}$. [1]

4.8 (i) ***Distribution of the time spent in each state***

In a continuous-time Markov jump process, the times spent in each state are exponentially distributed. [1]

(ii) ***Generator matrix***

If we measure times in minutes, the generator matrix (with zeros omitted) is:

$$\begin{array}{c|ccccc|} & W & A & M & S & H \\ W & -\frac{1}{15} & \frac{1}{15} & & & \\ A & & -\frac{1}{20} & \frac{3}{400} & \frac{1}{400} & \frac{1}{25} \\ M & & & -\frac{1}{30} & & \frac{1}{30} \\ S & & & & -\frac{1}{180} & \frac{1}{180} \\ H & & & & & \end{array}$$ [2]

If you work in hours, all these entries need to be multiplied by 60.

(iii)(a) ***Kolmogorov forward differential equations***

The general formula for the Kolmogorov forward differential equation in the time-homogeneous case is:

$$\frac{d}{dt}p_{ij}(t)=\sum_k p_{ik}(t)\,\mu_{kj}$$

Applying this, with $i=W$ and $j=M$, we get:

$$\frac{d}{dt}p_{WM}(t)=p_{WA}(t)\,\mu_{AM}+p_{WM}(t)\,\mu_{MM}=\tfrac{3}{400}p_{WA}(t)-\tfrac{1}{30}p_{WM}(t) \qquad [1]$$

Similarly:

$$\frac{d}{dt}p_{WA}(t)=\tfrac{1}{15}p_{WW}(t)-\tfrac{1}{20}p_{WA}(t) \qquad [1]$$

(iii)(b) ***Verify formula***

In order to check that the formula given in the question satisfies the differential equation just stated, we first need a formula for $p_{WW}(t)$. Since it is not possible to return to state W once it has been left, $p_{WW}(t)$ is the same as $p_{\overline{WW}}(t)$, which we can work out as:

$$p_{WW}(t)=p_{\overline{WW}}(t)=e^{-t/15} \qquad [½]$$

Substituting the formula given in the question for $p_{WA}(t)$ into the Kolmogorov equation, we see that:

$$LHS=\frac{d}{dt}p_{WA}(t)=\frac{d}{dt}\left(4e^{-t/20}-4e^{-t/15}\right)=-\tfrac{1}{5}e^{-t/20}+\tfrac{4}{15}e^{-t/15} \qquad [½]$$

and:

$$\begin{aligned}RHS&=\tfrac{1}{15}p_{WW}(t)-\tfrac{1}{20}p_{WA}(t)\\&=\tfrac{1}{15}e^{-t/15}-\tfrac{1}{20}\left(4e^{-t/20}-4e^{-t/15}\right)\\&=-\tfrac{1}{5}e^{-t/20}+\tfrac{4}{15}e^{-t/15}\end{aligned} \qquad [½]$$

So the differential equation is satisfied.

We also need to check the boundary condition. Substituting $t=0$ into the formula given, we get:

$$p_{WA}(0)=4e^{0}-4e^{0}=0$$

This is the correct value since the process cannot move from state W to state A in zero time. [½] (iii)(c) ***Derive an expression for*** $\boldsymbol{p_{WM}(t)}$

We can now use the formula for $p_{WA}(t)$ from part (iii)(b) in conjunction with the first differential equation from part (iii)(a) to find a formula for $p_{WM}(t)$. We have:

$$\frac{d}{dt}p_{WM}(t)=\tfrac{3}{400}p_{WA}(t)-\tfrac{1}{30}p_{WM}(t)$$

$$=\tfrac{3}{400}\left(4e^{-t/20}-4e^{-t/15}\right)-\tfrac{1}{30}p_{WM}(t)$$

$$=\tfrac{3}{100}\left(e^{-t/20}-e^{-t/15}\right)-\tfrac{1}{30}p_{WM}(t) \qquad [½]$$

We can solve this using an integrating factor. We first need to rearrange it in the form:

$$\frac{d}{dt}p_{WM}(t)+\tfrac{1}{30}p_{WM}(t)=\tfrac{3}{100}\left(e^{-t/20}-e^{-t/15}\right)$$

The integrating factor is:

$$\exp\left(\int\tfrac{1}{30}dt\right)=e^{t/30} \qquad [½]$$

Multiplying through by the integrating factor, we get:

$$e^{t/30}\frac{d}{dt}p_{WM}(t)+\tfrac{1}{30}e^{t/30}p_{WM}(t)=\tfrac{3}{100}\left(e^{-t/60}-e^{-t/30}\right)$$

So:

$$\frac{d}{dt}\left[e^{t/30}p_{WM}(t)\right]=\tfrac{3}{100}\left(e^{-t/60}-e^{-t/30}\right) \qquad [½]$$

Now we can integrate to get:

$$e^{t/30}p_{WM}(t)=\tfrac{3}{100}\left(-60e^{-t/60}+30e^{-t/30}\right)+c=-\tfrac{9}{5}e^{-t/60}+\tfrac{9}{10}e^{-t/30}+c \qquad [½]$$

When $t=0$, this becomes:

$$0=-\tfrac{9}{5}+\tfrac{9}{10}+c=-\tfrac{9}{10}+c \quad \Rightarrow c=\tfrac{9}{10} \qquad [½]$$

So we have:

$$e^{t/30}p_{WM}(t)=-\tfrac{9}{5}e^{-t/60}+\tfrac{9}{10}e^{-t/30}+\tfrac{9}{10}$$

Dividing through by the integrating factor gives us the final answer:

$$p_{WM}(t)=-\tfrac{9}{5}e^{-t/20}+\tfrac{9}{10}e^{-t/15}+\tfrac{9}{10}e^{-t/30} \qquad [½]$$

(iv)(a) ***Explain why*** $\boldsymbol{T_W = 15 + T_A}$

If a vehicle is currently in state W, it will wait 15 minutes on average in that state before moving to state A (definitely), after which it will wait on average a further time T_A before it can be driven home. So the average time T_W before it can be driven home is $15+T_A$. [1]

(iv)(b) ***Equations for*** $\boldsymbol{T_A}$***,*** $\boldsymbol{T_M}$ ***and*** $\boldsymbol{T_S}$

Using similar logic, a vehicle in state A will wait 20 minutes on average in that state before moving either to state H (with probability 0.8) or to state M (with probability 0.15) or to state S (with probability 0.05). So the corresponding equation is:

$$T_A = 20 + 0.8 \times 0 + 0.15T_M + 0.05T_S$$ [1]

Since we know that $T_M = 30$ and $T_S = 180$, this gives $T_A = 33.5$ minutes. [1]

(iv)(c) ***Calculate*** $\boldsymbol{T_W}$

Using the equation from part (iv)(a), we find that:

$$T_W = 15 + T_A = 15 + 33.5 = 48.5 \text{ minutes}$$ [1]

End of Part 1

What next?

1. Briefly **review** the key areas of Part 1 and/or re-read the **summaries** at the end of Chapters 1 to 4.
2. Ensure you have attempted some of the **Practice Questions** at the end of each chapter in Part 1. If you don't have time to do them all, you could save the remainder for use as part of your revision.
3. Attempt **Assignment X1**.

5

Time-inhomogeneous Markov jump processes

Syllabus objectives

3.3 Define and apply a Markov process.

3.3.1 State the essential features of a Markov process model.

3.3.3 Derive the Kolmogorov equations for a Markov process with time independent and time/age dependent transition intensities.

3.3.4 Solve the Kolmogorov equations in simple cases.

3.3.5 Describe simple survival models, sickness models and marriage models in terms of Markov processes and describe other simple applications.

3.3.6 State the Kolmogorov equations for a model where the transition intensities depend not only on age/time, but also on the duration of stay in one or more states.

3.3.7 Describe sickness and marriage models in terms of duration dependent Markov processes and describe other simple applications.

3.3.8 Demonstrate how Markov jump processes can be used as a tool for modelling and how they can be simulated.

0 Introduction

In this chapter we discuss time-inhomogeneous Markov jump processes. The transition probabilities $P(X_t = j \mid X_s = i)$ for a time-inhomogeneous process depend not only on the length of the time interval $[s,t]$, but also on the times s and t when it starts and ends. This is because the transition rates for a time-inhomogeneous process vary over time.

We start by discussing the important features of time-inhomogeneous processes. Then, just as we did for time-homogeneous processes in Chapter 4, we study the forward and backward Kolmogorov differential equations and occupancy probabilities. We go on to introduce the integrated forms of the Kolmogorov equations, and we look at some applications. Finally, we cover some modelling techniques for Markov jump processes and describe how the parameters of a Markov jump process can be estimated.

1 Features of time-inhomogeneous Markov jump processes

1.1 Chapman-Kolmogorov equations

The more general continuous-time Markov jump process $\{X_t, t \geq 0\}$ has transition probabilities:

$$p_{ij}(s,t) = P[X_t = j | X_s = i] \quad (s \leq t)$$

which obey a version of the *Chapman-Kolmogorov equations,* written in matrix form as:

$$P(s,t) = P(s,u)P(u,t) \qquad \text{for all } s < u < t$$

or equivalently:

$$p_{ij}(s,t) = \sum_{k \in S} p_{ik}(s,u)\, p_{kj}(u,t) \qquad \text{for all } s < u < t$$

Again, both upper and lower case P may be used to denote a transition probability.

1.2 Transition rates

Proceeding as in the time-homogeneous case, we obtain:

$$p_{ij}(s, s+h) = \begin{cases} h\,\mu_{ij}(s) + o(h) & \text{if } i \neq j \\ 1 + h\mu_{ii}(s) + o(h) & \text{if } i = j \end{cases}$$

Equivalently, we have:

$$\mu_{ij}(s) = \lim_{h \to 0} \frac{p_{ij}(s,s+h) - p_{ij}(s,s)}{h} = \lim_{h \to 0} \frac{p_{ij}(s,s+h) - \delta_{ij}}{h} = \left[\frac{\partial}{\partial t} p_{ij}(s,t)\right]_{t=s}$$

We see that the only difference between this case and the time-homogeneous case studied earlier is that the transition rates $\mu_{ij}(s)$ are allowed to change over time.

2 Kolmogorov's forward differential equations

Kolmogorov's forward equations may be derived.

To derive these equations, we consider the interval $(s,t+h)$ where h is a small amount. From the Chapman-Kolmogorov equations, we have:

$$p_{ij}(s,t+h)=\sum_{k\in S} p_{ik}(s,t)\,p_{kj}(t,t+h)$$

For small h, we know that:

$$p_{kj}(t,t+h)=h\mu_{kj}(t)+o(h) \qquad \text{provided } j\neq k$$

and:

$$p_{jj}(t,t+h)=1+h\mu_{jj}(t)+o(h)$$

So:

$$\begin{aligned} p_{ij}\left(s,t+h\right) &= \sum_{k\neq j} p_{ik}(s,t)\,h\mu_{kj}(t)+p_{ij}(s,t)\Big(1+h\mu_{jj}(t)\Big)+o(h) \\ &= p_{ij}(s,t)+\sum_{k\in S} p_{ik}(s,t)\,h\mu_{kj}(t)+o(h) \end{aligned}$$

Rearranging gives:

$$\frac{p_{ij}\left(s,t+h\right)-p_{ij}(s,t)}{h}=\sum_{k\in S} p_{ik}(s,t)\,\mu_{kj}(t)+\frac{o(h)}{h}$$

and letting $h\to 0$, we obtain:

$$\frac{\partial}{\partial t}p_{ij}(s,t)=\sum_{k\in S} p_{ik}(s,t)\,\mu_{kj}(t)$$

since $\lim\limits_{h\to 0}\dfrac{o(h)}{h}=0$.

This result is given on page 38 of the *Tables.* However, in the *Tables*, the notation $\sigma_{kj}(t)$ rather than $\mu_{kj}(t)$ is used to denote the force of transition from state k to state j at time t.

Kolmogorov's forward differential equations (time-inhomogeneous case)

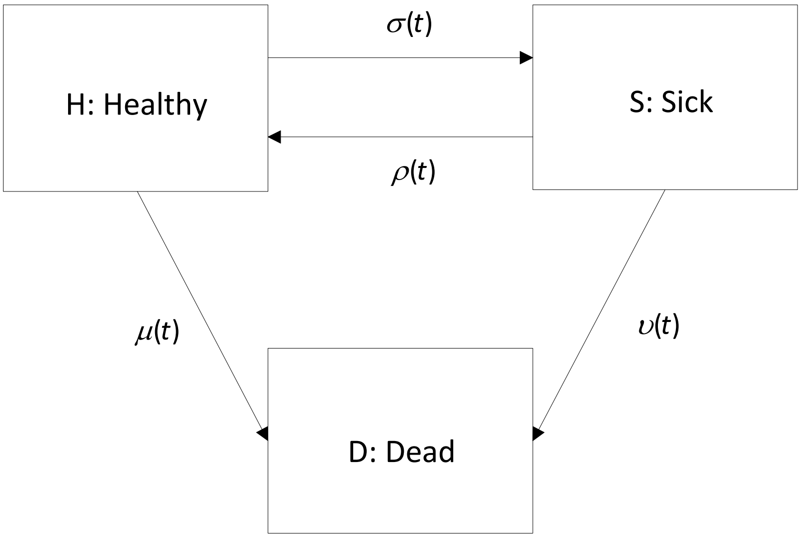
Written in matrix form these are:

$$\frac{\partial}{\partial t}\boldsymbol{P}(s,t)=\boldsymbol{P}(s,t)\boldsymbol{A}(t)$$

where *A*(*t*) is the matrix with entries $\mu_{ij}(t)$.

2.1 Time-inhomogeneous HSD model

We met the HSD model in Chapter 4. There we assumed that the transition rates were constant. However, in this chapter we will assume that they vary over time. The transition diagram is shown below.



Question

Write down Kolmogorov's forward differential equation for $p_{HD}(s,t)$.

Solution

The forward differential equation is:

$$\frac{\partial}{\partial t}p_{HD}(s,t)=p_{HH}(s,t)\mu_{HD}(t)+p_{HS}(s,t)\mu_{SD}(t)+p_{HD}(s,t)\mu_{DD}(t)$$

$$=p_{HH}(s,t)\mu(t)+p_{HS}(s,t)\upsilon(t)$$

since $\mu_{DD}=0$.

2.2 Non-standard forward equations

We may also want to construct differential equations for probabilities other than those of the form $p_{ij}(s,t)$. In particular, we are interested in the probability of remaining in state i throughout the time period (s,t). This is denoted by $p_{\overline{ii}}(s,t)$. It is different from $p_{ii}(s,t)$ since $p_{ii}(s,t)$ allows the possibility of leaving state i during the period. In these cases the standard forward equation is not applicable and we have to go back to first principles.

For example, suppose we are asked to derive the forward differential equation for $p_{\overline{SS}}(s,t)$ in the HSD model.

We start by considering the probability $p_{\overline{SS}}(s,t+h)$, where h is a small amount, and we condition on the state at time t to obtain the equation:

$$p_{\overline{SS}}(s,t+h)=p_{\overline{SS}}(s,t)\,p_{\overline{SS}}(t,t+h)$$

During the short interval of length *h*, the process either remains in *S*, moves from *S* to *H* or moves from *S* to *D*. We assume here that the probability of more than one move is very small (represented by the $o(h)$ term) so that:

$$p_{\overline{SS}}(t,t+h)+p_{SH}(t,t+h)+p_{SD}(t,t+h)+o(h)=1$$

Since we know that $p_{SH}(t,t+h)=h\rho(t)+o(h)$ and $p_{SD}(t,t+h)=h\upsilon(t)+o(h)$, we have:

$$p_{\overline{SS}}(t,t+h)=1-h\big(\rho(t)+\upsilon(t)\big)+o(h)$$

So:

$$p_{\overline{SS}}(s,t+h)=p_{\overline{SS}}(s,t)\Big[1-h\big(\rho(t)+\upsilon(t)\big)\Big]+o(h)$$

This can be rearranged to give:

$$\frac{p_{\overline{SS}}(s,t+h)-p_{\overline{SS}}(s,t)}{h}=-p_{\overline{SS}}(s,t)\big(\rho(t)+\upsilon(t)\big)+\frac{o(h)}{h}$$

Taking the limit as $h\to 0$ we obtain the differential equation:

$$\frac{\partial}{\partial t}p_{\overline{SS}}(s,t)=-p_{\overline{SS}}(s,t)\big(\rho(t)+\upsilon(t)\big)$$

since $\lim\limits_{h\to 0}\dfrac{o(h)}{h}=0$.

Equations of this formed can be solved using separation of variables.

Dividing both sides of the equation by $p_{\overline{SS}}(s,t)$ gives:

$$\frac{\dfrac{\partial}{\partial t}p_{\overline{SS}}(s,t)}{p_{\overline{SS}}(s,t)}=-\big(\rho(t)+\upsilon(t)\big)$$

This can also be written as:

$$\frac{\partial}{\partial t}\ln p_{\overline{SS}}(s,t)=-\big(\rho(t)+\upsilon(t)\big)$$

or equivalently, changing the variable from t to u:

$$\frac{\partial}{\partial u}\ln p_{\overline{SS}}(s,u)=-\big(\rho(u)+\upsilon(u)\big)$$

Integrating both sides with respect to u between the limits of s and t then gives:

$$\Big[\ln p_{\overline{SS}}(s,u)\Big]_{u=s}^{u=t}=-\int_s^t\big(\rho(u)+\upsilon(u)\big)du$$

Since $p_{\overline{SS}}(s,s)=1$ and $\ln 1=0$, this simplifies to:

$$\ln p_{\overline{SS}}(s,t)=-\int_s^t\big(\rho(u)+\upsilon(u)\big)du$$

Taking exponentials, we obtain the result:

$$p_{\overline{SS}}(s,t)=\exp\left[-\int_s^t\big(\rho(u)+\upsilon(u)\big)du\right]$$

It can similarly be shown that:

$$\frac{\partial}{\partial t}p_{\overline{HH}}(s,t)=-p_{\overline{HH}}(s,t)\big(\sigma(t)+\mu(t)\big)$$

and:

$$p_{\overline{HH}}(s,t)=\exp\left[-\int_s^t\big(\sigma(u)+\mu(u)\big)du\right]$$

3 Occupancy probabilities

We have just seen that $p_{\overline{HH}}(s,t)=\exp\left(-\int_s^t(\sigma(u)+\mu(u))\,du\right)$. This result can be generalised to give an expression for the probability of staying in any state i from time s to time t (known as the occupancy probability for state i). For any time-inhomogeneous Markov jump process, the probability that a process in state i at time s remains in state i until at least time t is given by:

$$\exp\left(-\int_s^t\left(\text{total force of transition out of state } i \text{ at time } u\right)du\right)$$

$$=\exp\left(-\int_0^{t-s}\left(\text{total force of transition out of state } i \text{ at time } s+u\right)du\right)$$

This important result is restated in the box below.

Occupancy probabilities for time-inhomogeneous Markov jump processes

For a time-inhomogeneous Markov jump process:

$$p_{\overline{ii}}(s,t)=\exp\left(-\int_s^t\lambda_i(u)\,du\right)=\exp\left(-\int_0^{t-s}\lambda_i(s+u)\,du\right)$$

where $\lambda_i(u)$ denotes the total force of transition out of state i at time u.

If the transition rates are constant (*ie* the process is time-homogeneous), the occupancy probabilities simplify to:

$$p_{\overline{ii}}(s,t)=e^{-\lambda_i(t-s)}$$

and the holding time in state i has an $Exp(\lambda_i)$ distribution. We saw this result in Chapter 4.

4 Kolmogorov's backward differential equations

As in the time-homogeneous case, we need to be able to derive and to write down Kolmogorov's backward differential equations in the time-inhomogeneous case.

Kolmogorov's backward differential equations (time-inhomogeneous case)

The matrix form of Kolmogorov's backward equations is:

$$\frac{\partial}{\partial s}\boldsymbol{P}(s,t) = -\boldsymbol{A}(s)\boldsymbol{P}(s,t)$$

It is still the case that:

$$\mu_{ii}(s) = -\sum_{j \neq i} \mu_{ij}(s)$$

Hence each row of the matrix $A(s)$ has zero sum.

Written in component form the equations are:

$$\frac{\partial}{\partial s} p_{ij}(s,t) = -\sum_{k \in S} \mu_{ik}(s)\, p_{kj}(s,t)$$

This result is given on page 38 of the *Tables.*

There are a couple of particular points to note here:

- we are now differentiating with respect to s rather than t
- there is a minus sign on the RHS.

We can derive the backward differential equations as follows. Start by considering the interval $(s-h,t)$ where h is a small amount. Then, using the Chapman-Kolmogorov equations, we have:

$$p_{ij}(s-h,t) = \sum_{k \in S} p_{ik}(s-h,s)\, p_{kj}(s,t)$$

Since h is small, we know that:

$$p_{ik}(s-h,s) = h\,\mu_{ik}(s-h) + o(h) \qquad \text{provided } k \neq i$$

and: $$p_{ii}(s-h,s) = 1 + h\,\mu_{ii}(s-h) + o(h)$$

So:

$$\begin{aligned} p_{ij}(s-h,t) &= \sum_{k \neq i} h\,\mu_{ik}(s-h)\, p_{kj}(s,t) + \big(1 + h\,\mu_{ii}(s-h)\big)\, p_{ij}(s,t) + o(h) \\ &= p_{ij}(s,t) + \sum_{k \in S} h\,\mu_{ik}(s-h)\, p_{kj}(s,t) + o(h) \end{aligned}$$

Rearranging gives:

$$\frac{p_{ij}(s-h,t)-p_{ij}(s,t)}{h}=\sum_{k\in S}\mu_{ik}(s-h)\,p_{kj}(s,t)+\frac{o(h)}{h}$$

or equivalently:

$$\frac{p_{ij}(s,t)-p_{ij}(s-h,t)}{h}=-\sum_{k\in S}\mu_{ik}(s-h)\,p_{kj}(s,t)+\frac{o(h)}{h}$$

Letting $h\to 0$ we obtain:

$$\frac{\partial}{\partial s}p_{ij}(s,t)=-\sum_{k\in S}\mu_{ik}(s)\,p_{kj}(s,t)$$

since $\lim\limits_{h\to 0}\dfrac{o(h)}{h}=0$.

We now consider at some examples of backward differential equations based on the time-inhomogeneous HSD model.

Using the general formula given above, the Kolmogorov backward differential equation for $p_{HH}(s,t)$ is given by:

$$\begin{aligned}\frac{\partial}{\partial s}p_{HH}(s,t)&=-\left[\mu_{HH}(s)p_{HH}(s,t)+\mu_{HS}(s)p_{SH}(s,t)+\mu_{HD}(s)p_{DH}(s,t)\right]\\&=-\left[-\left(\sigma(s)+\mu(s)\right)p_{HH}(s,t)+\sigma(s)p_{SH}(s,t)\right]\end{aligned}$$

since $p_{DH}(s,t)=0$. So:

$$\frac{\partial}{\partial s}p_{HH}(s,t)=\left(\sigma(s)+\mu(s)\right)p_{HH}(s,t)-\sigma(s)p_{SH}(s,t)$$

Similarly, the Kolmogorov backward differential equation for $p_{HS}(s,t)$ is given by:

$$\begin{aligned}\frac{\partial}{\partial s}p_{HS}(s,t)&=-\left[\mu_{HH}(s)p_{HS}(s,t)+\mu_{HS}(s)p_{SS}(s,t)+\mu_{HD}(s)p_{DS}(s,t)\right]\\&=-\left[-\left(\sigma(s)+\mu(s)\right)p_{HS}(s,t)+\sigma(s)p_{SS}(s,t)\right]\end{aligned}$$

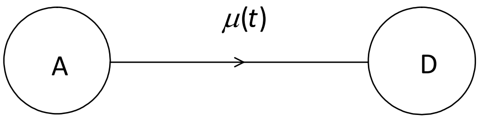
since $p_{DS}(s,t)=0$. So:

$$\frac{\partial}{\partial s}p_{HS}(s,t)=\left(\sigma(s)+\mu(s)\right)p_{HS}(s,t)-\sigma(s)p_{SS}(s,t)$$

The general theory of time-inhomogeneous Markov jump processes is rather too complicated to fall within the scope of the current syllabus, but the methods used can be illustrated by means of a number of practical examples.

5 Example – a two-state model

Consider the following *survival model*: transition from the alive state A to the dead state D takes place at rate $\mu_{AD}(t)$, which has been abbreviated to $\mu(t)$ here, since it is the only transition in the model.



In other words $\mu(t)$ is the force of mortality.

The theory of mortality functions, including the force of mortality, is discussed in more detail in Chapter 6. The two-state model is discussed in Chapter 3, where we develop the results without using the generator matrix.

Since $A(t)=\begin{pmatrix}-\mu(t) & \mu(t)\\ 0 & 0\end{pmatrix}$ the forward equations give:

$$\frac{\partial}{\partial t}p_{AA}(s,t)=-p_{AA}(s,t)\mu(t)$$

The solution corresponding to the initial condition $p_{AA}(s,s)=1$ is:

$$p_{AA}(s,t)=\exp\left(-\int_s^t \mu(x)\,dx\right)$$

This result should be familiar from Section 3 since $p_{AA}(s,t)=p_{\overline{AA}}(s,t)$.

Question

Write down the backward differential equation for $p_{AA}(s,t)$ and show that the solution of this equation is also $p_{AA}(s,t)=e^{-\int_s^t \mu(x)dx}$.

Solution

The backward differential equation for $p_{AA}(s,t)$ is:

$$\frac{\partial}{\partial s}p_{AA}(s,t)=-\left[-\mu(s)p_{AA}(s,t)\right]=\mu(s)p_{AA}(s,t)$$

Separating the variables gives:

$$\frac{\frac{\partial}{\partial s}p_{AA}(s,t)}{p_{AA}(s,t)}=\mu(s)$$

Now changing the variable from s to x, we have:

$$\frac{\partial}{\partial x}\ln p_{AA}(x,t) = \mu(x)$$

Integrating with respect to x between the limits of $x = s$ and $x = t$ gives:

$$\left[\ln p_{AA}(x,t)\right]_{x=s}^{x=t} = \int_s^t \mu(x)\,dx$$

But $p_{AA}(t,t) = 1$ and $\ln 1 = 0$. So we have:

$$-\ln p_{AA}(s,t) = \int_s^t \mu(x)\,dx$$

Moving the minus sign on to the RHS and taking exponentials gives the required result:

$$p_{AA}(s,t) = \exp\left(-\int_s^t \mu(x)\,dx\right)$$

Equivalently the probability for an individual aged s to survive for a further period of length at least w is:

$$\boldsymbol{{}_w p_s = p_{AA}(s,s+w) = \exp\left(-\int_s^{s+w} \mu(x)\,dx\right) = \exp\left(-\int_0^w \mu(s+y)\,dy\right)} \qquad \textbf{(5.1)}$$

Recall that ${}_w p_s$ denotes the probability that a person now aged s is still alive in w years' time.

Question

(i) A life aged 60 is subject to a constant force of mortality of 0.01 *pa*. Calculate the probability that the life survives to age 70.

(ii) Calculate the probability that a 25-year old with a constant force of mortality of 0.01 *pa* survives to age 35.

(iii) Comment on the answers.

Solution

(i) $${}_{10}p_{60} = \exp\left(-\int_{60}^{70} 0.01dt\right) = e^{-10\times 0.01} = 0.905$$

(ii) The probability ${}_{10}p_{25}$ is the same as ${}_{10}p_{60}$ if we assume that both lives are subject to the same force of mortality.

(iii) This is unrealistic. The force of mortality should vary with age.

This illustrates the need for time-dependent rates in mortality and many other actuarial models: a constant force of mortality μ would give rise to an age-independent survival probability ${}_{w}p_{s}$, an absurd result.

6 Residual holding times

As it stands, (5.1) is peculiar to the specific survival model under consideration; however, if properly reinterpreted it is but an instance of a general formula.

We have seen this already in Section 3, where we discussed occupancy probabilities. The general result is:

$$p_{\overline{ii}}(s,t)=\exp\left(-\int_s^t \lambda_i(u)\,du\right)=\exp\left(-\int_0^{t-s} \lambda_i(s+u)\,du\right)$$

For a general Markov jump process, $\{X_t, t \geq 0\}$, define the *residual holding time* R_s as the (random) amount of time between s and the next jump:

$$\{R_s > w, X_s = i\} = \{X_u = i, s \leq u \leq s+w\}$$

The residual holding time at time s is the amount of time after time s for which the process stays in the current state. The Core Reading equation above says that for the residual holding time at time s to be greater than w, given that the process is in state i at time s, the process must stay in state i for all times u between s and $s+w$.

Formula (5.1) gives the probability that $R_s > w$ given that the state at time s is A. In general one can prove:

$$P\left[R_s > w \mid X_s = i\right] = \exp\left(-\int_s^{s+w} \lambda_i(t)\,dt\right) \tag{5.2}$$

This result is similar to that shown in Section 7 of Chapter 4.

The probability $P\left[R_s > w \mid X_s = i\right]$ is the same as $p_{\overline{ii}}(s,s+w)$, *ie* it is the probability that the process stays in state i for at least another w time units.

Moreover, the characterisation of the state $X_s^+ = X_{s+R_s}$ to which the jump takes place is similar to the time-homogeneous case:

$$P\left[X_s^+ = j \mid X_s = i, R_s = w\right] = \frac{\mu_{ij}(s+w)}{\lambda_i(s+w)} \tag{5.3}$$

We can restate this result as follows.

Probability that the process goes into state *j* when it leaves state *i*

Given that the process is in state i at time s and it stays there until time $s+w$, the probability that it moves into state j when it leaves state i at time $s+w$ is:

$$\frac{\mu_{ij}(s+w)}{\lambda_i(s+w)} = \frac{\text{the force of transition from state } i \text{ to state } j \text{ at time } s+w}{\text{the total force out of state } i \text{ at time } s+w}$$

Question

Show that, for $w > 0$, the PDF of the random variable $R_s \mid X_s = i$ is given by:

$$f_{R_s|X_s=i}(w) = \lambda_i(s+w)\exp\left(-\int_s^{s+w} \lambda_i(u)\,du\right)$$

Solution

The PDF of any continuous random variable can be obtained from the CDF by differentiation. In this case the CDF is:

$$P\left[R_s \le w \mid X_s = i\right] = 1 - P\left[R_s > w \mid X_s = i\right] = 1 - \exp\left(-\int_s^{s+w} \lambda_i(u)\,du\right)$$

Differentiation of the exponential is straightforward, leaving the remaining problem of differentiating the integral with respect to *w*. Using the formula at the bottom of page 3 of the *Tables*, we see that:

$$\frac{\partial}{\partial w} \int_s^{s+w} \lambda_i(u)\,du = \lambda_i(s+w)$$

An alternative approach is the following. Suppose we know that if we integrate the function $\lambda_i(u)$ we get the function $\Lambda_i(u)$. Then:

$$\int_s^{s+w} \lambda_i(u)\,du = \left[\Lambda_i(u)\right]_s^{s+w} = \Lambda_i(s+w) - \Lambda_i(s)$$

If we differentiate this with respect to w, we get $\lambda_i(s+w)$ since $\lambda_i(u)$ is the derivative of $\Lambda_i(u)$. The $\Lambda_i(s)$ term doesn't contain any w's, so its derivative is 0.

Using the formula derived by either of the methods given above, it follows that:

$$\frac{\partial}{\partial w}\left(1 - \exp\left(-\int_s^{s+w} \lambda_i(u)\,du\right)\right) = -\left(\frac{\partial}{\partial w}\left(-\int_s^{s+w} \lambda_i(u)\,du\right)\right)\exp\left(-\int_s^{s+w} \lambda_i(u)\,du\right)$$

$$= \lambda_i(s+w)\exp\left(-\int_s^{s+w} \lambda_i(u)\,du\right)$$

We now have information on both the time that transitions take place and the states to which the transitions are made. By combining these we can calculate general transition probabilities, as described below. To do this, we condition on both:

- the residual holding time (a continuous variable, so we use integration over probability densities), and
- the current state (a discrete random variable, so we use summation over probabilities).

7 Integrated form of the Kolmogorov backward equations

The above is more than a neat picture for the behaviour of Markov jump processes: it is also a powerful computational tool. Indeed, conditioning on R_s and X_s^+ we have using the law of total probability:

$$p_{ij}(s,t) = P\left[X_t = j \mid X_s = i\right]$$

$$= \sum_{l \neq i} \int_0^{t-s} e^{-\int_s^{s+w} \lambda_i(u)du} \mu_{il}(s+w) P\left[X_t = j \mid X_s = i, R_s = w, X_s^+ = l\right] dw$$

and therefore:

$$p_{ij}(s,t) = \sum_{l \neq i} \int_0^{t-s} e^{-\int_s^{s+w} \lambda_i(u)du} \mu_{il}(s+w) p_{lj}(s+w,t) dw \qquad \textbf{(5.4)}$$

provided $j \neq i$.

This is the *integrated form of the backward equation*, as can be checked by differentiation with respect to s.

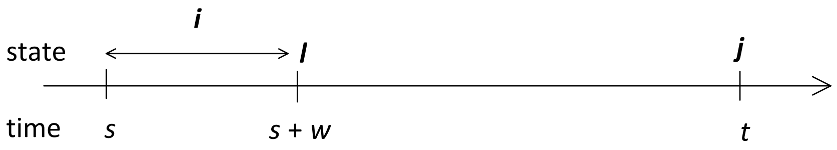
The formula may look intimidating but it conforms to intuition: since $j \neq i$, the process must jump out of i at some stage. By (5.2), the first jump after time s takes place at $s+w$ with probability density:

$$\lambda_i(s+w)\exp\left(-\int_s^{s+w} \lambda_i(u)du\right)$$

We saw this result in the previous question.

Backward integral equations always focus on the time of the first transition. Here we are thinking about the first transition occurring at time $s+w$.

By (5.3) the process jumps to l at time $s+w$ with probability $\dfrac{\mu_{il}(s+w)}{\lambda_i(s+w)}$. It then remains to effect a transition from l to j over the remaining time period $[s+w,t]$:



We can also reason as follows. The expression:

$$\exp\left(-\int_s^{s+w} \lambda_i(u)\,du\right)\mu_{il}(s+w)\,p_{lj}(s+w,t)$$

can be considered as the product of three factors:

- the probability of remaining in state i from time s to time $s+w$
- then making a transition to state l at time $s+w$
- and finally going from state l at time $s+w$ to state j at time t.

To take into account all the possible times at which the first transition can happen, we integrate with respect to w from $w=0$ to $t-s$. To take into account all possible intermediate states, we sum over all possible values of $l \neq i$.

The exponential term in the expression above can also be written as $p_{\overline{ii}}(s,s+w)$, so the backward integral equation can also be written as:

$$p_{ij}(s,t) = \sum_{l \neq i} \int_0^{t-s} p_{\overline{ii}}(s,s+w)\,\mu_{il}(s+w)\,p_{lj}(s+w,t)\,dw$$

for $j \neq i$.

Equation (5.4) gives a relationship between transition probabilities. To calculate them explicitly, however, we still need to solve the equations.

When $i = j$ there is an additional term $\exp\left(-\int_s^t \lambda_i(u)\,du\right)$ because the process can remain in state i throughout $[s,t]$.

Once again, we can write $\exp\left(-\int_s^t \lambda_i(u)\,du\right)$ as $p_{\overline{ii}}(s,t)$. So the integrated form of the backward equation for $p_{ii}(s,t)$ is:

$$p_{ii}(s,t) = \sum_{l \neq i} \int_0^{t-s} p_{\overline{ii}}(s,s+w)\,\mu_{il}(s+w)\,p_{li}(s+w,t)\,dw + p_{\overline{ii}}(s,t)$$

The first term on the right-hand side is the probability of leaving state i and returning to it. The second term is the probability of staying in state i from time s to time t.

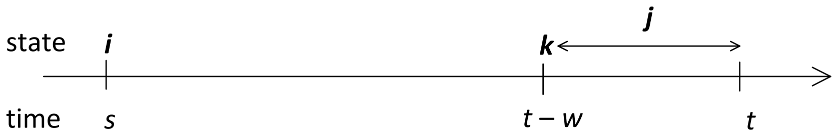
8 Integrated form of the Kolmogorov forward equations

If instead of considering the first jump after s one focuses on the last jump before t, one can obtain an intuitive derivation of the *integrated form of the forward equations.*

In the backward equation we thought about the time of the first transition as being $s+w$. For the forward equation we think about the time of the last transition as being $t-w$.

The forward equation when $i \neq j$ is:

$$p_{ij}(s,t) = \sum_{k \neq j} \int_0^{t-s} p_{ik}(s,t-w)\,\mu_{kj}(t-w)\,e^{-\int_{t-w}^{t} \lambda_j(u)du}\,dw \qquad (5.5)$$



Alternatively, this can be written as:

$$p_{ij}(s,t) = \sum_{k \neq j} \int_0^{t-s} p_{ik}(s,t-w)\,\mu_{kj}(t-w)\,p_{\overline{jj}}(t-w,t)\,dw$$

for $j \neq i$.

The factors in the integral are:

- the probability of going from state i at time s to state k at time $t-w$
- then making a transition from state k to state j at time $t-w$
- and staying in state j from time $t-w$ to time t.

Integrating over all possible values of w, namely 0 to $t-s$, and adding over all intermediate states $k \neq j$, we obtain the forward integral equation.

The forward integral equation for $p_{ii}(s,t)$ is:

$$p_{ii}(s,t) = \sum_{k \neq j} \int_0^{t-s} p_{ik}(s,t-w)\,\mu_{ki}(t-w)\,p_{\overline{ii}}(t-w,t)\,dw + p_{\overline{ii}}(s,t)$$

Here we've added on an extra term at the end to cover the possibility that the process stays in state i throughout the interval $[s,t]$.

Because of this intuitive interpretation, it shouldn't be too difficult to write out backward and forward equations in integrated form.

Derivation of the integrated form of the forward equations

For a full justification of this equation one needs to appeal to the properties of the *current holding time* C_t , namely the time between the last jump and t :

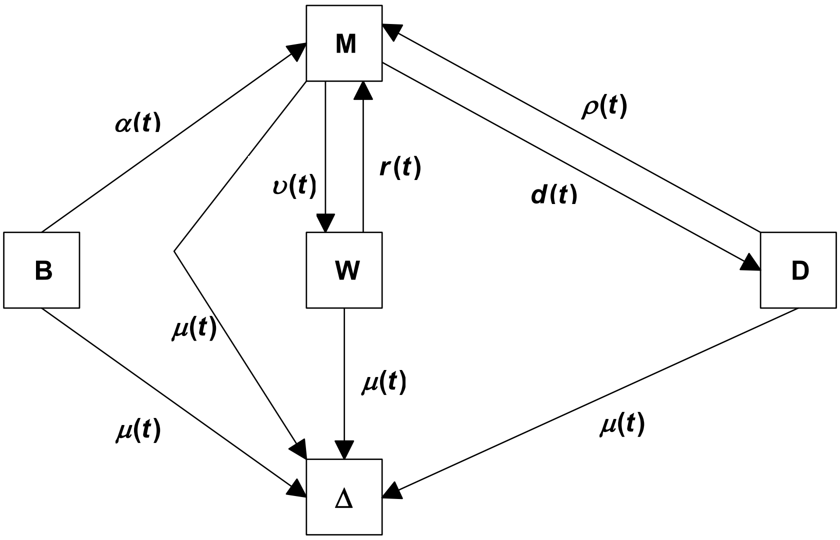
$$\{C_t \geq w, X_t = j\} = \{ X_u = j, t - w \leq u \leq t\}$$

or the length of time that the process has been in the current state. We consider the idea of current holding time in more detail in the next section.

9 Applications

9.1 Marriage

Describe the marital status of an individual as one of the following: bachelor (never married) (B), married (M), widowed (W), divorced (D), dead (Δ). We can define a Markov jump process on the state space {B, M, W, D, Δ } as illustrated below:

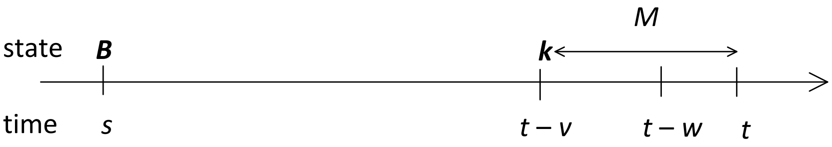


In the above, the death rate $\mu(t)$ has been taken to be independent of the marital status for simplicity.

The probability of being married at time t and of having been so for at least w given that you are a bachelor at time s is (assuming $w < t-s$):

$$\boldsymbol{P\left[X_t = M, C_t > w \mid X_s = B\right] = \int_w^{t-s} \left[p_{BB}(s,t-v)\alpha(t-v) + p_{BW}(s,t-v)r(t-v) \right.}$$

$$\boldsymbol{\left. + p_{BD}(s,t-v)\rho(t-v)\right] e^{-\int_{t-v}^{t}(\mu(u)+\upsilon(u)+d(u))du}\,dv}$$



where k is any of the states leading to M, namely B, W and D.

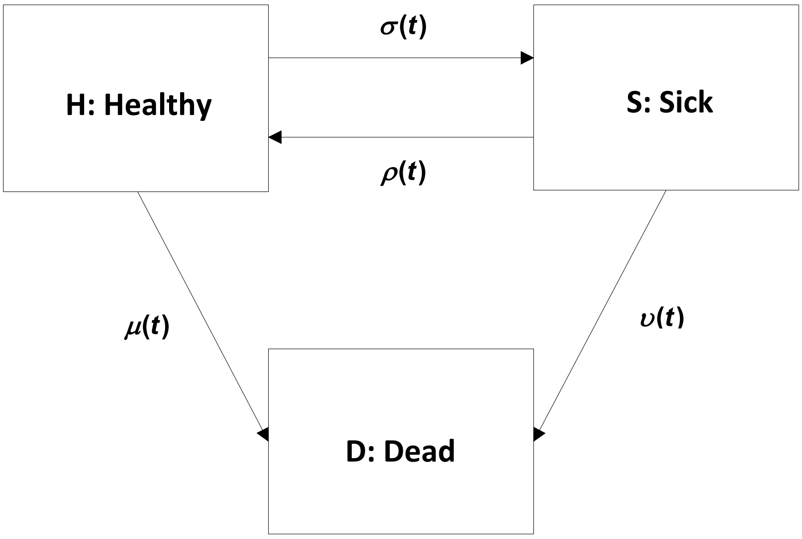
The integral on the RHS of this equation can also be written as:

$$\int_w^{t-s}\left[p_{BB}(s,t-v)\alpha(t-v)+p_{BW}(s,t-v)r(t-v)+p_{BD}(s,t-v)\rho(t-v)\right]p_{\overline{MM}}(t-v,t)\,dv$$

9.2 Sickness and death

Here we return to the HSD model with time-dependent forces of transition.

Describe the state of a person as 'healthy', 'sick' or 'dead'. For given time-dependent (*ie* age-dependent) transition rates, we can construct a Markov jump process with state space {H, S, D}:



The matrix $A(t)$ in Kolmogorov's equations is:

$$A(t)=\begin{pmatrix}-\sigma(t)-\mu(t) & \sigma(t) & \mu(t)\\ \rho(t) & -\rho(t)-\upsilon(t) & \upsilon(t)\\ 0 & 0 & 0\end{pmatrix}$$

In particular:

$$\lambda_H(t)=\sigma(t)+\mu(t),\quad \lambda_S(t)=\rho(t)+\upsilon(t)\quad \text{and}\quad \lambda_D=0$$

Remember that λ_i denotes the total force of transition out of state i.

The easiest probabilities to calculate are those of remaining continuously healthy or continuously sick over $[s,t]$. Using (5.2) these are:

$$P\left[R_s>t-s\,|\,X_s=H\right]=\exp\left[-\int_s^t\left(\sigma(u)+\mu(u)\right)du\right] \qquad (5.6)$$

and:

$$P\left[R_s>t-s\,|\,X_s=S\right]=\exp\left[-\int_s^t\left(\rho(u)+\upsilon(u)\right)du\right]$$

These probabilities can also be denoted as $p_{\overline{HH}}(s,t)$ and $p_{\overline{SS}}(s,t)$, respectively. They are not the same as $p_{HH}(s,t)$ and $p_{SS}(s,t)$, which include the possibility of changing state one or more times during the interval (but returning so as to be in the original state at time t).

Question

Describe and evaluate $P\left[R_s > t-s \,|\, X_s = D\right]$.

Solution

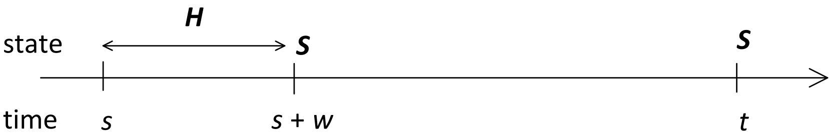
This is the probability of a life staying dead until at least time t, given that the life is dead at time s. It is 1.

The above equations can be used to give actual numerical values for the respective probabilities, assuming that we can evaluate the integrals. However, solving Kolmogorov's equations (*ie* evaluating transition probabilities) will not be possible in the general case of non-constant transition rates. Numerical methods do however exist that can give approximate solutions, but these are not included in the Subject CS2 syllabus.

We can also write down the integrated form of Kolmogorov's equations as below. Although this gives an expression for each transition probability, it does so only in terms of other *unknown* transition probabilities. In order to obtain actual transition probabilities we would still need to evaluate these integrals.

Transition probabilities can be related to each other as in (5.4) and (5.5). For instance:

$$\boldsymbol{p_{HS}(s,t) = \int_0^{t-s} p_{SS}(s+w,t)\,\sigma(s+w)\,e^{-\int_s^{s+w}(\sigma(u)+\mu(u))du}\,dw}$$



This is the integrated form of the backward equation for $P_{HS}(s,t)$, which can also be written as:

$$p_{HS}(s,t) = \int_0^{t-s} p_{\overline{HH}}(s,s+w)\,\sigma(s+w)\,p_{SS}(s+w,t)\,dw$$

Remember that in the backward equation we can think about the time of the first transition as being $s+w$, and for the forward equation we can think about the time of the last transition as being $t-w$.

Question

Give the forward version of the above equation.

Solution

The forward version is:

$$p_{HS}(s,t)=\int_0^{t-s} p_{HH}(s,t-w)\sigma(t-w)p_{\overline{SS}}(t-w,t)\,dw$$

The factors in this integral are:

- the probability of going from healthy at time s to healthy at time $t-w$
- then making a transition from healthy to sick at time $t-w$
- and finally staying in the sick state from time $t-w$ to time t.

Integrating over all possible values of w gives the integrated form of the forward equation.

To obtain the integrated form of the backward equation for $p_{HD}(s,t)$, we need to consider the two mutually exclusive events that the first transition from *H* is either to *S* or to *D*. The equation is:

$$p_{HD}(s,t)=\int_0^{t-s} p_{\overline{HH}}(s,s+w)\sigma(s+w)p_{SD}(s+w,t)\,dw+\int_0^{t-s} p_{\overline{HH}}(s,s+w)\mu(s+w)\,dw$$

Here we have used the fact that $p_{DD}(s+w,t)=1$.

Question

Write down the integrated form of the backward equation for $p_{SH}(s,t)$.

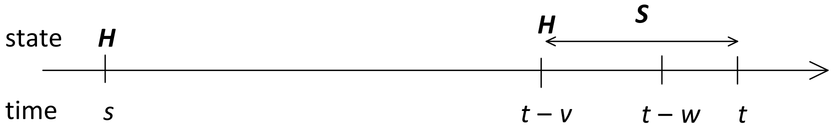
Solution

The equation is:

$$p_{SH}(s,t)=\int_0^{t-s} p_{\overline{SS}}(s,s+w)\,\rho(s+w)\,p_{HH}(s+w,t)\,dw$$

Extra conditions on residual or current holding times can be handled without difficulty. Consider for instance the probability of being sick at time t and of having been so for at least w, given that you are healthy at time s. This is:

$$\boldsymbol{P\left[X_t=S, C_t>w \mid X_s=H\right]=\int_w^{t-s} p_{HH}(s,t-v)\,\sigma(t-v)\,e^{-\int_{t-v}^{t}(\rho(u)+\upsilon(u))du}\,dv}$$



This equation can also be written as:

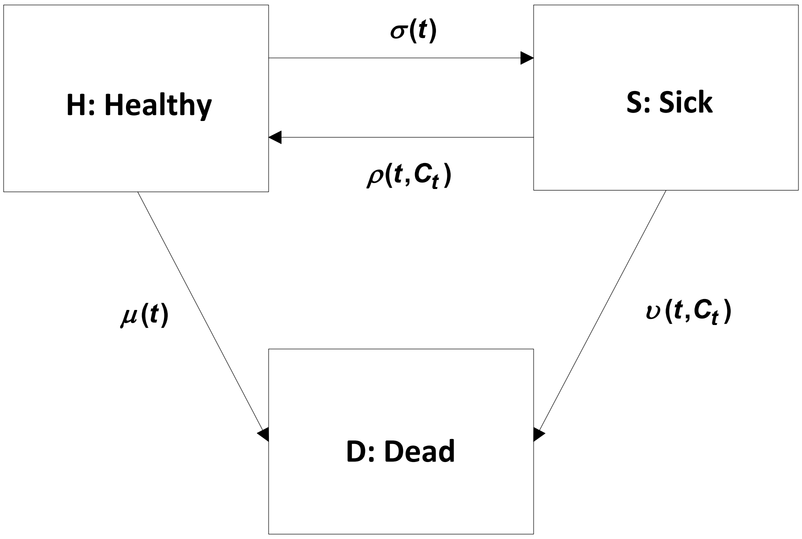
$$P\left[X_t = S, C_t > w \mid X_s = H\right] = \int_w^{t-s} p_{HH}(s, t-v)\,\sigma(t-v)\,p_{\overline{SS}}(t-v, t)\,dv$$

9.3 Sickness and death with duration dependence

In Section 9.2, the Markov property implies that:

$$\boldsymbol{P\left[X_t = H \mid X_s = S, C_s = w\right] = P\left[X_t = H \mid X_s = S\right]}$$

In other words, the duration of your current illness has no bearing on your future health prospects. In order to remove this undesirable feature, we modify the model by allowing the rates of transition out of *S* to depend on the current holding time C_t :



Question

Explain why this model hasn't made the transitions from the healthy state dependent on the holding time.

Solution

With most illnesses, sick people tend to follow a fairly predictable pattern of either recovering in roughly so many weeks, or getting worse and dying after such-and-such a time. So the probabilities of recovery and death will have a fairly definite pattern as a function of duration (for a specified illness, at least). With healthy people on the other hand, although they will in general become more likely to fall sick or die as they get older, there is no particular reason to expect a 40-year-old's chance of falling sick to depend on how long it is since (s)he was last sick, although it might do if some people are 'sickly' by nature.

This appears to take us outside the scope of this unit, as the value of C_t must now be incorporated into the state, so that the state space is not countable (*ie* discrete as opposed to continuous) any more. However, the framework from above can still be used provided that there is careful conditioning on the relevant current holding time.

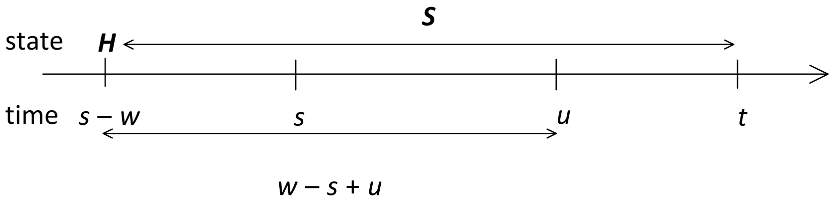
In fact, since the transition rates σ and μ do not depend on C_t the probability of remaining continuously healthy during $[s,t]$ is given by (5.6) as before.

This is because the overall rate out of *H* at time u is $\sigma(u)+\mu(u)$ so that:

$$p_{\overline{HH}}(s,t)=\exp\left[-\int_s^t\left(\sigma(u)+\mu(u)\right)du\right]=\exp\left[-\int_0^{t-s}\left(\sigma(s+u)+\mu(s+u)\right)du\right]$$

This is unaffected by the fact that recovery rates and mortality rates for a sick life depend on how long the life has been sick.

On the other hand, to calculate the probability of remaining continuously sick during $[s,t]$ given a current illness period $[s-w,s]$, one needs to update the values of ρ and υ as the illness progresses:



$$\boldsymbol{P\left[X_t=S,R_s>t-s\,|\,X_s=S,C_s=w\right]=\exp\left[-\int_s^t\left(\rho(u,w-s+u)+\upsilon(u,w-s+u)\right)du\right]}$$

If there were no duration dependence, this expression would simplify to:

$$p_{\overline{SS}}(s,t)=\exp\left[-\int_s^t\left(\rho(u)+\upsilon(u)\right)du\right]=\exp\left[-\int_0^{t-s}\left(\rho(s+u)+\upsilon(s+u)\right)du\right]$$

However, when there is duration dependence, it must be taken into account whenever ρ and υ occur. In the integral in the Core Reading, *u* denotes a time between *s* and *t*. For a given *u*, the duration is the length of time the person has been sick. Since at time *s* the duration is *w*, the duration at time u must be $w+(u-s)$. This explains the appearance of $w-s+u$ in the probability above. The Core Reading equation above can also be written as:

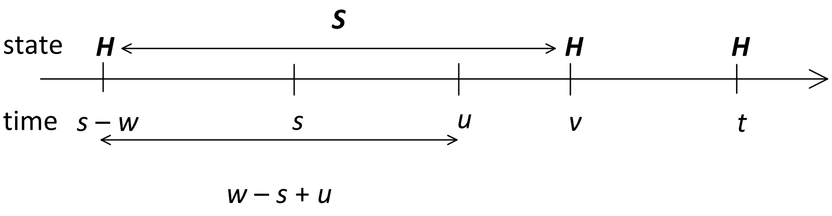
$$p_{\overline{S_wS}}(s,t)=\exp\left[-\int_0^{t-s}\left(\rho(s+u,w+u)+\upsilon(s+u,w+u)\right)du\right]$$

The subscript of w in the probability $p_{\overline{S_wS}}(s,t)$ indicates that the life has already been sick for w years at time s. So, if the life stays in the sick state up to time $s+u$, it will then have duration of sickness $w+u$. Some people find this form of the expression easier to work with. If we make the substitution $r=s+u$, we can see that the two versions are equivalent.

As a final example, the probability of being healthy at time t given that you are sick at time s with current illness duration w can be written as:

$$\boldsymbol{p_{S_wH}(s,t) = P\left[X_t = H \mid X_s = S, C_s = w\right]}$$

$$\boldsymbol{= \int_s^t e^{-\int_s^v (\rho(u,w-s+u)+\upsilon(u,w-s+u))du} \rho(v,w-s+v)\, p_{HH}(v,t)dv}$$



Again, this is the same as the formula without duration dependence, but with the transition rates ρ and υ modified as necessary. We can also write this expression as:

$$p_{S_wH}(s,t) = \int_s^t p_{\overline{S_wS}}(s,v)\rho(v,w-s+v)\, p_{HH}(v,t)dv$$

or, to be consistent with the formulation we have been using so far for backward integral equations:

$$p_{S_wH}(s,t) = \int_0^{t-s} p_{\overline{S_wS}}(s,s+v)\rho(s+v,w+v)\, p_{HH}(s+v,t)\,dv$$

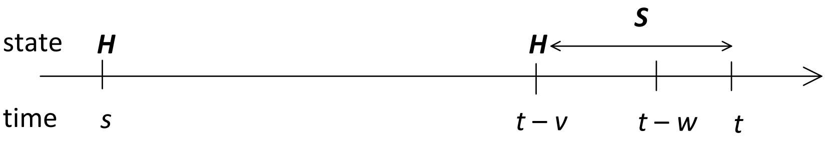
This is saying that the life remains sick throughout the time period s to $s+v$, then makes a transition to healthy at time $s+v$ and duration of sickness $w+v$, and finally goes from healthy at time $s+v$ to healthy at time t, though the life may be sick in between these times.

Question

Write down an integral expression for the probability of a life being sick at time t, having been so for at least w years, given that the life was healthy at time s.

Solution

The diagram for this situation is as follows:



In integral form:

$$P\left[X_t = S, C_t > w \mid X_s = H\right] = \int_w^{t-s} p_{HH}(s,t-v)\sigma(t-v)p_{\overline{S_0S}}(t-v,t)\,dv$$

where:

$$p_{\overline{S_0S}}\left(t-v,t\right) = \exp\left[-\int_{t-v}^{t}\left(\rho(u,u+v-t)+\upsilon(u,u+v-t)\right)du\right]$$

$$= \exp\left[-\int_0^{v}\left(\rho(r+t-v,r)+\upsilon(r+t-v,r)\right)dr\right]$$

The integral in v is from w because we are told that the current holding time at time t is at least w. For the integral in u, at time u the current holding time is $u-\left(t-v\right)=u+v-t$. We can show that the integral in u and the integral in r are equivalent by making the substitution $r=u+v-t$.

Question

Consider again the marriage model in Section 9.1, only now assume that the transition rate $d(t)$ depends on the current holding time. (So the chance of divorce depends on how long a person has been married.) Write down expressions for the probability that:

(i) a bachelor remains a bachelor throughout a period $\left[s,t\right]$

(ii) a person who gets married at time $s-w$ and remains married throughout $\left[s-w,s\right]$, continues to be married throughout $\left[s,t\right]$

(iii) a person is married at time t and has been so for at least time w, given that they were divorced at time $s<t-w$.

Solution

(i) ***Probability of staying in the bachelor state***

This is unaffected by the dependence on the current holding time. So:

$$p_{\overline{BB}}(s,t) = \exp\left[-\int_s^t\left(\alpha(u)+\mu(u)\right)du\right]$$

$$= \exp\left[-\int_0^{t-s}\left(\alpha(s+u)+\mu(s+u)\right)du\right]$$

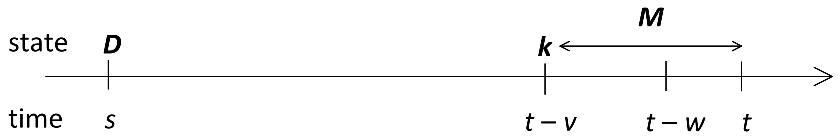
(ii) ***Probability of staying in the married state***

$$p_{\overline{M_wM}}(s,t) = \exp\left(-\int_s^t\left(d(u,u-s+w)+\upsilon(u)+\mu(u)\right)du\right)$$

$$= \exp\left(-\int_0^{t-s}\left(d(s+u,w+u)+\upsilon(s+u)+\mu(s+u)\right)du\right)$$

(iii) *Transition probability*

The diagram for this situation is as follows:



There are two possibilities for the state k, namely D and W, and both of these gives a contribution.

$$P\left[X_t = M, C_t > w \mid X_s = D\right]$$

$$= \int_w^{t-s} \left[p_{DD}(s,t-v)\rho(t-v) + p_{DW}(s,t-v)r(t-v)\right] p_{\overline{M_0 M}}(t-v,t)\,dv$$

The integral in v is from w because we are told that the current holding time at time t is at least w. The subscript of 0 on the M shows that the life is newly married at time $t-v$.

10 Modelling and simulation

This section is similar to the modelling and simulation section of Chapter 2, which dealt with Markov chains.

Modelling is discussed first. We deal with Poisson models initially, including time-inhomogeneous Poisson processes (as introduced below), before proceeding to more general homogeneous processes, and finally dealing with inhomogeneous processes.

A short discussion of simulation is then given.

10.1 Time-homogeneous Poisson process models

A time-homogeneous Poisson process has a single parameter λ. The estimation of this parameter given a collection of data is straightforward.

Example

An insurance office observes that m claims arrive in a total of T time units. If the company decides that a Poisson process model is appropriate, the most suitable estimate for λ would appear to be $\hat{\lambda} = m/T$. This intuitive estimate is confirmed by more formal procedures such as maximum likelihood estimation.

Having estimated the parameter, all that remains is to test goodness of fit.

It is a basic assumption here that a Poisson process is appropriate. If this is not a reasonable assumption then the fit may not be very good. So, if a goodness-of-fit test gives a result that would lead us to reject the null hypothesis, then an alternative model may be appropriate.

The test is carried out as follows.

Divide the total time T into k equal intervals. If the Poisson process model fits, the number of claims arriving in the k intervals should form a sequence of independent Poisson variates, each with mean $\lambda T/k$. There are two things to test here:

- **whether the distribution is Poisson and**
- **whether the observations are independent.**

A standard χ^2 goodness-of-fit test can be employed to determine whether the Poisson distribution fits.

Assuming that the fit is adequate, independence is probably best tested against the alternative that there is some form of serial dependence.

Monthly claims arriving may not be uncorrelated with the previous months, for example.

Tests for serial correlation are covered in Chapter 10.

10.2 Time-inhomogeneous Poisson process models

In some classes of business, such as insurance against storm damage, the intensity of arrival of claims may vary predictably with time, in the sense that the insurer can tell in advance that some time intervals will have more claim arrivals than other intervals of equal length. A suitable model here is the *time-inhomogeneous Poisson process*, for which the arrival rate of claims is a function $\lambda(t)$. In the given example λ will be periodic, with a period of one year.

It is impractical to attempt to estimate the value of $\lambda(t)$ separately for each value of t. A common procedure is to divide the whole time period up into pieces of a suitable size and to estimate the arrival rate separately for each piece.

Since t is continuous, we cannot hope to have enough data to make the former estimation procedure statistically significant. The same applies if the pieces are too small.

Thus data for a whole year may be divided into months, giving 12 estimated claim arrival rates. Tests of goodness of fit should be carried out for each month separately, but tests for serial correlation should use the whole data set at once.

For example, if we have several years of monthly data, then we could think of all the January months together as a time-homogeneous Poisson process with a certain fixed parameter λ. This could be tested for goodness of fit separately.

On the other hand, when testing for serial correlation, we need to test, for example, whether one month is correlated with the previous month, on average. We therefore use the whole data set at once.

10.3 Time-homogeneous Markov models

The structural analysis of Section 7 of Chapter 4 is of paramount importance when it comes to modelling continuous-time Markov jump processes. Recall that each visit to any given state i is of exponential duration with mean $1/\lambda_i$ and is independent of the durations of previous visits to that state and of the destination after the next jump. Further, the probability that the next transition is to state j is μ_{ij}/λ_i.

This suggests that it is feasible to separate the two estimation procedures.

- **First the λ_i may be estimated: look at the data for the durations of visits to state i and let $1/\hat{\lambda}_i$ be equal to the sample mean of this collection.**
- **Next proceed as in Chapter 2: let n_i be the number of completed visits to state i, n_{ij} the number of direct transitions from state i to state j, and set $\hat{p}_{ij} = n_{ij} / n_i$. Since p_{ij} is equal to μ_{ij}/λ_i, a sensible estimator for μ_{ij} is $\hat{\mu}_{ij} = \hat{\lambda}_i \hat{p}_{ij}$.**

We have already seen in Section 11 of Chapter 4 that the transition rate μ_{ij} is estimated by:

$$\hat{\mu}_{ij} = \frac{\text{number of transitions from state } i \text{ to state } j}{\text{total holding time in state } i}$$

This formula is equivalent to the one in the second bullet point above.

Also, λ_i can be estimated by $\hat{\lambda}_i = \sum_{j \neq i} \hat{\mu}_{ij}$.

Tests for goodness of fit are more problematical, if only because there is a vast collection of possible alternative hypotheses. It is reasonable to test whether the visits to a given state really are exponentially distributed: a χ^2 goodness-of-fit test will do this. It is also reasonable to test whether the jump chain really does exhibit the Markov property: see Chapter 2 for a discussion. But there are other implications of the Markov structure that should be tested and the procedure is not always clear.

For example, to derive a formal test as to whether the destination of a jump is independent of the duration of the previous holding time we would need to do something like this:

- **look at all visits to state i and classify them as long-duration, medium-duration or short-duration**
- **for each duration category, estimate the transition probabilities of the jump chain separately, giving estimates $\hat{p}_{ij}^{(L)}, \hat{p}_{ij}^{(M)}$ and $\hat{p}_{ij}^{(S)}$**
- **determine whether the differences between the sets of estimated transition probabilities are significant.**

However, it is by no means clear what test statistic could be employed or what its distribution might be. In practice the investigation of this question would be accomplished graphically: for each visit to state i, plot a point on a graph whose x-coordinate represents the duration of the visit, y-coordinate the destination of the next jump. If a pattern appears, reject the assumption of independence.

Other tests, such as testing whether the first visit to a given state is significantly longer than subsequent visits, are also best treated graphically.

10.4 Time-inhomogeneous Markov models

The structural decomposition of the time-homogeneous Markov model does not apply to the time-inhomogeneous case. The estimation of time-dependent transition rates, such as the force of mortality or age-dependent rate of recovery from sickness, is best treated within the context of the particular model being studied. This is covered in later units in this course.

10.5 Simulation

In order to simulate a process, random values of the random variables that are involved must be produced.

There are two approaches to the task of simulating a time-homogeneous Markov jump process. The first is an approximate method and the second exact.

Approximate method

Divide time into very short intervals of width h, say, where $h\mu_{ij}$ is much smaller than 1 for each i and j. The transition matrix $P(h)$ of the Markov chain has entries approximately given by:

$$p_{ij}^{*}(h) = \delta_{ij} + h\mu_{ij}$$

Using the techniques of Chapter 2 we may simulate a discrete-time Markov chain $\{Y_n, n \geq 0\}$ with these transition probabilities, then write $X_t = Y_{[t/h]}$.

For example, if $h = \dfrac{1}{100}$, we would simulate $\{Y_n, n \geq 0\}$ and define $X_t = Y_{[100t]}$, *ie*:

$$X_t = \begin{cases} Y_0 & \text{for } 0 \leq t < 0.01 \\ Y_1 & \text{for } 0.01 \leq t < 0.02 \\ Y_2 & \text{for } 0.02 \leq t < 0.03 \\ \vdots & \vdots \end{cases}$$

This simplistic method is not very satisfactory, as its long-term distribution may differ significantly from that of the process being modelled.

Since the probabilities being used are not exact, the errors introduced accumulate as time passes. In the long run they may be significant.

An improved version of this method is available, which uses the exact transition probabilities $p_{ij}(h)$ instead of $p_{ij}^{*}(h)$, but this naturally requires that the exact probabilities be calculated in advance. General techniques for such calculations, where not covered by this chapter, are beyond the scope of the syllabus.

Exact method

This takes advantage of the structural decomposition of the jump process. First simulate the jump chain of the process as a Markov chain with transition probabilities $p_{ij} = \mu_{ij}/\lambda_i$.

Once the path $\{\hat{X}_n : n = 0,1,\ldots\}$ has been generated, the holding times $\{T_n : n = 0,1,\ldots\}$ are a sequence of independent exponential random variables, T_n having rate parameter given by $\lambda_{\hat{X}_n}$.

We will describe how to use the exact method to simulate a sample path for a Health-Sickness-Death model with generator matrix:

$$\begin{array}{c} \begin{array}{ccc} H & S & D \end{array} \\ \begin{bmatrix} -0.5 & 0.4 & 0.1 \\ 0.6 & -0.8 & 0.2 \\ 0 & 0 & 0 \end{bmatrix} \end{array}$$

We assume that a policyholder begins in the healthy state.

The transition matrix of the Markov jump chain is:

$$\begin{array}{ccc} H & S & D \\ \left[\begin{array}{c} 0 \\ \frac{0.6}{0.8}=0.75 \\ 0 \end{array}\right. & \begin{array}{c} \frac{0.4}{0.5}=0.80 \\ 0 \\ 0 \end{array} & \left.\begin{array}{c} \frac{0.1}{0.5}=0.20 \\ \frac{0.2}{0.8}=0.25 \\ 1 \end{array}\right] \end{array}$$

Each probability is the ratio of the force between the two states and the total of the forces on paths leaving the initial state. Once the process enters state D it remains there for ever.

The holding time in the healthy state is $Exp(0.5)$ and the holding time in the sick state is $Exp(0.8)$.

The first step is to simulate the states occupied by the Markov jump chain.

Row 1 of the transition matrix is the conditional distribution of X_1 given that $X_0 = H$. We use Monte Carlo simulation to generate a simulated value for X_1.

If the simulated value is D, then the simulation of the sample path is complete because the process never leaves state D. If the simulated value is S, then we use row 2 of the transition matrix, which is the conditional distribution of X_2 given that $X_1 = S$, to simulate a value for X_2.

This process is repeated to simulate additional values of the Markov jump chain.

The second step is to simulate the holding times corresponding to the states in the simulated Markov jump chain.

The holding times for each occupancy of state H will be simulated from an $Exp(0.5)$ distribution. We use Monte Carlo simulation to generate these values. The same method can be used to generate $Exp(0.8)$ random variables for each holding time in state S.

By adding up the holding times to match the states simulated from the Markov jump chain, we will obtain the simulated times at which the Markov process jumps between states.

Time-inhomogeneous processes

Given the transition rates of a time-inhomogeneous Markov chain and given the state X_t at time t, it is in principle possible to determine the density function of the time until the next transition and the destination of the next jump: see Section 9 for examples. This means that standard simulation techniques can be deployed to generate an exact simulation of the process.

In practice, however, such a procedure is cumbersome in the extreme, unless the number of states is very small, and a more usual approach is to use the approximate method outlined above. The exact transition probabilities $p_{ij}(t,t+h)$ will seldom be to hand, meaning that the less satisfactory approximate values $p^*_{ij}(t,t+h) = \delta_{ij} + h\mu_{ij}(t)$ must be used instead.

The method is acceptable for short-term simulations but is unreliable in the long term.

As above, the errors introduced will accumulate so that the long-term simulation is not acceptable.

Chapter 5 Summary

Chapman-Kolmogorov equations

$$p_{ij}(s,t) = \sum_k p_{ik}(s,u)p_{kj}(u,t)$$

Here u is any intermediate time between s and t (possibly equal to s or t) that is convenient for the calculation.

Transition rates

$$\mu_{ij}(s) = \left[\frac{\partial}{\partial t} p_{ij}\left(s,t\right)\right]_{t=s} = \lim_{h \to 0} \frac{p_{ij}(s,s+h) - p_{ij}(s,s)}{h}$$

This is equivalent to:

$$p_{ij}(s,s+h) = \begin{cases} h\mu_{ij}\left(s\right) + o(h) & \text{if } i \neq j \\ 1 + h\mu_{ii}\left(s\right) + o(h) & \text{if } i = j \end{cases}$$

for small h.

Generator matrix

The generator matrix is the matrix of transition rates $\mu_{ij}(t)$. It is usually denoted by $A(t)$.

Each row of the generator matrix $A(t)$ sums to zero since $\mu_{ii}(t) = -\sum_{j \neq i} \mu_{ij}(t)$.

Backward and forward differential equations (time-inhomogeneous case)

Forward: $$\frac{\partial}{\partial t} p_{ij}(s,t) = \sum_k p_{ik}(s,t)\mu_{kj}(t)$$

$$\frac{\partial}{\partial t} P\left(s,t\right) = P\left(s,t\right)A(t) \quad \text{(matrix form)}$$

Backward: $$\frac{\partial}{\partial s} p_{ij}(s,t) = -\sum_k \mu_{ik}(s)p_{kj}(s,t)$$

$$\frac{\partial}{\partial s} P\left(s,t\right) = -A(s)P\left(s,t\right) \qquad \text{(matrix form)}$$

Occupancy probabilities

The probability of remaining in state i throughout the interval (s,t) is:

$$p_{\overline{ii}}(s,t)=\exp\left(-\int_0^{t-s}\lambda_i(s+u)\,du\right)=\exp\left(-\int_s^t\lambda_i(u)\,du\right)$$

where $\lambda_i(u)$ is the total force of transition out of state i at time u.

Probability that the process goes into state *j* when it leaves state *i*

Given that the process is in state i at time s and it stays there until time $s+w$, the probability that it moves into state j when it leaves state i at time $s+w$ is:

$$\frac{\mu_{ij}(s+w)}{\lambda_i(s+w)}=\frac{\text{the force of transition from state } i \text{ to state } j \text{ at time } s+w}{\text{the total force out of state } i \text{ at time } s+w}$$

Backward and forward integral equations

Backward: $$p_{ij}(s,t)=\sum_{l\neq i}\int_0^{t-s}p_{\overline{ii}}(s,s+w)\,\mu_{ik}(s+w)p_{kj}(s+w,t)\,dw \qquad i\neq j$$

The backward equation is obtained by considering the timing and nature of the first jump after time s. The duration spent in this initial state before jumping to another state (state k say) is denoted by w. The integral reflects the three stages involved:

1. remaining in state i from time s to time $s+w$
2. jumping from state i to state k at time $s+w$
3. moving from state k at time $s+w$ to state j at time t (possibly visiting other states along the way).

We then consider the possible values of w to obtain limits of 0 and $t-s$ for the integral, and we sum over all possible intermediate states k.

When $i=j$, the equation is:

$$p_{ii}(s,t)=\sum_{l\neq i}\int_0^{t-s}p_{\overline{ii}}(s,s+w)\,\mu_{ik}(s+w)\,p_{ki}(s+w,t)\,dw+p_{\overline{ii}}(s,t)$$

The extra term here is to account for the possibility of staying in state i from time s to time t.

Forward: $$p_{ij}(s,t)=\sum_{k\neq j}\int_0^{t-s} p_{ik}(s,t-w)\,\mu_{kj}(t-w)\;p_{\overline{jj}}(t-w,t)\,dw \qquad i\neq j$$

The forward equation is obtained by considering the timing and nature of the last jump before time t. The duration then spent in this final state (state j) before time t is denoted by w. The integral reflects the three stages involved:

(1) moving from state i at time s to state k at time $t-w$ (possibly visiting other states along the way)

(2) jumping from state k to state j at time $t-w$

(3) remaining in state j from time $t-w$ to time t.

We then consider the possible values of w to obtain limits of 0 and $t-s$ for the integral, and sum over all possible intermediate states k.

When $i=j$, the equation is:

$$p_{ii}(s,t)=\sum_{k\neq j}\int_0^{t-s} p_{ik}(s,t-w)\,\mu_{ki}(t-w)\;p_{\overline{ii}}(t-w,t)\,dw+p_{\overline{ii}}(s,t)$$

The extra term here is to account for the possibility of staying in state i from time s to time t.

Integral equations can be adjusted to deal with transition rates that are duration-dependent, *ie* transition rates that depend on the holding time in the current state.

The practice questions start on the next page so that you can keep all the chapter summaries together for revision purposes.

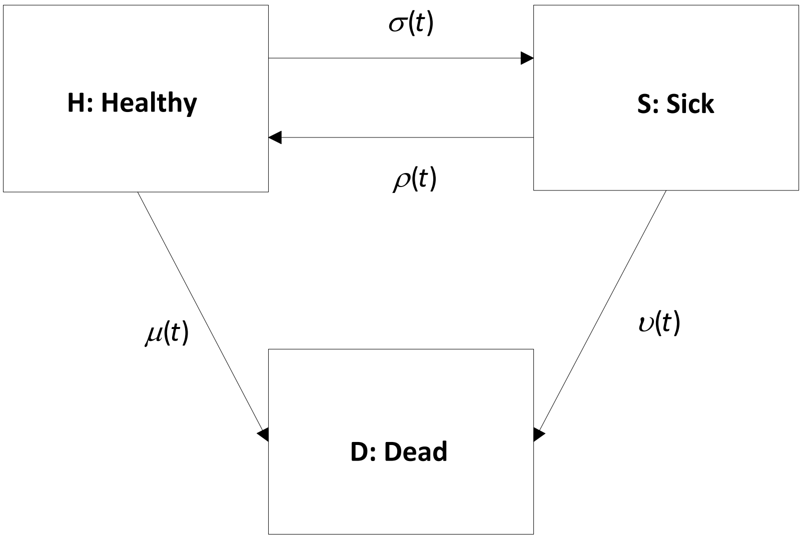
Chapter 5 Practice Questions

5.1 Derive the differential equation $\frac{\partial}{\partial t} p_{\overline{HH}}(s,t) = -p_{\overline{HH}}(s,t)\big(\sigma(t)+\mu(t)\big)$.

5.2 A Markov jump process is used to model sickness and death. Four states are included, namely H, S_1, S_2 and D, which represent healthy, sick, terminally sick and dead, respectively. We are told that the people who are terminally sick never recover and die at a rate of $1.03(1.01)^t$ where t is their age in years.

Calculate the probability that a terminally sick 50-year-old dies within a year.

5.3 In a Markov jump process model of sickness and death there are three states: healthy (H), sick (S) and dead (D). The transition graph is shown below. Let X_t denote the state of the process at time *t*.



(i) Write down the generator matrix at time t.

(ii) Define the residual holding time, R_s.

(iii) Given that a life is sick at time *s*, give an expression for the probability it remains sick for a further period of at least *w*.

(iv) State the probability density function of R_s given that $X_s = S$.

(v) Given that a transition from *H* takes place at time t, give an expression for the probability that it is to *S*.

(vi) Give the integral form of the Kolmogorov backward equation for $p_{SD}(s,t)$, the probability that an individual who is sick at time *s* is dead at time *t*.

(vii) Explain your formula in (vi) by general reasoning.

5.4 A 3-state time-homogeneous Markov jump process is determined by the following matrix of transition rates:

$$A=\begin{pmatrix} -3 & 2 & 1 \\ 0 & -2 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

The distribution at time 0 is $\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$. Determine the distribution at time 1.

5.5 An investigator wishes to construct a multiple decrement model of the mortality of a population, subdivided by the following causes of death.

Cause 1: Cancer

Cause 2: Heart disease

Cause 3: All other causes

You are given the following definitions:

μ_x^i is the force of mortality due to cause i ($i=1,2,3$) at exact age x

${}_uq_x^i$ is the probability that a life at exact age x dies due to cause i ($i=1,2,3$) before reaching exact age $x+u$ ($u\geq 0$)

${}_up_x$ is the probability that a life aged exactly x is still alive at exact age $x+u$ ($u>0$)

You may assume that $\dfrac{{}_uq_x^i}{u}\to\mu_x^i \text{ as } u\to 0$.

(i) Derive an expression for ${}_tp_x$ ($t\geq 0$), in terms of the forces of mortality, using only the above functions in your derivation.

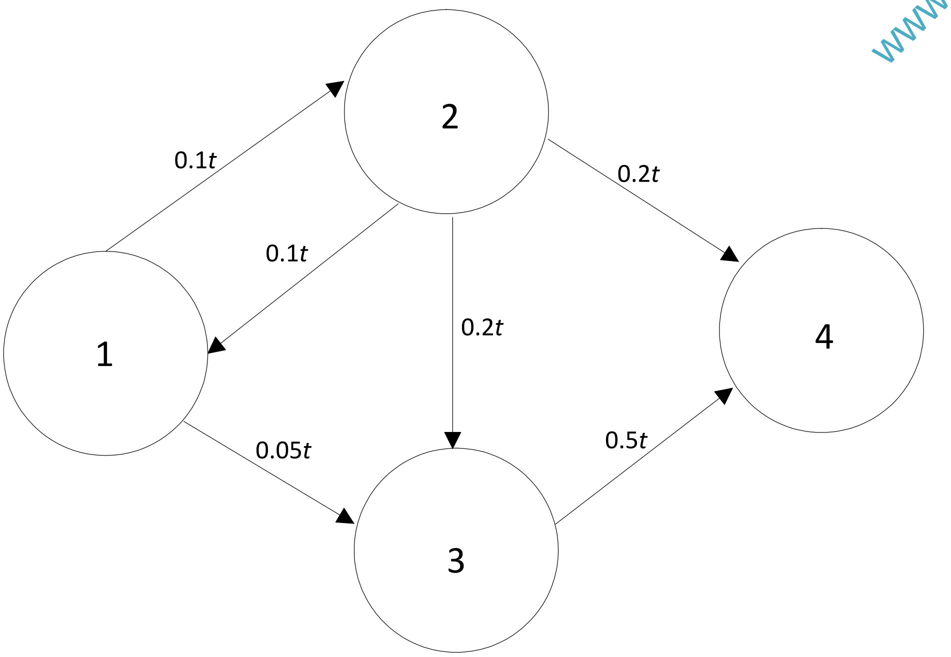
(ii) Write down an integral expression for q_x^i in terms of ${}_tp_x$ and the appropriate force(s) of mortality. (Note that $q_x^i={}_1q_x^i$.)

(iii) Assuming that the force of mortality from each cause i is a constant μ^i between integer ages x and $x+1$, show that:

$$q_x^i=\frac{\mu^i}{\sum_{i=1}^{3}\mu^i}\times q_x \quad \text{where} \quad q_x=1-p_x$$

(Note that $p_x={}_1p_x$.)

5.6 Consider the following time-inhomogeneous Markov jump process with transition rates as shown below:



(i) Write down the generator matrix at time t .

(ii) Write down the Kolmogorov backward differential equations for $P_{33}(s,t)$ and $P_{13}(s,t)$.

(iii) Using the technique of separation of variables, or otherwise, show that the solution of the differential equation for $P_{33}(s,t)$ is:

$$P_{33}(s,t) = e^{-0.25\left(t^2 - s^2\right)}$$

(iv) Show that the probability that the process visits neither state 2 nor state 4 by time t , given that it starts in state 1 at time 0, is:

$$\frac{8}{7}e^{-0.075t^2} - \frac{1}{7}e^{-0.25t^2}$$

(v) State the limiting value as $t \to \infty$ of the probability in (iv). Explain why this must be the case for this particular model.

5.7 Exam style

A time-inhomogeneous Markov jump process has state space {A, B} and the transition rate for switching between states equals $2t$, regardless of the state currently occupied, where t is time.

The process starts in state A at $t = 0$.

(i) Calculate the probability that the process remains in state A until at least time s. [2]

(ii) Show that the probability that the process is in state B at time T, and that it is in the first visit to state B, is given by $T^2 \times e^{-T^2}$. [3]

(iii) (a) Sketch the probability function given in (ii).

(b) Give an explanation of the shape of the probability function.

(c) Calculate the time at which it is most likely that the process is in its first visit to state B.

[6]

[Total 11]

5.8 Exam style

An illness-death model has three states:

1 = healthy
2 = sick
3 = dead

Let ${}_tp_x^{ij}$ denote the probability that a life in state i at age x is in state j at age $x+t$ and let μ_{x+t}^{ij} denote the force of transition from state i to state j at age $x+t$.

(i) Draw and label a diagram showing the three states and the transition intensities between them. [2]

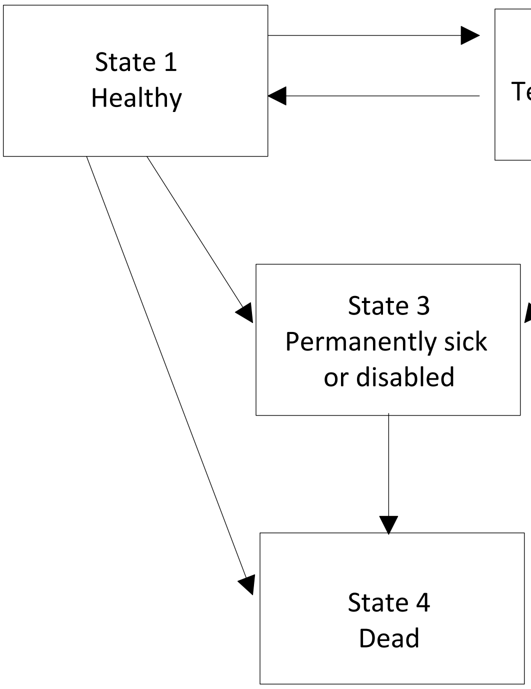
(ii) Show, from first principles, that in this illness-death model:

$$\frac{\partial}{\partial t}\,{}_tp_x^{12} = {}_tp_x^{11}\mu_{x+t}^{12} - {}_tp_x^{12}\mu_{x+t}^{21} - {}_tp_x^{12}\mu_{x+t}^{23}$$ [6]

[Total 8]

5.9 Exam style

The following diagram represents a four-state Markov model.



The force of transition from state i to state j $(i \neq j)$ at age x is denoted by μ_x^{ij}, and the probability that a life, who is in state i when aged x, will be in state j at age $x+t$ is ${}_tp_x^{ij}$.

(i) Derive from first principles a differential equation for ${}_tp_x^{23}$, stating all assumptions made. [5]

(ii) Given that, for $x = 40, 41$:

$${}_1p_x^{12} = 0.03, {}_1p_x^{13} = 0.002, {}_1p_x^{14} = 0.001,$$

$${}_1p_x^{21} = 0.4, {}_1p_x^{23} = 0.1, {}_1p_x^{24} = 0.01 \text{ and } {}_1p_x^{34} = 0.3$$

calculate ${}_2p_{40}^{13}$. [2]

(iii) An insurance company issues a combined sickness, disability and assurance contract that provides the following benefits:

- an income payable while the policyholder is temporarily sick or disabled; and
- a lump sum payable either on becoming permanently sick or disabled, or on death.

The contract terminates as soon as the lump sum has been paid.

Explain how the model could be simplified for the purpose of modelling the claims process involved. State how your answer to (i) would be altered as a result of this change. (You are not required to derive this result from first principles). [2]

[Total 9]

Chapter 5 Solutions

5.1 We consider $p_{\overline{HH}}(s,t+h)$ where h is a small amount and condition on the state at time t. In this case, no transition out of H is possible, so:

$$p_{\overline{HH}}(s,t+h)=p_{\overline{HH}}(s,t)\,p_{\overline{HH}}(t,t+h)$$

However, during the short interval from time t to time $t+h$, the process either remains in *H*, changes from *H* to *S* or changes from *H* to *D*. We assume here that the probability of more than one change is very small, *ie* it is an $o(h)$ function. So:

$$p_{\overline{HH}}(t,t+h)+p_{HS}(t,t+h)+p_{HD}(t,t+h)+o(h)=1$$

Since we know that $p_{HS}(t,t+h)=h\sigma(t)+o(h)$ and $p_{HD}(t,t+h)=h\mu(t)+o(h)$ this gives:

$$p_{\overline{HH}}(t,t+h)=1-h\big(\sigma(t)+\mu(t)\big)+o\big(h\big)$$

Therefore:

$$p_{\overline{HH}}(s,t+h)=p_{\overline{HH}}(s,t)\big[1-h\big(\sigma(t)+\mu(t)\big)\big]+o(h)$$

This can be rearranged to give:

$$\frac{p_{\overline{HH}}(s,t+h)-p_{\overline{HH}}(s,t)}{h}=-p_{\overline{HH}}(s,t)\big(\sigma(t)+\mu(t)\big)+\frac{o(h)}{h}$$

Letting $h\to 0$ gives:

$$\frac{\partial}{\partial t}p_{\overline{HH}}(s,t)=-p_{\overline{HH}}(s,t)\big(\sigma(t)+\mu(t)\big)$$

since $\lim\limits_{h\to 0}\dfrac{o(h)}{h}=0$.

5.2 The probability of surviving the year is:

$$p_{\overline{S_2S_2}}(50,51)=\exp\left(-\int_{50}^{51}1.03\big(1.01\big)^t\,dt\right)$$

Noting that $\big(1.01\big)^t=e^{t\ln 1.01}$, this is an exponential integral:

$$\int_{50}^{51}1.03\big(1.01\big)^t\,dt=1.03\left[\frac{\big(1.01\big)^t}{\ln 1.01}\right]_{50}^{51}=1.70242$$

So the probability of dying within the year is:

$$1-e^{-1.70242}=0.818$$

5.3 (i) ***Generator matrix***

The generator matrix at time t is:

$$A(t)=\begin{pmatrix} -\sigma(t)-\mu(t) & \sigma(t) & \mu(t) \\ \rho(t) & -\rho(t)-\upsilon(t) & \upsilon(t) \\ 0 & 0 & 0 \end{pmatrix}$$

(ii) ***Residual holding time***

The residual holding time at time s is the random variable representing the remaining time until the next jump.

(iii) ***Occupancy probability***

The required expression is:

$$P(R_s > w \mid X_s = S) = e^{-\int_s^{s+w}(\rho(u)+\upsilon(u))du}$$

(iv) ***Probability density function***

The PDF is:

$$f_{R_s}(w) = \left(\rho(s+w)+\upsilon(s+w)\right)e^{-\int_s^{s+w}(\rho(u)+\upsilon(u))du}, \quad w>0$$

This can be obtained by differentiating the CDF:

$$P(R_s \le w \mid X_s = S) = 1 - e^{-\int_s^{s+w}(\rho(u)+\upsilon(u))du}$$

(v) ***Conditional probability of transition to state S***

This is the ratio of the force of transition from *H* to *S* to the total force of transition out of *H*:

$$\frac{\mu_{HS}(t)}{\mu_{HS}(t)+\mu_{HD}(t)} = \frac{\sigma(t)}{\sigma(t)+\mu(t)}$$

(vi) ***Integral form of the Kolmogorov backward equation***

Considering the two possible destination states after the first transition, we obtain:

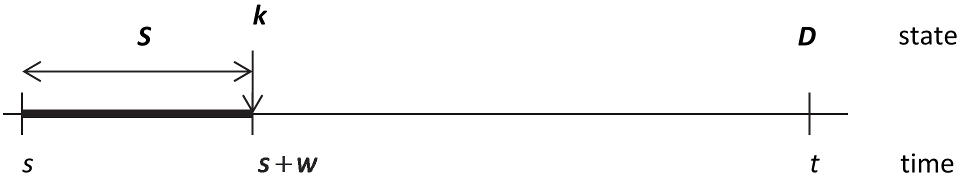
$$p_{SD}(s,t) = \int_0^{t-s} e^{-\int_s^{s+w}(\rho(u)+\upsilon(u))du}\upsilon(s+w)p_{DD}(s+w,t)dw$$
$$+\int_0^{t-s} e^{-\int_s^{s+w}(\rho(u)+\upsilon(u))du}\rho(s+w)p_{HD}(s+w,t)dw$$

Since $p_{DD}(s+w,t)=1$, this simplifies slightly to:

$$p_{SD}(s,t)=\int_0^{t-s} e^{-\int_s^{s+w}(\rho(u)+\upsilon(u))du}\,\upsilon(s+w)\,dw$$

$$+\int_0^{t-s} e^{-\int_s^{s+w}(\rho(u)+\upsilon(u))du}\,\rho(s+w)\,p_{HD}(s+w,t)\,dw$$

(vii) ***General reasoning explanation***



The backward equation is constructed by conditioning on the first transition time. Let $s+w$ be the time of the first transition from state *S*. Let this transition be to state *k*, which can be either *H* or *D*.

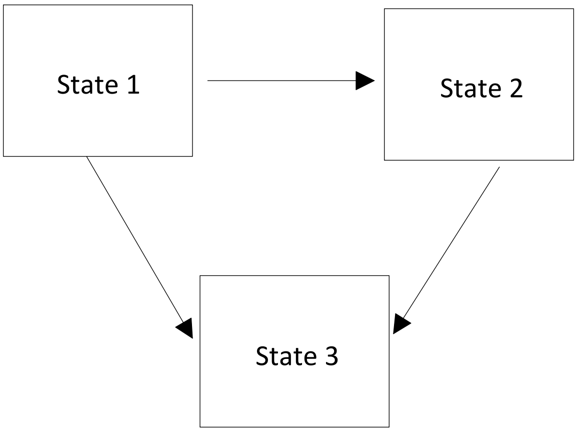
There steps to consider are as follows:

- The process is in state S from time s to time $s+w$. This has probability:

$$p_{\overline{SS}}(s,s+w)=e^{-\int_s^{s+w}(\rho(u)+\upsilon(u))du}$$

- At time $s+w$, the process makes a transition from state S to state k. So we multiply by the force of transition $\mu_{Sk}(s+w)$.
- Finally, we need the probability $p_{kD}(s+w,t)$ for going from state k to state *D*.
- Multiplying these together and integrating over the possible times for w, we obtain the given expression. We can simplify using the fact that $p_{DD}(s,t)=1$.

5.4 A diagram may help you to see what is going on here. Let's call the states 1, 2 and 3. The transition diagram is as follows:



We need to find the matrix of transition probabilities, $P(t)$, and then calculate:

$$\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)P(1)$$

State 3 in the diagram above resembles the dead state, since once we enter state 3 we cannot leave it.

We have:

$$P_{31}(t)=P_{32}(t)=0 \text{ and } P_{33}(t)=1$$

Also $P_{21}(t)=0$.

Since the only path from 2 to 2 is to stay there throughout:

$$P_{22}(t)=P_{\overline{22}}(t)=e^{-\lambda_2 t}=e^{-2t}$$

and $P_{23}(t)=1-e^{-2t}$.

Similarly $P_{11}(t)=P_{\overline{11}}(t)=e^{-3t}$.

To calculate $P_{12}(t)$ we can use the integral form of the Kolmogorov equation. (This is slightly quicker to deal with than the differential form.) If we use the backward form we have:

$$P_{12}(t)=\int_0^t P_{\overline{11}}(w)\mu_{12}P_{22}(t-w)\,dw=\int_0^t e^{-3w}\times 2\times e^{-2(t-w)}dw$$

$$=2e^{-2t}\left[\frac{e^{-w}}{-1}\right]_0^t=2e^{-2t}\left(1-e^{-t}\right)$$

It follows that:

$$P_{13}(t)=1-e^{-3t}-2e^{-2t}\left(1-e^{-t}\right)=1+e^{-3t}-2e^{-2t}$$

Finally:

$$\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)P(1)=\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)\begin{pmatrix} e^{-3} & 2e^{-2}\left(1-e^{-1}\right) & 1+e^{-3}-2e^{-2} \\ 0 & e^{-2} & 1-e^{-2} \\ 0 & 0 & 1 \end{pmatrix}$$

$$=(0.0166,0.1021,0.8813)$$

5.5 (i) ***Derivation***

Consider the probability ${}_{u+h}p_x$ where h is a small amount. From the Markov assumption, we know that:

$${}_{u+h}p_x = {}_u p_x \times {}_h p_{x+u}$$

The probability that an individual survives a period is one minus the probability that it dies. So:

$${}_h p_{x+u} = 1-\sum_{i=1}^{3} {}_h q^i_{x+u}$$

We know that:

$${}_h q^i_{x+u} = h\mu^i_{x+u} + o(h)$$

Substituting this into the first expression gives:

$${}_h p_{x+u} = 1-h\sum_{i=1}^{3} \mu^i_{x+u} + o(h)$$

Hence:

$${}_{u+h}p_x = {}_u p_x \left(1-h\sum_{i=1}^{3}\mu^i_{x+u}\right)+o(h)$$

Rearranging, we see that:

$$\frac{{}_{u+h}p_x - {}_u p_x}{h} = {}_u p_x\left(-\sum_{i=1}^{3}\mu^i_{x+u}+\frac{o(h)}{h}\right)$$

Letting h tend to zero gives:

$$\frac{\partial}{\partial u}\,{}_u p_x = -\,{}_u p_x \sum_{i=1}^{3} \mu_{x+u}^{i}$$

since $\dfrac{o(h)}{h} \to 0$ as $h \to 0$.

Dividing both sides by ${}_u p_x$, we have:

$$\frac{\partial}{\partial u}\log {}_u p_x = -\sum_{i=1}^{3} \mu_{x+u}^{i}$$

Integrating with respect to u between $u=0$ and $u=t$, we see that:

$$\log {}_t p_x - \log {}_0 p_x = -\int_0^t \sum_{i=1}^{3} \mu_{x+u}^{i}\,du$$

Since $\log {}_0 p_x = \log 1 = 0$, we have:

$${}_t p_x = \exp\left(-\int_0^t \sum_{i=1}^{3} \mu_{x+u}^{i}\,du\right)$$

(ii) ***Formula***

q_x^i is the probability that an individual aged x leaves through cause i during the coming year. Expressed as an integral, this is:

$$q_x^i = \int_0^1 {}_t p_x \mu_{x+t}^i dt$$

since an individual who leaves through cause i during the year must survive all decrements up to some time t in the range $0<t<1$, and then must leave through cause i at time t. Integrating over all possible values of t gives the required probability.

(iii) ***Proof***

Combining the formulae derived in parts (i) and (ii), we have:

$$q_x^i = \int_0^1 {}_t p_x\, \mu_{x+t}^i dt = \int_0^1 \exp\left(-\int_0^t \sum_{i=1}^{3} \mu_{x+u}^{i}\,du\right)\mu_{x+t}^i dt$$

Since the force of mortality is constant over the year, this is:

$$q_x^i = \mu^i \int_0^1 \exp\left(-\int_0^t \sum_{i=1}^{3} \mu^i\,du\right)dt = \mu^i \int_0^1 \exp\left(-t\sum_{i=1}^{3} \mu^i\right)dt$$

Integrating gives:

$$q_x^i = \frac{\mu^i}{\sum\limits_{i=1}^{3} \mu^i}\left[-\exp\left(-t\sum_{i=1}^{3} \mu^i\right)\right]_0^1 = \frac{\mu^i}{\sum\limits_{i=1}^{3} \mu^i}\left[1-\exp\left(-\sum_{i=1}^{3} \mu^i\right)\right]$$

This simplifies to:

$$q_x^i = \frac{\mu^i}{\sum\limits_{i=1}^{3} \mu^i}\left[1-p_x\right] = \frac{\mu^i}{\sum\limits_{i=1}^{3} \mu^i} \times q_x$$

5.6 (i) ***Generator matrix at time t***

At time t we have:

$$A(t) = \begin{pmatrix} -0.15t & 0.1t & 0.05t & 0 \\ 0.1t & -0.5t & 0.2t & 0.2t \\ 0 & 0 & -0.5t & 0.5t \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

As always, the rows of the generator matrix sum to 0.

(ii) ***Backward differential equations***

The matrix form of the backward differential equations is:

$$\frac{\partial}{\partial s} P(s,t) = -A(s)P(s,t)$$

Since this model is time-inhomogeneous and we're asked for the backward differential equation, we are differentiating with respect to s.

For this model:

$$\frac{\partial}{\partial s} P_{33}(s,t) = -\left[-0.5s\,P_{33}(s,t)\right] = 0.5s\,P_{33}(s,t)$$

and:

$$\frac{\partial}{\partial s} P_{13}(s,t) = -\left[-0.15s\,P_{13}(s,t) + 0.1s\,P_{23}(s,t) + 0.05s\,P_{33}(s,t)\right]$$

$$= 0.15s\,P_{13}(s,t) - 0.1s\,P_{23}(s,t) - 0.05s\,P_{33}(s,t)$$

(iii) ***Solving the differential equation***

Separating the variables gives:

$$\frac{\frac{\partial}{\partial s}P_{33}(s,t)}{P_{33}(s,t)}=0.5s$$

and changing the variable from s to u:

$$\frac{\partial}{\partial u}\ln P_{33}(u,t)=0.5u$$

Integrating both sides with respect to u between the limits of $u=s$ and $u=t$, we get:

$$\left[\ln P_{33}(u,t)\right]_s^t=\int_s^t 0.5u\,du=\left[0.25u^2\right]_s^t$$

ie:

$$\ln P_{33}(t,t)-\ln P_{33}(s,t)=0.25\left(t^2-s^2\right)$$

However, since $P_{33}(t,t)=1$ and $\ln 1=0$, we have:

$$-\ln P_{33}(s,t)=0.25\left(t^2-s^2\right)$$

The expression above can be rearranged to give:

$$P_{33}(s,t)=e^{-0.25\left(t^2-s^2\right)}$$

(iv) ***Probability of having visited neither state 2 nor state 4 by time t***

There are two possible ways for the process, which started in state 1 at time 0, to have visited neither state 2 nor state 4 by time t. *These are:*

1. *the process has stayed in state 1 throughout the time interval* $[0,t]$, *or*

2. *for some* s, $0<s<t$, *the process has stayed in state 1 throughout the time interval* $[0,s)$, *jumped into state 3 at time* s, *and stayed in state 3 throughout the time interval* $(s,t]$.

So the probability that we require is the sum of the probabilities of events 1 and 2 above.

Event 1

The probability that the process stays in state 1 throughout the time interval $[0,t]$ is:

$$P_{\overline{11}}(0,t)=\exp\left(-\int_0^t 0.15s\,ds\right)=\exp\left[-0.075s^2\right]_0^t=e^{-0.075t^2}$$

Event 2

The probability that for some s, $0<s<t$, the process stays in state 1 throughout the time interval $[0,s)$, jumps into state 3 at time s, and stays in state 3 throughout the time interval $(s,t]$ is:

$$\int_0^t P_{\overline{11}}(0,s)\,\mu_{13}(s)\,P_{\overline{33}}(s,t)\,ds$$

From above:

$$P_{\overline{11}}(0,s)=e^{-0.075s^2}$$

Also, since a return to state 3 is impossible, we know from (iii):

$$P_{\overline{33}}(s,t)=P_{33}(s,t)=e^{-0.25\left(t^2-s^2\right)}$$

So the probability of event 2 is:

$$\int_0^t e^{-0.075s^2}\;0.05s\,e^{-0.25\left(t^2-s^2\right)}\,ds=0.05e^{-0.25t^2}\int_0^t s\,e^{0.175s^2}\,ds$$

Making the substitution $u=0.175s^2$ (so that $du=0.35s\,ds$), the integral on the RHS above becomes:

$$\int_0^{0.175t^2}\frac{e^u}{0.35}du=\left[\frac{e^u}{0.35}\right]_0^{0.175t^2}=\frac{1}{0.35}\left(e^{0.175t^2}-1\right)$$

So the probability of event 2 is:

$$0.05e^{-0.25t^2}\times\frac{1}{0.35}\left(e^{0.175t^2}-1\right)=\frac{1}{7}\left(e^{-0.075t^2}-e^{-0.25t^2}\right)$$

Hence the probability that the process has visited neither state 2 nor state 4 by time t is:

$$e^{-0.075t^2}+\frac{1}{7}\left(e^{-0.075t^2}-e^{-0.25t^2}\right)=\frac{8}{7}e^{-0.075t^2}-\frac{1}{7}e^{-0.25t^2}$$

(v) ***Limiting value***

As $t\to\infty$, the probability in (iv) tends to 0.

This must be the case for this particular model because eventually the process will end up in state 4, which is an absorbing state. In other words the probability of visiting state 4 by time t tends to 1 as $t\to\infty$. So the probability of not having visited state 4 tends to 0.

5.7 *This is Subject CT4, September 2005, Question A7.*

(i) ***Probability that the process remains in state A until at least time s***

The probability that the process remains in state A until at least time s, given that it started in state A at time 0, is:

$$p_{\overline{AA}}(0,s) = \exp\left(-\int_0^s \mu_{AB}(t)\,dt\right) = \exp\left(-\int_0^s 2t\,dt\right) = \exp\left(\left[-t^2\right]_0^s\right) = e^{-s^2} \qquad [2]$$

(ii) ***Proof***

The probability that the process is in state B at time T and that it is in the first visit to state B can be expressed in integral form as follows:

$$\int_0^T p_{\overline{AA}}(0,s)\ \mu_{AB}(s)\ p_{\overline{BB}}(s,T)\,ds \qquad [1]$$

This expression is constructed using the following reasoning:

- *Pick a point in time between 0 and T, call it s, and assume that the process stays in state A up to time s. This gives us the factor $p_{\overline{AA}}(0,s)$.*
- *Now suppose that there is a transition from state A to state B at time s. This gives us the factor $\mu_{AB}(s)$.*
- *Then we need the process to stay in state B from time s to time T. This gives us the factor $p_{\overline{BB}}(s,T)$.*
- *Finally, we integrate over all possible times s when the first transition could happen, ie from $s=0$ up to T.*

Now, from part (i) we know that:

$$p_{\overline{AA}}(0,s) = e^{-s^2}$$

Also:

$$\mu_{AB}(s) = 2s$$

and:

$$p_{\overline{BB}}(s,T) = \exp\left(-\int_s^T \mu_{BA}(t)\,dt\right) = \exp\left(-\int_s^T 2t\,dt\right) = \exp\left(\left[-t^2\right]_s^T\right) = e^{-(T^2-s^2)} \qquad [1]$$

So the probability that the process is in state B at time T, and it is in the first visit to state B, is:

$$\int_0^T e^{-s^2}\, 2s\, e^{-(T^2-s^2)}\,ds = e^{-T^2}\int_0^T 2s\,ds = e^{-T^2}\left[s^2\right]_0^T = T^2 e^{-T^2} \qquad [1]$$

as required.

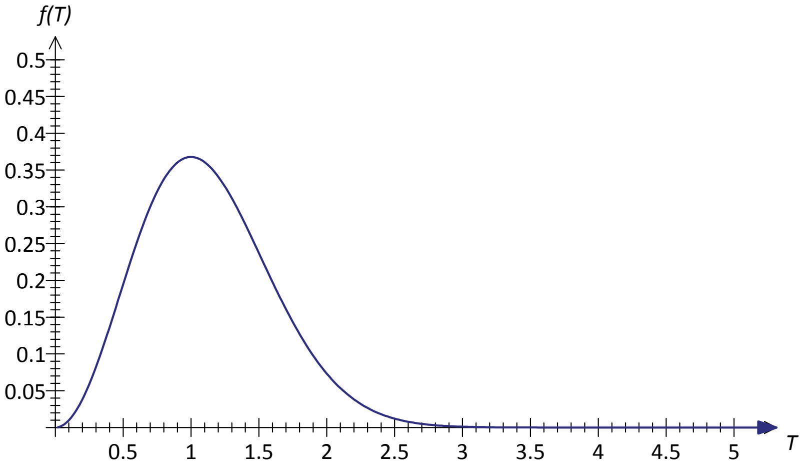
(iii)(a) ***Sketch of the probability function***

The function $f(T)=T^2e^{-T^2}$ *will tend to 0 as* $T\to\infty$ *because the exponential term will dominate the polynomial term. Also* $f(0)=0$. *Differentiating* f *we get:*

$$f'(T)=2Te^{-T^2}-2T^3e^{-T^2}=2Te^{-T^2}\left(1-T^2\right)$$

This derivative is equal to 0 when $T=1$. *(We are only considering positive values of* T *here.) These calculations should help you to sketch the graph.*

The function $f(T)=T^2e^{-T^2}$ is shown below:



[3]

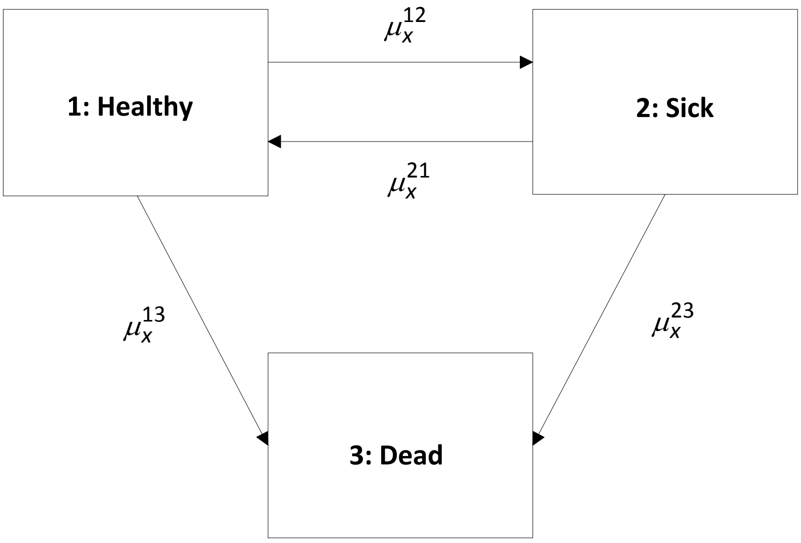
(iii)(b) ***Explanation of the shape of the probability function***

The graph increases at first, due to the increasing force of transition out of state A. It then reaches a peak and starts to decrease because the increasing force of transition out of state B means that the process is less likely to still be in its first visit to state B. [2]

(iii)(c) ***Time at which it is most likely that the process is in its first visit to state B***

From our calculations in part (iii)(a), and the graph above, we see that the time at which it is most likely that the process is in its first visit to state B is time 1. [1]

5.8 (i) ***Diagram of three-state model***



[2]

(ii) ***Derivation of partial differential equation***

Consider the interval from age x to age $x+t+h$, where h is a small amount. By the Markov property, we have:

$$ {}_{t+h}p_x^{12} = {}_tp_x^{11}\,{}_hp_{x+t}^{12} + {}_tp_x^{12}\,{}_hp_{x+t}^{22} $$ [1]

However, using the assumption about the transition rates, we can write:

$$ {}_hp_{x+t}^{12} = h\mu_{x+t}^{12} + o(h) $$ [1]

and:

$$ {}_hp_{x+t}^{22} = 1 - {}_hp_{x+t}^{21} - {}_hp_{x+t}^{23} = 1 - h\mu_{x+t}^{21} - h\mu_{x+t}^{23} + o(h) $$ [1]

So:

$$ {}_{t+h}p_x^{12} = {}_tp_x^{11}\,h\mu_{x+t}^{12} + {}_tp_x^{12}\left(1 - h\mu_{x+t}^{21} - h\mu_{x+t}^{23}\right) + o(h) $$ [1]

We can rearrange this equation to get:

$$ \frac{{}_{t+h}p_x^{12} - {}_tp_x^{12}}{h} = {}_tp_x^{11}\,\mu_{x+t}^{12} - {}_tp_x^{12}\left(\mu_{x+t}^{21} + \mu_{x+t}^{23}\right) + \frac{o(h)}{h} $$ [1]

Finally, letting $h \to 0$ gives:

$$ \frac{\partial}{\partial t}\,{}_tp_x^{12} = {}_tp_x^{11}\,\mu_{x+t}^{12} - {}_tp_x^{12}\left(\mu_{x+t}^{21} + \mu_{x+t}^{23}\right) $$

since $\frac{o(h)}{h} \to 0$ as $h \to 0$. [1]

5.9 (i) ***Differential equation***

Using the Markov assumption, which says that the probabilities of being found in any state at any future age depend only on the ages involved and the current state occupied … [½]

… we can write:

$$_{t+h}p_x^{23} = {}_tp_x^{21}\,{}_hp_{x+t}^{13} + {}_tp_x^{22}\,{}_hp_{x+t}^{23} + {}_tp_x^{23}\,{}_hp_{x+t}^{33} + {}_tp_x^{24}\,{}_hp_{x+t}^{43}$$ [1]

According to the law of total probability: ${}_hp_{x+t}^{33} = 1 - {}_hp_{x+t}^{34}$ [½]

Assuming that, for $i \neq j$ and small h, ${}_hp_{x+t}^{ij} = h\mu_{x+t}^{ij} + o(h)$ [½]

… where $\lim\limits_{h\to 0+} \dfrac{o(h)}{h} = 0$ [½]

… and noting that ${}_tp_x^{43} = 0$ [½]

… we have ${}_{t+h}p_x^{23} = {}_tp_x^{21}\,h\mu_{x+t}^{13} + {}_tp_x^{22}\,h\mu_{x+t}^{23} + {}_tp_x^{23}\left(1 - h\mu_{x+t}^{34}\right) + o(h)$. [½]

So:

$$\frac{{}_{t+h}p_x^{23} - {}_tp_x^{23}}{h} = {}_tp_x^{21}\,\mu_{x+t}^{13} + {}_tp_x^{22}\,\mu_{x+t}^{23} - {}_tp_x^{23}\,\mu_{x+t}^{34} + \frac{o(h)}{h}$$ [½]

and:

$$\frac{\partial}{\partial t}\,{}_tp_x^{23} = \lim_{h\to 0+} \frac{{}_{t+h}p_x^{23} - {}_tp_x^{23}}{h} = {}_tp_x^{21}\,\mu_{x+t}^{13} + {}_tp_x^{22}\,\mu_{x+t}^{23} - {}_tp_x^{23}\,\mu_{x+t}^{34}$$ [½]

(ii) ***Calculate probability***

$${}_2p_{40}^{13} = \left({}_1p_{40}^{11}\right)\left({}_1p_{41}^{13}\right) + \left({}_1p_{40}^{12}\right)\left({}_1p_{41}^{23}\right) + \left({}_1p_{40}^{13}\right)\left({}_1p_{41}^{33}\right)$$ [1]

$$= (1 - 0.03 - 0.002 - 0.001) \times 0.002 + 0.03 \times 0.1 + 0.002 \times (1 - 0.3) = 0.006334$$ [1]

(iii) ***Simplified model?***

The transition from permanently sick and disabled to dead is not necessary for the modelling of the claims process, because this transition has no effect on the incidence of any claim payments. (A lump sum claim is payable only on transitions from state 1 to 3, 1 to 4, 2 to 3 or 2 to 4; no payment is made on transition from state 3 to 4.) [1]

The revised differential equation is:

$$\frac{\partial}{\partial t}\,{}_tp_x^{23} = {}_tp_x^{21}\,\mu_{x+t}^{13} + {}_tp_x^{22}\,\mu_{x+t}^{23}$$ [1]

6

Survival models

Syllabus objectives

4.1 Explain the concept of survival models.

4.1.1 Describe the model of lifetime or failure time from age x as a random variable.

4.1.2 State the consistency condition between the random variable representing lifetimes from different ages.

4.1.3 Define the distribution and density functions of the random future lifetime, the survival function, the force of mortality or hazard rate, and derive relationships between them.

4.1.4 Define the actuarial symbols ${}_tp_x$ and ${}_tq_x$ and derive integral formulae for them.

4.1.5 State the Gompertz and Makeham laws of mortality.

4.1.6 Define the curtate future lifetime from age x and state its probability function.

4.1.7 Define the symbols e_x and $\overset{\circ}{e}_x$ and derive an approximate relation between them. Define the expected value and variance of the complete and curtate future lifetimes and derive expressions for them.

0 Introduction

In this chapter we will discuss a model of random lifetimes where we treat the future lifetime of an individual as a continuous random variable. From this simple starting point we will derive many useful results that are the building blocks of actuarial work relating to human mortality.

Although we will mostly study the lifetime model in the context of human mortality, the theory can equally be applied to other problems, such as:

- analysing the lengths of time that surviving individuals hold insurance policies – here mortality is replaced by 'withdrawal'.
- analysing the lengths of time that surviving individuals remain healthy – here mortality is replaced by 'sickness'.

1 A simple model of survival

1.1 Future lifetime

The starting point for a simple mathematical model of survival is the observation that the future lifetime of a person (called a 'life' in actuarial work) is not known in advance. Further, we observe that lifetimes range from 0 to in excess of 100 years. A natural assumption therefore is that the future lifetime of a given life is a random variable.

Assumption

The future lifetime of a new-born person is a random variable, denoted T, which is continuously distributed on an interval $[0,\omega]$ where $0<\omega<\infty$.

The maximum age ω is called the *limiting age*.

Typical values of ω for practical work are in the range 100–120. The possibility of survival beyond age ω is excluded by the model for convenience and simplicity.

When Jeanne Calment died in France on 4 August 1997, she was 122 years and 164 days old. According to the *Guinness Book of Records*, this is the highest authenticated age ever recorded.

Centenarians surviving beyond their 113th year are extremely rare.

Distribution function and survival function of a new-born life

$F(t)=P[T\le t]$ **is the distribution function of T.**

$S(t)=P[T>t]=1-F(t)$ **is the survival function of T.**

$S(t)$ is known as the survival function of T because it represents the probability of a new-born person surviving to age t.

In insurance contexts, we will not be dealing with new-born babies, so we need to extend the notation to deal with older individuals.

We often need to deal with ages greater than zero. To meet this need, we define T_x to be the future lifetime after age x, of a life who survives to age x, for $0\le x\le\omega$. Note that $T_0=T$.

Distribution function and survival function of a life aged *x*

For $0\le x\le\omega$:

$F_x(t)=P[T_x\le t]$ **is the distribution function of T_x**

$S_x(t)=P[T_x>t]=1-F_x(t)$ **is the survival function of T_x**

For example, the probability that a 40-year old dies before reaching age 100 is given by $F_{40}(60)$.

Question

Explain what $S_{29}(36)$ represents.

Solution

$S_{29}(36)$ represents the probability of an individual currently aged 29 reaching age 65, *ie* living for at least another 36 years.

For consistency with T, the distribution function of the random variable T_x $(0 \le x \le \omega)$ must satisfy the following relationships:

$$\boldsymbol{F_x(t) = P[\,T_x \le t\,] = P[\,T \le x + t \mid T > x\,] = \frac{F(x+t) - F(x)}{S(x)}}$$

This expression comes from the definition of conditional probabilities. $P(A|B)$ represents the probability of event A given that event B has occurred and:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

So:

$$P(T \le x+t \,|\, T > x) = \frac{P(x < T \le x+t)}{P(T > x)} = \frac{F(x+t) - F(x)}{S(x)}$$

1.2 Probabilities of death and survival

We now introduce the notation used by actuaries for probabilities of death and survival.

Actuarial notation for survival and death probabilities

$$\boldsymbol{{}_tq_x = F_x(t)}$$

$$\boldsymbol{{}_tp_x = 1 - {}_tq_x = S_x(t)}$$

So, ${}_{60}q_{40}$ represents the probability that a 40-year old dies before reaching age 100, and ${}_5p_{37}$ represents the probability that a 37-year old lives for at least another 5 years.

Question

Explain which of ${}_5p_{34}$ and ${}_7p_{33}$ is larger.

Solution

The probability of surviving from age 33 to 40 must be less than the probability of surviving from 34 to 39 since the first survival period includes the second, as well as the additional risk of dying between ages 33 and 34 and between ages 39 and 40. Hence ${}_5p_{34} > {}_7p_{33}$.

It is convenient in much actuarial work to use a time unit of one year. When this is the case, so that $t = 1$, we omit the '*t*' from these probabilities. That is, we define:

$$q_x = {}_1q_x \qquad \text{and} \qquad p_x = {}_1p_x$$

q_x and ${}_tq_x$ are called *rates of mortality*.

So:

${}_tq_x$ is the probability that a life now aged *x* dies within *t* years

q_x is the probability that a life now aged *x* dies within 1 year

${}_tp_x$ is the probability that a life now aged *x* is still alive after *t* years

p_x is the probability that a life now aged *x* is still alive after 1 year

These are actually probabilities. Do not confuse them with transition rates.

1.3 The force of mortality μ_x

A quantity that plays a central role in a survival model is the *force of mortality* (which is more widely known as the hazard rate in statistics).

We denote the force of mortality at age x $(0 \le x < \omega)$ by μ_x , and define it as:

$$\mu_x = \lim_{h \to 0^+} \frac{1}{h} \times P[T \le x + h \,|\, T > x]$$

We will always suppose that the limit exists.

The interpretation of μ_x is very important.

The force of mortality μ_x is an *instantaneous* measure of mortality at age *x*. It is the continuous equivalent of the discrete quantity q_x .

The probability $P[\,T \le x + h \,|\, T > x\,]$ is (from the definitions above) $F_x(h) = {}_hq_x$.

For small h , we can ignore the limit and write:

$${}_hq_x \approx h.\mu_x$$

In other words, the probability of death in a *short* time h after age x is roughly proportional to h , the constant of proportionality being μ_x .

The intuitive way of thinking of the force of mortality is in terms of the expected number of deaths in a very large population. The expected number of deaths during a short time interval of length h years in a very large population consisting of n individuals aged exactly x is $n \times \mu_x \times h$.

We could estimate the value of μ_{50} by taking a very large group of people, all aged exactly 50, and counting how many died during the next hour. We could then work out the proportion of the group that had died, and express this as an annual rate by multiplying by 24×365. This figure would give the value of μ_{50} (very nearly).

The figure would not be exact for a number of reasons, *eg*:

1. The actual number of deaths we observe will differ from the expected number because of statistical fluctuations and the fact that people die in 'whole units'.
2. We have ignored leap years. Assuming 365.25 days in a year would give a more 'accurate' answer.
3. We have used a period of 1 hour. The force of mortality is an *instantaneous* measure, so we need to take the limit of 1 hour, 1 minute, 1 second …

As well as these theoretical reasons, there are practical reasons why we could not do this. For example, there are only around 2,000 babies born each day in the whole of the UK. So, if we take 'a life aged exactly 50' to mean a life whose 50th birthday is on the day in question, we will have somewhat fewer than 2,000 people in our group (because some people will have died before age 50). Since this is a relatively small number of people, it is very unlikely that any of them at all would die during the next hour.

We could actually have defined the force of mortality in two ways: either by thinking in terms of a new-born baby (as we did at the start of this section) or by thinking in terms of a person who has already reached the age in question.

Equivalent definitions of force of mortality

For $x \geq 0$ and $t > 0$, we could define the force of mortality μ_{x+t} in two ways:

(1) $$\mu_{x+t} = \lim_{h \to 0^+} \frac{1}{h} \times P[\,T \leq x+t+h \mid T > x+t\,]$$

(2) $$\mu_{x+t} = \lim_{h \to 0^+} \frac{1}{h} \times P[\,T_x \leq t+h \mid T_x > t\,]$$

It is an easy exercise to show from the definitions that these are equal. We will often use μ_{x+t} for a fixed age x and $0 \leq t < \omega - x$.

1.4 Survival probabilities

The definition of $S_x(t)$ leads to an important relationship:

$$S_x(t) = P[\,T_x > t\,] = P[\,T > x+t \mid T > x\,] = \frac{P[\,T > x+t\,]}{P[\,T > x\,]} = \frac{S(x+t)}{S(x)}$$

This can be expressed in actuarial notation as:

$$ {}_t p_x = \frac{{}_{x+t}p_0}{{}_x p_0} $$

Therefore, for any age x and for $s > 0, t > 0$:

$$ {}_{s+t}p_x = \frac{{}_{x+s+t}p_0}{{}_x p_0} = \frac{{}_{x+s}p_0}{{}_x p_0} \times \frac{{}_{x+s+t}p_0}{{}_{x+s}p_0} = {}_s p_x \times {}_t p_{x+s} $$

Similarly,

$$ {}_{s+t}p_x = {}_t p_x \times {}_s p_{x+t} $$

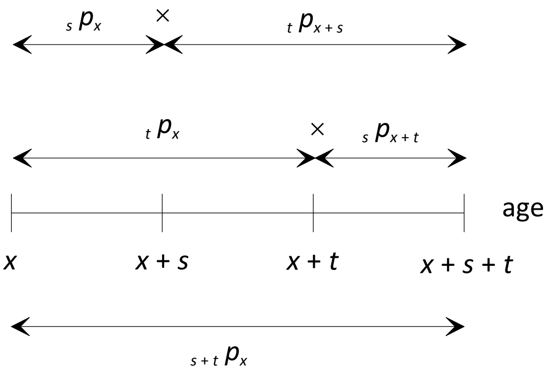
In words, the probability of surviving for time $(s+t)$ after age x is given by multiplying:

1. **the probability of surviving for time s, and**
2. **the probability of then surviving for a further time t**

or by multiplying:

1. **the probability of surviving for time t, and**
2. **the probability of *then* surviving for a further time s.**

This is illustrated below:



The order in which we consider the two periods is irrelevant.

This is the consistency condition referred to in syllabus objective 4.1.2.

1.5 The probability density function of T_x

The distribution function of T_x is $F_x(t)$, by definition. We also want to know its probability density function (PDF).

Denote this by $f_x(t)$, and recall that:

$$f_x(t) = \frac{d}{dt} F_x(t)$$

Then:

$$\begin{aligned}
f_x(t) &= \frac{d}{dt} P[T_x \le t] \\
&= \lim_{h \to 0^+} \frac{1}{h} \times (P[T_x \le t+h] - P[T_x \le t]) \\
&= \lim_{h \to 0^+} \frac{P[T \le x+t+h \mid T > x] - P[T \le x+t \mid T > x]}{h} \\
&= \lim_{h \to 0^+} \frac{P[T \le x+t+h] - P[T \le x] - (P[T \le x+t] - P[T \le x])}{S(x) \times h} \\
&= \lim_{h \to 0^+} \frac{P[T \le x+t+h] - P[T \le x+t]}{S(x) \times h}
\end{aligned}$$

Now multiply and divide by $S(x+t)$ and we have:

$$\begin{aligned}
f_x(t) &= \frac{S(x+t)}{S(x)} \times \lim_{h \to 0^+} \frac{1}{h} \frac{P[T \le x+t+h] - P[T \le x+t]}{S(x+t)} \\
&= S_x(t) \times \lim_{h \to 0^+} \frac{1}{h} P[T \le x+t+h \mid T > x+t] \\
&= S_x(t) \times \mu_{x+t}
\end{aligned}$$

or, in actuarial notation, for a fixed age x between 0 and ω:

$$f_x(t) = {}_t p_x \mu_{x+t} \qquad (0 \le t < \omega - x)$$

This is one of the most important results concerning survival models.

Let's summarise the model we have introduced.

Summary of model

T_x is the (random) future lifetime after age x.

It is, by assumption, a continuous random variable taking values in $[0, \omega - x]$.

Its distribution function is $F_x(t) = {}_tq_x$.

Its probability density function is $f_x(t) = {}_tp_x \mu_{x+t}$.

The force of mortality is interpreted by the approximate relationship:

$${}_hq_x \approx h.\mu_x \text{ (for small } h\text{)}$$

The survival functions $S_x(t)$ or ${}_tp_x$ satisfy the relationship:

$${}_{s+t}p_x = {}_sp_x \times {}_tp_{x+s} = {}_tp_x \times {}_sp_{x+t} \qquad \text{for any } s > 0, t > 0$$

1.6 Life table functions

A life table is a table showing the expected number that will survive to each age in a hypothetical group of lives. For the English Life Tables No15 (Males) given on pages 68 and 69 of the *Tables*, the table starts at age 0 with 100,000 lives. l_x denotes the expected number of lives at age x and d_x denotes the expected number of deaths between the ages of x and $x+1$. l_x and d_x can be used to calculate survival and death probabilities as follows:

$$d_x = l_x - l_{x+1}$$

$$p_x = \frac{l_{x+1}}{l_x} \qquad q_x = 1 - p_x = 1 - \frac{l_{x+1}}{l_x} = \frac{l_x - l_{x+1}}{l_x} = \frac{d_x}{l_x}$$

$${}_tp_x = \frac{l_{x+t}}{l_x} \qquad {}_tq_x = 1 - {}_tp_x = 1 - \frac{l_{x+t}}{l_x} = \frac{l_x - l_{x+t}}{l_x}$$

Values are tabulated only for integer ages. If we require a value at a non-integer age, we must make an assumption about how mortality varies between integer ages. For example, we could assume that deaths occur uniformly between integer ages or that the force of mortality is constant between integer ages.

Question

Below is an extract from English Life Table No15 (Males):

Age, x	l_x
58	88,792
59	87,805

Estimate $l_{58.25}$ assuming a uniform distribution of deaths between exact ages 58 and 59.

Solution

There are:

$$88,792 - 87,805 = 987$$

deaths expected between the ages of 58 and 59. Assuming that these are uniformly distributed throughout the year of age, the number of deaths expected between the ages of 58 and 58.25 is:

$$\frac{987}{4} = 246.75$$

So the expected number of lives at age 58.25 is:

$$88,792 - 246.75 = 88,545.25$$

Although the same number of people are dying each quarter, under the uniform distribution of deaths assumption, the surviving population at the start of each quarter is decreasing. So this assumption implies that the force of mortality is increasing over the year of age $(58,59)$.

The following formula is also useful.

Uniform distribution of deaths assumption

If deaths are uniformly distributed between the ages of x and $x+1$, it follows that:

$${}_tq_x = t\,q_x$$

for $0 \le t \le 1$.

This result can be proved as follows. Assuming that deaths are uniformly distributed between exact ages x and $x+1$, we have (by linear interpolation):

$$l_{x+t} = (1-t)l_x + t\,l_{x+1}$$

for $0 \le t \le 1$.

So:

$${}_tq_x = 1 - \frac{l_{x+t}}{l_x} = 1 - \frac{(1-t)l_x + t\,l_{x+1}}{l_x} = \frac{t\,l_x - t\,l_{x+1}}{l_x} = t(1-p_x) = t\,q_x$$

1.7 Initial and central rates of mortality

q_x is called an *initial* rate of mortality, because it is the probability that a life alive at exact age x (the initial time) dies before exact age $x+1$.

Since $q_x = \dfrac{d_x}{l_x}$, it is the number of deaths over the year of age x to $x+1$ divided by the number alive at the start of that year.

An alternative often used (especially in demography) is the *central* rate of mortality, denoted m_x.

Central rate of mortality

$$\boldsymbol{m_x = \frac{q_x}{\int_0^1 {}_tp_x \, dt}}$$

Another formula for m_x is given on page 121 of the *Tables*:

$$m_x = \frac{d_x}{\int_0^1 l_{x+t} \, dt}$$

This formula shows that the central rate of mortality at age x can be thought of as the number of deaths over the year of age x to $x+1$ divided by the average number of lives alive over the year of age x and $x+1$.

When we think about averaging, we usually have in mind an expression of the form:

$$\frac{1}{n}\sum_{i=1}^{n}$$

ie we sum n terms and divide by n. This is averaging in a discrete sense.

Similarly, the average of a continuous function, $g(t)$ say, over the interval $t=0$ to $t=n$ is:

$$\frac{1}{n}\int_0^n g(t)\,dt$$

Setting $n=1$, we see that the average of value of $g(t)$ over the interval $t=0$ to $t=1$ is:

$$\int_0^1 g(t)\,dt$$

So the average value of l_{x+t} over the year of age x to $x+1$ is:

$$\int_0^1 l_{x+t}\,dt$$

Dividing the numerator and denominator of $\dfrac{d_x}{\int_0^1 l_{x+t}\,dt}$ by l_x gives:

$$m_x = \frac{d_x / l_x}{\int_0^1 l_{x+t} / l_x \, dt} = \frac{q_x}{\int_0^1 {}_tp_x \, dt}$$

as before.

The quantity m_x is the probability of dying between exact ages x and $x+1$ per person-year lived between exact ages x and $x+1$; the denominator $\int_0^1 {}_tp_x \, dt$ is interpreted as the expected amount of time spent alive between ages x and $x+1$ by a life alive at age x, and the numerator is the probability of that life dying between exact ages x and $x+1$.

There is another interpretation of m_x. The probability q_x can be represented by the integral $q_x = \int_0^1 {}_tp_x \, \mu_{x+t} \, dt$, which means that:

$$m_x = \frac{\int_0^1 {}_tp_x \, \mu_{x+t} \, dt}{\int_0^1 {}_tp_x \, dt}$$

So m_x is a weighted average of the force of mortality over the next year of age. The weighting factors are the survival probabilities. m_x is a measure of the rate of mortality over the year from exact age x to exact age $x+1$, whereas the force of mortality μ_x is a measure of the *instantaneous* rate of mortality at exact age x.

m_x is useful when the aim is to project numbers of deaths, given the number of lives alive in age groups; this is one of the basic components of a population projection. In practice the age groups used in population projection are often broader than one year, so the definition of m_x has to be suitably adjusted.

In this course, we consider how to estimate the mortality of a particular population using data from an investigation. Historically, actuaries tended to use the data to estimate m_x rather than μ_x or q_x.

Historically, m_x was estimated by statistics of the form:

$$\frac{\textbf{Number of deaths}}{\textbf{Total time spent alive and at risk}}$$

called ‘occurrence-exposure rates’.

More recently, these statistics have been used to estimate the force of mortality rather than m_x, because in that context they have a solid basis in terms of a probabilistic model. However, if μ_{x+t} is a constant, μ, between ages x and $x+1$, then:

$$m_x = \frac{q_x}{\int_0^1 {}_tp_x \, dt} = \frac{\int_0^1 {}_tp_x \, \mu \, dt}{\int_0^1 {}_tp_x \, dt} = \mu$$

So there is still a close connection.

It is important to understand the relationship between these three measures of mortality.

Question

Consider the statement 'm_x can never be less than q_x.' Explain whether this is true or false.

Solution

The denominator $\int_0^1 {}_tp_x \, dt \le 1$, so $m_x \ge q_x$. So the statement is true.

2 Expected future lifetime

2.1 Complete expectation of life

The expected future lifetime after age x, which is referred to by demographers as the expectation of life at age x, is defined as $E[T_x]$. It is denoted $\mathring{e}_x$.

The symbol $\mathring{e}_x$ is read as '*e*-circle-*x*' and is tabulated in some of the actuarial tables we will be using.

The next bit of derivation of the formula for $\mathring{e}_x$ uses the following result, which we will see again in Section 3.3:

$$ {}_tp_x\,\mu_{x+t} = f_x(t) = \frac{\partial}{\partial t}F_x(t) = \frac{\partial}{\partial t}\,{}_tq_x = \frac{\partial}{\partial t}(1-{}_tp_x) = -\frac{\partial}{\partial t}\,{}_tp_x $$

It also uses the definition of the expected value of a continuous random variable:

$$ E(Y) = \int_y y\,f(y)\,dy $$

where $f(y)$ is the PDF of the random variable Y.

By definition:

$$ \begin{aligned} \mathring{e}_x &= \int_0^{\omega-x} t.{}_tp_x\mu_{x+t}dt \\ &= \int_0^{\omega-x} t\,(-\frac{\partial}{\partial t}\,{}_tp_x)\,dt \\ &= -\left[t\,{}_tp_x\right]_0^{\omega-x} + \int_0^{\omega-x} {}_tp_x\,dt \qquad \textbf{(integrating by parts)} \\ &= \int_0^{\omega-x} {}_tp_x\,dt \end{aligned} $$

since the term in square brackets is zero for both $t=0$ and $t=\omega-x$.

The formula for $\mathring{e}_x$ is often written more simply as follows.

Complete expectation of life

$$ \mathring{e}_x = E(T_x) = \int_0^{\infty} {}_tp_x\,dt $$

This formula holds since ${}_tp_x = 0$ for $t > \omega - x$.

According to the ELT15 (Males) life table, the complete expectations of life for a new-born baby boy, a 21-year-old male and a 70-year-old male are 73.413 years, 53.497 years and 11.187 years, respectively.

The figures illustrate the fact that $\overset{\circ}{e}_0 \neq 21 + \overset{\circ}{e}_{21}$. For equality, the probability of dying before age 21 would have to be zero. Although this probability is quite low, it is greater than zero. So $\overset{\circ}{e}_0 < 21 + \overset{\circ}{e}_{21}$.

2.2 Curtate expectation of life

To define the *curtate expectation of life*, we first need to define K_x, the *curtate future lifetime* of a life age x.

Curtate future lifetime random variable

The curtate future lifetime of a life age x is:

$$K_x = [T_x]$$

where the square brackets denote the integer part. In words, K_x is equal to T_x rounded down to the integer below.

So, the curtate future lifetime K_x of a life aged exactly x is the whole number of years lived after age x.

Clearly K_x is a discrete random variable, taking values on the integers $0, 1, 2, \ldots [\omega - x]$.

The probability distribution of K_x is easy to write down using the definitions of Section 1 of this chapter.

$$\begin{aligned} P[K_x = k] &= P[k \le T_x < k+1] \\ &= P[k < T_x \le k+1] \qquad (*) \\ &= {}_k p_x \, q_{x+k} \end{aligned}$$

We also use the symbol ${}_{k|}q_x$ to represent $P(K_x = k)$. It is read as 'k deferred q_x', and we can think about this as deferring the event of death until the year that begins in k years from now.

Note that switching the inequalities at step (*) requires an assumption about T_x. It is enough to suppose that $F_x(t)$ is continuous in t. We will not discuss this further here.

We now define the *curtate expectation of life*, denoted e_x, by:

$$e_x = E[K_x]$$

Then:

$$\begin{aligned} e_x &= \sum_{k=0}^{[\omega-x]} k.{}_k p_x.q_{x+k} \\ &= {}_1 p_x.q_{x+1} \\ &\quad + {}_2 p_x.q_{x+2} + {}_2 p_x.q_{x+2} \\ &\quad + {}_3 p_x.q_{x+3} + {}_3 p_x.q_{x+3} + {}_3 p_x.q_{x+3} \\ &\quad + \cdots \end{aligned}$$

$$= \sum_{k=1}^{[\omega-x]} \sum_{j=k}^{[\omega-x]} {}_j p_x.q_{x+j} \qquad \textbf{(summing columns)}$$

$$= \sum_{k=1}^{[\omega-x]} {}_k p_x$$

The last step follows since:

$$\sum_{j=k}^{[\omega-x]} {}_j p_x \, q_{x+j} = \sum_{j=k}^{[\omega-x]} P(K_x = j) = P(K_x \geq k)$$

This is the probability of dying at *any* time after age $x+k$, which is the same as ${}_k p_x$.

The formula for e_x is often written more simply as follows.

Curtate expectation of life

$$e_x = E(K_x) = \sum_{k=1}^{\infty} {}_k p_x$$

This formula holds since ${}_k p_x = 0$ for $k > [\omega - x]$.

Question

Show algebraically that $e_x = p_x(1+e_{x+1})$.

Solution

We have:

$$\begin{aligned} e_x &= p_x + {}_2 p_x + {}_3 p_x + \cdots \\ &= p_x \left(1 + p_{x+1} + {}_2 p_{x+1} + \cdots\right) \\ &= p_x(1+e_{x+1}) \end{aligned}$$

Intuitively, this is saying that the life expectancy for a life now aged x is one year more than the life expectancy when the life reaches age $x+1$, provided that the life does survive to age $x+1$.

According to the AM92 ultimate mortality table, the curtate expectation of life for a 21-year-old male is 57.481 years and the curtate expectation of life for a 70-year-old male is 13.023 years.

2.3 The relationship between the complete and curtate expectations of life

We have two simple formulae:

$$\mathring{e}_x = \int_0^{\omega-x} {}_t p_x \, dt$$

$$e_x = \sum_{k=1}^{[\omega-x]} {}_k p_x$$

The complete and curtate expectations of life are related by the approximate equation:

$$\mathring{e}_x \approx e_x + ½$$

To see this, define $J_x = T_x - K_x$ to be the random lifetime after the highest *integer* age to which a life age *x* survives.

Approximately, $E[J_x] = ½$, but $E[T_x] = E[K_x] + E[J_x]$ so $\mathring{e}_x \approx e_x + ½$ as stated.

When stating that $E[J_x] = ½$, we are assuming that deaths occur half way between birthdays, on average.

Question

Using ELT15 (Males) mortality, approximate the curtate expectation of life for:

(a) a new-born baby

(b) a 21-year-old actuarial student

(c) a 70-year-old pensioner.

Solution

(a) $e_0 \cong \mathring{e}_0 - 0.5 = 72.913$ years

(b) $e_{21} \cong \mathring{e}_{21} - 0.5 = 52.997$ years

(c) $e_{70} \cong \mathring{e}_{70} - 0.5 = 10.687$ years

2.4 Future lifetimes – variance

It is easy to write down the variances of the complete and curtate future lifetimes:

$$\text{var}[T_x] = \int_0^{\omega-x} t^2 \, {}_tp_x \, \mu_{x+t} \, dt - \mathring{e}_x^2$$

$$\text{var}[K_x] = \sum_{k=0}^{[\omega-x]} k^2 \, {}_kp_x \, q_{x+k} - e_x^2$$

but these do not simplify neatly as the expected values do.

It is not particularly useful to know the variance of future lifetimes. However, it *is* useful to be able to find the variance of financial functions (*eg* the profits from a life insurance policy or the cost of providing a benefit from a pension scheme) based on future lifetimes. This information would enable us to quantify the likely variation in profits *etc*.

2.5 Uses of the expectation of life

The expectation of life is often used as a measure of the standard of living and health care in a given country.

Here are some examples of average life expectancy at birth in different countries (2015):

45-50	Chad
50-55	Afghanistan, Namibia, Nigeria
55-60	Angola, Sierra Leone, Zimbabwe
60-65	Cambodia, Kenya, South Africa
65-70	Myanmar, India, Pakistan
70-75	Bangladesh, Brazil, North Korea, Russia
75-80	Barbados, China, Denmark, Hungary, USA
80-85	Germany, Israel, South Korea, UK
85-90	Monaco

The data come from the CIA World Factbook.

3 Some important formulae

3.1 Introduction

In this section we give two important formulae, one for $_tq_x$ and one for $_tp_x$.

These formulae will provide a useful link between $_tq_x$, $_tp_x$ and μ_x.

3.2 A formula for $_tq_x$

The first follows from the result that $f_x(t) = {}_tp_x\mu_{x+t}$. We have:

$$_tq_x = F_x(t) = \int_0^t f_x(s)ds = \int_0^t {}_sp_x\mu_{x+s}ds$$

This formula is easy to interpret. For each time s, between 0 and t, the integrand is the product of:

(i) $_sp_x$, the probability of surviving to age $x+s$, and

(ii) $\mu_{x+s}\,ds$, which is approximately equal to $_{ds}q_{x+s}$, the probability of dying just after age $x+s$.

Since it is impossible to die at more than one time, we simply add up, or in the limit integrate, all these mutually exclusive probabilities.

This result is not usually used to calculate $_tq_x$ from $_tp_x$ and μ_x since if we knew $_tp_x$ we could calculate $_tq_x$ directly. However, the result does allow us to derive a very important relationship between $_tp_x$ and μ_x.

3.3 A formula for $_tp_x$

The formula for $_tp_x$ follows from the solution of the following equation:

$$\frac{\partial}{\partial s}{}_sp_x = -\frac{\partial}{\partial s}{}_sq_x = -f_x(s) = -\,{}_sp_x\mu_{x+s}$$

This is the Kolmogorov forward differential equation for $_sp_x$, which we met in Chapter 3.

(You will see why we have used s as the variable in a moment.)

To solve this, note that:

$$\frac{\partial}{\partial s}\log {}_sp_x = \frac{\frac{\partial}{\partial s}{}_sp_x}{{}_sp_x}$$

so that the above equation can be rewritten as:

$$\frac{\partial}{\partial s}\log {}_sp_x = -\mu_{x+s}$$

Here we are using the separation of variables technique, which we also met in Chapter 3.

We are using s rather than t as the variable so that we can integrate this relationship between the limits of 0 and t without causing confusion.

Hence:

$$\int_0^t \frac{\partial}{\partial s} \log {}_s p_x \, ds = -\int_0^t \mu_{x+s} \, ds + c$$

where *c* is some constant of integration.

Actually, we don't need to include the constant of integration here because we've put limits on both integrals. So c will turn out to be zero.

The left-hand side is:

$$\left[\log {}_s p_x\right]_0^t = \log {}_t p_x \qquad \text{(since } {}_0 p_x = 1\text{)}$$

so taking exponentials of both sides gives:

$${}_t p_x = \exp\left\{-\int_0^t \mu_{x+s} \, ds + c\right\}$$

Now since ${}_0 p_x = 1$, we must have $c = 0$ (since $e^0 = 1$)**, so finally:**

$${}_t p_x = \exp\left\{-\int_0^t \mu_{x+s} \, ds\right\}$$

3.4 Summary

To summarise, we have derived the following *very important* results.

Integral expressions

$${}_t q_x = \int_0^t {}_s p_x \, \mu_{x+s} \, ds \qquad \textbf{(6.1)}$$

$${}_t p_x = \exp\left\{-\int_0^t \mu_{x+s} \, ds\right\} \qquad \textbf{(6.2)}$$

4 Simple parametric survival models

Several survival models are in common use in which the random variable denoting future lifetime has a distribution expressed in terms of a small number of parameters. Perhaps the simplest is the exponential model, in which the hazard is constant:

$$\mu_x = \mu$$

It follows from (6.2) above that in the exponential model:

$$_tp_x = S_x(t) = \exp\left\{-\int_0^t \mu\, ds\right\} = \exp\left\{-\left[\mu s\right]_0^t\right\} = \exp(-\mu t)$$

and hence that:

$$_tq_x = 1 - {_tp_x} = 1 - \exp(-\mu t)$$

For example, if μ_x takes the constant value 0.001 between ages 25 and 35, then the probability that a life aged exactly 25 will survive to age 35 is:

$$_{10}p_{25} = \exp\left(-\int_0^{10} 0.001\, dt\right) = e^{-0.01} = 0.99005$$

We can use R to simulate values from an exponential distribution, plot its PDF, and calculate probabilities and percentiles.

Suppose we have an exponential distribution with parameter $\lambda = 0.5$. The R code for simulating 100 values is given by:

```
rexp(100,rate=0.5)
```

The PDF is obtained by `dexp(x, rate=0.5)` **and is useful for graphing. For example:**

```
plot(seq(0:5000),dexp(seq(0:5000), rate=0.5),type="l")
```

To calculate probabilities for a continuous distribution we use the CDF which is obtained by `pexp`**. For example, to calculate $P(X \le 2) = 0.6321206$ we use the R code:**

```
pexp(2,rate=0.5)
```

Similarly, the quantiles can be calculated with `qexp`**.**

A simple extension to the exponential model is the Weibull model, in which the survival function $S_x(t)$ is given by the two-parameter formula:

$$S_x(t) = \exp\left[-\alpha t^{\beta}\right] \tag{6.3}$$

Recall that $S_x(t) = 1 - F_x(t)$, where $F_x(t) = P\left(T_x \le t\right)$. The CDF of the Weibull distribution is given on page 15 of the *Tables.*

Since:

$$\mu_{x+t} = -\frac{\partial}{\partial t}\log[S_x(t)]$$

we see that:

$$\mu_{x+t} = -\frac{\partial}{\partial t}[-\alpha t^{\beta}] = -[-\alpha\beta t^{\beta-1}] = \alpha\beta t^{\beta-1}$$

Different values of the parameter β can give rise to a hazard that is monotonically increasing or monotonically decreasing as t increases, or in the specific case where $\beta = 1$, a hazard that is constant, since if $\beta = 1$:

$$\alpha\beta t^{\beta-1} = \alpha.1.t^{0} = \alpha$$

This can be seen also from the expression for $S_x(t)$ (6.3), from which it is clear that, when $\beta = 1$, the Weibull model is the same as the exponential model.

We can adjust the R code given above for an exponential distribution to calculate corresponding quantities for a Weibull distribution.

The R code for simulating a random sample of 100 values from the Weibull distribution with $c = 2$ and $\gamma = 0.25$ is:

```
rweibull(100, 0.25, 2^(-1/0.25))
```

R uses a different parameterisation for the scale parameter, *c*.

Similarly, the PDF, CDF and quantiles can be obtained using the R functions `dweibull`, `pweibull` **and** `qweibull`**.**

Alternatively, we could redefine them from first principles as follows:

```
rweibull <- function(n,c,g){
rp <- (log(1-runif(n))/c)^(1/g)
rp}

dweibull <- function(x,c,g){
c*g*x^(g-1)*exp(-c*x^g)}

pweibull <- function(q,c,g){
1-exp(-c*x^g)}

qweibull <- function(p,c,g){
q <- (log(1-p)/c)^(1/g)
q}
```

5 The Gompertz and Makeham laws of mortality

The Gompertz and Makeham laws of mortality are two further examples of parametric survival models. They can be expressed as follows:

Gompertz' Law: $\mu_x = Bc^x$ **(6.4)**

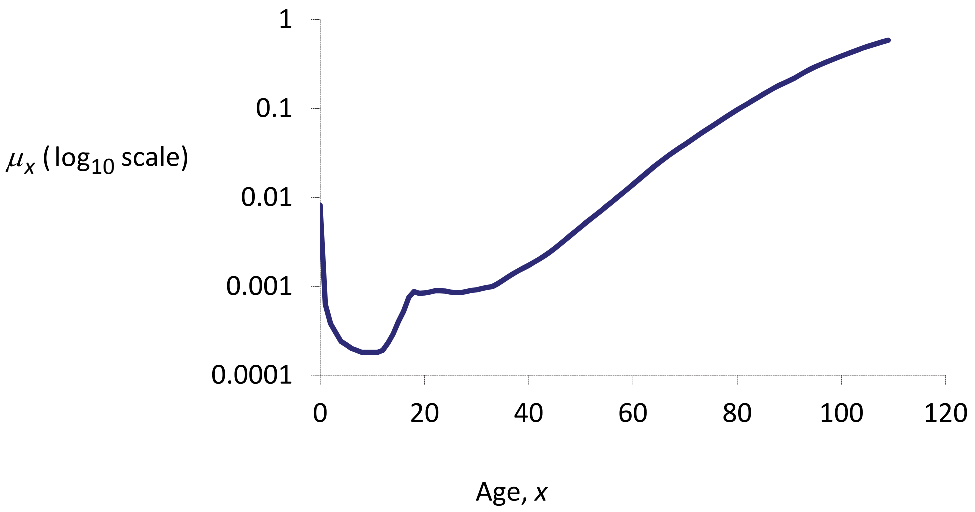
Makeham's Law: $\mu_x = A + Bc^x$

These formulae are given on page 32 of the *Tables*.

Gompertz' Law is an exponential function, and it is often a reasonable assumption for middle ages and older ages.

Makeham's Law incorporates a constant term, which is sometimes interpreted as an allowance for accidental deaths, not depending on age.

The rationale behind the laws is based on an observation made by Benjamin Gompertz in the early 1800s. When μ_x is plotted on a logarithmic scale against age, the graph often appears to follow a straight line for a large part of the age range. We can see this from the graph of μ_x in the diagram below (for ages above 35 or so).



μ_x on $\log_{10}$ scale (ELT15 (Males) Mortality Table)

5.1 Calculating the parameter values

If a life table is known to follow Gompertz' Law, the parameters B and c can be determined given the values of μ_x at any two ages. In the case of a life table following Makeham's Law, the parameters A, B and c can be determined given the values of μ_x at any three ages.

Question

For a force of mortality μ_x that is known to follow Gompertz' Law, calculate the parameters B and c if $\mu_{50} = 0.017609$ and $\mu_{55} = 0.028359$.

Solution

We have:

$$\frac{\mu_{55}}{\mu_{50}}=\frac{Bc^{55}}{Bc^{50}}=c^5 \Rightarrow c=\left(\frac{0.028359}{0.017609}\right)^{1/5}=1.1$$

and:

$$B=\frac{\mu_{50}}{c^{50}}=\frac{0.017609}{1.1^{50}}=0.00015$$

5.2 Survival probabilities

Survival probabilities $_tp_x$ can be found using:

$$_tp_x=\exp\left(-\int_0^t \mu_{x+s}\,ds\right)$$

Gompertz' Law

In the case of Gompertz' Law:

$$_tp_x=g^{c^x(c^t-1)}$$

where:

$$g=\exp\left(\frac{-B}{\log c}\right)$$

Here we are using 'log' to mean natural log, *ie* $\log_e$ or $\ln$.

This result can be derived as follows. Under Gompertz' Law, we have:

$$_tp_x=\exp\left(-\int_0^t \mu_{x+s}\,ds\right)=\exp\left(-\int_0^t Bc^{x+s}\,ds\right)$$

We can write c^{x+s} as $c^xe^{s\ln c}$, so:

$$\int_0^t Bc^{x+s}\,ds=\int_0^t Bc^xe^{s\ln c}\,ds=\frac{Bc^x}{\ln c}\Big[e^{s\ln c}\Big]_0^t=\frac{Bc^x}{\ln c}\Big[c^s\Big]_0^t=\frac{Bc^x}{\ln c}(c^t-1)$$

If we introduce the auxiliary parameter g defined by $\ln g=-B/\ln c$, then:

$$-\int_0^t Bc^{x+s}\,ds=(\ln g)\,c^x(c^t-1)$$

Hence:

$$ {}_tp_x = \exp\left[(\ln g)c^x(c^t-1)\right] = (e^{\ln g})^{c^x(c^t-1)} = g^{c^x(c^t-1)} $$

Question

A mortality table, which obeys Gompertz' Law for older ages, has:

$$\mu_{70} = 0.025330 \text{ and } \mu_{90} = 0.126255$$

Calculate the probability that a life aged 60 will survive for 20 years.

Solution

If the table follows Gompertz' Law then $\mu_x = Bc^x$ and:

$$\frac{\mu_{90}}{\mu_{70}} = \frac{Bc^{90}}{Bc^{70}} = c^{20} = \frac{0.126255}{0.025330}$$

$$\Rightarrow c = \left(\frac{0.126255}{0.025330}\right)^{\frac{1}{20}} = 1.083629$$

So:

$$B = 0.126255 \times (1.083629)^{-90} = 9.16196 \times 10^{-5}$$

$$g = \exp\left\{-\frac{B}{\ln c}\right\} = \exp\left\{-\frac{9.16196\times10^{-5}}{\ln 1.083629}\right\} = 0.998860$$

and:

$$ {}_{20}p_{60} = g^{c^{60}(c^{20}-1)} = (0.998860)^{493.4052} = 0.56958$$

Makeham's Law

In the case of Makeham's Law:

$$ {}_tp_x = s^t g^{c^x(c^t-1)} $$

where:

$$g = \exp\left(\frac{-B}{\log c}\right) \quad \text{and} \quad s = \exp(-A)$$

We will use these laws in Chapter 11, *Methods of graduation*.

Gompertz-Makeham family

More generally, we can model the force of mortality using one of the Gompertz-Makeham family of curves. This family consists of functions of the form:

$$\text{GM}(r,s)=\alpha_1+\alpha_2 t+\alpha_3 t^2+\cdots+\alpha_r t^{r-1}$$
$$+\exp\left\{\alpha_{r+1}+\alpha_{r+2} t+\alpha_{r+3} t^2+\cdots+\alpha_{r+s} t^{s-1}\right\}$$

where $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{r+s}$ are constants which do not depend on *t*.

This form of the Gompertz–Makeham family of curves is the one that is used most widely. However it does not match the form given on page 32 of the *Tables*. The form given in the *Tables* is:

$$\mu_x=\text{GM}(r,s)=poly_1(t)+\exp\left\{poly_2(t)\right\}$$

where t is a linear function of x and $poly_1(t)$ and $poly_2(t)$ are polynomials of degree r and s respectively.

The R base system does not have a command to simulate the Gompertz distribution.

In the package `flexsurv`**, the command** `rgompertz` **will simulate a Gompertz distribution. The commands** `dgompertz`**,** `pgompertz` **and** `qgompertz` **will generate the density, distribution function and quantiles respectively.**

The command `hgompertz` **generates the hazard, and** `Hgompertz` **the cumulative hazard.**

Note that in these commands the parameters of the Gompertz distribution are to be specified as 'shape' and 'rate'. If the shape is α and the rate is β, then Gompertz's Law (6.4) may be written:

$$\mu_x=\beta e^{\alpha x}$$

In terms of the notation used in (6.4) above, we have:

shape $=\log c$

rate $=B$

For example, if the force of mortality at age 30 years is 0.001 and mortality at ages over 30 is governed by the Gompertz law with a shape parameter equal to 0.01, the force of mortality or hazard at age 60 can be calculated using:

```
hgompertz(30, shape = 0.01, rate = 0.001)
```

to be 0.00135.

Chapter 6 Summary

Modelling mortality

We can model mortality by assuming that future lifetime is a continuous random variable taking values between 0 and some limiting age ω. From this starting point, we can calculate probabilities of survival $({}_tp_x)$ and death $({}_tq_x)$ for an individual initially aged x over a period of t years.

Death and survival probabilities

$${}_tq_x = F_x(t) = P[T_x \le t]$$

$${}_tp_x = 1 - {}_tq_x = S_x(t) = 1 - F_x(t) = P[T_x > t]$$

$${}_{t+s}p_x = {}_tp_x \times {}_sp_{x+t} = {}_sp_x \times {}_tp_{x+s}$$

Force of mortality

The force of mortality μ_x is the instantaneous rate of mortality at age x. It is defined by the equation:

$$\mu_x = \lim_{h \to 0^+} \frac{1}{h} \times P[T \le x + h \mid T > x]$$

We also have the following results about μ_x:

$$\mu_x = \lim_{h \to 0+} \frac{1}{h} \times {}_hq_x \qquad \text{so } {}_hq_x \approx h.\mu_x \quad \text{(for small } h\text{)}$$

$${}_tq_x = \int_0^t {}_sp_x \, \mu_{x+s} \, ds$$

$${}_tp_x = \exp\left\{ -\int_0^t \mu_{x+s} \, ds \right\}$$

Life table functions

l_x is the expected number of survivors at age x

d_x is the expected number of deaths between the ages of x and $x+1$

Central rate of mortality

The central rate of mortality is given by:

$$m_x = \frac{d_x}{\int_0^1 l_{x+t}\,dt} = \frac{q_x}{\int_0^1 {}_tp_x\,dt} = \frac{\int_0^1 {}_tp_x\,\mu_{x+t}\,dt}{\int_0^1 {}_tp_x\,dt}$$

Complete future lifetime random variable

The PDF of the complete future lifetime random variable T_x is given by:

$$f_x(t) = \frac{d}{dt}F_x(t) = {}_tp_x\mu_{x+t} \qquad (0 \le t < \omega - x)$$

The expected value of T_x, sometimes called the *complete expectation of life*, is:

$$\overset{\circ}{e}_x = E[T_x] = \int_0^\infty {}_tp_x\,dt$$

Curtate future lifetime random variable

K_x is defined to be the integer part of T_x.

The probability function of K_x is given by:

$$P\left(K_x = k\right) = {}_kp_x\,q_{x+k} = {}_{k|}q_x$$

The expected value of K_x, sometimes called the *curtate expectation of life*, is:

$$e_x = E[K_x] = \sum_{k=1}^{\infty} {}_kp_x$$

If deaths occur on average halfway between birthdays, then:

$$\overset{\circ}{e}_x \approx e_x + ½$$

Exponential model

In the exponential model, the hazard rate (or force of mortality) is constant. So:

$${}_tp_x = e^{-\mu t}$$

Weibull model

In the Weibull model:

$${}_t p_x = \exp\left(-\alpha t^{\beta}\right)$$

$$\mu_{x+t} = \alpha\beta t^{\beta-1}$$

Different values of β can give rise to a hazard that is monotonically increasing or decreasing. In the case when $\beta = 1$, the Weibull model is the same as the exponential model.

Gompertz' law

$$\mu_x = B\,c^x$$

$${}_t p_x = g^{c^x (c^t - 1)} \quad \text{where} \quad g = \exp\left(\frac{-B}{\log c}\right)$$

Makeham's law

$$\mu_x = A + B\,c^x$$

$${}_t p_x = s^t g^{c^x (c^t - 1)} \quad \text{where} \quad g = \exp\left(\frac{-B}{\log c}\right) \quad \text{and} \quad s = \exp(-A)$$

Both Gompertz' and Makeham's laws include an exponential term, which makes them particularly useful for middle and older ages.

Gompertz-Makeham family

The Gompertz-Makeham family consists of curves of the form:

$$\begin{aligned}\text{GM}(r,s) = {} & \alpha_1 + \alpha_2\,t + \alpha_3\,t^2 + \cdots + \alpha_r\,t^{r-1} \\ & + \exp\left\{\alpha_{r+1} + \alpha_{r+2}\,t + \alpha_{r+3}\,t^2 + \cdots + \alpha_{r+s}\,t^{s-1}\right\}\end{aligned}$$

where $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{r+s}$ are constants that do not depend on t.

The practice questions start on the next page so that you can keep the chapter summaries together for revision purposes.

Chapter 6 Practice Questions

6.1 If $\mu_x = 0.01908 + 0.001(x-70)$ for $x \geq 55$, calculate ${}_5q_{60}$.

6.2 Consider the following expressions:

I $\displaystyle\sum_{k=1}^{[\omega-x]} {}_kp_x$

II $\displaystyle\sum_{k=0}^{[\omega-x]} k \; {}_kp_x$

III $\displaystyle\int_0^{\omega-x} {}_tp_x \, \mu_{x+t} \, dt$

State which of these are correct expressions for calculating the curtate expectation of life for a life aged exactly x. Explain your answers.

6.3 Mortality of a group of lives is assumed to follow Gompertz' law. Calculate μ_x for a 30-year old and a 70-year old, given that μ_x is 0.003 for a 50-year old and 0.01 for a 60-year old.

6.4 Express q_{30}, e_{30} and ${}_5p_{35}$ in terms of probabilities of the random variable K_{30}, which represents the curtate future lifetime of a life aged exactly 30.

6.5 Calculate the exact values of the complete and curtate expectation of life for a newborn animal subject to a constant force of mortality of 0.05 per annum.

6.6 The 'Very-ruthless Management Consultancy Company' pays very high wages but also has a very high failure rate, both from sackings and through people leaving. A life table for a typical new recruit (with durations measured in years) is given below:

Duration	*No of lives*
0	100,000
1	72,000
2	51,000
3	36,000
4	24,000
5	15,000
6	10,000
7	6,000
8	2,500
9	0

75 graduates started working at the company on 1 September this year. Calculate the expected number of complete years that a graduate will complete with the company.

6.7 Given that $e_{50} = 30$ and $\mu_{50+t} = 0.005$ for $0 \leq t \leq 1$, calculate the value of e_{51}.

6.8 Exam style

Describe the difference between the following assumptions about mortality between any two ages, x and y ($y > x$):

- uniform distribution of deaths
- constant force of mortality.

In your answer, explain the shape of the survival function between ages x and y under each of the two assumptions. [2]

6.9 Exam style

In a certain population, the force of mortality is given by:

	μ_x
$60 < x \le 70$	0.01
$70 < x \le 80$	0.015
$x > 80$	0.025

Calculate the probability that a life aged exactly 65 will die between exact ages 80 and 83. [3]

6.10 Exam style

The mortality of a certain species of furry animal has been studied. It is known that at ages over five years the force of mortality μ is constant, but the variation in mortality with age below five years of age is not understood. Let the proportion of furry animals that survive to exact age five years be ${}_5p_0$.

(i) Show that, for furry animals that die at ages over five years, the average age at death in years is $\dfrac{5\mu+1}{\mu}$. [1]

(ii) Obtain an expression, in terms of μ and ${}_5p_0$, for the proportion of all furry animals that die between exact ages 10 and 15 years. [3]

A new investigation of this species of furry animal revealed that 30 per cent of those born survived to exact age 10 years and 20 per cent of those born survived to exact age 15 years.

(iii) Calculate μ and ${}_5p_0$. [3]

[Total 7]

Chapter 6 Solutions

6.1 The probability that a 60-year old will survive for 5 years is:

$$
\begin{aligned}
{}_5p_{60} &= \exp\left(-\int_0^5 \mu_{60+t}dt\right) \\
&= \exp\left(-\int_0^5 [\,0.01908+0.001(t-10)\,]\,dt\right) \\
&= \exp\left(-\int_0^5 [0.00908+0.001t]\,dt\right) \\
&= \exp\left(-\left[0.00908t+0.0005t^2\right]_0^5\right) \\
&= \exp(-0.00908\times 5-0.0005\times 25) \\
&= 0.94374
\end{aligned}
$$

So:

$$ {}_5q_{60} = 1-0.94374 = 0.05626 $$

6.2 I is correct. The sum is the total of the probabilities that the life survives to the end of each future year, which gives the expected curtate future lifetime.

II is not correct. It would be right if ${}_kp_x$ was replaced by $P(K_x = k)$ *ie* ${}_kp_x\,q_{x+k}$.

III is not correct. The integral gives the probability of dying. It is also the integral of $f_x(t)$ over all possible values of the future lifetime. So its value is 1.

6.3 Gompertz' Law is $\mu_x = Bc^x$.

Therefore $0.003 = Bc^{50}$ and $0.01 = Bc^{60}$.

Dividing these equations:

$$ \frac{Bc^{60}}{Bc^{50}} = c^{10} = \frac{0.01}{0.003} = 3.33333 $$

$$ \Rightarrow c = 1.128 \quad \text{and} \quad B = 7.29\times 10^{-6} $$

This gives Gompertz' Law as $\mu_x = 7.29\times 10^{-6}\times 1.128^x$.

So $\mu_{30} = 0.00027$ and $\mu_{70} = 0.033$.

6.4 In terms of probabilities involving K_{30} :

$$q_{30} = P(K_{30} = 0)$$

$$e_{30} = E(K_{30}) = \sum_{k=0}^{\infty} kP(K_{30} = k)$$

and:

$${}_5p_{35} = P(K_{30} \geq 10 \mid K_{30} \geq 5) = \frac{P(K_{30} \geq 10)}{P(K_{30} \geq 5)}$$

6.5 The complete expectation of life is:

$$\overset{\circ}{e}_0 = \int_0^{\infty} {}_tp_0 \, dt = \int_0^{\infty} e^{-0.05t} dt = \frac{1}{0.05} = 20 \text{ years}$$

The curtate expectation of life can be calculated exactly as follows:

$$e_0 = \sum_{k=1}^{\infty} {}_kp_0 = \sum_{k=1}^{\infty} e^{-0.05k} = \frac{e^{-0.05}}{1-e^{-0.05}} = 19.504 \text{ years}$$

6.6 The curtate expectation of life is:

$$\sum_{k=1}^{8} {}_kp_0 = \frac{72,000}{100,000} + \frac{51,000}{100,000} + \cdots + \frac{2,500}{100,000} = 2.165 \text{ years}$$

6.7 We can calculate the value of e_{51} using the formula:

$$e_{50} = p_{50}\left(1 + e_{51}\right)$$

Since the force of mortality is constant between the ages of 50 and 51:

$$p_{50} = e^{-0.005}$$

So:

$$30 = e^{-0.005}\left(1 + e_{51}\right) \Rightarrow e_{51} = 29.15 \text{ years}$$

6.8 *This is Subject CT4, September 2009, Question 1.*

Uniform distribution of deaths (UDD) assumption

If deaths are uniformly distributed between the ages of x and y, then the number of lives in the population decreases linearly between the ages of x and y.

The survival function is a linearly decreasing function of t. [1]

Constant force of mortality assumption

This assumption says that μ_{x+t} is equal to some constant μ for all t between 0 and $\omega - x$.

In general:

$$S_x(t) = {}_tp_x = \exp\left(-\int_0^t \mu_{x+s}\,ds\right)$$

Under the constant force assumption, this simplifies to:

$$S_x(t) = e^{-t\mu}$$

This is an exponentially decreasing function of t. [1]

6.9 We need to calculate:

$${}_{15|3}q_{65} = {}_{15}p_{65} \times {}_3q_{80} = {}_5p_{65} \times {}_{10}p_{70} \times (1 - {}_3p_{80}) \qquad [1]$$

We have:

$${}_5p_{65} = e^{-5(0.01)} = e^{-0.05} \qquad [½]$$

$${}_{10}p_{70} = e^{-10(0.015)} = e^{-0.15} \qquad [½]$$

and: $${}_3p_{80} = e^{-3(0.025)} = e^{-0.075} \qquad [½]$$

So:

$${}_{15|3}q_{65} = e^{-0.05} \times e^{-0.15} \times (1 - e^{-0.075}) = 0.0592 \qquad [½]$$

6.10 *This is Subject CT4, April 2013, Question 4.*

(i) ***Average age at death***

Since the lives have a constant future force of mortality μ, their future lifetimes have an $Exp(\mu)$ distribution and their expected future lifetime is $1/\mu$. However, they have already lived for 5 years. So their average age at death will be $5 + 1/\mu$, which can be written in the equivalent form $\dfrac{5\mu+1}{\mu}$.

(ii) ***Proportion that die between ages 10 and 15 years***

The proportion of animals that will die between ages 10 and 15 is ${}_{10}p_0 - {}_{15}p_0$.

For ages 5 and above, the force of mortality takes a constant value μ, and we have:

$${}_tp_x = \exp\left(-\int_0^t \mu_{x+s}ds\right) = \exp\left(-\int_0^t \mu\,ds\right) = e^{-\mu t}$$

So, splitting the age range at age 5, we have:

$$
\begin{aligned}
{}_{10}p_0 - {}_{15}p_0 &= {}_5p_0 \times {}_5p_5 - {}_5p_0 \times {}_{10}p_5 \\
&= {}_5p_0\left({}_5p_5 - {}_{10}p_5\right) \\
&= {}_5p_0\left(e^{-5\mu} - e^{-10\mu}\right)
\end{aligned}
$$

(iii) **Calculate μ and ${}_5p_0$**

We are now told that:

$$ {}_{10}p_0 = {}_5p_0 e^{-5\mu} = 0.3 \quad \text{and} \quad {}_{15}p_0 = {}_5p_0 e^{-10\mu} = 0.2 $$

Dividing the first of these equations by the second:

$$ e^{5\mu} = \frac{0.3}{0.2} = 1.5 \Rightarrow \mu = \tfrac{1}{5}\ln 1.5 = 0.08109 $$

From the first equation, we then have:

$$ {}_5p_0 = 0.3e^{5\mu} = 0.3 \times \frac{0.3}{0.2} = \frac{0.3^2}{0.2} = 0.45 $$

7

Estimating the lifetime distribution function

Syllabus objectives

4.2 Describe estimation procedures for lifetime distributions.

4.2.1 Describe the various ways in which lifetime data might be censored.

4.2.2 Describe the estimation of the empirical survival function in the absence of censoring, and what problems are introduced by censoring.

4.2.3 Describe the Kaplan-Meier (or product limit) estimator of the survival function in the presence of censoring, compute it from typical data and estimate its variance.

4.2.4 Describe the Nelson-Aalen estimator of the cumulative hazard rate in the presence of censoring, compute it from typical data and estimate its variance.

0 Introduction

In Chapter 6 we introduced T, the continuous random variable representing future lifetime. In this chapter, we will see how to use observations from an investigation to obtain an empirical estimate (*ie* one based on observation) of the distribution function, $F(t)=P(T\leq t)$. We will consider the statistical properties of the estimator so that we can measure its variance and construct confidence intervals. We will also need to bear in mind that data may be incomplete in practice.

The Core Reading refers to the decrement of interest as 'death'. ('Decrement' here means a method of leaving a population.) The models can easily be extended to the analysis of any decrements, *eg* sickness or mechanical breakdown.

Parts of this chapter are based on the paper 'An Actuarial Survey of Statistical Models for Decrement and Transition Data' by A S Macdonald, BAJ 2 (1996), by kind permission of the editor of BAJ; and on pp. 67–74 of A Hinde, *Demographic Methods* (London, Arnold, 1998) by permission of Dr Hinde.

1 Questions of inference

We now turn to statistical inference. Given some mild conditions on the distribution of T, we can obtain all information by estimating $F(t)$, $S(t)$, $f(t)$ or μ_t for all $t \geq 0$.

In other words, we can derive $F(t)$, $S(t)$, $f(t)$ and μ_t from any one of these items.

Question

State the fundamental relationships that link $F(t)$, $S(t)$, $f(t)$ and μ_t.

Solution

The relationships are:

$$S(t)=1-F(t), \qquad f(t)=\frac{d}{dt}F(t), \qquad f(t)=S(t)\mu_t, \qquad \mu_t=-\frac{S'(t)}{S(t)}$$

1.1 Estimating the lifetime distribution

The simplest experiment would be to observe a large number of new-born lives. The proportion alive at age $t > 0$ would furnish an estimate of $S(t)$. The estimate would be a step function, and the larger the sample the closer to a smooth function we would expect it to be. For use in applications it could be smoothed further.

For example, a life insurance company would prefer to base its premium calculations on a smooth estimate to ensure that the premiums change gradually from one age to the next without sudden jumps.

We need not assume that T is a member of any parametric family; this is a *non-parametric* approach to estimation. You will recognise this as the empirical distribution function of T.

Under a *non-parametric* approach, we make no prior assumptions about the shape or form of the distribution. Under a *parametric* approach, we assume that the distribution belongs to a certain family (*eg* normal or exponential) and use the data to estimate the appropriate parameters (*eg* mean and variance).

Statistical results can be derived *theoretically* (from first principles) or *empirically* (from observation). In this chapter we will use data to calculate the empirical distribution function $F(t)$.

Clearly, there are some practical problems:

- **Even if a satisfactory group of lives could be found, the experiment would take about 100 years to complete.**
- **The observational plan requires us to observe the deaths of all the lives in the sample. In practice many would be lost to the investigation, for one reason or another, and to exclude these from the analysis might bias the result. The statistical term for this problem is *censoring*. All we know in respect of some lives is that they died after a certain age.**

So censoring results in the loss of data. Depending on the nature of the censoring mechanism, it can also result in the introduction of bias into the mortality rates. This would occur if informative censoring were present – see the next section.

Question

Explain why lives might be 'lost to the investigation' if we are carrying out:

(a) a national investigation into the rate of death from natural causes

(b) a study of the mortality of life insurance policyholders.

Solution

(a) If we are interested only in natural causes of death, some lives will be 'lost' to the investigation through accidents, crime, terrorism, suicide *etc*.

Even if we are interested in all causes of death, we may lose track of some lives through data collection problems, *eg* changes of address or emigration.

(b) With life office policyholders the main reason for 'losing' people is when policyholders cancel their policies and withdraw from the group.

An 'observational plan' is just the framework for a mortality investigation. Amongst other things, it will specify the start and end date of the investigation and the category (or categories) of lives to be included in the study.

The experiment described above, in which we observe a large number of newborn lives, would provide detailed information on the lifetimes of these individuals. However, this information may only be useful as a retrospective measure of mortality patterns. This is because the level and shape of mortality rates would probably have changed significantly over time.

Such an experiment would therefore not provide a clear indication of future levels of mortality (or even very recent levels), which is the information that we are most interested in. For example, 100 years ago people in industrialised countries were dying of diseases that are no longer significant today.

In medical statistics, where the lifetimes are often shorter, non-parametric estimation is very important.

In this chapter we show how the experiment above can be amended to allow for censoring. Otherwise, we must use a different observational plan, and base inference on data gathered over a shorter time, *eg* 3 or 4 years.

A consequence is that we no longer observe the same cohort throughout their joint lifetimes, so we might not be sampling from the same distribution. It might be sensible to widen the model assumption, so that the mortality of lives born in year *y* is modelled by a random variable T^y, for example. In practice we usually divide the investigation up into single years of age. We return to investigations like these in Chapter 8.

Observing lives between (say) integer ages x and $x+1$, and limiting the period of investigation, are also forms of censoring. Censoring might still occur at unpredictable times – by lapsing a life policy, for example – but survivors will certainly be lost to observation at a known time, either on attaining age $x+1$ or when the investigation ends.

2 Censoring mechanisms

Data are *censored* if we do not know the exact values of each observation but we do have information about the value of each observation in relation to one or more bounds. For example, we may know that an individual's lifetime exceeded 20 years because the individual was still alive at age 20 when the investigation closed, but we have no further information about the remaining lifetime.

Censoring is the key feature of survival data (indeed survival analysis might be defined as the analysis of censored data) and the mechanisms that give rise to censoring play an important part in statistical inference. Censoring is present when we do not observe the exact length of a lifetime, but observe only that its length falls within some interval. This can happen in several ways.

Right censoring

Data are right censored if the censoring mechanism cuts short observations in progress. An example is the ending of a mortality investigation before all the lives being observed have died. Persons still alive when the investigation ends are right censored – we know only that their lifetimes exceed some value.

Right censoring also occurs when:

- life insurance policyholders surrender their policies
- active lives of a pension scheme retire
- endowment assurance policies mature.

Left censoring

Data are left censored if the censoring mechanism prevents us from knowing when entry into the state that we wish to observe took place. An example arises in medical studies in which patients are subject to regular examinations. Discovery of a condition tells us only that the onset fell in the period since the previous examination; the time elapsed since onset has been left censored.

Left censoring occurs, for example:

- when estimating functions of exact age and we don't know the exact date of birth
- when estimating functions of exact policy duration and we don't know the exact date of policy entry
- when estimating functions of the duration since onset of sickness and we don't know the exact date of becoming sick.

Interval censoring

Data are interval censored if the observational plan only allows us to say that an event of interest fell within some interval of time. An example arises in actuarial investigations, where we might know only the calendar year of death. Both right and left censoring can be seen as special cases of interval censoring.

Further examples of interval censoring include the following situations:

- when we only know the calendar year of withdrawal
- when estimating functions of exact age and we only know that deaths were aged '*x* nearest birthday' at the date of death
- when we know the calendar *date* of death and we know the calendar *year* of birth. This is an example of left censoring (and therefore interval censoring). Another way of viewing this situation is to say that we actually have data grouped by 'age next birthday at the 1 January prior to death'. Since we only know that the lifetime falls within a certain range, this is an example of interval censoring.

In actuarial investigations, right-censoring is the most common form of censoring encountered.

Random censoring

Suppose that the time C_i (say) at which observation of the ith lifetime is censored is a random variable. Suppose that T_i is the (random) lifetime of the ith life. Then the observation will be censored if $C_i < T_i$. In such a situation, censoring is said to be random.

Random censoring arises when individuals may leave the observation by a means other than death, and where the time of leaving is not known in advance.

Examples of random censoring include:

- life insurance withdrawals
- emigration from a population
- members of a company pension scheme may leave voluntarily when they move to another employer.

Random censoring is a special case of right censoring.

The case in which the censoring mechanism is a second decrement of interest gives rise to multiple decrement models.

For example, suppose that lives can leave a pension scheme through death, age retirement or withdrawal. We can estimate the rates of decrement for all three causes of decrement by using a multiple decrement model. Multiple decrement models are studied in detail in Subject CM1.

Type I censoring

If the censoring times $\{C_i\}$ are known in advance (a degenerate case of random censoring) then the mechanism is called 'Type I censoring'.

Type I censoring is therefore another special case of right censoring. Type I censoring occurs, for example:

- when estimating functions of exact age and we stop following individuals once they have reached their 60th birthday
- when lives retire from a pension scheme at normal retirement age (if normal retirement age is a predetermined exact age)
- when estimating functions of policy duration and we only observe individuals up to their 10th policy anniversary
- when measuring functions of duration since having a particular medical operation and we only observe people for a maximum of 12 months from the date of their operation.

Lives censored at the end of an investigation period might also be considered as an example of Type I censoring.

Type II censoring

If observation is continued until a predetermined number of deaths has occurred, then 'Type II censoring' is said to be present. This can simplify the analysis, because then the number of events of interest is non-random.

An example of Type II censoring is:

- when a medical trial is ended after 100 lives on a particular course of treatment have died.

Many actuarial investigations are characterised by a combination of random and Type I censoring, for example, in life office mortality studies where policies rather than lives are observed, and observation ceases either when a policy lapses (random censoring) or at some predetermined date marking the end of the period of investigation (Type I censoring).

Type I and Type II censoring are most frequently met with in the design of medical survival studies.

Informative and non-informative censoring

Censoring is non-informative if it gives no information about the lifetimes $\{T_i\}$.

This just means that the mortality of the lives that remain in the at-risk group is the same as the mortality of the lives that have been censored.

In the case of random censoring, the independence of each pair T_i, C_i is sufficient to ensure that the censoring is non-informative. Informative censoring is more difficult to analyse, essentially because the resulting likelihoods cannot usually be factorised.

Recall that, when we are dealing with events that are statistically independent, the likelihood function representing all the events is the product of the likelihood functions for each individual event. This greatly simplifies the mathematics required in the analysis.

Examples of informative censoring include:

- Withdrawal of life insurance policies, because these are likely to be in better average health than those who do not withdraw. So the mortality rates of the lives that remain in the at-risk group are likely to be higher than the mortality rates of the lives that surrendered their policies.
- Ill-health retirements from pension schemes, because these are likely to be in worse than average health than the continuing members. So the mortality rates of those who remain in the pension scheme are likely to be lower than the mortality rates of the lives that left through ill-health retirement.

An example of non-informative censoring is:

- the end of the investigation period (because it affects all lives equally, regardless of their propensity to die at that point).

It is obvious that the observational plan is likely to introduce censoring of some kind, and consideration should be given to the effect on the analysis in specifying the observational plan. Censoring might also depend on the results of the observations to date. For example, if strong enough evidence accumulates during the course of a medical experiment, the investigation might be ended prematurely, so that the better treatment can be extended to all the subjects under study, or the inferior treatment withdrawn.

Question

An investigation is carried out into the mortality rates of married male accountants. A group of 10,000 married male accountants is selected at random on 1 January 2016. Each member of the sample group supplies detailed personal information as at 1 January 2016 including name, address and date of birth. The same information is collected as at each 1 January in the years 2017, 2018, 2019 and 2020. The investigation closes in 2020.

Describe the ways in which the available data for this investigation may be censored.

Solution

There will be *left censoring* of all lives that change marital status from single (or divorced or widowed) to married during the investigation. We only know that the change of status occurred since the previous set of information was collected.

There will be *interval censoring* if the exact date of death is unknown, *eg* if only the calendar year of death is known.

There will be *random censoring* of all lives that change marital status from married to divorced or widowed, or give up accountancy, and consequently no longer qualify as participants in the mortality investigation. There will also be random censoring of all lives from whom data cannot be collected.

There will be *right censoring* of all lives that survive until the end of the investigation in 2020.

3 The Kaplan-Meier (product-limit) model

3.1 Introduction

In this section we develop the empirical distribution function to allow for censoring.

This is the distribution function derived from the data.

We will consider lifetimes as a function of time t without mention of a starting age x. The following could be applied equally to new-born lives, to lives aged x at outset, or to lives with some property in common at time $t = 0$, for example diagnosis of a medical condition. Medical studies are often based on time since diagnosis or time since the start of treatment, and if the patient's age enters the analysis it is usually as an explanatory variable in a regression model.

For example, we may be interested in measuring mortality amongst patients suffering from a highly virulent tropical disease. The future lifetime of a sufferer will depend on many factors. The age of the patient may be an important factor (*eg* the rate of deterioration may be quicker amongst older patients) but it may not be the sole determinant. It may be appropriate to model the lifetime as starting at the time of diagnosis. (In actuarial terminology, 'duration' is the dominant factor here.)

We will look at regression models in Chapter 8.

Although the notation in this section looks quite complicated, the numerical calculations are quite intuitive.

3.2 Assumptions and notation

Suppose we observe a population of n lives in the presence of non-informative right censoring, and suppose we observe m deaths.

By assuming that the type of censoring present is non-informative, we are assuming that the mortality of those lives remaining in the group under observation is not systematically higher or lower than the mortality of the lives that have been censored.

If informative censoring is present and we ignore it, then the resulting estimates of the distribution and survival functions will be biased.

If informative censoring is present and we allow for it, then the lifetimes and censoring times will no longer be independent. This means that the likelihood function, which is made up of joint probabilities and probability density functions, can no longer be written as the product of simple probabilities and the resulting algebra will be very complicated.

So we proceed with the assumption that any censoring present is non-informative. Bear in mind though that the results of any model are only as reliable as the assumptions on which the model is based.

We now define the rest of the notation.

Let $t_1 < t_2 < ... < t_k$ be the ordered times at which deaths were observed. We do not assume that $k = m$, so more than one death might be observed at a single failure time.

In other words, two or more lives may die on the same day.

Suppose that d_j deaths are observed at time t_j $(1 \le j \le k)$ so that $d_1 + d_2 + ... + d_k = m$.

Observation of the remaining $n - m$ lives is censored.

In other words, we don't try to track some of these remaining lives throughout the investigation.

Suppose that c_j lives are censored between times t_j and t_{j+1} $(0 \le j \le k)$, where we define $t_0 = 0$ and $t_{k+1} = \infty$ to allow for censored observations after the last observed failure time; then $c_0 + c_1 + \ ... \ + \ c_k \ = \ n - m$.

So, c_j represents the number of lives that are removed from the investigation between times t_j and t_{j+1} for a reason other than the decrement we are investigating.

The *Kaplan-Meier estimator* of the survivor function adopts the following conventions.

(a) The hazard of experiencing the event is zero at all durations except those where an event actually happens in our sample.

(b) The hazard of experiencing the event at any particular duration, t_j, when an event takes place is equal to $\dfrac{d_j}{n_j}$, where d_j is the number of individuals experiencing the event at duration t_j and n_j is the risk set at that duration (that is, the number of individuals still at risk of experiencing the event just prior to duration t_j).

So if we observed 2 deaths out of 10 lives at risk, the hazard would be equal to $\dfrac{2}{10}$.

(c) Persons that are censored are removed from observation at the duration at which censoring takes place, save that persons who are censored at a duration where events also take place are assumed to be censored immediately after the events have taken place (so that they are still at risk at that duration).

In other words, if any of the individuals are observed to be censored at the same time as one of the deaths, the convention is to treat the censoring as if it happened shortly afterwards, *ie* the deaths are assumed to have occurred first.

We will use this notation for the rest of the chapter. Let's look at an example to illustrate how it can be applied.

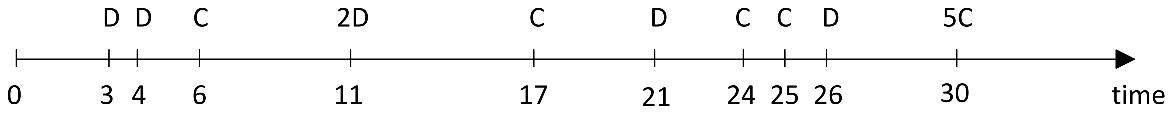
Example

Suppose that a group of 15 laboratory rats are injected with a new drug. They are observed over the next 30 days.

The following events occur:

Day	Event
3	Rat 4 dies from effects of drug.
4	Rat 13 dies from effects of drug.
6	Rat 7 gnaws through bars of cage and escapes.
11	Rats 6 and 9 die from effects of drug.
17	Rat 1 killed by other rats.
21	Rat 10 dies from effects of drug.
24	Rat 8 freed during raid by animal liberation activists.
25	Rat 12 accidentally freed by journalist reporting earlier raid.
26	Rat 5 dies from effects of drug.
30	Investigation closes.

This information is illustrated in the timeline below. In this diagram we use the notation D to represent death from the effects of the drug and C to represent censoring.



The death on day 17 is not directly related to the effects of the drug, so it is an example of random right censoring.

Using the notation defined above we have:

Number of lives under investigation, $n=15$

Number of drug-related rat deaths observed, $m=6$

Number of times at which deaths were observed: $k=5$

Times at which deaths are observed: $t_1=3, t_2=4, t_3=11, t_4=21, t_5=26$

Number of deaths observed at each failure time: $d_1=1, d_2=1, d_3=2, d_4=1, d_5=1$

Number of lives that didn't die because of the drug: $n-m=15-6=9$

Number of lives censored: $c_0=0, c_1=0, c_2=1, c_3=1, c_4=2, c_5=5$

Number of lives alive and at risk at time t_j: $n_1=15, n_2=14, n_3=12, n_4=9, n_5=6$

Note that $\sum_{j=0}^{k} c_j = n-m$.

Effectively, what we are doing is partitioning duration into very small intervals such that at the vast majority of such intervals no events occur. There is no reason to suppose, given the data that we have, that the risk of the event happening is anything other than zero at those intervals where no events occur. We have no evidence *in our data* to suppose anything else.

For those very small intervals in which events do occur, we suppose that the hazard is constant (*ie* piecewise exponential) within each interval, but that it can vary between intervals.

Recall that, if $\mu_{x+t} = \mu$, the survival function is given by:

$$S_x(t) = {}_t p_x = e^{-\mu t}$$

So the Core Reading means that the survival function is exponential over each short interval during which the force of mortality (or hazard) is constant.

We estimate the hazard within the interval containing event time t_j as:

$$\hat{\lambda}_j = \frac{d_j}{n_j}$$

Of course, effectively this formula is being used for all the other intervals as well, but as $d_j = 0$ in all these intervals, the hazard will be zero.

It is possible to show that this estimate arises as a maximum likelihood estimate. The likelihood of the data can be written:

$$\prod_{j=1}^{k} \lambda_j^{d_j} (1-\lambda_j)^{n_j - d_j}$$

The proof of this formula is beyond the syllabus.

This is proportional to a product of independent binomial likelihoods, so that the maximum is attained by setting:

$$\hat{\lambda}_j = \frac{d_j}{n_j} \qquad (1 \le j \le k)$$

Question

Given that the likelihood function can be written as:

$$L = \prod_{j=1}^{k} \lambda_j^{d_j} (1-\lambda_j)^{n_j - d_j}$$

show that the maximum likelihood estimate of λ_j is $\dfrac{d_j}{n_j}$ for $j = 1, 2, \ldots, k$.

(You may assume that the estimates are maxima.)

Solution

The log-likelihood is:

$$\ln L = \sum_{j=1}^{k}\left[d_j \ln \lambda_j + \left(n_j - d_j\right)\ln\left(1-\lambda_j\right)\right]$$

Differentiating with respect to λ_1:

$$\frac{\partial \ln L}{\partial \lambda_1} = \frac{d_1}{\lambda_1} - \frac{n_1 - d_1}{1-\lambda_1}$$

Setting this equal to 0:

$$\frac{d_1}{\lambda_1} = \frac{n_1 - d_1}{1-\lambda_1} \quad \Rightarrow d_1 - d_1\lambda_1 = n_1\lambda_1 - d_1\lambda_1$$

$$\Rightarrow d_1 = n_1\lambda_1$$

$$\Rightarrow \lambda_1 = \frac{d_1}{n_1}$$

We are told to assume this is a maximum. So we have $\hat{\lambda}_1 = \frac{d_1}{n_1}$ and it similarly follows that

$\hat{\lambda}_j = \frac{d_j}{n_j}$ for $j = 2, 3, .., k$.

3.3 Extending the force of mortality to discrete distributions

It is convenient to extend to discrete distributions the definition of force of mortality (or hazard) given in Chapter 6 for continuous distributions.

Discrete hazard function

Suppose $F(t)$ has probability masses at the points $t_1, t_2, \ldots, t_k$.

Then define:

$$\lambda_j = P\left[\, T = t_j \,|\, T \geq t_j \right] \quad (1 \leq j \leq k) \qquad \textbf{(7.1)}$$

This is called the *discrete hazard function*.

(We use the symbol λ to avoid confusion with the usual force of mortality.)

Intuitively, we can think of λ_j as the probability that a given individual dies on day t_j, given that they were still alive at the start of that day.

Question

Butterflies of a certain species have short lives. After hatching, each butterfly experiences a lifetime defined by the following probability distribution:

Lifetime (days)	**Probability**
1	0.10
2	0.30
3	0.25
4	0.20
5	0.15

Calculate λ_j for $j=1,2,...,5$ (to 3 decimal places) and sketch a graph of the discrete hazard function.

Solution

We have:

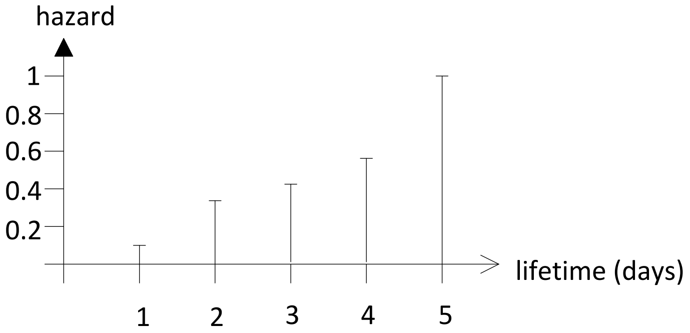
$$\lambda_j = P\left[T = t_j \middle| T \ge t_j \right] = \frac{P[T = t_j]}{P[T \ge t_j]}$$

So:

$$\lambda_1 = \frac{0.1}{1} = 0.100 \qquad \lambda_2 = \frac{0.3}{0.9} = 0.333 \qquad \lambda_3 = \frac{0.25}{0.6} = 0.417$$

$$\lambda_4 = \frac{0.2}{0.35} = 0.571 \qquad \lambda_5 = \frac{0.15}{0.15} = 1.000$$

and a graph of the discrete hazard function is given below.



3.4 Calculating the Kaplan-Meier estimate of the survival function

If we assume that T has a discrete distribution then:

$$1-F(t)=\prod_{t_j\le t}(1-\lambda_j)$$

Since $1-F(t)=S(t)$, we can estimate the survival function using the formula:

$$\hat{S}(t)=\prod_{t_j\le t}\left(1-\hat{\lambda}_j\right)$$

This is the *Kaplan-Meier estimate*.

To compute the Kaplan-Meier estimate of the survivor function, $\hat{S}(t)$, we simply multiply the survival probabilities within each of the intervals up to and including duration t.

The survival probability at time t_j is estimated by:

$$1-\hat{\lambda}_j=\frac{n_j-d_j}{n_j}=\frac{\text{number of survivors}}{\text{number at risk}}$$

So we have the following formula.

Kaplan-Meier estimate of the survival function

$$\hat{S}(t)=\prod_{t_j\le t}\left(1-\hat{\lambda}_j\right)=\prod_{t_j\le t}\left(\frac{n_j-d_j}{n_j}\right)$$

Because the Kaplan-Meier estimate involves multiplying up survival probabilities, it is sometimes called the *product limit estimate.* In effect, we choose finer and finer partitions of the time axis, and estimate $\left(1-F(t)\right)$ as the product of the probabilities of surviving each sub-interval. Then, with the above definition of the discrete force of mortality (7.1), we obtain the Kaplan-Meier estimate as the mesh of the partition tends to zero. This is the origin of the name 'product-limit' estimate, by which the Kaplan-Meier estimate is sometimes known.

Note that the Kaplan-Meier estimate of the survivor function is constant after the last duration at which an event is observed to occur. It is not defined at durations longer than the duration of the last censored observation.

Only those at risk at the observed lifetimes $\{t_j\}$ contribute to the estimate. It follows that it is unnecessary to start observation on all lives at the same time or age; the estimate is valid for data truncated from the left, provided the truncation is non-informative in the sense that entry to the study at a particular age or time is independent of the remaining lifetime. (Note that left truncation is not the same as left censoring.)

Left censoring occurs when the exact time of entry into a particular state is unknown. All that is known about the time of entry is that it occurred before a particular date. This means that we don't know exactly when to start counting duration from.

Left truncation occurs when only the events (*eg* deaths) that happen after a particular time are observed.

As mentioned in Section 2, examples of left censoring include the following situations:

- when estimating functions of exact age and we don't know the exact date of birth;
- when estimating functions of exact policy duration and we don't know the exact date of policy entry;
- when estimating functions of the duration since onset of sickness and we don't know the exact date of becoming sick.

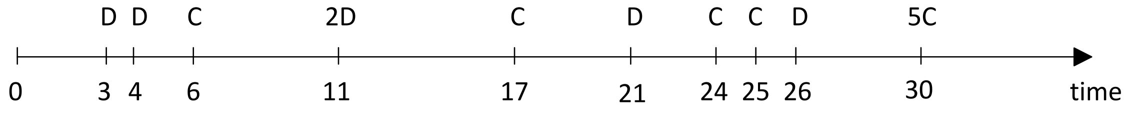
Examples of left censoring do *not* include:

- when estimating functions of exact age and we 'lose' the information from before the start of the investigation period, or before the entry date of a policy, *etc*. These are examples of *left truncation* and do *not* affect our ability to measure the exact duration of individuals from their dates of birth.

Example

Let's now return to the rats example to see how the estimation actually works in practice. There were 15 rats under observation at the start of the trial. The results of the observation and the timeline are repeated below for convenience.

Day	Event
3	Rat 4 dies from effects of drug.
4	Rat 13 dies from effects of drug.
6	Rat 7 gnaws through bars of cage and escapes.
11	Rats 6 and 9 die from effects of drug.
17	Rat 1 killed by other rats.
21	Rat 10 dies from effects of drug.
24	Rat 8 freed during raid by animal liberation activists.
25	Rat 12 accidentally freed by journalist reporting earlier raid.
26	Rat 5 dies from effects of drug.
30	Investigation closes. Remaining rats hold street party.



The calculation of the Kaplan-Meier estimate is set out in the table below:

j	t_j	d_j	n_j	$\hat{\lambda}_j = d_j / n_j$	$1-\hat{\lambda}_j = \dfrac{n_j - d_j}{n_j}$	$\prod_{k=1}^{j}\left(1-\hat{\lambda}_k\right)$
1	3	1	15	1/15	14/15	14/15
2	4	1	14	1/14	13/14	13/15
3	11	2	12	2/12	10/12	13/18
4	21	1	9	1/9	8/9	52/81
5	26	1	6	1/6	5/6	130/243

From the final column in the table, the Kaplan-Meier estimate of the survival function is:

$$\hat{S}(t) = \begin{cases} 1 & \text{for } 0 \le t < 3 \\ {}^{14}\!/_{15} & \text{for } 3 \le t < 4 \\ {}^{13}\!/_{15} & \text{for } 4 \le t < 11 \\ {}^{13}\!/_{18} & \text{for } 11 \le t < 21 \\ {}^{52}\!/_{81} & \text{for } 21 \le t < 26 \\ {}^{130}\!/_{243} & \text{for } 26 \le t < 30 \end{cases} = \begin{cases} 1 & \text{for } 0 \le t < 3 \\ 0.93333 & \text{for } 3 \le t < 4 \\ 0.86667 & \text{for } 4 \le t < 11 \\ 0.72222 & \text{for } 11 \le t < 21 \\ 0.64198 & \text{for } 21 \le t < 26 \\ 0.53498 & \text{for } 26 \le t < 30 \end{cases}$$

and the Kaplan-Meier estimate of the distribution function is:

$$\hat{F}(t) = 1 - \hat{S}(t) = \begin{cases} 0 & \text{for } 0 \le t < 3 \\ {}^{1}\!/_{15} & \text{for } 3 \le t < 4 \\ {}^{2}\!/_{15} & \text{for } 4 \le t < 11 \\ {}^{5}\!/_{18} & \text{for } 11 \le t < 21 \\ {}^{29}\!/_{81} & \text{for } 21 \le t < 26 \\ {}^{113}\!/_{243} & \text{for } 26 \le t < 30 \end{cases} = \begin{cases} 0 & \text{for } 0 \le t < 3 \\ 0.06667 & \text{for } 3 \le t < 4 \\ 0.13333 & \text{for } 4 \le t < 11 \\ 0.27778 & \text{for } 11 \le t < 21 \\ 0.35802 & \text{for } 21 \le t < 26 \\ 0.46502 & \text{for } 26 \le t < 30 \end{cases}$$

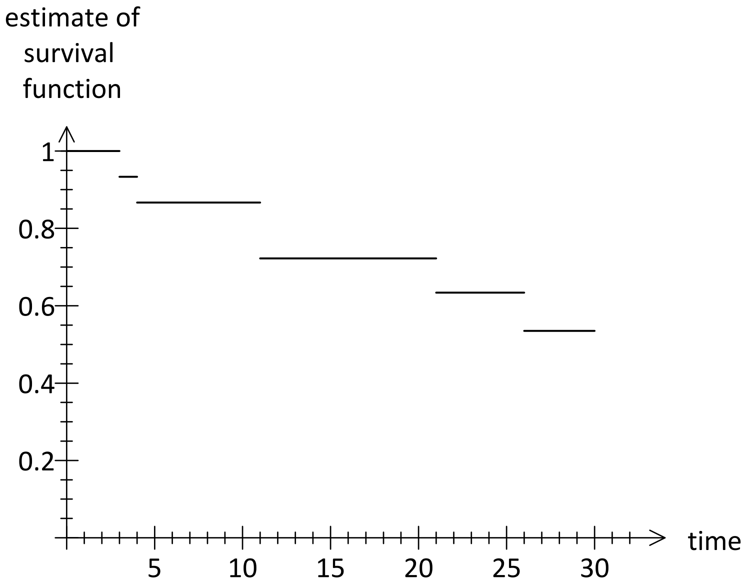
The estimate of the distribution function never reaches 1 because some rats are still alive at the end of the investigation. The estimate will only ever reach 1 if we design an experiment in which observation continues until the last life dies.

Also, since this experiment lasts for a period of 30 days, we are only able to estimate the survival function and distribution function up to time 30 days.

3.5 A graphical approach

Rather than using a table and formulae, we could carry out the Kaplan-Meier calculations using the following graphical approach. The graph of $\hat{S}(t)$ is a step function, starting at 1 and stepping down every time a death occurs.

The graph of $\hat{S}(t)$ for the rats data is given below.



To specify $\hat{S}(t)$, we need to work out the height of each of the steps.

We know that $\hat{S}(t)$ starts at 1 and remains constant until the first death, which occurs at time 3. So:

$$\hat{S}(t) = 1 \text{ for } 0 \le t < 3$$

Just before time 3, there were 15 rats under observation. One rat died at time 3. Given that a death occurred at time 3, the probability of any given rat surviving past time 3 is $^{14}\!/_{15}$.

The figure of $^{14}\!/_{15}$ corresponds to $1-\hat{\lambda}_1$ in the notation of the Kaplan-Meier model.

As the next death does not occur until time 4, we have:

$$\hat{S}(t) = {^{14}\!/_{15}} = 0.93333 \text{ for } 3 \le t < 4$$

One more rat died at time 4. There were 14 rats under observation just before time 4. So the probability that any given rat, which was alive just before time 4, is still alive at time 4 is $^{13}\!/_{14}$.

If we treat survival in non-overlapping time intervals as independent, then the probability of any given rat surviving past time 4 is:

$$P(\text{does not die at time 3}) \times P(\text{does not die at time 4}) = {^{14}\!/_{15}} \times {^{13}\!/_{14}} = {^{13}\!/_{15}} = 0.86667$$

As the next death does not occur until time 11, it follows that:

$$\hat{S}(t) = {^{13}\!/_{15}} = 0.86667 \text{ for } 4 \le t < 11$$

Just before time 11, there were 12 rats under observation (since one of the 13 still alive at time 4 was censored at time 6). Two rats died at time 11. So the probability that any given rat, which was alive just before time 11, is still alive at time 11 is ${}^{10}\!/_{12}$. Furthermore, the probability that any given rat survives past time 11 is:

$$ {}^{14}\!/_{15} \times {}^{13}\!/_{14} \times {}^{10}\!/_{12} = {}^{13}\!/_{18} = 0.72222 $$

As the next death does not occur until time 21, it follows that:

$$ \hat{S}(t) = {}^{13}\!/_{18} = 0.72222 \text{ for } 11 \le t < 21 $$

Continuing in this way, we obtain:

$$ \hat{S}(t) = \begin{cases} 1 & \text{for } 0 \le t < 3 \\ {}^{14}\!/_{15} & \text{for } 3 \le t < 4 \\ {}^{13}\!/_{15} & \text{for } 4 \le t < 11 \\ {}^{13}\!/_{18} & \text{for } 11 \le t < 21 \\ {}^{52}\!/_{81} & \text{for } 21 \le t < 26 \\ {}^{130}\!/_{243} & \text{for } 26 \le t < 30 \end{cases} = \begin{cases} 1 & \text{for } 0 \le t < 3 \\ 0.93333 & \text{for } 3 \le t < 4 \\ 0.86667 & \text{for } 4 \le t < 11 \\ 0.72222 & \text{for } 11 \le t < 21 \\ 0.64198 & \text{for } 21 \le t < 26 \\ 0.53498 & \text{for } 26 \le t < 30 \end{cases} $$

and hence:

$$ \hat{F}(t) = 1 - \hat{S}(t) = \begin{cases} 0 & \text{for } 0 \le t < 3 \\ {}^{1}\!/_{15} & \text{for } 3 \le t < 4 \\ {}^{2}\!/_{15} & \text{for } 4 \le t < 11 \\ {}^{5}\!/_{18} & \text{for } 11 \le t < 21 \\ {}^{29}\!/_{81} & \text{for } 21 \le t < 26 \\ {}^{113}\!/_{243} & \text{for } 26 \le t < 30 \end{cases} = \begin{cases} 0 & \text{for } 0 \le t < 3 \\ 0.06667 & \text{for } 3 \le t < 4 \\ 0.13333 & \text{for } 4 \le t < 11 \\ 0.27778 & \text{for } 11 \le t < 21 \\ 0.35802 & \text{for } 21 \le t < 26 \\ 0.46502 & \text{for } 26 \le t < 30 \end{cases} $$

as before.

In the package '`survival`', the function `survfit()` is used to find the Kaplan-Meier estimate of the survival function.

R code:

```
survfit(formula, conf.int = 0.95, conf.type = "log")
```

In this code '`formula`' is a survival object. With right-censored data, a survival object may be created with the R command:

```
Surv(time, delta)
```

Here '`time`' is a vector containing the times to the event of censoring, and '`delta`' is a 0/1 vector denoting whether the individual was censored (0) or experienced the event (1).

4 Comparing lifetime distributions

Since Kaplan-Meier estimates are often used to compare the lifetime distributions of two or more populations – for example, in comparing medical treatments – their statistical properties are important. Approximate formulae for the variance of $\tilde{F}(t)$ are available.

We're using $\tilde{F}(t)$ to denote the estimator of the distribution function at time t and $\hat{F}(t)$ to represent our estimate. Recall that an *estimator* is a random variable. So its value depends on the outcome of some experiment, and it has a statistical distribution. An *estimate* is a number. It is the value taken by an estimator, given a particular set of sample data.

Greenwood's formula (proof not required):

$$\mathrm{var}\left[\tilde{F}(t)\right] \approx \left(1-\hat{F}(t)\right)^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j-d_j)}$$

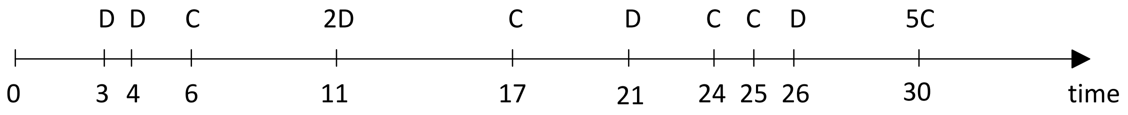
is reasonable over most t, but might tend to understate the variance in the tails of the distribution.

This formula is given on page 33 of the *Tables.*

Similarly, $\tilde{S}(t)$ denotes the estimator of the survival function at time t. Since $\tilde{S}(t) = 1-\tilde{F}(t)$, it follows that:

$$\mathrm{var}\left[\tilde{S}(t)\right] = \mathrm{var}\left[1-\tilde{F}(t)\right] = \mathrm{var}\left[\tilde{F}(t)\right]$$

Now let's return once more to the rats example. Recall that there were 15 rats under observation at the start of the trial. The timeline is repeated below for convenience.



We have already seen that the Kaplan-Meier estimate of the survival function is:

$$\hat{S}(t) = \begin{cases} 1 & \text{for } 0 \le t < 3 \\ {}^{14}\!/_{15} & \text{for } 3 \le t < 4 \\ {}^{13}\!/_{15} & \text{for } 4 \le t < 11 \\ {}^{13}\!/_{18} & \text{for } 11 \le t < 21 \\ {}^{52}\!/_{81} & \text{for } 21 \le t < 26 \\ {}^{130}\!/_{243} & \text{for } 26 \le t < 30 \end{cases}$$

Suppose we now want to estimate the variance of $\tilde{S}(20)$. According to Greenwood's formula:

$$\operatorname{var}\left[\tilde{S}(20)\right] \approx \left(\hat{S}(20)\right)^2 \sum_{t_j \le 20} \frac{d_j}{n_j(n_j - d_j)}$$

$$= \left(\frac{13}{18}\right)^2 \left(\frac{1}{15 \times 14} + \frac{1}{14 \times 13} + \frac{2}{12 \times 10}\right)$$

$$= 0.0140$$

Here we are summing over the values of t_j that are less than or equal to 20. So we include the deaths at times 3, 4, and 11 in the sum. The deaths at times 21 and 26 are not relevant when estimating $\operatorname{var}\left[\tilde{S}(20)\right]$ as they occur after time 20.

Confidence intervals

Recall that maximum likelihood estimators are asymptotically normally distributed. So, if the sample size is large, we can calculate an approximate 95% confidence interval for a survival probability using the formula:

$$\hat{S}(t) \pm 1.96 \sqrt{\operatorname{var}\left[\tilde{S}(t)\right]}$$

5 The Nelson-Aalen model

The Kaplan-Meier model is not the only non-parametric approach that can be used to estimate the distribution function. Like the Kaplan-Meier estimate, the Nelson-Aalen estimate is based on an assumption of non-informative censoring. So knowing when individuals are censored must not give us any extra information about their future lifetimes.

However, instead of using the $\hat{\lambda}_j$ values to estimate the survival probabilities via the Kaplan-Meier formula:

$$\hat{S}(t) = \prod_{t_j \le t} (1 - \hat{\lambda}_j)$$

we use them to estimate the integrated (or cumulative) hazard function.

5.1 The integrated hazard function

An alternative non-parametric approach is to estimate the integrated hazard.

This is denoted by Λ (capital λ) and is defined as follows:

$$\boldsymbol{\Lambda_t = \int_0^t \mu_s \, ds + \sum_{t_j \le t} \lambda_j}$$

where the integral deals with the continuous part of the distribution and the sum with the discrete part. (Since this methodology was developed by statisticians, the term 'integrated hazard' is in universal use, and 'integrated force of mortality' is almost never seen.)

The estimate of Λ_t can then be used to estimate $S(t)$ and $F(t)$. The integrated hazard is a function of t and is sometimes also written as $\Lambda(t)$.

To help see where Λ_t comes from, consider the probability of surviving one year in two populations. Suppose that the hazard operates continuously in the first population, so that:

$$p_0^{(1)} = \exp\left[-\int_0^1 \mu_s \, ds\right]$$

Suppose also that the hazard operates discretely in the second population, at time ½ say. Then:

$$p_0^{(2)} = 1 - \lambda_{½}$$

where $\lambda_{x+½}$ is the expected proportion of people dying at exact age $x + ½$.

If both types of hazard were to occur in the same population, then the total survival probability is:

$$p_0 = p_0^{(1)} \times p_0^{(2)} = \exp\left[-\int_0^1 \mu_s \, ds\right] \times \left(1 - \lambda_{½}\right)$$

If we extend this analysis to t years and assume that we have discrete hazards λ_j operating at exact times t_j, then:

$$S(t) = {}_t p_0 = \exp\left[-\int_0^t \mu_s\, ds\right] \times \prod_{t_j \le t} \left(1 - \lambda_j\right)$$

Now, using the approximation $e^x \approx 1 + x$ for small x, we have:

$$S(t) = \exp\left[-\int_0^t \mu_s\, ds\right] \times \prod_{t_j \le t} e^{-\lambda_j}$$

$$= \exp\left[-\int_0^t \mu_s\, ds - \sum_{t_j \le t} \lambda_j\right]$$

$$= \exp\left[-\Lambda_t\right]$$

As λ_j is the proportion of people dying at exact time t_j, we can estimate λ_j using $\hat{\lambda}_j = \dfrac{d_j}{n_j}$.

Empirically (*ie* in real life), hazards (such as death) that we theorise as operating continuously, can only occur discretely. So the continuous part of Λ_t disappears and we are left with $\hat{\Lambda}_t = \sum_{t_j \le t} \dfrac{d_j}{n_j}$ as our estimate of the integrated hazard.

5.2 Calculating Nelson-Aalen estimates

The Nelson-Aalen estimate of the integrated hazard is:

$$\boldsymbol{\hat{\Lambda}_t = \sum_{t_j \le t} \frac{d_j}{n_j}}$$

Once we have estimated the integrated hazard, we can estimate the survival function.

Nelson-Aalen estimate of the survival function

The Nelson-Aalen estimate of the survival function is:

$$\hat{S}(t) = \exp\left[-\hat{\Lambda}_t\right]$$

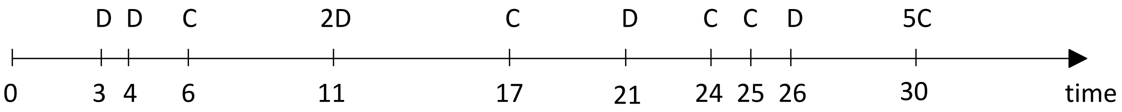
where $\hat{\Lambda}_t = \sum_{t_j \le t} \dfrac{d_j}{n_j}$.

The Nelson-Aalen estimate of the distribution function can then be calculated as:

$$\hat{F}(t) = 1 - \hat{S}(t) = 1 - \exp\left[-\hat{\Lambda}_t\right]$$

Example

To illustrate the calculations, let's return once more to the rats example. Recall that there were 15 rats under observation at the start of the trial. The timeline is repeated below for convenience.



The Nelson-Aalen estimate of the integrated hazard is:

$$\hat{\Lambda}_t = \sum_{t_j \le t} \frac{d_j}{n_j} = \begin{cases} 0 & \text{for } 0 \le t < 3 \\ {}^1\!/_{15} = 0.0667 & \text{for } 3 \le t < 4 \\ {}^1\!/_{15} + {}^1\!/_{14} = 0.1381 & \text{for } 4 \le t < 11 \\ {}^1\!/_{15} + {}^1\!/_{14} + {}^2\!/_{12} = 0.3048 & \text{for } 11 \le t < 21 \\ {}^1\!/_{15} + {}^1\!/_{14} + {}^2\!/_{12} + {}^1\!/_{9} = 0.4159 & \text{for } 21 \le t < 26 \\ {}^1\!/_{15} + {}^1\!/_{14} + {}^2\!/_{12} + {}^1\!/_{9} + {}^1\!/_{6} = 0.5825 & \text{for } 26 \le t < 30 \end{cases}$$

and the Nelson-Aalen estimate of the survival function is:

$$\hat{S}(t) = e^{-\hat{\Lambda}_t} = \begin{cases} 1 & \text{for } 0 \le t < 3 \\ e^{-0.0667} = 0.9355 & \text{for } 3 \le t < 4 \\ e^{-0.1381} = 0.8710 & \text{for } 4 \le t < 11 \\ e^{-0.3048} = 0.7373 & \text{for } 11 \le t < 21 \\ e^{-0.4159} = 0.6598 & \text{for } 21 \le t < 26 \\ e^{-0.5825} = 0.5585 & \text{for } 26 \le t < 30 \end{cases}$$

Corresponding to Greenwood's formula for the variance of the Kaplan-Meier estimator, there is a formula for the variance of the Nelson-Aalen estimator:

$$\textbf{var}\,[\tilde{\Lambda}_t] \approx \sum_{t_j \le t} \frac{d_j\left(n_j - d_j\right)}{n_j^3}$$

This formula gives the variance of the integrated hazard estimator, $\tilde{\Lambda}_t$, not the variance of $\tilde{F}(t)$. It is given on page 33 of the *Tables*.

For the rats example:

$$\text{var}[\tilde{\Lambda}_{20}] \approx \sum_{t_j \le 20} \frac{d_j(n_j - d_j)}{n_j^3} = \frac{1\times 14}{15^3} + \frac{1\times 13}{14^3} + \frac{2\times 10}{12^3} = 0.0205$$

This variance formula can be used when constructing approximate confidence intervals for the integrated hazard. If the data set is large, then an approximate 95% confidence interval for Λ_t is:

$$\hat{\Lambda}_t \pm 1.96\sqrt{\text{var}[\tilde{\Lambda}_t]}$$

The endpoints of this confidence interval can then be substituted into the formula $S(t) = e^{-\Lambda_t}$ to obtain an approximate 95% confidence interval for $S(t)$.

5.3 Relationship between the Kaplan-Meier and Nelson-Aalen estimates

The connection between the Kaplan-Meier and Nelson-Aalen estimates is discussed below.

The Kaplan-Meier estimate can be approximated in terms of $\hat{\Lambda}_t$.

Recall that the Kaplan-Meier estimate of the distribution function is:

$$\boldsymbol{\hat{F}(t) = 1 - \prod_{t_j \le t}\left(1 - \frac{d_j}{n_j}\right)}$$

To avoid confusion between the Kaplan-Meier and Nelson-Aalen estimates, we will now denote the Kaplan-Meier estimate of the distribution function by $\hat{F}_{KM}(t)$, and the Nelson-Aalen estimate of the distribution function by $\hat{F}_{NA}(t)$, so that:

$$\hat{F}_{KM}(t) = 1 - \prod_{t_j \le t}\left(1 - \frac{d_j}{n_j}\right)$$

Using the approximation $e^x \approx 1 + x$ for small x, and replacing x by $-\frac{d_j}{n_j}$, we have:

$$\boldsymbol{\hat{F}_{KM}(t) \approx 1 - \exp\left(-\sum_{t_j \le t} \frac{d_j}{n_j}\right) = 1 - \exp(-\hat{\Lambda}_t) = \hat{F}_{NA}(t)}$$

6 Parametric estimation of the survival function

An alternative approach to estimating the survival function proceeds as follows:

- **assume a functional form for the survival function $S(t)$**
- **express $S(t)$ and the hazard $h(t)$ in terms of the parameters of the chosen function**
- **estimate the parameters by maximum likelihood.**

Unless the functional form chosen is very simple, estimation will involve the solution of several simultaneous equations and must be done iteratively.

Possible simple functional forms include the exponential and Weibull distributions, and Gompertz' Law, which are all described in Chapter 6.

For the exponential distribution:

$$S(t) = P(T \geq t) = e^{-\mu t}$$

and for the Weibull distribution:

$$S(t) = \exp\left[-\alpha t^{\beta}\right]$$

These can be obtained from the formulae for distribution functions given in the *Tables* using the result $S(t) = 1 - F(t)$. Gompertz' Law, which states that:

$$\mu_x = Bc^x$$

is also given in the *Tables* (on page 32).

For many processes, such as human mortality, it turns out that no simple functional form can describe human mortality at all ages. However, for estimation purposes this is not a problem, since we can divide the age range into small sections, estimate the chosen function for each section (the parameters for each section will be different) and then 'chain' the sections together to create a life table for the whole age (or duration) range with which we are concerned (see Section 6.2 below).

Life tables were introduced in Chapter 6 and are studied in detail in Subject CM1.

6.1 Maximum likelihood estimation

We illustrate maximum likelihood estimation by considering the exponential hazard, which has one parameter, μ.

In other words, we are considering the case when the future lifetime random variable T has an exponential distribution with parameter μ. This is equivalent to assuming that the force of mortality is constant.

Consider only the single year of age between exact ages x and $x+1$.

We follow a sample of n independent lives from exact age x until the first of the following things happens:

(a) their death between exact ages x and $x+1$

(b) they withdraw from the investigation between exact ages x and $x+1$

(c) their $(x+1)$th birthday.

The Core Reading means that each life stops being observed when the earliest of the 3 events described above happens to that life.

Cases (b) and (c) are treated as censored at either the time of withdrawal, or exact age $x+1$ respectively.

Assume that the hazard of death (or force of mortality) is constant between ages x and $x+1$ and takes the unknown value μ. We ask the question: what is the most likely value of μ given the data in our investigation? Assume that we measure duration in years since a person's xth birthday.

Consider first those lives in category (a), who die before exact age $x+1$. Suppose there are k of these.

Take the first of these, and suppose that he or she died at duration t_1. Given only the data on this life, the value of μ that is most likely is the value that maximises the probability that he or she actually dies at duration t_1.

The probability that Life 1 will actually die at duration t_1 is equal to $f(t_1)$, where $f(t)$ is the probability density function of T. So the value of μ that we need is the value that maximises $f(t_1)$.

For the exponential distribution:

$$f(t_1) = \mu e^{-\mu t_1}$$

However, in the investigation, we have more than one life that died. Suppose a second life died at duration t_2. The probability of this happening is $f(t_2)$, and the joint probability that Life 1 died at duration t_1 and Life 2 died at duration t_2 is $f(t_1)f(t_2)$. Given just these two lives, the value of μ we need will be that which maximises $f(t_1)f(t_2)$.

If we now consider all the k lives that died, then the value of μ we want is that which maximises:

$$\prod_{\text{all lives which died}} f(t_i)$$

This product is the probability of observing the data we actually did observe.

It can also be written as:

$$\prod_{\text{deaths}} \mu e^{-\mu t_i} = \mu^k \exp\left(-\mu \sum_{i=1}^{k} t_i\right)$$

We are summing from $i=1$ to k because we are assuming that there are k deaths.

But what of the lives that were censored? Their experience must also be taken into account.

Consider the first censored life, and suppose he or she was censored at duration t_{k+1}. All we know about this person is that he or she was still alive at duration t_{k+1}. The probability that a life will still be alive at duration t_{k+1} is $S(t_{k+1})$.

We are using a subscript of $k+1$ because we are assuming there are k deaths and we are labelling the first censored life as Life $k+1$.

For the exponential distribution, we have:

$$S(t_{k+1}) = e^{-\mu t_{k+1}}$$

Considering all the censored lives, the probability of observing the data we do observe is:

$$\prod_{\text{all censored lives}} S(t_i)$$

This can also be written as:

$$\prod_{\text{all censored lives}} e^{-\mu t_i} = \exp\left(-\mu \sum_{i=k+1}^{n} t_i\right)$$

since we have n lives altogether and the censored ones are those labelled Life $k+1$ up to Life n.

Now, putting the deaths and the censored cases together, we can write down the probability of observing all the data we actually observe – both censored lives and those that died. This probability is:

$$\prod_{\text{all censored lives}} S(t_i) \prod_{\text{all lives which died}} f(t_i)$$

This is called the likelihood of the data.

For the exponential hazard model, the likelihood function can also be written as:

$$L = \mu^k \exp\left(-\mu \sum_{i=1}^{k} t_i\right) \exp\left(-\mu \sum_{i=k+1}^{n} t_i\right) = \mu^k \exp\left(-\mu \sum_{i=1}^{n} t_i\right)$$

Question

Determine the likelihood function when the future lifetime random variable follows the Weibull distribution with parameters α and β.

Solution

Under the Weibull model:

$$S(t_i) = e^{-\alpha t_i^{\beta}}$$

and:

$$f(t_i) = \alpha \beta t_i^{\beta-1} e^{-\alpha t_i^{\beta}}$$

So, in this case the likelihood function is:

$$\begin{aligned} L &= \prod_{\text{censored lives}} S(t_i) \prod_{\text{deaths}} f(t_i) \\ &= \prod_{\text{censored lives}} e^{-\alpha t_i^{\beta}} \prod_{\text{deaths}} \alpha \beta t_i^{\beta-1} e^{-\alpha t_i^{\beta}} \\ &= \alpha^k \beta^k \left(\prod_{\text{deaths}} t_i^{\beta-1} \right) \exp\left(-\alpha \sum_{\text{all lives}} t_i^{\beta} \right) \end{aligned}$$

where k is the observed number of deaths.

The maximum likelihood estimate of the parameter μ, which we denote by $\hat{\mu}$, is the value that maximises this likelihood.

To obtain $\hat{\mu}$, define a variable δ_i such that:

δ_i = 1 if life i died

δ_i = 0 if life i was censored

Then, in the general case, the likelihood can be written:

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Now, since $f(t) = S(t)h(t)$, or equivalently $f(t) = {}_t p_x \, \mu_{x+t}$, **the likelihood can also be written:**

$$L = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i)$$

We now substitute the chosen functional form into this equation to express the likelihood in terms of the parameter μ. This produces:

$$L = \prod_{i=1}^{n} \mu^{\delta_i} \exp(-\mu t_i)$$

This is equivalent to the expression:

$$L = \mu^k \exp\left(-\mu \sum_{i=1}^{n} t_i\right)$$

given above, bearing in mind that we have observed k deaths out of the sample of n lives.

Noting that whatever value of μ maximises L will also maximise the logarithm of L, we first take the logarithm of L:

$$\log L = \sum_{i=1}^{n} \delta_i \log \mu - \sum_{i=1}^{n} \mu t_i$$

We differentiate this with respect to μ to give:

$$\frac{\partial \log L}{\partial \mu} = \frac{\sum_{i=1}^{n} \delta_i}{\mu} - \sum_{i=1}^{n} t_i$$

Setting this equal to zero produces:

$$\frac{\sum_{i=1}^{n} \delta_i}{\mu} = \sum_{i=1}^{n} t_i$$

so that:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}$$

or equivalently:

$$\hat{\mu} = \frac{k}{\sum_{i=1}^{n} t_i}$$

where k is the total number of deaths from the n lives.

We can check that this is a maximum by noting that:

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{\sum_{i=1}^{n} \delta_i}{\mu^2}$$

This must be negative, as both numerator and denominator are necessarily positive (unless we have no deaths at all in our data, in which case the maximum likelihood estimate of the hazard is 0).

Since $\sum_{i=1}^{n}\delta_i$ is just the total number of deaths in our data, and $\sum_{i=1}^{n}t_i$ is the total time that the lives in the data are exposed to the risk of death, our maximum likelihood estimate of the force of mortality (or hazard) is just deaths divided by exposed to risk, which is intuitively sensible.

This is the same estimate for μ as the one we obtain from the two-state Markov model in Chapter 3. In that chapter we use the notation d to represent the observed number of deaths and v to represent the total waiting time. Here we have $v=\sum_{i=1}^{n}t_i$.

For parametric distributions with more than one parameter, maximum likelihood estimation of the parameters involves the solution of simultaneous equations, the number of simultaneous equations being equal to the number of parameters to be estimated. These equations often require iterative methods.

6.2 Using the estimates for different age ranges

If we repeat the exercise for other years of age, we can obtain a series of estimates for the different hazards in each year of age.

Suppose that the maximum likelihood estimate of the constant force during the single year of age from x to $x+1$ is $\hat{\mu}_x$. Then the probability that a person alive at exact age x will still be alive at exact age $x+1$ is just $S_x(1)$. Given the constant force, then:

$$\hat{S}_x(1)=\exp(-\hat{\mu}_x)$$

This is the maximum likelihood estimate of the survival function at time 1.

To work out the probability that a person alive at exact age x will survive to exact age $x+2$ we note that this probability is equal to:

$$\hat{S}_x(1)\hat{S}_{x+1}(1)=\exp(-\hat{\mu}_x)\exp(-\hat{\mu}_{x+1})$$

Therefore:

$$\hat{S}_x(1)\hat{S}_{x+1}(1)=\hat{S}_x(2)=\exp[-(\hat{\mu}_x+\hat{\mu}_{x+1})]$$

In general, therefore:

$$\hat{S}_x(m)={}_m\hat{p}_x=\exp\left(-\sum_{j=0}^{m-1}\hat{\mu}_{x+j}\right)$$

By 'chaining' together the probabilities in this way, we can evaluate probabilities over any relevant age range.

Question

If $\hat{\mu}_{60}=0.01$, $\hat{\mu}_{61}=0.02$ and $\hat{\mu}_{62}=0.03$, estimate the values of p_{60}, ${}_2p_{60}$ and ${}_3p_{60}$.

Solution

The estimates of the survival probabilities are:

$$\hat{p}_{60} = e^{-\hat{\mu}_{60}} = e^{-0.01} = 0.99005$$

$$_2\hat{p}_{60} = e^{-\left(\hat{\mu}_{60}+\hat{\mu}_{61}\right)} = e^{-\left(0.01+0.02\right)} = e^{-0.03} = 0.97045$$

$$_3\hat{p}_{60} = e^{-\left(\hat{\mu}_{60}+\hat{\mu}_{61}+\hat{\mu}_{62}\right)} = e^{-\left(0.01+0.02+0.03\right)} = e^{-0.06} = 0.94176$$

The chapter summary starts on the next page so that you can keep all the chapter summaries together for revision purposes.

Chapter 7 Summary

Estimating the future lifetime distribution

We can derive many useful functions from the lifetime distribution $F(t)=P[T \leq t]$. However, $F(t)$ is typically unknown and must be estimated from data.

A non-parametric approach is one in which we do not pre-constrain the form of the distribution function before analysing the data.

Censored data

Data for some lives may be censored. The main types of censoring (which are not mutually exclusive) are:

- right censoring
- left censoring
- interval censoring
- random censoring
- informative censoring
- non-informative censoring
- Type I censoring
- Type II censoring.

Censored data must be accounted for in the likelihood function. They tend to make the maximisation procedure more complicated.

Kaplan-Meier model

The Kaplan-Meier (or product-limit) estimate $\hat{F}_{KM}(t)$ of the lifetime distribution is a step function with jumps at each observed death. It is calculated with reference to the number and timing of deaths and the number of lives alive at each point. To calculate $\hat{F}_{KM}(t)$, we first need to estimate the discrete hazard function.

Discrete hazard function

The discrete hazard function is defined by:

$$\lambda_j = P\left[T = t_j \middle| T \geq t_j \right] \qquad (1 \leq j \leq k)$$

where t_j denotes the j th observed lifetime. It is estimated by:

$$\hat{\lambda}_j = \frac{d_j}{n_j} \; (1 \leq j \leq k)$$

Kaplan-Meier estimate of the survival function

$$\hat{S}_{KM}(t) = \prod_{t_j \le t} \left(1 - \hat{\lambda}_j\right) = \prod_{t_j \le t} \left(\frac{n_j - d_j}{n_j}\right)$$

Variance of the Kaplan-Meier estimator

We can estimate the variance of the Kaplan-Meier estimator so that we can compare lifetime distributions of two or more populations and construct confidence intervals.

Greenwood's formula

$$\text{var}\left[\tilde{S}(t)\right] = \text{var}\left[\tilde{F}(t)\right] \approx \left(1 - \hat{F}(t)\right)^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

Nelson-Aalen model

An alternative non-parametric approach is the Nelson-Aalen method. For this method we need to estimate the integrated hazard.

Integrated (or cumulative) hazard

The integrated hazard is given by:

$$\Lambda_t = \int_0^t \mu_s ds + \sum_{t_j \le t} \lambda_j$$

If we know (or can estimate) the integrated hazard function, then we can obtain (an estimate of) the survival function using the result $S(t) = e^{-\Lambda_t}$.

Nelson-Aalen estimate of the integrated hazard

$$\hat{\Lambda}_t = \sum_{t_j \le t} \frac{d_j}{n_j}$$

Nelson-Aalen estimate of the survival function

$$\hat{S}(t) = e^{-\hat{\Lambda}_t} = \exp\left(-\sum_{t_j \le t} \frac{d_j}{n_j}\right)$$

Variance of the Nelson-Aalen estimator of the integrated hazard

$$\text{var}\left(\tilde{\Lambda}_t\right) \approx \sum_{t_j \le t} \frac{d_j\left(n_j - d_j\right)}{n_j^3}$$

Parametric estimation of the survival function

The survival function can also be estimated by assuming that the future lifetime random variable belongs to a particular family of distributions and estimating the parameters of the distribution using maximum likelihood. The general likelihood function is of the form:

$$\prod_{\substack{\text{censored}\\\text{lives}}} S(t_i) \prod_{\text{deaths}} f(t_i)$$

The practice questions start on the next page so that you can keep the chapter summaries together for revision purposes.

Chapter 7 Practice Questions

7.1 A chef specialising in the manufacture of fluffy meringues uses a *Whiskmatic* disposable electric kitchen implement. The *Whiskmatic* is rather unreliable and often breaks down, so the chef is in the habit of replacing the implement in use at a given time, shortly before an important social function or after making the 1,000th fluffy meringue with that implement.

The following times until mechanical failure (no asterisk) or replacement whilst in working order (asterisk) were observed (measured in days of use):

17, 13, 15*, 7*, 21, 18*, 5, 18, 6*, 22, 19*, 15, 4, 11, 14*, 18, 10, 10, 8*, 17

(i) State the values n, m, k, t_j, d_j, c_j and n_j for these data, assuming that censoring occurs just after the failures were observed.

(ii) Calculate the Kaplan-Meier estimate of the *Whiskmatic* survival function.

(iii) Using Greenwood's formula, estimate $\text{var}\left[\tilde{S}(16)\right]$.

(iv) Calculate the Nelson-Aalen estimate of the cumulative hazard function using the given data values.

(v) Use the given data values to estimate $\text{var}[\tilde{\Lambda}_{16}]$.

7.2 You have been asked to investigate whether the rate of ill-health retirement of the employees of a large company varies with their duration of employment.

The company's records show:

- the date on which an employee was hired
- the calendar year in which they retired, if an employee left employment as a result of ill-health retirement
- the date of retirement, if an employee reached the normal retirement age of 65
- the date of leaving, if an employee left the company for any other reason.

In the context of this investigation consider the following types of censoring and in each case:

- describe the nature of the censoring
- state whether or not that type of censoring is present in these data
- if that particular type of censoring is present, explain how it arises.

(a) Left censoring

(b) Right censoring

(c) Interval censoring

(d) Informative censoring.

7.3 A clinical trial is being carried out to test the effectiveness of a new drug. Sixty patients were involved in the trial, which followed them for 2 years from the start of their treatment. The following data show the period in complete months from the start of treatment to the end of observation for those patients who died or withdrew from the trial before the end of the 2-year period.

Deaths: 8, 10, 10, 16, 20

Withdrawals: 2, 6, 9, 16, 18, 22, 22

(i) Calculate the Kaplan-Meier estimate of the survival function.

(ii) Construct an approximate 95% confidence interval for the probability that a patient survives for at least 18 months after the start of the drug treatment.

7.4 A life insurance company has carried out a mortality investigation. It followed a sample of independent policyholders aged between 50 and 55 years. Policyholders were followed from their 50th birthday until they died, withdrew from the investigation while still alive, or reached their 55th birthday (whichever of these events occurred first).

Exam style

(i) Describe the types of censoring that are present in this investigation. [2]

(ii) An extract from the data for 12 policyholders is shown in the table below. Use these data values to calculate the Nelson-Aalen estimate of the survival function.

Life	**Last age at which life was observed (years and months)**		**Reason for exit**
1	50	9	Died
2	51	3	Withdrew
3	51	6	Died
4	51	6	Died
5	51	6	Withdrew
6	52	9	Withdrew
7	53	3	Withdrew
8	54	3	Died
9	54	6	Died
10	55	0	Reached age 55
11	55	0	Reached age 55
12	55	0	Reached age 55

[5]

(iii) Determine an approximate 95% confidence interval for $S(t)$ for all values of t, $0 \le t < 5$.

[7]

[Total 14]

7.5 Exam style The following data relate to 12 patients who had an operation that was intended to correct a life-threatening condition, where time 0 is the start of the period of the investigation:

Patient number	Time of operation (in weeks)	Time observation ended (in weeks)	Reason observation ended
1	0	120	Censored
2	0	68	Death
3	0	40	Death
4	4	120	Censored
5	5	35	Censored
6	10	40	Death
7	20	120	Censored
8	44	115	Death
9	50	90	Death
10	63	98	Death
11	70	120	Death
12	80	110	Death

You can assume that censoring was non-informative with regard to the survival of any individual patient.

(i) Compute the Nelson-Aalen estimate of the cumulative hazard function, $\Lambda(t)$, where t is the time since having the operation. [6]

(ii) Using the results of part (i), deduce an estimate of the survival function for patients who have had this operation. [2]

(iii) Estimate the probability of a patient surviving for at least 70 weeks after undergoing the operation. [1]

[Total 9]

7.6

Exam style

A medical study was carried out between 1 January 2011 and 1 January 2016, to assess the survival rates of cancer patients. The patients all underwent surgery during 2011 and then attended 3-monthly check-ups throughout the study. The following data were collected.

For those patients who died during the study exact dates of death were recorded as follows:

Patient	Date of surgery	Date of death
A	1 April 2011	1 August 2015
B	1 April 2011	1 October 2011
C	1 May 2011	1 March 2012
D	1 September 2011	1 August 2013
E	1 October 2011	1 August 2012

For those patients who survived to the end of the study:

Patient	Date of surgery
F	1 February 2011
G	1 March 2011
H	1 April 2011
I	1 June 2011
J	1 September 2011
K	1 September 2011
L	1 November 2011

For those patients with whom the hospital lost contact before the end of the investigation:

Patient	Date of surgery	Date of last check-up
M	1 February 2011	1 August 2013
N	1 June 2011	1 March 2012
O	1 September 2011	1 September 2015

(i) Explain whether and where each of the following types of censoring is present in this investigation:

(a) type I censoring

(b) interval censoring; and

(c) informative censoring. [3]

(ii) Calculate the Kaplan-Meier estimate of the survival function for these patients. State any assumptions that you make. [7]

(iii) Hence estimate the probability that a patient will die within 4 years of surgery. [1]
[Total 11]

7.7 Exam style

A study of the mortality of 12 laboratory-bred insects was undertaken. The insects were observed from birth until either they died or the period of study ended, at which point those insects still alive were treated as censored.

The following table shows the Kaplan-Meier estimate of the survival function, based on data from the 12 insects.

t (weeks)	$S(t)$
$0 \le t < 1$	1.0000
$1 \le t < 3$	0.9167
$3 \le t < 6$	0.7130
$6 \le t$	0.4278

(i) Calculate the number of insects dying at durations 3 and 6 weeks. [6]

(ii) Calculate the number of insects whose history was censored. [1]
[Total 7]

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 7 Solutions

7.1 (i) ***Notation***

The original observations are:

17, 13, 15*, 7*, 21, 18*, 5, 18, 6*, 22, 19*, 15, 4, 11, 14*, 18, 10, 10, 8*, 17

These can be re-ordered to obtain:

4, 5, 6*, 7*, 8*, 10, 10 , 11, 13, 14*, 15, 15*, 17, 17, 18, 18, 18*, 19*, 21, 22

Number of 'lives' under investigation, $n=20$

Number of *Whiskmatic* failures observed, $m=13$

Number of times at which failures were observed: $k=10$

Times at which failures are observed:

$$t_1=4,\ t_2=5,\ t_3=10,\ t_4=11,\ t_5=13,\ t_6=15,\ t_7=17,\ t_8=18,\ t_9=21,\ t_{10}=22$$

Number of failures observed at each failure time:

$$d_1=1,\ d_2=1,\ d_3=2,\ d_4=1,\ d_5=1,\ d_6=1,\ d_7=2,\ d_8=2,\ d_9=1,\ d_{10}=1$$

Number of remaining lives = $n-m=20-13=7$

Number of lives censored:

$$c_0=0,\ c_1=0,\ c_2=3,\ c_3=0,\ c_4=0,\ c_5=1,\ c_6=1,\ c_7=0,\ c_8=2,\ c_9=0,\ c_{10}=0$$

Number of lives alive and at risk at t_j :

$$n_1=20,\ n_2=19,\ n_3=15,\ n_4=13,\ n_5=12,\ n_6=10,\ n_7=8,\ n_8=6,\ n_9=2,\ n_{10}=1$$

(ii) *Kaplan-Meier estimate of survival function*

We have:

j	t_j	d_j	n_j	$\hat{\lambda}_j = d_j / n_j$	$1-\hat{\lambda}_j = \dfrac{n_j - d_j}{n_j}$	$\prod_{k=1}^{j}\left(1-\hat{\lambda}_k\right)$
1	4	1	20	1/20	19/20	19/20
2	5	1	19	1/19	18/19	9/10
3	10	2	15	2/15	13/15	39/50
4	11	1	13	1/13	12/13	18/25
5	13	1	12	1/12	11/12	33/50
6	15	1	10	1/10	9/10	297/500
7	17	2	8	2/8	6/8	891/2,000
8	18	2	6	2/6	4/6	297/1,000
9	21	1	2	1/2	1/2	297/2,000
10	22	1	1	1/1	0/1	0

So the Kaplan-Meier estimate of the survival function is:

$$\hat{S}(t) = \begin{cases} 1 & \text{for } 0 \le t < 4 \\ {}^{19}\!/_{20} & \text{for } 4 \le t < 5 \\ {}^{9}\!/_{10} & \text{for } 5 \le t < 10 \\ {}^{39}\!/_{50} & \text{for } 10 \le t < 11 \\ {}^{18}\!/_{25} & \text{for } 11 \le t < 13 \\ {}^{33}\!/_{50} & \text{for } 13 \le t < 15 \\ {}^{297}\!/_{500} & \text{for } 15 \le t < 17 \\ {}^{891}\!/_{2,000} & \text{for } 17 \le t < 18 \\ {}^{297}\!/_{1,000} & \text{for } 18 \le t < 21 \\ {}^{297}\!/_{2,000} & \text{for } 21 \le t < 22 \\ 0 & \text{for } t \ge 22 \end{cases} = \begin{cases} 1 & \text{for } 0 \le t < 4 \\ 0.95 & \text{for } 4 \le t < 5 \\ 0.9 & \text{for } 5 \le t < 10 \\ 0.78 & \text{for } 10 \le t < 11 \\ 0.72 & \text{for } 11 \le t < 13 \\ 0.66 & \text{for } 13 \le t < 15 \\ 0.594 & \text{for } 15 \le t < 17 \\ 0.4455 & \text{for } 17 \le t < 18 \\ 0.297 & \text{for } 18 \le t < 21 \\ 0.1485 & \text{for } 21 \le t < 22 \\ 0 & \text{for } t \ge 22 \end{cases}$$

(iii) *Variance using Greenwood's formula*

Greenwood's formula gives

$$\text{var}\left[\tilde{S}(16)\right] \approx \left(1-\hat{F}(16)\right)^2 \sum_{t_j \le 16} \frac{d_j}{n_j(n_j - d_j)} = \left(\hat{S}(16)\right)^2 \sum_{t_j \le 16} \frac{d_j}{n_j(n_j - d_j)}$$

We have:

j	t_j	d_j	n_j	$\dfrac{d_j}{n_j(n_j-d_j)}$
1	4	1	20	$\dfrac{1}{20\times 19}$
2	5	1	19	$\dfrac{1}{19\times 18}$
3	10	2	15	$\dfrac{2}{15\times 13}$
4	11	1	13	$\dfrac{1}{13\times 12}$
5	13	1	12	$\dfrac{1}{12\times 11}$
6	15	1	10	$\dfrac{1}{10\times 9}$

So:

$$\text{var}\left[\tilde{S}(16)\right]\approx\left(\frac{297}{500}\right)^2\left(\frac{1}{20\times 19}+\frac{1}{19\times 18}+\frac{2}{15\times 13}+\frac{1}{13\times 12}+\frac{1}{12\times 11}+\frac{1}{10\times 9}\right)$$

$$=(0.594)^2\times 0.04091$$

$$=0.01443$$

(iv) ***Nelson-Aalen estimate of cumulative hazard function***

The Nelson-Aalen estimate of the cumulative hazard function is:

$$\hat{\Lambda}_t=\begin{cases}0 & \text{for } 0\le t<4\\ \frac{1}{20}=0.05 & \text{for } 4\le t<5\\ \frac{1}{20}+\frac{1}{19}=0.1026 & \text{for } 5\le t<10\\ \frac{1}{20}+\frac{1}{19}+\frac{2}{15}=0.2360 & \text{for } 10\le t<11\\ \frac{1}{20}+\frac{1}{19}+\frac{2}{15}+\frac{1}{13}=0.3129 & \text{for } 11\le t<13\\ \frac{1}{20}+\frac{1}{19}+\frac{2}{15}+\frac{1}{13}+\frac{1}{12}=0.3962 & \text{for } 13\le t<15\\ \frac{1}{20}+\frac{1}{19}+\frac{2}{15}+\frac{1}{13}+\frac{1}{12}+\frac{1}{10}=0.4962 & \text{for } 15\le t<17\\ \frac{1}{20}+\frac{1}{19}+\frac{2}{15}+\frac{1}{13}+\frac{1}{12}+\frac{1}{10}+\frac{2}{8}=0.7462 & \text{for } 17\le t<18\\ \frac{1}{20}+\frac{1}{19}+\frac{2}{15}+\frac{1}{13}+\frac{1}{12}+\frac{1}{10}+\frac{2}{8}+\frac{2}{6}=1.0796 & \text{for } 18\le t<21\\ \frac{1}{20}+\frac{1}{19}+\frac{2}{15}+\frac{1}{13}+\frac{1}{12}+\frac{1}{10}+\frac{2}{8}+\frac{2}{6}+\frac{1}{2}=1.5796 & \text{for } 21\le t<22\\ \frac{1}{20}+\frac{1}{19}+\frac{2}{15}+\frac{1}{13}+\frac{1}{12}+\frac{1}{10}+\frac{2}{8}+\frac{2}{6}+\frac{1}{2}+\frac{1}{1}=2.5796 & \text{for } t\ge 22\end{cases}$$

(v) ***Variance of estimator of integrated hazard***

Using the formula from page 33 of the *Tables*:

$$\text{var}\,[\tilde{\Lambda}_{16}] \approx \sum_{t_j \le 16} \frac{d_j\left(n_j - d_j\right)}{n_j^3} = \frac{1\times 19}{20^3} + \frac{1\times 18}{19^3} + \frac{2\times 13}{15^3} + \frac{1\times 12}{13^3} + \frac{1\times 11}{12^3} + \frac{1\times 9}{10^3} = 0.03353$$

7.2 (a) ***Left censoring***

Data in this study would be left censored if the censoring mechanism prevented us from knowing when an employee joined the company.

This is not present because the exact date of joining is given.

(b) ***Right censoring***

Data in this study would be right censored if the censoring mechanism cuts short observations in progress, so that we are not able to discover if and when an employee retired as a result of ill health.

Here there is right censoring of those lives who leave employment before their normal retirement date for reasons other than ill health.

(c) ***Interval censoring***

Data in this study would be interval censored if the observational plan only allows us to say that the duration of employment at the date of ill-health retirement fell within some interval of time (and does not allow us to find the exact duration of employment).

Here we know the calendar year of ill-health retirement and the date of employment, so we will know that the duration of employment falls within a one-year interval. Interval censoring is present.

(d) ***Informative censoring***

Censoring in this study would be informative if the censoring event divided individuals into two groups whose subsequent experience of ill-health retirement was thought to be different.

Here the censoring event of leaving the company might be suspected to be informative. Those who leave are more likely to be in better health (less likely to have retired on ill-health grounds had they remained in employment) because they probably left to take another (perhaps better paid and more responsible) job for which they may have been required to pass a medical examination. Similarly, those not resigning their jobs are more likely to retire on ill-health grounds. Informative censoring is present if these groups have different subsequent experience.

7.3 (i) ***Kaplan-Meier estimate of the survival function***

Let t denote time measured in months from the start of treatment. The Kaplan-Meier estimate of the survival function is a step function that starts at 1 and steps down every time a death is observed.

We start with 60 lives, and two of them are censored before the first death, which occurs at time 8. So:

$$\hat{S}(t) = 1 \text{ for } 0 \le t < 8$$

Assuming that the censoring at time 16 occurs after the death, we have:

j	t_j	n_j	d_j	$\hat{\lambda}_j = \dfrac{d_j}{n_j}$	$1 - \hat{\lambda}_j$
1	8	58	1	$\frac{1}{58}$	$\frac{57}{58} = 0.98276$
2	10	56	2	$\frac{2}{56}$	$\frac{54}{56} = 0.96429$
3	16	54	1	$\frac{1}{54}$	$\frac{53}{54} = 0.98148$
4	20	51	1	$\frac{1}{51}$	$\frac{50}{51} = 0.98039$

The Kaplan-Meier estimate of the survival function is then:

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \hat{\lambda}_j\right) = \begin{cases} 1 & \text{for } 0 \le t < 8 \\ 0.98276 & \text{for } 8 \le t < 10 \\ 0.94766 & \text{for } 10 \le t < 16 \\ 0.93011 & \text{for } 16 \le t < 20 \\ 0.91187 & \text{for } 20 \le t < 24 \end{cases}$$

(ii) ***95% confidence interval***

An approximate 95% confidence interval for $S(18)$ is:

$$\hat{S}(18) \pm 1.96\sqrt{\operatorname{var}\left(\tilde{S}(18)\right)}$$

From (i):

$$\hat{S}(18) = 0.93011$$

The variance term can be calculated using Greenwood's formula:

$$\operatorname{var}\left(\tilde{S}(18)\right) = \left(\hat{S}(18)\right)^2 \sum_{t_j \le 18} \frac{d_j}{n_j\left(n_j - d_j\right)}$$

$$= 0.93011^2 \left[\frac{1}{58 \times 57} + \frac{2}{56 \times 54} + \frac{1}{54 \times 53}\right]$$

$$= 0.001136$$

So the required confidence interval is:

$$0.93011 \pm 1.96\sqrt{0.001136} = (0.8640, 0.9962)$$

7.4 (i) ***Types of censoring present***

Right censoring is present since we don't know the exact future lifetime for the lives that withdrew or left the investigation at age 55. (Right censoring is a special case of interval censoring.) [½]

Random censoring occurs since we don't know the withdrawal times in advance. [½]

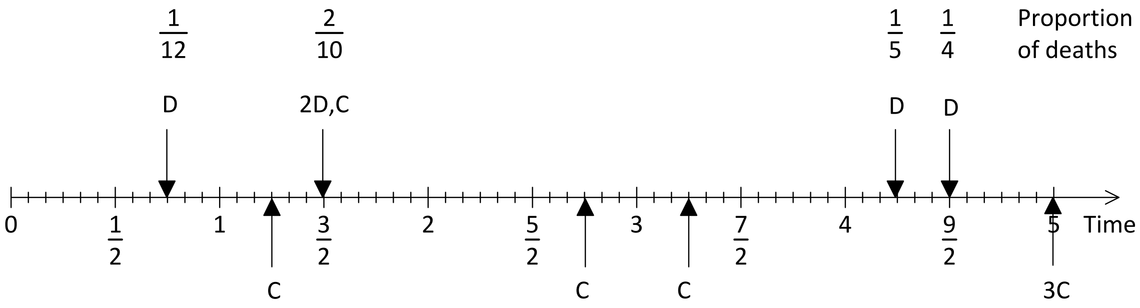
Type I censoring occurs since lives that survive to age 55 are certain to be censored at that age. [½]

Non-informative censoring is also present since the withdrawals give us no information about the future mortality of the lives remaining in the investigation. [½]

(ii) ***Nelson-Aalen estimate of the survival function***

Suppose that time is measured in years from age 50 and the withdrawal of life 5 occurs immediately after the deaths of lives 3 and 4.

A timeline of the data is shown below:



The Nelson-Aalen estimate of the survival function is:

$$\hat{S}(t) = e^{-\hat{\Lambda}_t}$$

where:

$$\hat{\Lambda}_t = \sum_{t_j \le t} \frac{d_j}{n_j} = \begin{cases} 0 & \text{for } 0 \le t < \frac{9}{12} \\ \frac{1}{12} = 0.08333 & \text{for } \frac{9}{12} \le t < 1\frac{6}{12} \\ \frac{1}{12} + \frac{2}{10} = 0.28333 & \text{for } 1\frac{6}{12} \le t < 4\frac{3}{12} \\ \frac{1}{12} + \frac{2}{10} + \frac{1}{5} = 0.48333 & \text{for } 4\frac{3}{12} \le t < 4\frac{6}{12} \\ \frac{1}{12} + \frac{2}{10} + \frac{1}{5} + \frac{1}{4} = 0.73333 & \text{for } 4\frac{6}{12} \le t < 5 \end{cases}$$

[3]

So:

$$\hat{S}(t)=\begin{cases}1 & \text{for } 0\le t<\frac{9}{12}\\ 0.92004 & \text{for } \frac{9}{12}\le t<1\frac{6}{12}\\ 0.75327 & \text{for } 1\frac{6}{12}\le t<4\frac{3}{12}\\ 0.61672 & \text{for } 4\frac{3}{12}\le t<4\frac{6}{12}\\ 0.48031 & \text{for } 4\frac{6}{12}\le t<5\end{cases}$$

[2]

(iii) ***95% confidence interval***

An approximate 95% confidence interval for the integrated hazard function is:

$$\hat{\Lambda}_t \pm 1.96\sqrt{\text{var}\left[\tilde{\Lambda}_t\right]}$$

where:

$$\text{var}\left[\tilde{\Lambda}_t\right]\approx\sum_{t_j\le t}\frac{d_j\left(n_j-d_j\right)}{n_j^3}=\begin{cases}0 & \text{for } 0\le t<\frac{9}{12}\\ 0.00637 & \text{for } \frac{9}{12}\le t<1\frac{6}{12}\\ 0.02237 & \text{for } 1\frac{6}{12}\le t<4\frac{3}{12}\\ 0.05437 & \text{for } 4\frac{3}{12}\le t<4\frac{6}{12}\\ 0.10124 & \text{for } 4\frac{6}{12}\le t<5\end{cases}$$

[2]

An approximate 95% confidence interval for the integrated hazard function is then:

$$\begin{array}{ll} 0 & \text{for } 0\le t<\frac{9}{12}\\ \left(-0.07305,\ 0.23971\right) & \text{for } \frac{9}{12}\le t<1\frac{6}{12}\\ \left(-0.00979,\ 0.57645\right) & \text{for } 1\frac{6}{12}\le t<4\frac{3}{12}\\ \left(0.02633,\ 0.94034\right) & \text{for } 4\frac{3}{12}\le t<4\frac{6}{12}\\ \left(0.10969,\ 1.35697\right) & \text{for } 4\frac{6}{12}\le t<5\end{array}$$

[2]

The integrated hazard must always be a positive number, so we truncate the estimated confidence intervals to reflect this, giving:

0	for $0 \le t < \frac{9}{12}$	
$(0,\ 0.23971)$	for $\frac{9}{12} \le t < 1\frac{6}{12}$	
$(0,\ 0.57645)$	for $1\frac{6}{12} \le t < 4\frac{3}{12}$	
$(0.02633,\ 0.94034)$	for $4\frac{3}{12} \le t < 4\frac{6}{12}$	
$(0.10969,\ 1.35697)$	for $4\frac{6}{12} \le t < 5$	[2]

This truncation will also ensure that the survival probabilities will be between 0 and 1. An approximate 95% confidence interval for the survival function is:

1	for $0 \le t < \frac{9}{12}$	
$(0.78685, 1)$	for $\frac{9}{12} \le t < 1\frac{6}{12}$	
$(0.56189,\ 1)$	for $1\frac{6}{12} \le t < 4\frac{3}{12}$	
$(0.39050,\ 0.97401)$	for $4\frac{3}{12} \le t < 4\frac{6}{12}$	
$(0.25744,\ 0.89611)$	for $4\frac{6}{12} \le t < 5$	[1]

7.5 *This is Subject 104, September 2000, Question 10.*

(i) ***Computing the Nelson-Aalen estimate***

The first thing to do is to rewrite the data in terms of the duration since having the operation (call this t), as follows (D = died; C = censored):

Patient number	**Duration *t***	**Event**
6	30	D
12	30	D
5	30	C
10	35	D
3	40	D
9	40	D
11	50	D
2	68	D
8	71	D
7	100	C
4	116	C
1	120	C

[2]

Assuming that lives who were censored at any time t were at risk of death at time t, we can calculate the required statistics as follows:

j	jth time of death t_j	Number available to die at time t_j n_j	Number of deaths at time t_j d_j	$\hat{\lambda}_j = \dfrac{d_j}{n_j}$	Estimate of cumulative hazard function $\hat{\Lambda}(t)$	Values of t to which $\hat{\Lambda}(t)$ applies
0	0	12			0	$0 \le t < 30$
1	30	12	2	0.1667	0.1667	$30 \le t < 35$
2	35	9	1	0.1111	0.2778	$35 \le t < 40$
3	40	8	2	0.2500	0.5278	$40 \le t < 50$
4	50	6	1	0.1667	0.6944	$50 \le t < 68$
5	68	5	1	0.2000	0.8944	$68 \le t < 71$
6	71	4	1	0.2500	1.1444	$71 \le t < 120$

[4]

The Nelson-Aalen estimate, and the range of t to which it applies, are shown in the last two columns of the above table.

(ii) ***Estimating the survival function***

This is calculated using:

$$\hat{S}(t) = \exp\left[-\hat{\Lambda}(t)\right]$$

The results are shown in the following table:

$\hat{S}(t)$	Value of t to which $\hat{S}(t)$ applies
1	$0 \le t < 30$
0.8465	$30 \le t < 35$
0.7575	$35 \le t < 40$
0.5899	$40 \le t < 50$
0.4994	$50 \le t < 68$
0.4088	$68 \le t < 71$
0.3184	$71 \le t < 120$

[2]

(iii) ***Survival probability***

The probability of surviving for at least 70 weeks from the operation is $S(70)$. From the table in part (ii), we see that $\hat{S}(70) = 0.4088$. [1]

7.6 *This is Subject CT4, April 2007, Question 8 (with the dates changed).*

(i)(a) ***Type I censoring***

Type I censoring occurs when the censoring times are known in advance. It is present in this investigation since we knew in advance that all lives still in the investigation on 1 January 2016 were going to be censored on that date. [1]

(i)(b) ***Interval censoring***

Interval censoring occurs when the observational plan only allows us to say that the deaths fell within some interval of time. Here we know the exact duration at the time of death for Patients A to E. So there is no interval censoring in respect of these patients. However, if Patient M, N or O had died between the last check-up and the first missed check-up, this would be an example of interval censoring. In this case, the only information we would have about the duration at death would be that it fell within a particular 3-month period. [1]

Right censoring is a special case of interval censoring. It occurs when the censoring mechanism cuts short observations in progress. If contact had been lost with Patients M, N and O for a reason other than death, then this would be an example of right censoring. Right censoring also occurs at the end of the investigation since there are patients who are still alive at that time and all we know about the lifetimes of these lives is that they are greater than some known value. [1]

(i)(c) ***Informative censoring***

Informative censoring occurs when the censoring mechanism provides some information about the future lifetimes. It is not likely to be present here. [1]

[Maximum 3]

(ii) ***Kaplan-Meier estimate of the survival function***

We assume that:

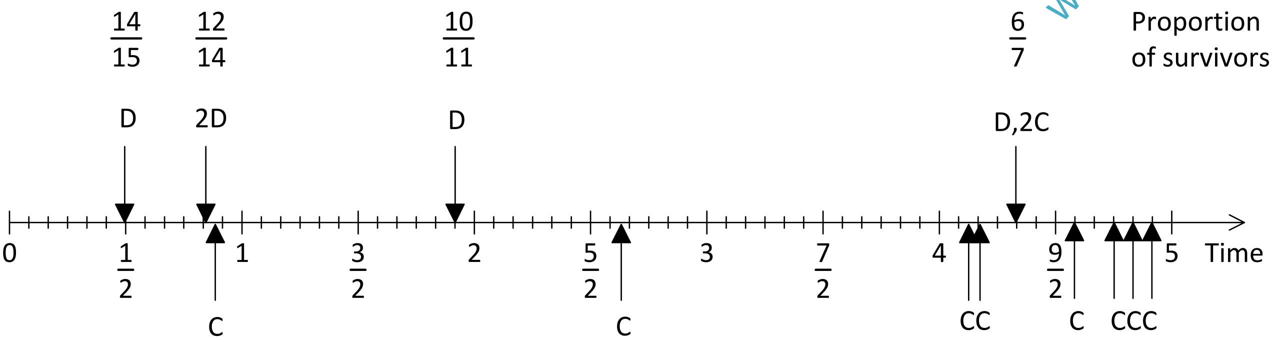
- the lives in the investigation are independent with respect to mortality and all follow the same model of mortality [½]
- the censoring is non-informative [½]
- the patients with whom contact is lost are censored half-way through the 3-month period in which contact with them was lost [½]
- at duration 4 years, 4 months, the death of Patient A occurred before Patients J and K were censored. [½]

For each life, duration (*ie* time since surgery) at exit is shown below.

Patient	Duration at exit	Reason for exit
A	4 years, 4 months	Death
B	6 months	Death
C	10 months	Death
D	1 year, 11 months	Death
E	10 months	Death
F	4 years, 11 months	Censored
G	4 years, 10 months	Censored
H	4 years, 9 months	Censored
I	4 years, 7 months	Censored
J	4 years, 4 months	Censored
K	4 years, 4 months	Censored
L	4 years, 2 months	Censored
M	2 years, 7½ months	Censored
N	10½ months	Censored
O	4 years, 1½ months	Censored

[2]

These events are shown in the timeline below, together with the proportion of survivors at each death time:



The Kaplan-Meier estimate of the survival function is:

$$\hat{S}_{KM}(t)=\begin{cases}1 & \text{for } 0\le t<\tfrac{6}{12}\\ \tfrac{14}{15} & \text{for } \tfrac{6}{12}\le t<\tfrac{10}{12}\\ \tfrac{14}{15}\times\tfrac{12}{14} & \text{for } \tfrac{10}{12}\le t<1\tfrac{11}{12}\\ \tfrac{14}{15}\times\tfrac{12}{14}\times\tfrac{10}{11} & \text{for } 1\tfrac{11}{12}\le t<4\tfrac{4}{12}\\ \tfrac{14}{15}\times\tfrac{12}{14}\times\tfrac{10}{11}\times\tfrac{6}{7} & \text{for } 4\tfrac{4}{12}\le t<5\end{cases}$$

$$=\begin{cases}1 & \text{for } 0\le t<\tfrac{6}{12}\\ \tfrac{14}{15} & \text{for } \tfrac{6}{12}\le t<\tfrac{10}{12}\\ \tfrac{4}{5} & \text{for } \tfrac{10}{12}\le t<1\tfrac{11}{12}\\ \tfrac{8}{11} & \text{for } 1\tfrac{11}{12}\le t<4\tfrac{4}{12}\\ \tfrac{48}{77} & \text{for } 4\tfrac{4}{12}\le t<5\end{cases}$$ [3]

where t is measured in time in years since surgery.

Since our first observed death is at time $t=\tfrac{6}{12}$, the first part of the estimated survival function is:

$$\hat{S}_{KM}(t)=1 \qquad \textit{for } 0\le t<\tfrac{6}{12}$$

At time $t=\tfrac{6}{12}$ there are 15 patients in the at-risk group and 1 of them dies at this time. So we estimate the survival probability to be $1-\tfrac{1}{15}=\tfrac{14}{15}$, and this stays constant until the next observed death time, ie until time $t=\tfrac{10}{12}$. So the second part of the estimated survival function is:

$$\hat{S}_{KM}(t)=\tfrac{14}{15} \qquad \textit{for } \tfrac{6}{12}\le t<\tfrac{10}{12}$$

At time $t = {}^{10}\!/_{12}$ there are 14 patients in the at-risk group. (We started with 15 but 1 died at time $t = {}^{6}\!/_{12}$.) Out of these 14, 2 die. So we estimate the probability of not dying ***at*** *time $t = {}^{10}\!/_{12}$ to be $1 - {}^{2}\!/_{14} = {}^{12}\!/_{14}$, and the probability of still being alive* ***after*** *time $t = {}^{10}\!/_{12}$ to be ${}^{14}\!/_{15} \times {}^{12}\!/_{14}$. (To be alive after time $t = {}^{10}\!/_{12}$, a life must have not died at time $t = {}^{6}\!/_{12}$* ***and*** *not died at time $t = {}^{10}\!/_{12}$.) Our estimate of the survival probability stays constant until the next observed death time, ie time $t = 1{}^{11}\!/_{12}$.*

So the third part of the estimated survival function is:

$$\hat{S}_{KM}(t) = {}^{14}\!/_{15} \times {}^{12}\!/_{14} = {}^{12}\!/_{15} = {}^{4}\!/_{5} \quad \textit{for } {}^{10}\!/_{12} \le t < 1{}^{11}\!/_{12}$$

The rest of the function follows in a similar way.

Alternatively, we could assume that Patients M, N and O were censored on the dates of their last check-ups. This gives durations at censoring of 2 years 6 months, 9 months and 4 years, respectively. With this assumption, the Kaplan-Meier estimate of the survival function is:

$$\hat{S}_{KM}(t) = \begin{cases} 1 & \text{for } 0 \le t < {}^{6}\!/_{12} \\ {}^{14}\!/_{15} & \text{for } {}^{6}\!/_{12} \le t < {}^{10}\!/_{12} \\ {}^{14}\!/_{15} \times {}^{11}\!/_{13} & \text{for } {}^{10}\!/_{12} \le t < 1{}^{11}\!/_{12} \\ {}^{14}\!/_{15} \times {}^{11}\!/_{13} \times {}^{10}\!/_{11} & \text{for } 1{}^{11}\!/_{12} \le t < 4{}^{4}\!/_{12} \\ {}^{14}\!/_{15} \times {}^{11}\!/_{13} \times {}^{10}\!/_{11} \times {}^{6}\!/_{7} & \text{for } 4{}^{4}\!/_{12} \le t < 5 \end{cases}$$

$$= \begin{cases} 1 & \text{for } 0 \le t < {}^{6}\!/_{12} \\ {}^{14}\!/_{15} & \text{for } {}^{6}\!/_{12} \le t < {}^{10}\!/_{12} \\ {}^{154}\!/_{195} & \text{for } {}^{10}\!/_{12} \le t < 1{}^{11}\!/_{12} \\ {}^{28}\!/_{39} & \text{for } 1{}^{11}\!/_{12} \le t < 4{}^{4}\!/_{12} \\ {}^{8}\!/_{13} & \text{for } 4{}^{4}\!/_{12} \le t < 5 \end{cases}$$

(iii) ***Probability that a patient will die within 4 years of surgery***

From (ii), the Kaplan-Meier estimate of this death probability is:

$$\hat{F}_{KM}(4) = 1 - \hat{S}_{KM}(4) = 1 - {}^{8}\!/_{11} = {}^{3}\!/_{11} = 0.27273 \qquad [1]$$

7.7 *This is Subject CT4, April 2005, Question B5.*

(i) ***Number of insects dying at durations 3 and 6 weeks***

The estimated survival function has 'steps' at times 1, 3 and 6. This means that deaths can only occur at these times.

The estimate of the discrete hazard at time 1, $\hat{\lambda}_1$, is:

$$1 - 0.9167 = 0.0833 = \frac{1}{12} \qquad [½]$$

There are 12 insects exposed to the risk of death at time 0. We are told implicitly that no additional insects join the study after time 0, so the number of insects exposed to the risk of death immediately before time 1 must be $n_1 \leq 12$. The number of deaths at time 1 can only be a positive integer, $d_1 = 1, 2, \ldots$. The only feasible values are $n_1 = 12$ and $d_1 = 1$. [1]

The estimate of the discrete hazard at time 3, $\hat{\lambda}_2$, is given by:

$$\left(1 - \frac{1}{12}\right)\left(1 - \hat{\lambda}_2\right) = 0.7130 \qquad [½]$$

So:

$$\hat{\lambda}_2 = \frac{d_2}{n_2} = 0.2222 = \frac{2}{9} \qquad [½]$$

There are 11 insects exposed to the risk of death at time 1. The number of insects exposed to the risk of death immediately before time 3 must be $n_2 \leq 11$. The number of deaths at time 3 can only be a positive integer, $d_2 = 1, 2, \ldots$. The only feasible values are $n_2 = 9$ and $d_2 = 2$. [1]

The estimate of the discrete hazard at time 6, $\hat{\lambda}_3$, is given by:

$$\left(1 - \frac{1}{12}\right)\left(1 - \frac{2}{9}\right)\left(1 - \hat{\lambda}_3\right) = 0.4278 \qquad [½]$$

So:

$$\hat{\lambda}_3 = \frac{d_3}{n_3} = 0.4000 = \frac{2}{5} \qquad [½]$$

There are 7 insects exposed to the risk of death at time 3. The number of insects exposed to the risk of death immediately before time 6 must be $n_3 \leq 7$. The number of deaths at time 6 can only be a positive integer, $d_3 = 1, 2, \ldots$. The only feasible values are $n_3 = 5$ and $d_3 = 2$. [1]

In summary, 1 insect died at time 1, 2 insects died at time 3, and 2 insects died at time 6. [½]

(ii) ***Number of insects whose history was censored***

12 insects were observed and 5 died. So 7 were censored. [1]

End of Part 2

What next?

1. Briefly **review** the key areas of Part 2 and/or re-read the **summaries** at the end of Chapters 5 to 7.

2. Ensure you have attempted some of the **Practice Questions** at the end of each chapter in Part 2. If you don't have time to do them all, you could save the remainder for use as part of your revision.

3. Attempt **Assignment X2**.

8

Proportional hazards models

Syllabus objectives

4.2 Describe estimation procedures for lifetime distributions.

4.2.5 Describe models for proportional hazards, and how these models can be used to estimate the impact of covariates on the hazard.

4.2.6 Describe the Cox model for proportional hazards, derive the partial likelihood estimate in the absence of ties and state the asymptotic distribution of the partial likelihood estimator.

0 Introduction

The true level of mortality for an individual is unknown in practice. In order to estimate it, we can carry out an investigation and make statistical inferences based on the observed data.

One of the main problems is *heterogeneity*. The population may include lives with very different characteristics, *eg* males and females, smokers and non-smokers. In such circumstances we will observe an average mortality rate over the population as a whole. It would be more informative to split the population into *homogeneous* subgroups of individuals with similar characteristics (*eg* male smokers, female non-smokers) and identify the level of mortality experienced by members of each subgroup.

In this chapter we will consider:

- how to incorporate in a model the different factors (called *covariates*) that are used to split the population into subgroups
- *proportional hazards models*, where the formula incorporates an adjustment to reflect the characteristics of each particular individual
- *fully parametric models*, where the hazard rate is a simple function of some time period t, and the limitations of these models
- the *Cox model*, which is a particular type of proportional hazards model.

This chapter is based on the paper 'An Actuarial Survey of Statistical Models for Decrement and Transition Data' by A S Macdonald, BAJ 2 (1996), by kind permission of the editor of BAJ.

1 Covariates and proportional hazards models

1.1 Covariates

Estimates of the lifetime distribution, whether parametric or non-parametric, are limited in their ability to deal with some important questions in survival analysis, such as the effect of *covariates* on survival.

A covariate is any quantity recorded in respect of each life, such as age, sex, type of treatment, level of medication, severity of symptoms and so on. If the covariates partition the population into a small number of homogeneous groups, it is possible to compare Kaplan-Meier or other estimates of the survivor function in respect of each population, but a more direct and transparent method is to construct a model in which the effects of the covariates on survival are modelled directly: a regression model. In this section, we will assume that the values of the covariates in respect of the ith life are represented by a $1\times p$ vector, z_i.

The vector notation in this chapter requires some care. The Core Reading uses the same notation for both vectors and scalars. The notation in the ActEd material is consistent with the Core Reading. When trying questions on this topic, you might want to use the notation $\underset{\sim}{z}$ to denote the vector of covariates.

The covariates can be:

- direct measurements (*eg* age or weight)
- indicator or dummy variables (*eg* 0 for a male and 1 for a female or 0 for new treatment and 1 for placebo)
- a quantitative interpretation of a qualitative measurement (*eg* severity of symptoms from 0 to 5 with 0 representing no symptoms and 5 representing extreme severity).

For example, the vector z_i might be (sex, age, weight, symptoms). If the third life is a 68-year-old male (with dummy variable 0), weighing 74kg, with mild symptoms of the condition under investigation (graded as 1 on a scale from 0 to 5), then we would have $z_3=(0,68,74,1)$.

1.2 Proportional hazards models

The most widely used regression model in recent years has been the *proportional hazards* model. Proportional hazards (PH) models can be constructed using both parametric and non-parametric approaches to estimating the effect of duration on the hazard function.

In PH models the hazard function for the ith life, $\lambda_i(t,z_i)$, may be written:

$$\lambda_i(t,z_i)=\lambda_0(t)g(z_i)$$

where $\lambda_0(t)$ is a function *only* of duration t, and $g(z_i)$ is a function *only* of the covariate vector. (In keeping with statistical habit, we denote hazards by λ rather than μ.) Here, $\lambda_0(t)$ is the hazard for an individual with a covariate vector equal to zero. It is called the *baseline hazard*.

We will see later that in this type of model, when the covariates all have value zero, the function $g(z_i)$ will equal 1.

Models can be specified in which the effect of covariates changes with duration:

$$\boldsymbol{\lambda_i(t, z_i) = \lambda_0(t) g(z_i, t)}$$

but because the hazard no longer factorises into two terms, one depending only on duration and the other depending only on the covariates, these are not PH models.

We will have a look in more detail in Section 2.3 at how the proportional element works.

They are also both more complex to interpret and more computer-intensive to estimate.

2 Fully parametric models

2.1 Parametric models for the hazard function

In a fully parametric PH model, the strong assumption is made that the lifetime distribution, and hence the hazard, belongs to a given family of parametric distributions, and the regression problem is reduced to estimating the parameters from the data.

Recall from Chapter 6 that the PDF of the future lifetime random variable T_x is:

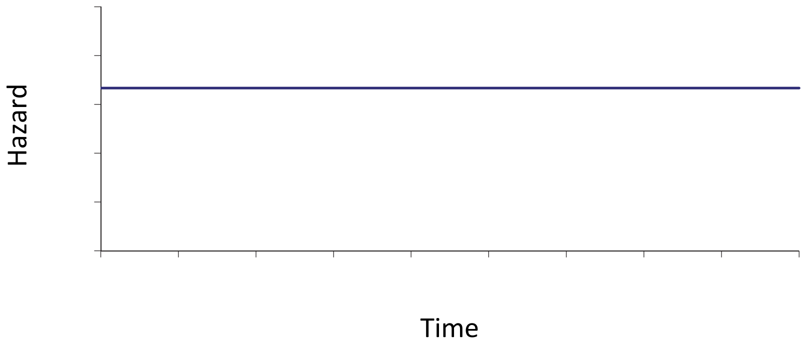
$$f_x(t) = {}_tp_x\,\mu_{x+t} = \mu_{x+t}\exp\left(-\int_0^t \mu_{x+s}\,ds\right)$$

The 'hazard' referred to in the Core Reading is just the force of mortality. The hazard function (from age x) may also be written as $h(t)$ or as $h_x(t)$.

Distributions commonly used are the exponential (constant hazard), Weibull (monotonic hazard), Gompertz-Makeham (exponential hazard) and log-logistic ('humped' hazard).

The general shapes of the more commonly used distributions are illustrated below.

(a) ***Exponential (constant hazard)***



Question

Explain why the constant hazard model is described as 'exponential'.

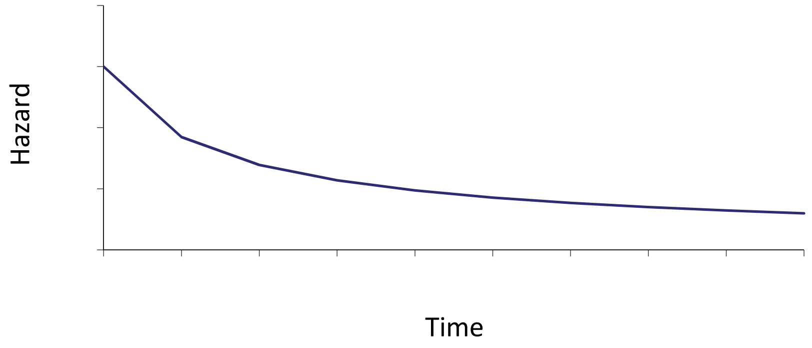
Solution

Under the constant hazard model with hazard rate λ, the distribution function of the future lifetime of a life aged x is:

$$F_x(t) = {}_tq_x = 1-\exp\left(-\int_0^t \mu_{x+s}\,ds\right) = 1-\exp\left(-\int_0^t \lambda\,ds\right) = 1-e^{-\lambda t} \qquad (t \geq 0)$$

This is the distribution function of an $Exp(\lambda)$ random variable.

(b) ***Weibull (monotonically decreasing hazard)***



The PDF of the Weibull distribution is:

$$f(t)=c\gamma t^{\gamma-1}\exp(-ct^{\gamma}) \qquad (c>0,\ \gamma>0,\ t>0)$$

and its CDF is:

$$F(t)=1-\exp(-ct^{\gamma}) \qquad (c>0,\ \gamma>0,\ t>0)$$

The Weibull model can also be used for a monotonically increasing hazard.

Question

Write down the hazard function for the Weibull distribution. State the values of γ for which this is:

(a) decreasing

(b) constant

(c) increasing.

Solution

Since $f_x(t)={}_tp_x\mu_{x+t}$ and $F_x(t)=P(T_x\le t)={}_tq_x=1-{}_tp_x$, it follows that:

$$\mu_{x+t}=\frac{f_x(t)}{1-F_x(t)}$$

So the hazard function for the Weibull distribution is:

$$h(t)=\frac{c\gamma t^{\gamma-1}\exp(-ct^{\gamma})}{\exp(-ct^{\gamma})}=c\gamma t^{\gamma-1} \qquad (t>0)$$

Differentiating this gives:

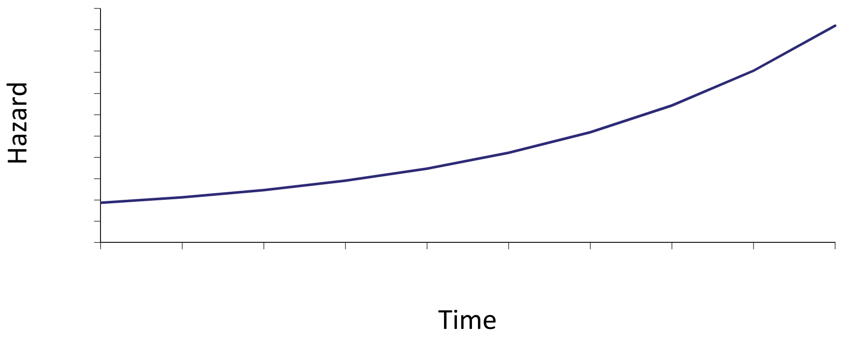
$$h'(t) = c\gamma(\gamma-1)t^{\gamma-2}$$

(a) The derivative is negative if $\gamma < 1$. So the hazard function is decreasing if $\gamma < 1$.

(b) The derivative is 0 if $\gamma = 1$. So the hazard function is constant if $\gamma = 1$.

(c) The derivative is positive if $\gamma > 1$. So the hazard function is increasing if $\gamma > 1$.

(c) ***Gompertz-Makeham (exponential hazard)***



Question

State Makeham's law for the force of mortality.

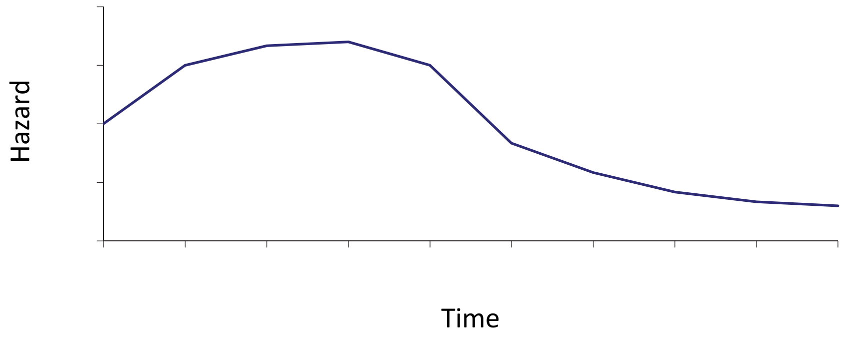
Solution

Makeham's law for the force of mortality is:

$$\mu_x = A + Bc^x$$

for some parameters A, B and c. This is an exponential hazard since the variable x appears as the power.

(d) ***Log-logistic hazard ('humped' hazard)***



The log-logistic hazard function is:

$$h(t)=\frac{\gamma(t/\theta)^{\gamma}}{t\left[1+(t/\theta)^{\gamma}\right]}$$

Question

Give an example of a situation in which the hazard function may be expected to follow each of the following distributions:

(i) exponential

(ii) decreasing Weibull

(iii) Gompertz-Makeham

(iv) log-logistic.

Solution

(i) The constant hazard model (exponential) could reflect the hazard for an individual who remains in good health. The level of hazard would reflect the risk of death from unnatural causes, *eg* accident or murder.

(ii) The decreasing hazard model (decreasing Weibull) could reflect the hazard for patients recovering from major heart surgery. The level of hazard is expected to fall as the time since the operation increases.

(iii) The exponentially increasing hazard model (Gompertz-Makeham) could reflect the hazard for leukaemia sufferers who are not responding to treatment. The severity of the condition and the level of hazard increase with the survival time. Over longer time periods, the Gompertz-Makeham model could be suitable for describing the increasing chance of death from natural causes as age increases. (We saw this in Chapter 6.)

(iv) The humped hazard (log-logistic) could reflect a hazard for patients with a disease that is most likely to cause death during the early stages. As the initial condition becomes more severe, the level of hazard increases. But once patients have survived the period of highest risk, the level of hazard decreases.

2.2 Other applications of these models

The same distributions are often used as loss distributions with insurance claims data, but censored observations complicate the likelihoods considerably and numerical methods are usually required. For the distributions above, the likelihoods can be written down (though not always solved) explicitly.

We will consider loss distributions in more detail in Chapter 15. The problem of censored claims data is discussed in Chapter 18.

2.3 Use of parametric models

Parametric models can be used with a homogeneous population (the one-sample case) as described in Section 6 of Chapter 7, or can be fitted to a moderate number of homogeneous groups, in which case confidence intervals for the fitted parameters give a test of differences between the groups which should be better than non-parametric procedures.

A parametric PH model using the Gompertz distribution might be specified as follows. The Gompertz hazard is:

$$\lambda(t) = Bc^t$$

with two parameters B and c. If we let the value of the parameter B depend on the covariate vector z_i:

$$B = \exp(\beta z_i^T)$$

where β is a $1 \times p$ vector of *regression coefficients*, then through the scalar product βz_i^T the influence of each factor in z_i enters the hazard multiplicatively. (Note that the '*T*' denotes the transpose of the vector z_i, not a lifetime.)

We then have the PH model:

$$\lambda_i(t, z_i) = c^t \exp(\beta z_i^T)$$

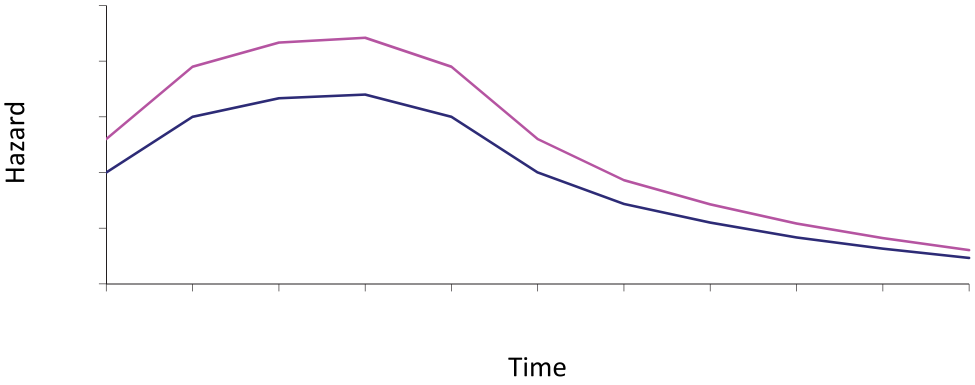
Actuaries are frequently interested in both the baseline hazard and the effect of the covariates. As long as numerical methods are available to maximise the full likelihood (and find the information matrix), which nowadays should not be a problem, it is not difficult to specify any baseline hazard required and to estimate all the parameters simultaneously, *ie* those in the baseline hazard and the regression coefficients.

Under PH models, the hazards of different lives with covariate vectors z_1 and z_2 are in the same proportion at all times:

$$\frac{\lambda(t,z_1)}{\lambda(t,z_2)}=\frac{\exp(\beta z_1^T)}{\exp(\beta z_2^T)}$$

Hence the name *proportional hazards* model.

The following graph shows the hazard for two lives under a proportional hazards model. The hazard functions are the same shape. The ratio of the hazard rates is constant at all times.



Moreover, the specification above ensures that the hazard is always positive and gives a linear model for the log-hazard:

$$\log\lambda_i(t,z_i)=t\log c+\beta z_i^T$$

which is very convenient in theory and practice.

However, fully parametric models are difficult to apply without foreknowledge of the form of the hazard function. Moreover, in many medical applications answers to questions depend mainly on estimating the regression coefficients. The baseline hazard is relatively unimportant. For these reasons, an alternative semi-parametric approach, originally proposed by D R Cox in 1972, has become popular.

The main problem of using a parametric approach to analyse observed survival times is that if an inappropriate family of parametric distributions is chosen, the hazard function will be the wrong shape. Whilst regression parameters can be chosen to maximise the likelihood for the observed data, the model will not be suitable for estimation. Typically, we do not know the form of the distribution before analysing the data.

The hazard function could also be the wrong shape if the population comprises several heterogeneous subgroups.

3 The Cox proportional hazards model

3.1 Introduction

The general formula for the Cox proportional hazards model is given below.

Cox proportional hazards (PH) model

The Cox PH model proposes the following form of hazard function for the *i* th life:

$$\lambda(t;z_i)=\lambda_0(t)\exp(\beta z_i^T)$$

$\lambda_0(t)$ is the *baseline hazard.*

So the hazard for a life with covariates z_i is proportional to the baseline hazard, the proportionality factor being the exponential term $\exp(\beta z_i^T)$.

If the covariates for the i th life are $z_i=\left(X_{i1},X_{i2},\dots,X_{ip}\right)$ and the vector of regression parameters is $\beta=\left(\beta_1,\beta_2,\dots,\beta_p\right)$ then $\exp(\beta z_i^T)=\exp\left(\sum_{j=1}^{p}\beta_j X_{ij}\right)$. We are assuming that the entries in the vector z_i are positive as they reflect observed quantities, *eg* age or height.

If the j th regression parameter is positive, the hazard rate (*eg* the force of mortality) increases with the j th covariate, *ie* there is a positive correlation between hazard rate and covariate. For example, if obese individuals are more likely to suffer from major heart disease, we would expect to find the regression parameter associated with the covariate representing weight to be positive.

If the j th regression parameter is negative, the hazard rate decreases with the j th covariate, *ie* there is a negative correlation between hazard rate and covariate. For example, if individuals who drink a high volume of non-alcoholic liquids are less likely to suffer from liver disease, we would expect to find the regression parameter associated with the covariate representing non-alcoholic liquid intake to be negative.

If the magnitude of the j th regression parameter is large, the hazard rate is significantly affected by the j th covariate, *ie* there is a strong correlation (positive or negative) between hazard rate and covariate. If the magnitude of the j th regression parameter is small, the hazard rate is not significantly affected by the j th covariate, *ie* there is a weak correlation between hazard rate and covariate.

The significance of each covariate can be tested statistically.

Question

Suppose that, in a Cox model, the covariates for the i th observed life are (56, 183, 40) representing (age last birthday at the start of the study, height in *cm*, daily dose of drug A in *mg*).

Using the regression parameters β = (0.0172, 0.0028, –0.0306), give a formula for $\lambda(t;z_i)$ in terms of $\lambda_0(t)$.

Solution

Here we have:

$$\begin{aligned}\lambda(t;z_i) &= \lambda_0(t)\times\exp(56\times 0.0172+183\times 0.0028+40\times -0.0306)\\ &= \lambda_0(t)\times\exp(0.2516)\\ &= \lambda_0(t)\times 1.286\end{aligned}$$

More generally, if the covariates for the i th life are $z_i=\left(X_{i1},X_{i2},\dots,X_{ip}\right)$ and the vector of regression parameters is $\beta=\left(\beta_1,\beta_2,\dots,\beta_p\right)$, then the ratio of the hazards of lives with covariate vectors z_1 and z_2 is:

$$\frac{\lambda(t;z_1)}{\lambda(t;z_2)}=\frac{\exp(\beta z_1^T)}{\exp(\beta z_2^T)}=\frac{\exp\left(\sum_{j=1}^{p}\beta_j X_{1j}\right)}{\exp\left(\sum_{j=1}^{p}\beta_j X_{2j}\right)}=\exp\left(\sum_{j=1}^{p}\beta_j\left(X_{1j}-X_{2j}\right)\right)$$

It is important to realise that this ratio is constant, *ie* it is independent of t.

3.2 The utility of the Cox model

The utility of this model arises from the fact that the general 'shape' of the hazard function for all individuals is determined by the baseline hazard, while the exponential term accounts for differences between individuals. So, if we are not primarily concerned with the precise form of the hazard, but with the effects of the covariates, we can ignore $\lambda_0(t)$ and estimate β from the data irrespective of the shape of the baseline hazard. This is termed a *semi-parametric* approach.

In other words, by estimating the vector of parameters β, we can use the Cox model to compare the relative forces of mortality of two lives (or two homogeneous groups of lives). However, we cannot estimate the absolute force of mortality for an individual without first estimating the baseline hazard.

So useful and flexible has this proved, that the Cox model now dominates the literature on survival analysis, and it is probably the tool to which a statistician would turn first for the analysis of survival data.

Question

An investigation is being carried out into the survival times of patients who have just undergone heart surgery at one of 3 city hospitals – A, B or C. The following data have been recorded for each patient:

$$Z_1 = \begin{cases} 1 & \text{for males} \\ 0 & \text{for females} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if patient attended Hospital B} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_3 = \begin{cases} 1 & \text{if patient attended Hospital C} \\ 0 & \text{otherwise} \end{cases}$$

The force of mortality at time t (measured in days since the operation was performed) is being modelled by an equation of the form $\lambda(t) = \lambda_0(t)e^{\beta Z^T}$. The estimated parameter values are:

$$\hat{\beta}_1 = 0.031 \qquad \hat{\beta}_2 = -0.025 \qquad \hat{\beta}_3 = 0.011$$

Use this model to compare the force of mortality for a female patient who attended Hospital A with that of:

(i) a female patient who attended Hospital B

(ii) a male patient who attended Hospital C.

Solution

(i) According to the model, the force of mortality at time t for a female who attended Hospital A is:

$$\lambda_{female,A}(t) = \lambda_0(t)$$

and the force of mortality at time t for a female who attended Hospital B is:

$$\lambda_{female,B}(t) = \lambda_0(t)e^{-0.025}$$

The ratio of these two quantities is:

$$\frac{\lambda_{female,A}(t)}{\lambda_{female,B}(t)} = e^{0.025} = 1.0253$$

So we estimate that the force of mortality for a female who attended Hospital A is 2.53% higher than that of a female who attended Hospital B.

(ii) Similarly, the force of mortality for a male who attended Hospital C is:

$$\lambda_{male,C}(t) = \lambda_0(t)e^{0.031+0.011} = \lambda_0(t)e^{0.042}$$

So: $$\frac{\lambda_{female,A}(t)}{\lambda_{male,C}(t)} = e^{-0.042} = 0.9589$$

ie we estimate that the force of mortality for a female who attended Hospital A is 4.11% lower than that of a male who attended Hospital C.

In the question above, we had to use 3 dummy variables (*ie* 3 Z's): one for gender (which has 2 categories), and two for hospital (which has 3 categories). In general, for a covariate that has n categories, we will need $n-1$ dummy variables.

The group of lives for whom all the dummy variables are 0 is called the baseline group. In the example above, the baseline group is females who attended Hospital A.

3.3 Summary

Before we look at the mathematics underlying the Cox model, it is useful to summarise the material we have covered so far. Understanding the 'big picture' will help you to understand the mathematics without getting bogged down in the detail.

The Cox model is a popular mathematical model for the analysis of survival data. Although the model cannot help us to identify the *absolute* level of mortality of a population, it can help us to identify the factors that influence the *relative* levels of mortality between members of the population.

Under the Cox model, we assume that each individual's mortality is proportional to some general mortality function, called the baseline hazard. (This is why it is also a *proportional hazards* model.) What makes the Cox model so flexible is that we do not have to make any assumptions about the shape of this baseline hazard before looking at the data. This helps us to avoid the potential pitfall of trying to fit data to an incompatible model ('a square peg in a round hole').

The constant of proportionality for each individual depends on certain measurable quantities called *covariates*. These may be quantitative (*eg* age) or qualitative (*eg* severity of symptoms of a certain illness).

What we don't yet know is to what extent an individual's covariates affect that individual's mortality. This is the subject of the next section.

4 Estimating the regression parameters

The unknown regression parameters provide the link between an individual's (measurable) covariates and the (unknown) level of the individual's mortality. We will now consider how to estimate the regression parameters, $\beta=\left(\beta_1,\beta_2,\dots,\beta_p\right)$.

4.1 The partial likelihood

To estimate β in the Cox model it is usual to maximise the *partial likelihood*. The partial likelihood estimates the regression coefficients but avoids the need to estimate the baseline hazard. Moreover, since (remarkably) it behaves essentially like an ordinary likelihood, it furnishes all the statistical information needed for standard inference on the regression coefficients.

Let $R(t_j)$ denote the set of lives which are at risk just before the jth observed lifetime, and for the moment assume that there is only one death at each observed lifetime, that is $d_j=1 \quad (1\le j\le k)$.

Partial likelihood

The partial likelihood is:

$$L(\beta)=\prod_{j=1}^{k}\frac{\exp(\beta z_j^T)}{\sum_{i\in R(t_j)}\exp(\beta z_i^T)}$$

Intuitively, each observed lifetime contributes the probability that the life observed to die should have been the one out of the $R(t_j)$ lives at risk to die, conditional on one death being observed at time t_j.

So the contribution to the partial likelihood from the first death is the force of mortality for the first life to die divided by the total force of mortality for the lives in the at-risk group just prior to this event.

We can also describe this within the framework of Markov jump processes, which are covered in Chapters 3-5. Suppose that the lives are labelled Life 1, Life 2, …, Life N. Before the first death is observed, the process is in the state where all the lives are still alive and in the at-risk group. Let's call this State 0. Suppose that the first person to die is Life i. When this death occurs, the process jumps into the state where everyone except the Life i is still alive. Let's call this State i.

Then the probability that the process jumps into State i when it leaves State 0 is:

$$\frac{\text{the force of transition from State 0 to State } i}{\text{the total force of transition out of State 0}}$$

Now the force of transition from State 0 to State i is the force of mortality for Life i, and the total force of transition out of State 0 is the sum of the forces of mortality for everyone in the at-risk group. So if Life i were observed to die at age x, the contribution made to the partial likelihood in respect of this death would be:

$$\frac{\mu_i(x)}{\mu_1(x)+\mu_2(x)+\cdots+\mu_N(x)}$$

where $\mu_j(x)$ is the force of mortality for Life j at age x.

A contribution is made to the partial likelihood every time a death is observed, and the partial likelihood is obtained by multiplying all these contributions together.

Note that the baseline hazard cancels out and the partial likelihood depends only on the order in which deaths are observed. (The name 'partial' likelihood arises because those parts of the full likelihood involving the times at which deaths were observed and what was observed between the observed deaths are thrown away.)

The form of the partial likelihood gives the *comparative* risk of a *particular* individual dying, given that a death occurs.

For example, if the first life to die was the tallest individual in the population and the i th covariate is height, then we may infer that height has a significant influence on mortality. In terms of the Cox model, we may infer that the value of β_i is positive.

Of course, our inferences should be based on *all* the observed deaths. By maximising this partial likelihood, our estimates of the regression parameters will be based on the *order* in which the deaths occurred. After all, the model seeks to identify the factors that influence mortality rates and hence increase or reduce the chance of an untimely death.

Question

A group of six lives was observed over a period of time as part of a mortality investigation. Each of the lives was under observation at all ages from age 55 until they died or were censored. The table below shows the sex, age at exit and reason for exit from the investigation.

Life	Sex	Age at exit	Reason for exit
1	M	56	death
2	F	62	censored
3	F	63	death
4	M	66	death
5	M	67	censored
6	M	67	censored

The following model has been suggested for the force of mortality:

$$\mu(x \mid Z=z)=\mu_0(x)e^{\beta z}$$

where:

- x denotes age
- $\mu_0(x)$ is the baseline hazard
- $z=0$ for males and $z=1$ for females.

Write down the partial likelihood for these observations using the model above.

Solution

Since there are three ages at which deaths occur, the partial likelihood will be the product of three terms – one in respect of each death.

The contribution to the partial likelihood from the first death is:

$$\frac{\mu_1(56)}{\mu_1(56)+\mu_2(56)+\cdots+\mu_6(56)}$$

where $\mu_i(y)$ is the force of mortality of the i th life at age y. In other words, we take the force of mortality for the life that dies at the youngest age and divide it by the total force of mortality for those alive at that age.

Under the given model, this is:

$$\frac{\mu_0(56)}{\mu_0(56)+\mu_0(56)e^{\beta}+\mu_0(56)e^{\beta}+\mu_0(56)+\mu_0(56)+\mu_0(56)}=\frac{1}{4+2e^{\beta}}$$

Similarly, the contribution of the second death to the partial likelihood is:

$$\frac{\mu_3(63)}{\mu_3(63)+\mu_4(63)+\mu_5(63)+\mu_6(63)}=\frac{e^{\beta}}{e^{\beta}+3}$$

Finally, the contribution of the third death to the partial likelihood is:

$$\frac{\mu_4(66)}{\mu_4(66)+\mu_5(66)+\mu_6(66)}=\frac{1}{3}$$

Multiplying these three terms together, we obtain the partial likelihood:

$$L=\frac{1}{4+2e^{\beta}}\times\frac{e^{\beta}}{e^{\beta}+3}\times\frac{1}{3}=\frac{Ce^{\beta}}{(e^{\beta}+2)(e^{\beta}+3)}$$

where C is a constant.

4.2 Maximising the partial likelihood

Maximisation of this expression has to proceed numerically, and most statistics packages have procedures for fitting a Cox model.

Maximisation of this partial likelihood will yield our maximum likelihood estimate of the regression parameters and hence provide a link between measurable covariates and mortality (or hazard) rates. The maximisation process is complicated and often cannot be achieved directly. It may be carried out by an iterative numerical technique such as the Newton-Raphson method, which uses repeated calculations to refine the choice of regression parameters until the maximum is found to a sufficient degree of accuracy.

In the last question, there was only one covariate and the partial likelihood function was:

$$L=\frac{Ce^{\beta}}{(e^{\beta}+2)(e^{\beta}+3)}$$

In this case it is straightforward to work out the maximum likelihood estimate of the parameter β. Taking logs gives:

$$\log L=\log C+\beta-\log(e^{\beta}+2)-\log(e^{\beta}+3)$$

Differentiating with respect to β:

$$\frac{d\log L}{d\beta}=1-\frac{e^{\beta}}{e^{\beta}+2}-\frac{e^{\beta}}{e^{\beta}+3}$$

Setting this equal to 0:

$$\frac{(e^{\beta}+2)(e^{\beta}+3)-(e^{\beta}+3)e^{\beta}-(e^{\beta}+2)e^{\beta}}{(e^{\beta}+2)(e^{\beta}+3)}=0$$

$$\Rightarrow e^{2\beta}+5e^{\beta}+6-e^{2\beta}-3e^{\beta}-e^{2\beta}-2e^{\beta}=0$$

$$\Rightarrow -e^{2\beta}+6=0$$

$$\Rightarrow 2\beta=\log 6$$

$$\Rightarrow \beta=\tfrac{1}{2}\log 6$$

Differentiating the partial log-likelihood a second time (using the quotient rule) gives:

$$\frac{d^2\log L}{d\beta^2}=-\frac{e^{\beta}(e^{\beta}+2)-e^{2\beta}}{(e^{\beta}+2)^2}-\frac{e^{\beta}(e^{\beta}+3)-e^{2\beta}}{(e^{\beta}+3)^2}$$

$$=-\frac{2e^{\beta}}{(e^{\beta}+2)^2}-\frac{3e^{\beta}}{(e^{\beta}+3)^2}<0$$

So the maximum likelihood estimate of β is $\hat{\beta}=\tfrac{1}{2}\log 6=0.896$.

Question

Suppose that from the investigation in the previous question we now have the following additional data:

Life	Sex	Age at exit	Reason for exit
7	M	56	censored
8	F	62	censored

Explain how these extra data values affect the contribution to the partial likelihood from the first death.

Solution

We include both these lives in the at-risk group at age 56, using the same assumption that we met when working with the Kaplan-Meier and Nelson-Aalen models, *ie* that censoring occurs immediately after a death at the same age. So we now have 5 males and 3 females at risk at age 56, and the contribution to the partial likelihood from the first death is $\dfrac{1}{5+3e^{\beta}}$.

In practice there might be ties in the data, that is:

(a) **some $d_j > 1$; or**

(b) **some observations are censored at an observed lifetime.**

It is usual to deal with (b) by including the lives on whom observation was censored at time t_j in the risk set $R(t_j)$, effectively assuming that censoring occurs just after the deaths were observed.

Breslow's approximation

Accurate calculation of the partial likelihood in case (a) is messy, since all possible combinations of d_j deaths out of the $R(t_j)$ at risk at time t_j ought to contribute, and an approximation due to Breslow is often used, namely:

$$L(\beta)=\prod_{j=1}^{k}\frac{\exp(\beta s_j^T)}{\left(\sum_{i\in R(t_j)}\exp(\beta z_i^T)\right)^{d_j}}$$

where s_j is the sum of the covariate vectors *z* of the d_j lives observed to die at time t_j.

So, if two lives – A and B, say – are observed to die at time t_j, we assume that the contribution to the partial likelihood from A's death is $\dfrac{\mu_A(t_j)}{\sum_{i \in R(t_j)} \mu_i(t_j)}$ and the contribution to the partial likelihood from B's death is $\dfrac{\mu_B(t_j)}{\sum_{i \in R(t_j)} \mu_i(t_j)}$. Lives A and B are both included in the at-risk group $R(t_j)$ in each denominator. This is illustrated in the question below.

Question

An investigation was carried out into the survival times (measured in months) of patients in hospital following liver transplants. The covariates are z_{1i} = 0 for placebo, 1 for treatment X, and z_{2i} = weight of patient (measured in *kg*).

The observed lifetimes (with weights in brackets) were as follows:

Placebo	***Treatment X***
3 (83)	6*(58)
9 (68)	11(73)
14 (75)	14(68)
16 (86)	14* (49)

Observations with an asterisk represent censored observations.

Using Breslow's assumption, determine the contribution to the partial likelihood that is made by the deaths at time 14.

Solution

Just before time 14, there were four lives at risk. The total force of mortality for these four lives at time 14 is:

$$\mu_0(14)e^{75\beta_2} + \mu_0(14)e^{\beta_1+68\beta_2} + \mu_0(14)e^{\beta_1+49\beta_2} + \mu_0(14)e^{86\beta_2}$$

where $\mu_0(t)$ denotes the baseline hazard at time t, measured in months since the transplant operation.

The individual forces of mortality for the two lives that die at time 14 are:

$$\mu_0(14)e^{75\beta_2} \quad \text{and} \quad \mu_0(14)e^{\beta_1+68\beta_2}$$

So the contribution to the partial likelihood from the deaths that occur at time 14 is:

$$\frac{\mu_0(14)e^{75\beta_2}}{\mu_0(14)e^{75\beta_2}+\mu_0(14)e^{\beta_1+68\beta_2}+\mu_0(14)e^{\beta_1+49\beta_2}+\mu_0(14)e^{86\beta_2}}$$

$$\times\frac{\mu_0(14)e^{\beta_1+68\beta_2}}{\mu_0(14)e^{75\beta_2}+\mu_0(14)e^{\beta_1+68\beta_2}+\mu_0(14)e^{\beta_1+49\beta_2}+\mu_0(14)e^{86\beta_2}}$$

$$=\frac{\mu_0(14)e^{75\beta_2}\times\mu_0(14)e^{\beta_1+68\beta_2}}{\left[\mu_0(14)e^{75\beta_2}+\mu_0(14)e^{\beta_1+68\beta_2}+\mu_0(14)e^{\beta_1+49\beta_2}+\mu_0(14)e^{86\beta_2}\right]^2}$$

$$=\frac{e^{\beta_1+143\beta_2}}{\left[e^{75\beta_2}+e^{\beta_1+68\beta_2}+e^{\beta_1+49\beta_2}+e^{86\beta_2}\right]^2}$$

since all the baseline hazard terms cancel.

4.3 Properties of the partial likelihood

As mentioned earlier, the partial likelihood behaves much like a full likelihood; it yields an estimator for β which is asymptotically (multivariate) normal and unbiased, and whose asymptotic variance matrix can be estimated by the inverse of the observed information matrix.

Recall that $\tilde{\beta}$ is an unbiased estimator of β if $E(\tilde{\beta})=\beta$. The word 'asymptotically' means as the sample size tends to ∞.

The *efficient score* function, namely the vector function:

$$\boldsymbol{u}(\beta)=\left(\frac{\partial\log L(\beta)}{\partial\beta_1},\dots,\frac{\partial\log L(\beta)}{\partial\beta_p}\right)$$

plays an important part; in particular solving $\boldsymbol{u}(\hat{\beta})=0$ furnishes the maximum likelihood estimate $\hat{\beta}$.

The observed information matrix $I(\hat{\beta})$ is then the negative of the $p\times p$ matrix of second partial derivatives:

$$I(\beta)_{ij}=-\frac{\partial^2\log L(\beta)}{\partial\beta_i\,\partial\beta_j}\qquad(1\le i,j\le p)$$

evaluated at $\hat{\beta}$.

The variance matrix is the symmetric matrix C, whose i, j th entry is equal to $\operatorname{cov}\left(\tilde{\beta}_i, \tilde{\beta}_j\right)$. The above Core Reading is saying that, asymptotically:

$$C = \left[I(\hat{\beta})\right]^{-1} = -\begin{bmatrix} \left.\dfrac{\partial^2 \ln L}{\partial \beta_1^2}\right|_{\beta=\hat{\beta}} & \left.\dfrac{\partial^2 \ln L}{\partial \beta_1 \partial \beta_2}\right|_{\beta=\hat{\beta}} & \cdots & \left.\dfrac{\partial^2 \ln L}{\partial \beta_1 \partial \beta_p}\right|_{\beta=\hat{\beta}} \\ \left.\dfrac{\partial^2 \ln L}{\partial \beta_2 \partial \beta_1}\right|_{\beta=\hat{\beta}} & \left.\dfrac{\partial^2 \ln L}{\partial \beta_2^2}\right|_{\beta=\hat{\beta}} & \cdots & \left.\dfrac{\partial^2 \ln L}{\partial \beta_2 \partial \beta_p}\right|_{\beta=\hat{\beta}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\dfrac{\partial^2 \ln L}{\partial \beta_p \partial \beta_1}\right|_{\beta=\hat{\beta}} & \left.\dfrac{\partial^2 \ln L}{\partial \beta_p \partial \beta_2}\right|_{\beta=\hat{\beta}} & \cdots & \left.\dfrac{\partial^2 \ln L}{\partial \beta_p^2}\right|_{\beta=\hat{\beta}} \end{bmatrix}^{-1}$$

The algebra simplifies considerably when we consider the one-parameter case.

One-parameter case

For a model with only one covariate β, say, we calculate the maximum partial likelihood estimate of β by solving the equation:

$$\frac{d \ln L}{d\beta} = 0$$

We can also estimate the variance of the maximum partial likelihood estimator $\tilde{\beta}$ using the approximation:

$$\operatorname{var}\left(\tilde{\beta}\right) \approx \left.\left(-\frac{d^2 \ln L}{d\beta^2}\right)^{-1}\right|_{\beta=\hat{\beta}}$$

This is the estimated value of the Cramér-Rao lower bound.

Question

For the scenario described in the question in Section 4.1 (without the extra two lives), we have seen that:

$$L = \frac{Ce^{\beta}}{(e^{\beta}+2)(e^{\beta}+3)}$$

$$\frac{d \log L}{d\beta} = 1 - \frac{e^{\beta}}{e^{\beta}+2} - \frac{e^{\beta}}{e^{\beta}+3}$$

and: $$\frac{d^2 \log L}{d\beta^2} = -\frac{2e^{\beta}}{(e^{\beta}+2)^2} - \frac{3e^{\beta}}{(e^{\beta}+3)^2}$$

We have also calculated the value of $\hat{\beta}$ (the maximum likelihood estimate of β) to be $0.5\log 6$.

Use this information to construct an approximate 95% confidence interval for β and explain what can be inferred from this.

Solution

If $\tilde{\beta}$ is the maximum partial likelihood estimator of β, then the asymptotic variance of $\tilde{\beta}$ is given by:

$$\text{var}\left(\tilde{\beta}\right)=\left[I(\beta)\right]^{-1}_{\beta=\hat{\beta}}$$

$$=\left[\frac{2e^{\hat{\beta}}}{(e^{\hat{\beta}}+2)^2}+\frac{3e^{\hat{\beta}}}{(e^{\hat{\beta}}+3)^2}\right]^{-1}$$

$$=\left[\frac{2\sqrt{6}}{(\sqrt{6}+2)^2}+\frac{3\sqrt{6}}{(\sqrt{6}+3)^2}\right]^{-1}$$

$$=2.02062$$

So the asymptotic standard error is $\sqrt{2.02062}=1.4215$.

As $\tilde{\beta}$ is asymptotically normally distributed, an approximate 95% confidence interval for β is:

$$\hat{\beta}\pm 1.96\sqrt{\text{var}\left(\tilde{\beta}\right)}=\tfrac{1}{2}\ln 6\pm\left(1.96\times 1.4215\right)=\left(-1.890,\ 3.682\right)$$

Since this interval contains the value 0, we conclude on the basis of these data values that sex is not a significant covariate.

A useful feature of most computer packages for fitting a Cox model is that the information matrix evaluated at $\hat{\beta}$ is usually produced as a by-product of the fitting process (it is used in the Newton-Raphson algorithm) so standard errors of the components of $\hat{\beta}$ are available. These are helpful in evaluating the fit of a particular model.

5 Model fitting

5.1 Assessing the effect of the covariates

In a practical problem, several possible explanatory variables might present themselves, and part of the modelling process is the selection of those that have significant effects. Therefore criteria are needed for assessing the effects of covariates, alone or in combination.

A common criterion is the *likelihood ratio statistic*. Suppose we need to assess the effect of adding further covariates to the model. In general, suppose we fit a model with p covariates, and another model with $p+q$ covariates, which include the p covariates of the first model.

Each is fitted by maximising a likelihood; let L_p and L_{p+q} be the maximised log-likelihoods of the first and second models respectively.

Likelihood ratio test

The likelihood ratio statistic is then:

$$-2(L_p - L_{p+q})$$

and it has an asymptotic χ^2 distribution on q degrees of freedom, under the hypothesis that the extra q covariates have no effect in the presence of the original p covariates.

This result is given on page 23 of the *Tables*.

The null hypothesis for the likelihood ratio test is:

$$H_0 : \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0$$

ie the extra covariates are not significant.

The test statistic is:

$$-2\left(\ln L_p - \ln L_{p+q}\right)$$

where the log-likelihoods are evaluated using the optimised parameter values, *ie* the maximum partial likelihood estimates. (The Core Reading is using L rather than $\ln L$ to denote a *log*-likelihood here.)

The null hypothesis is rejected at the 5% significance level if the value of the test statistic is greater than the upper 5% point of χ_q^2.

This likelihood ratio test is a one-tailed test, as adding extra covariates to a model will increase the log-likelihood and hence improve the fit. This means that the value of the test statistic will always be positive.

If the value of the test statistic is small, then adding in the extra covariates does not improve the fit very much. If, however, the value of the test statistic is large (*ie* greater than the upper 5% point of χ_q^2), then we conclude that the inclusion of the extra covariates significantly improves the fit and so is worthwhile.

Strictly this statistic is based upon full likelihoods, but when fitting a Cox model it is used with partial likelihoods.

For example, suppose we have considered a model for the effect of hypertension on survival, in which z_i has two components, with the level of $z_i^{(1)}$ representing sex and the level of $z_i^{(2)}$ representing blood pressure.

Suppose we want to test the hypothesis that cigarette smoking has no effect, allowing for sex and blood pressure.

Then we could define an augmented covariate vector $z_i' = (z_i^{(1)}, z_i^{(2)}, z_i^{(3)})$ in which $z_i^{(3)}$ is a factor (say, 0 for non-smoker and 1 for smoker) and refit the model.

The likelihood ratio statistic $-2(L_2 - L_3)$ then has an asymptotic χ^2 distribution on 1 degree of freedom, under the null hypothesis (which is that the new parameter $\beta_3 = 0$).

In practice, the likelihood ratio statistic would be calculated numerically using a statistical computer package.

5.2 Building models

The likelihood ratio statistic is the basis of various model-building strategies, in which:

(a) we start with the *null model* (one with no covariates) and add possible covariates one at a time; or

(b) we start with a *full model* which includes all possible covariates, and then try to eliminate those of no significant effect.

In addition, it is necessary to test for *interactions* between covariates, in case their effects should depend on the presence or absence of each other in the same way as described in Subject CS1.

The likelihood ratio statistic is a standard tool in model selection; for example it was used in the UK to choose members of a Gompertz-Makeham family of functions for parametric graduations (see Chapter 11).

We can extend a model to test for *interactions* between covariates.

For example, suppose that a study is carried out to ascertain the link between the mortality of pensioners and socio-economic group. The survival times are to be modelled using a Cox regression model, which is to include allowance for two other influences on mortality – sex and smoking status. The model is to be used to test for two-way interaction between socio-economic group and the other factors.

The model might be specified as:

$$\lambda(x;z_i)=\lambda_0(x)\exp(\beta z_i^T)=\lambda_0(x)\exp\left(\sum_{j=1}^{p}\beta_j\, z_{ij}\right)$$

where x denotes age, and the covariates of the model for the i th life are:

z_{i1} = socio-economic group from 0 (low) to 4 (high)

z_{i2} = sex (0 for male, 1 for female)

z_{i3} = smoking status (0 for smoker, 1 for non-smoker)

$z_{i4}=z_{i1}\times z_{i2}$

$z_{i5}=z_{i1}\times z_{i3}$

We want to test for interaction between socio-economic group and the other factors. The null hypothesis for this test is:

$H_0:\beta_4=\beta_5=0$, *ie* there is no interaction

To perform this test, we would fit a model with the first 3 covariates (z_{i1},z_{i2},z_{i3}) and another model with all 5 covariates $(z_{i1},\dots,z_{i5})$. Each model is fitted by maximising the partial likelihood, using an appropriate statistics package. Let $\ln L_3$ and $\ln L_5$ be the maximised log-likelihoods of the 3-parameter and 5-parameter models respectively.

The likelihood ratio statistic is then $-2(\ln L_3-\ln L_5)$. Under the null hypothesis, this has an asymptotic χ_2^2 distribution. If the likelihood ratio statistic exceeds the upper 5% point of χ_2^2, then the null hypothesis should be rejected.

In the R package `survival`, the command `coxph()` fits a Cox proportional hazards model to the supplied data.

R code:

```
coxph(formula)
```

The argument `formula` will be similar to that used when fitting a linear model via `lm()` (see Subject CS1) except that the response variable will be a survival object instead of a vector.

5.3 Using the results

After fitting the model and analysing the likelihood ratio statistics, we can make inferences about how each covariate affects mortality. This information can be used in many different ways:

- The model may be used to assess the efficacy of a new medical treatment for patients. The treatment would be represented by a covariate, which may be a quantitative measure of dose or an indicator, *eg* 0 for placebo, 1 for treatment.
- A life insurance company may wish to know how certain covariates affect mortality, so that it can charge premiums that accurately reflect the risk for an individual, *eg* higher premiums for smokers. However, an insurance company will be restricted to covariates that can be collected easily and reliably from potential policyholders. (We will return to this idea in Chapter 9, when we discuss heterogeneity within a population.)

The Cox model can provide an estimate of the relative level of an individual's mortality in comparison to the baseline hazard. By making certain assumptions about the shape and level of the baseline hazard, we can then estimate the absolute level of an individual's mortality.

The chapter summary starts on the next page so that you can keep all the chapter summaries together for revision purposes.

Chapter 8 Summary

Covariates

A covariate is any quantity recorded in respect of each life, such as age, sex, type of treatment, level of medication, severity of symptoms and so on.

Proportional hazards (PH) models

In a proportional hazards model the hazard function for the i th life, $\lambda_i(t;z_i)$, may be written as:

$$\lambda_i(t;z_i) = \lambda_0(t)g(z_i)$$

The baseline hazard $\lambda_0(t)$ is a function *only* of the duration t and $g(z_i)$ is a function *only* of the covariate vector z_i.

The hazards of different lives are in the same proportion at all times. This proportion depends on the values of the covariates recorded for each life, but not on the baseline hazard.

Fully parametric models

Fully parametric models assume a lifetime distribution based on a statistical distribution whose parameters must then be determined.

Commonly used distributions include:

- the exponential distribution (constant hazard)
- the Weibull distribution (monotonic hazard)
- the Gompertz-Makeham formula (exponential hazard)
- the log-logistic distribution (‘humped’ hazard).

The Cox PH model

The Cox model is a semi-parametric proportional hazards model under which the force of mortality (or hazard function) for an individual life is given by:

$$\lambda(t;z_i) = \lambda_0(t)\exp(\beta z_i^T)$$

The force of mortality is proportional to the baseline hazard $\lambda_0(t)$.

The Cox model is a proportional hazards model because the hazards of different lives are in the same proportion at all times. This proportion depends on the values of the covariates recorded for each life, but not on the baseline hazard

This proportion depends on the values of the covariates recorded for each life and the values of the regression parameters β :

$$\frac{\lambda(t;z_1)}{\lambda(t;z_2)}=\frac{\exp(\beta z_1^T)}{\exp(\beta z_2^T)}=\text{constant}$$

It can be used to investigate the effect of different factors on mortality. The data collected for each life in the investigation must include information about the covariates, which may be qualitative or quantitative.

Fitting the regression parameters

The regression parameters are estimated by maximising the partial likelihood:

$$L(\beta)=\prod_{j=1}^{k}\frac{\exp(\beta z_j^T)}{\sum\limits_{i\in R(t_j)}\exp(\beta z_i^T)}$$

Solving the equation:

$$u(\beta)=\left(\frac{\partial\log L(\beta)}{\partial\beta_1},\dots,\frac{\partial\log L(\beta)}{\partial\beta_p}\right)=0$$

gives the maximum partial likelihood estimates of $\beta_1,\beta_2,\dots,\beta_p$. We denote these estimates by $\hat{\beta}_1,\hat{\beta}_2,\dots,\hat{\beta}_p$. The maximisation procedure is usually carried out using a computer.

Breslow's approximation to the partial likelihood

If there are ties in the data, *ie* the death times are not distinct, then Breslow's approximation to the partial likelihood can be used:

$$L(\beta)=\prod_{j=1}^{k}\frac{\exp(\beta s_j^T)}{\left(\sum\limits_{i\in R(t_j)}\exp(\beta z_i^T)\right)^{d_j}}$$

Distribution of the maximum partial likelihood estimators of the regression parameters

The maximum partial likelihood estimator of the vector of parameters β, which we denote by $\tilde{\beta}$, has the following asymptotic properties:

- It has an asymptotic multivariate normal distribution
- It is asymptotically unbiased
- Its variance matrix is equal to the inverse of the observed information matrix, *ie* the inverse of the negative of the matrix of second derivatives of the log-likelihood, evaluated at the point $\hat{\beta}$.

So an approximate 95% confidence interval for β_j (the j th parameter) is:

$$\hat{\beta}_j \pm 1.96\sqrt{\text{var}\left(\tilde{\beta}_j\right)}$$

Model testing

We can compare two models using a likelihood ratio test. Suppose we want to compare a model with p covariates against an extended model with an extra q covariates.

The null hypothesis for this test is:

$$H_0 : \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0$$

The test statistic is:

$$-2\left(\ln L_p - \ln L_{p+q}\right)$$

where $\ln L_p$ and $\ln L_{p+q}$ denote the maximised log-likelihoods of the models with p and $p+q$ covariates, respectively.

If the null hypothesis is true, then the test statistic should be a realisation of a χ_q^2 random variable. So we reject the null hypothesis at the 5% significance level if the value of the test statistic is greater than the upper 5% point of χ_q^2.

The practice questions start on the next page so that you can keep the chapter summaries together for revision purposes.

Chapter 8 Practice Questions

8.1 You want to use a Cox regression model to estimate the force of mortality for a group of endowment assurance policyholders. You propose using a model that takes account of duration (*ie* the time that has elapsed since the policy was issued) and the age and sex of the policyholder. You start by investigating the model:

$$\mu(x,z_1,z_2)=\mu_0(x)e^{\beta_1 z_1+\beta_2 z_2}$$

where :

x denotes the age of the policyholder

$$z_1=\begin{cases}0 & \text{if the duration is less than 1 year}\\ 1 & \text{if the duration is at least 1 year}\end{cases}$$

$$z_2=\begin{cases}0 & \text{for males}\\ 1 & \text{for females}\end{cases}$$

You have estimated the values of the parameters β_1 and β_2, and have obtained the following results:

Covariate	**Parameter**	**Standard error**
Duration	0.416	0.067
Sex	–0.030	0.017

(i) State the class of policyholders to which the baseline hazard refers.

(ii) Explain whether the duration covariate is significant in determining mortality.

(iii) Compare the force of mortality for a new female policyholder to that of a male policyholder of the same age, who took out a policy 2 years ago.

8.2 Exam style (i) Explain what is meant by a proportional hazards model. [3]

(ii) Outline three reasons why the Cox proportional hazards model is widely used in empirical work. [3]

[Total 6]

8.3 Exam style

The Cox proportional hazards model is to be used to model the rate at which students leave a certain profession before qualification. Assuming they stay in the profession, students will qualify three years after joining the profession. In the fitted model, the hazard depends on the time, t, since joining the profession and three covariates. The covariates, their categories and the fitted parameters for each category are shown in the table below:

Covariate	Possibility	Parameter
Size of employer	Large	0
	Small	0.4
Degree studied	None	0.3
	Science	–0.1
	Arts	0.2
	Other	0
Location	London	0
	Other UK	–0.3
	Overseas	0.4

(i) Defining clearly all the terms you use, write down an expression for the hazard function in this model. [3]

(ii) State the class of students that is most likely to proceed to qualification under this model, and that which is least likely. [2]

(iii) A student who has been in the profession for one year moves from a 'small' employer to a 'large' employer. Express the probability that he will qualify with the 'large' employer P_L in terms of the probability that he would have qualified if he had stayed with the 'small' employer P_S, all other factors being equal. [2]

[Total 7]

8.4 Exam style

A study has been undertaken into the effect of a new treatment on the survival times of patients suffering from a tropical disease. The following model has been fitted:

$$h_i(t) = h_0(t)\ \exp(\underline{\beta}^T \underline{z})$$

where $h_i(t)$ is the hazard at time *t*, where *t* is the time since treatment

$h_0(t)$ is the baseline hazard at time *t*

$\underline{z}$ is a vector of covariates, where:

z_1 = period from diagnosis to treatment in years

z_2 = 0 if existing treatment given, 1 if new treatment given

z_3 = 0 if female, 1 if male

$\underline{\beta}$ is a vector of parameters, where:

$\beta_1 = 0.5$

$\beta_2 = 0.01$

$\beta_3 = -0.05$

(i) State the group of lives to which the baseline hazard applies. [1]

(ii) For a male who was given the new treatment 6 months after diagnosis:

(a) Write down the hazard function, in terms of $h_0(t)$ only.

(b) Express the survival function, in terms of $h_0(t)$ only. [3]

(iii) For a female given the new treatment at the time of diagnosis, the probability of survival for 5 years is 0.75. Calculate the probability that the male in (ii) will survive 5 years. [3]

[Total 7]

8.5 Exam style

(i) Compare the advantages and disadvantages of fully parametric models and the Cox regression model for assessing the impact of covariates on survival. [3]

You have been asked to investigate the impact of a set of covariates, including age, sex, smoking, region of residence, educational attainment and amount of exercise undertaken, on the risk of heart attack. Data are available from a prospective study which followed a set of several thousand persons from an initial interview until their first heart attack, or until their death from a cause other than a heart attack, or until 10 years had elapsed since the initial interview (whichever of these occurred first).

(ii) State the types of censoring present in this study, and explain how each arises. [2]

(iii) Describe a criterion which would allow you to select those covariates which have a statistically significant effect on the risk of heart attack, when controlling the other covariates of the model. [4]

Suppose your final model is a Cox model which has three covariates: age (measured in age last birthday minus 50 at the initial interview), sex (male = 0, female = 1) and smoking (non-smoker = 0, smoker = 1), and that the estimated parameters are:

Age	0.01
Sex	–0.4
Smoking	0.5
Sex × smoking	–0.25

where 'sex × smoking' is an additional covariate formed by multiplying the two covariates 'sex' and 'smoking'.

(iv) Describe the final model's estimate of the effect of sex and of smoking behaviour on the risk of heart attack. [3]

(v) Use the results of the model to determine how old a female smoker must be at the initial interview to have the same risk of heart attack as a male non-smoker aged 50 years at the initial interview. [3]

[Total 15]

Chapter 8 Solutions

8.1 (i) ***Class of policyholders to which baseline hazard refers***

The baseline hazard refers to male endowment assurance policyholders, who took out their policies less than one year ago.

(ii) ***Is duration significant?***

An approximate 95% confidence interval for the duration parameter is:

$$0.416 \pm (1.96 \times 0.067) = (0.285,\ 0.547)$$

As this interval does not contain 0, we conclude that the duration covariate is significant in determining mortality.

(iii) ***Comparison of forces of mortality***

According to the model, the force of mortality for a new female policyholder aged x is $\mu_0(x)e^{-0.030}$; the force of mortality for a male policyholder at the same age who took out his policy 2 years ago is $\mu_0(x)e^{0.416}$. Since:

$$\frac{\mu_0(x)e^{-0.030}}{\mu_0(x)e^{0.416}} = e^{-0.446} = 0.640$$

the model implies that the force of mortality for the female is 36% less than the force of mortality for the male.

You could also say that the force of mortality for the male is 56% higher than the force of mortality for the female.

8.2 *This is Subject CT4, April 2015, Question 3.*

(i) ***Proportional hazards models***

Proportional hazards models are used to describe the hazard rate of individuals where this depends on both duration (the time since a specified event) and other covariates. [½]

The hazard rate for each individual consists of a baseline hazard, which is a component that depends only on the duration, multiplied by a function that depends only on the values of the covariates for the individual. [1]

The model is ‘proportional’ because the hazard rate for each individual always remains in the same proportion to the baseline hazard (and hence also to other individuals). [1]

The baseline hazard rate corresponds to an individual with all covariates equal to zero. [½]

(ii) ***Advantages of the Cox model***

The Cox regression model allows us to compare individuals with different covariates (*eg* males and females) without needing to consider the form of the baseline hazard rates. [1]

The Cox model is a commonly used model and reliable software is available for carrying out the required calculations. [½]

The exponential function ensures that the hazard rate is always positive. [½]

It is a semi-parametric model, so the baseline hazard rate does not need to be specified in advance. [1]

8.3 *This is Subject 104, April 2003, Question 5.*

(i) ***Hazard function***

The hazard function for leaving the profession is given by:

$$\lambda(t,\mathbf{Z})=\lambda_0(t)\exp\left[0.4Z_1+0.3Z_2-0.1Z_3+0.2Z_4-0.3Z_5+0.4Z_6\right] \qquad [1]$$

where:

$\lambda_0(t)=$ baseline hazard at time *t* since entry into profession

$\mathbf{Z}=\left(Z_1,Z_2,Z_3,Z_4,Z_5,Z_6\right)$

$Z_1=$ 1 if small employer, 0 if not

$Z_2=$ 1 if no degree, 0 if not

$Z_3=$ 1 if science degree, 0 if not

$Z_4=$ 1 if arts degree, 0 if not

$Z_5=$ 1 if location = UK except London, 0 if not

$Z_6=$ 1 if location = overseas, 0 if not [2]

(ii) ***Most and least likely to qualify***

The students most likely to qualify are those with the lowest hazard function, *ie* those for which $Z_1=0$, $Z_2=0$, $Z_3=1$, $Z_4=0$, $Z_5=1$ and $Z_6=0$. So the students most likely to qualify are those who work for large employers, have science degrees and work in the UK but outside London. [1]

The least likely to qualify are those for which $Z_1=1$, $Z_2=1$, $Z_3=0$, $Z_4=0$, $Z_5=0$ and $Z_6=1$, *ie* those who work for small employers, have no degrees and who work overseas. [1]

(iii) ***Probability of qualifying***

The probability that a student who has been in the profession for one year will qualify is:

$$\exp\left(-\int_1^3\lambda(t,\mathbf{z})\,dt\right)$$

We can think of this as the probability that the student will 'survive', ie avoid leaving the profession, from time 1 to time 3.

If the student works for a large employer, the probability is:

$$P_L = \exp\left(-\int_1^3 \lambda_0(t) e^{0.3Z_2 - 0.1Z_3 + 0.2Z_4 - 0.3Z_5 + 0.4Z_6}\, dt\right)$$

$$= \exp\left[-e^{0.3Z_2 - 0.1Z_3 + 0.2Z_4 - 0.3Z_5 + 0.4Z_6} \int_1^3 \lambda_0(t)\, dt\right] \qquad [½]$$

If the student works for a small employer, the probability is:

$$P_S = \exp\left(-\int_1^3 \lambda_0(t) e^{0.4 + 0.3Z_2 - 0.1Z_3 + 0.2Z_4 - 0.3Z_5 + 0.4Z_6}\, dt\right)$$

$$= \exp\left[-e^{0.4 + 0.3Z_2 - 0.1Z_3 + 0.2Z_4 - 0.3Z_5 + 0.4Z_6} \int_1^3 \lambda_0(t)\, dt\right]$$

$$= \exp\left[-e^{0.4} e^{0.3Z_2 - 0.1Z_3 + 0.2Z_4 - 0.3Z_5 + 0.4Z_6} \int_1^3 \lambda_0(t)\, dt\right]$$

$$= \left\{\exp\left[-e^{0.3Z_2 - 0.1Z_3 + 0.2Z_4 - 0.3Z_5 + 0.4Z_6} \int_1^3 \lambda_0(t)\, dt\right]\right\}^{\exp(0.4)}$$

$$= \left(P_L\right)^{\exp(0.4)} \qquad [1]$$

The second last equality follows from the result $e^{AB} = \left(e^B\right)^A$.

So: $P_L = \left(P_S\right)^{\exp(-0.4)} = \left(P_S\right)^{0.67032}$ [½]

8.4 *This is Subject 104, September 2004, Question 3.*

(i) ***Group of lives to which baseline hazard applies***

Lives who are:

- treated immediately following diagnosis, $z_1 = 0$
- who receive the existing treatment, $z_2 = 0$
- who are female, $z_3 = 0$. [1]

(ii)(a) ***Hazard function for male life who received the new treatment six months after diagnosis***

We use the model parameters we are given, together with the values of the regression variables for this life:

$$z_1 = \frac{1}{2}\text{ year}, \quad z_2 = 1 \text{ for the new treatment} \quad z_3 = 1 \text{ for a male life} \qquad [1]$$

Then:

$$h(t)=h_0(t)\exp\left\{0.5\times\frac{1}{2}+0.01\times1-0.05\times1\right\}=h_0(t)e^{0.21}$$ [1]

(ii)(b) ***Survival function for male life who received the new treatment six months after diagnosis***

The survival function is:

$$S(t)=\exp\left\{-\int_{s=0}^{t} h(s)ds\right\}$$

$$=\exp\left\{-\int_{s=0}^{t} h_0(s)e^{0.21}ds\right\}$$

$$=\exp\left\{-e^{0.21}\int_{s=0}^{t} h_0(s)ds\right\}$$ [1]

(iii) ***Probability that the life in (ii) will survive for five years***

We use the information given about the female life to determine an expression for the baseline hazard. We can then use this expression to evaluate the probability for the male life.

For a female life given the new treatment at the time of diagnosis we can write:

$$h_f(t)=h_0(t)\exp\{0.5\times0+0.01\times1-0.05\times0\}=h_0(t)e^{0.01}$$ [½]

Then:

$$S_f(5)=\exp\left\{-\int_{s=0}^{5} h(s)ds\right\}=\exp\left\{-\int_{s=0}^{5} h_0(s)e^{0.01}ds\right\}$$

$$=\exp\left\{-e^{0.01}\int_{s=0}^{5} h_0(s)ds\right\}=\left(\exp\left\{-\int_{s=0}^{5} h_0(s)ds\right\}\right)^{e^{0.01}}$$

$$=0.75$$ [1]

Rearranging this result gives:

$$\exp\left\{-\int_{s=0}^{5} h_0(s)ds\right\}=(0.75)^{e^{-0.01}}$$ [½]

Then using the result from (ii)(b) for $t=5$ we can write:

$$S_m(5)=\exp\left\{-e^{0.21}\int_{s=0}^{5}h_0(s)\,ds\right\}=\left(\exp\left\{-\int_{s=0}^{5}h_0(s)\,ds\right\}\right)^{e^{0.21}}$$ [½]

Finally substitution gives:

$$S_m(5)=\left(\exp\left\{-\int_{s=0}^{5}h_0(s)\,ds\right\}\right)^{e^{0.21}}=\left((0.75)^{e^{-0.01}}\right)^{e^{0.21}}=(0.75)^{e^{0.20}}=0.7037$$ [½]

8.5 *This is Subject CT4, September 2007, Question 10.*

(i) ***Fully parametric models versus Cox regression model***

The Cox regression model is an example of a semi-parametric approach, in which we do not pre-constrain the precise form of the hazard function. It has been the most widely used regression model in recent years and is an example of a proportional hazards model. [1]

Parametric models can be used with a homogeneous population or can be fitted to a moderate number of homogeneous groups, in which case confidence intervals for the fitted parameters give a test of differences between the groups which should be better than non-parametric procedures. [1]

However, fully parametric models are difficult to apply without foreknowledge of the form of the hazard function, which might be the very object of the study. For this reason a semi-parametric approach can be more popular. [1]

(ii) ***Censoring present in this study***

Right censoring and Type I censoring are present at the end of the investigation. [1]

Random censoring is present since death from a cause other than heart attack can occur at any time. [1]

(iii) ***Criterion – likelihood ratio test***

A common criterion is the *likelihood ratio test*. Suppose we need to assess the effect of adding further covariates to the model. For example, suppose we fit a model with p covariates, and another model with $p+q$ covariates (which include the p covariates of the first model). [1]

Each model is fitted by maximising a likelihood. Let $\ln L_p$ and $\ln L_{p+q}$ be the maximised log-likelihoods of the first and second models respectively. [½]

The null hypothesis for this test is:

$$H_0:\beta_{p+1}=\beta_{p+2}=\cdots=\beta_{p+q}=0$$

ie the extra covariates are not significant. [1]

The likelihood ratio statistic is:

$$-2\left(\ln L_p - \ln L_{p+q}\right)$$

where the log-likelihoods are calculated using the maximum partial likelihood estimates. This has an asymptotic χ^2 distribution, with q degrees of freedom, under the null hypothesis. [1]

The null hypothesis will be rejected at the 5% significance level if the value of the test statistic is greater than the upper 5% point of χ_q^2. [½]

(iv) ***Estimate of the effect of sex and smoking behaviour on the risk of heart attack***

The final model is:

$$\lambda(t;z_i) = \lambda_0(t)\exp(\beta z_i^T) = \lambda_0(t)\exp\left(0.01 z_{i1} - 0.4 z_{i2} + 0.5 z_{i3} - 0.25 z_{i4}\right)$$

where $z_{i4} = z_{i3} \times z_{i2}$.

The value of $\beta_2 = -0.4$ will decrease the hazard function for the i th life if the sex is female, $z_{i2} = 1$. This implies that, according to the model, females have a lower risk of heart attack. [1]

The value of $\beta_3 = 0.5$ will increase the hazard function for the i th life if the smoker status is 'smoker', $z_{i3} = 1$. This implies that, according to the model, smokers have a higher risk of heart attack. [1]

The value of $\beta_4 = -0.25$ will decrease the hazard function for the i th life if the life is both female and a smoker, *ie* if $z_{i4} = 1$. This implies that, according to the model, whilst female smokers have a higher risk of heart attack than female non-smokers, smoking has a much more detrimental effect on males than it does on females. [1]

(v) ***How old a female smoker must be***

A male 50-year old non-smoker has the baseline hazard function:

$$\lambda(t) = \lambda_0(t)\exp\left(0\right) = \lambda_0(t) \qquad [½]$$

A female smoker has the hazard function:

$$\lambda(t) = \lambda_0(t)\exp\left(0.01 z_{i1} - 0.4 + 0.5 - 0.25\right) = \lambda_0(t)\exp\left(0.01 z_{i1} - 0.15\right) \qquad [½]$$

For these two hazard functions to be the same, we require:

$$0.01 z_{i1} - 0.15 = 0$$

ie: $z_{i1} = 15$ [1]

So, according to the model, a female smoker must be 65 to have the same risk of heart attack as a male non-smoker aged 50. [1]

9

Exposed to risk

Syllabus objectives

4.4 Estimate transition intensities dependent on age (exact or census).

4.4.1 Explain the importance of dividing the data into homogeneous classes, including subdivision by age and sex.

4.4.2 Describe the principle of correspondence and explain its fundamental importance in the estimation procedure.

4.4.3 Specify the data needed for the exact calculation of a central exposed to risk (waiting time) depending on age and sex.

4.4.4 Calculate a central exposed to risk given the data in 4.4.3.

4.4.5 Explain how to obtain estimates of transition probabilities.

4.4.6 Explain the assumptions underlying the census approximation of waiting times.

4.4.7 Explain the concept of the rate interval.

4.4.8 Develop census formulae given age at birthday where the age may be classified as next, last, or nearest relative to the birthday as appropriate, and the deaths and census data may use different definitions of age.

4.4.9 Specify the age to which estimates of transition intensities or probabilities in 4.4.8 apply.

0 Introduction

In this chapter we will take a closer look at how to calculate mortality rates from our observed data. This might at first sight appear to be a very simple task. All we need to do is to count the number of deaths at each age occurring during a specified *observation period* and use the estimators derived in earlier chapters of this course to obtain a set of mortality rates for the relevant ages.

Basically, that is all that's involved. However, there are a couple of complications that we need to overcome.

First, the multiple-state and Poisson models are based on the assumption that the force of mortality μ_x is constant over a year of age, whereas we know intuitively that it is not.

The second problem relates to data. It may be that the data that a life insurance company can provide are not classified according to age in precisely the way we would like. If this is the case, we will need to group the data according to an age 'label' appropriate to the form of the available data. In order to estimate mortality rates at different ages, we will need to decide what age is implied by our arbitrary age label. Additionally, the data may be incomplete for the task ahead.

1 Calculating the exposed to risk

We have seen how the central exposed to risk arises in a probabilistic model of mortality.

Recall from Chapter 3 that the central exposed to risk is another name for the total waiting time. This quantity features in both the two-state Markov model and the Poisson model.

In this chapter we consider some problems of a computational nature, concerning the approximation of exposed to risk from incomplete exposure data.

The central exposed to risk (or waiting time) is a very natural quantity, intrinsically observable even if observation may be incomplete in practice – that is, just record the time spent under observation by each life. Note that this is so even if lives are observed for only part of the year of age $[x, x+1]$, for whatever reason.

The central exposed to risk carries through unchanged to arbitrarily complicated multiple-decrement or multiple-state models. As we shall see, it can easily be approximated in terms of the kind of incomplete observations that are typically available in insured lives investigations.

2 Homogeneity

2.1 The problem of heterogeneity

The multiple-state and Poisson models and analyses are based on the assumption that we can observe groups of *identical* lives (or at least lives whose mortality characteristics are the same).

Such a group is said to be *homogeneous.*

In practice, this is never possible.

Even if we were to limit the scope of a mortality investigation to people of a specified age and a specified sex (*eg* females aged 25), there would still be a wide variety of lives – smokers and non-smokers, healthy people and ill people, rocket scientists and actuarial students. A group of lives with different characteristics is said to be *heterogeneous*.

As a result of this heterogeneity, our estimate of the mortality rate would be the estimate of the *average* rate over the whole group of lives. We could use the estimate to predict the rate of mortality for a similar group of lives but it would not provide an accurate estimate of the probability of death for any single individual. This could be a particular problem for an insurance company that wishes to set premiums that accurately reflect the risk of each individual policyholder.

For example, consider a country in which 50% of the population are smokers. If $\mu_{40} = 0.001$ for non-smokers and $\mu_{40} = 0.002$ for smokers, then a mortality investigation based on the entire population may lead us to the estimate $\hat{\mu}_{40} = 0.0015$. An insurance company that calculates its premiums using this average figure would overcharge non-smokers and undercharge smokers.

Question

Comment on the suggestion that although the situation above is inherently unfair, it is of no real consequence to the insurance company since the average premiums will be sufficient to cover the claims.

Solution

A company that charges the same premium rate to lives that present different risks (*ie* to smokers and non-smokers) is in an unstable position. Its premium rate will be based on the aggregate expected risk of its applicants for insurance, assuming a certain mix of high risk and low risk lives. The office will tend to lose low risk business to its competitors if they are charging different premium rates to high and low risk lives, and will itself attract high risk business, so that its aggregate premium rate will be inadequate to meet the actual claim cost. This is called *anti-selection*. The office will then make losses, which will ultimately threaten solvency.

The company can avoid this anti-selection only by charging different premium rates appropriate to the different levels of risk presented by the applicants. This is the process of risk classification. The avoidance of anti-selection is therefore one of its key advantages, leading to improved financial stability for the insurer and a reduced risk of insolvency.

Throughout the course we have acknowledged that mortality varies with age. This is an example of heterogeneity within a population. In this section, we extend the argument by looking briefly at the other factors affecting individual lives that can cause their underlying mortality to differ.

2.2 The solution

We can subdivide our data according to characteristics known, from experience, to have a significant effect on mortality. This ought to reduce the heterogeneity of each class so formed, although much will probably remain.

Among the factors in respect of which life insurance mortality statistics are often sub-divided are:

(a) **Sex**

(b) **Age (as we have assumed throughout)**

(c) **Type of policy (which often reflects the reason for insuring)**

(d) **Smoker/non-smoker status**

(e) **Level of underwriting** (*eg* have they undergone a medical examination?)

(f) **Duration in force.**

Others that might be used are:

(g) **Sales channel**

(h) **Policy size**

(i) **Occupation of policyholder**

(j) **Known impairments**

(k) **Postcode/geographical location**

(l) **Marital status.**

This information will be available from the *proposal form*, which the individual must complete when applying for insurance.

'Known impairments' simply refers to any existing medical conditions that the individual has.

If sufficient data were available, we could use the Cox regression model (Chapter 8) to identify the relevant factors.

Question

Explain how the following factors may influence mortality rates:

(i) sales channel (consider a mailshot to selected existing policyholders and an advert in a popular tabloid national newspaper)

(ii) occupation of policyholder (consider a deep-sea diver, a high-street newspaper vendor and an actuary).

Solution

(i) The sales channel will determine the section of the population targeted by the insurance company. For example, an advert in a popular tabloid national newspaper will typically be read by the lower socio-economic groups within the population. Different sections of society experience very different rates of mortality. The mortality experienced by the lower socio-economic groups within the population is likely to be significantly heavier than existing policyholders who have been selected according to favourable lifestyle and medical history criteria.

(ii) Occupation can influence mortality rates directly (*eg* deep-sea divers suffer a high rate of accidental death whilst performing their job) and indirectly (*eg* actuaries may have access to company medical schemes, which will help to identify and cure medical problems before they become life threatening). Other occupations may only be carried out by a specific subsection of the population (*eg* high-street newspaper vendors may typically be old people whose health prevents them from doing a more active job).

Two key points are:

- **Sub-division cannot be carried out unless the relevant information is collected, generally on the proposal form. Sometimes factors for which there is strong external evidence of an effect on mortality cannot be used because (for example) proposal forms have been kept short for marketing or administrative reasons.**

  Some insurance products are marketed on the strength of the simplicity and brevity of the application process, since some people may be put off by having to provide information relating to their lifestyle and medical history *etc*.

- **Even in quite large investigations, sub-division using many factors results in much smaller populations in each class, making the statistics more difficult. A balance must be struck between obtaining more and more homogeneity, and retaining large enough populations to make analysis possible.**

  The finer the subdivision of the data, the less credible the results of the analysis.

3 The principle of correspondence

Mortality investigations based on estimation of $\mu_{x+½}$ at individual ages must bring together two different items of data: deaths and exposures. It is self-evident that these should be *defined consistently*, or their ratios are meaningless. Care is sometimes needed, however, because these data are often obtained from different sources in the life office. For example, death data might be obtained from the claims file, while exposure data might be obtained from the premium collection file. There is no guarantee that these use the same definition of the policyholders' ages.

In a large insurance company the payment of claims and the collection of premiums will be handled by different departments who may use different databases or computer systems.

A precise statement of what we mean by 'defined consistently' is given by the *principle of correspondence*.

Principle of correspondence

A life *alive* at time t should be included in the exposure at age x at time t if and only if, were that life to die immediately, he or she would be counted in the death data d_x at age x .

This seems almost a triviality, but it is very important and useful.

This means that, when we are calculating crude estimates of mortality rates, we should try to ensure that the age definition used in the numerator (the number of deaths) is the same as the age definition used in the denominator (the exposed to risk).

Although this may seem obvious at first glance, we will see that the principle of correspondence is particularly important when we specify the ages of policyholders by definitions other than 'age last birthday'. Other definitions that may be used include:

- age next birthday
- age nearest birthday.

We will consider some examples of different age definitions later in this chapter.

4 Exact calculation of the central exposed to risk

4.1 Working with complete data

The procedure for the exact calculation of E_x^c is obvious:

- **record all dates of birth**
- **record all dates of entry into observation**
- **record all dates of exit from observation**
- **compute E_x^c.**

If we add to the data above the cause of the cessation of observation, we have d_x as well, and we have finished.

The central exposed to risk E_x^c for a life with age label x is the time from Date A to Date B where:

Date A is the latest of: the date of reaching age label x

the start of the investigation and

the date of entry

Date B is the earliest of: the date of reaching age label $x+1$

the end of the investigation and

the date of exit (for whatever reason)

Question

If the age label is 'age nearest birthday', give the exact age at which a life attains age label x.

Solution

Under this definition, a life attains age label x at exact age $x-½$.

Note that:

- The calculation takes account of *all* movements into and out of the population (not just deaths).
- *All* decrements contribute a fraction of a year in the year of exit and increments contribute a fraction of a year in the year of entry.
- The central exposed to risk is independent of the cause of exit under consideration.
- It is usual to assume an average of 365¼ days in a year in order to convert days of exposure to years.

Although exact exposed to risk calculations are messy to do by hand, they can be done very easily on a computer (*eg* using the date functions on a spreadsheet) if we have the required information for all lives.

Conventions are often needed to define whether the day of entry or day of exit contributes to the total exposed to risk. We do not count both days.

Example

Suppose that a mortality investigation covers the period 1 January 2015 to 31 December 2017. In this investigation, the age label used is 'age last birthday'. The table below gives information about three males involved in the investigation.

Life	Date of birth	Date of joining	Date of exit	Reason for exit
A	25.04.83	01.01.15	30.10.16	Death
B	01.07.83	12.09.16	–	–
C	04.09.82	22.07.17	04.12.17	Withdrawal

We can use these data values to determine the range of dates for which these lives contribute to E_x^c at each age where they make a contribution. We will assume that the day of entry counts in the exposed to risk but the day of exit does not.

Life A joins the investigation at age 31 last birthday. His periods of contribution to the central exposed to risk are as follows:

E_{31}^c	01.01.15 to 24.04.15
E_{32}^c	25.04.15 to 24.04.16
E_{33}^c	25.04.16 to 29.10.16

Life B joins the investigation at age 33 last birthday. So his contributions are:

E_{33}^c	12.09.16 to 30.06.17
E_{34}^c	01.07.17 to 31.12.17

Life C joins the investigation at age 34 last birthday. So his contributions are:

E_{34}^c	22.07.17 to 03.09.17
E_{35}^c	04.09.17 to 03.12.17

Question

Now suppose that we are using the age label ‘age next birthday’. Give the range of dates for which the lives in the table above contribute to E_{34}^c.

Solution

Life A has age label ‘34 next birthday’ from 25.04.16 to 24.04.17. But Life A dies on 30.10.16, so his contribution to E_{34}^c is from 25.04.16 to 29.10.16. (His contribution to the central exposed to risk at age 34 next birthday is the same as his contribution to the central exposed to risk at age 33 last birthday.)

Life B contributes to E_{34}^c from 12.09.16 to 30.06.17.

Life C makes no contribution to E_{34}^c based on the age label ‘age next birthday’.

4.2 Working with incomplete data

All of the remainder of this chapter is about approximate procedures when the data above have *not* been recorded. We will deal with two questions:

- **What happens when the dates of entry to and exit from observation have not been recorded? (Section 5)**
- **What happens if the definition of age does not correspond exactly to the age interval x to $x+1$ (for integer x)? (Section 6)**

5 Census approximations to the central exposed to risk

In this section we will consider how to calculate E_x^c approximately when the exact dates of entry to and exit from observation have not been recorded.

5.1 The available data

Suppose that we have death data of the form:

d_x = total number of deaths age x last birthday during calendar years $K, K+1, \ldots, K+N$

That is, we have observations over $N+1$ calendar years of all deaths between ages x and $x+1$.

However, instead of the times of entry to and exit from observation of each life being known, we have instead only the following *census* data:

$P_{x,t}$ = Number of lives under observation aged x last birthday at time t where $t = 1$ January in calendar years $K, K+1, \ldots, K+N, K+N+1$

This is in fact similar to the way in which data are submitted to the CMI.

The CMI is the Continuous Mortality Investigation, which collects data from insurance companies in order to create standard mortality tables.

It is often quite convenient for companies to submit a total of policies in force on a date such as 1 January.

Companies may not take the time to calculate the number of policies in force every day because this information would be of limited use. However, each insurance company is likely to perform an annual actuarial valuation to assess its financial position. The number of policies in force on the annual valuation date (usually 1 January in the UK) would be calculated and recorded as part of the valuation process.

5.2 The census approximation to E_x^c

Now define $P_{x,t}$ to be the number of lives under observation, aged x last birthday, at *any* time t. Note that:

$$E_x^c = \int_K^{K+N+1} P_{x,t}\,dt$$

During any short time interval $(t, t+dt)$ there will be $P_{x,t}$ lives each contributing a fraction of a year dt to the exposure. So, integrating $P_{x,t} \times dt$ over the observation period gives the total central exposed to risk for this age. In other words, E_x^c is the area under the $P_{x,t}$ 'curve' between $t = K$ and $t = K+N+1$.

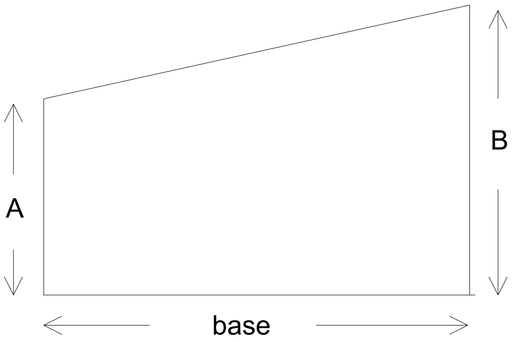
The problem is that we do not know the value of $P_{x,t}$ for all *t*, so we cannot work out the exact value of the integral.

We have the values of $P_{x,t}$ only if t is a 1 January (a *census* date), so we must estimate E_x^c from the given census data. The problem reduces to estimating an integral, given the integrand at just a few points (in this case, integer spaced calendar times). This is a routine problem in numerical analysis.

The simplest approximation, and the one most often used, is that $P_{x,t}$ is *linear* between census dates, leading to the *trapezium* approximation.

The area of a trapezium is:

base × ½ (length of side A + length of side B)



In this case:

- the base of the trapezium is equal to 1, *ie* the period between census dates
- the length of side A is $P_{x,t}$, the number of policies in force at the *start* of the year (at time *t*)
- the length of side B is $P_{x,t+1}$, the number of policies in force at the *end* of the year (at time $t+1$).

Using the trapezium approximation:

$$E_x^c = \int_K^{K+N+1} P_{x,t}\, dt \cong \sum_{t=K}^{K+N} ½(P_{x,t} + P_{x,t+1})$$

This is the method used by the CMI. It is easily adapted to census data available at more or less frequent intervals, or at irregular intervals.

Example

To illustrate how the census approximation works, we will now use it to estimate E_{55}^c from the following data:

Calendar year	Population aged 55 last birthday on 1 January
2015	46,233
2016	42,399
2017	42,618
2018	42,020

Using the notation $P_{55,t}$ to denote the number of lives in the population at time t aged 55 last birthday, and measuring time in years from 1 January 2015, the central exposed to risk at age 55 last birthday for the 3-year period from 1 January 2015 to 1 January 2018 is:

$$E_{55}^c = \int_0^3 P_{55,t}\,dt$$

Now, assuming that $P_{55,t}$ is linear between the census dates, we have:

$$E_{55}^c = \frac{1}{2}\left[P_{55,0} + P_{55,1}\right] + \frac{1}{2}\left[P_{55,1} + P_{55,2}\right] + \frac{1}{2}\left[P_{55,2} + P_{55,3}\right]$$

$$= \frac{1}{2}P_{55,0} + P_{55,1} + P_{55,2} + \frac{1}{2}P_{55,3}$$

$$= \frac{1}{2} \times 46,233 + 42,399 + 42,618 + \frac{1}{2} \times 42,020$$

$$= 129,143.5$$

Question

The disreputable insurance company *Honest Sid's Mutual* had mixed fortunes in the year 2018. At both the start and the end of the year 547 policies were in force in respect of policyholders aged 40 last birthday, but these figures do not tell the whole story.

There was adverse publicity early in the year linking the company's investment managers with a gambling syndicate. As a result, many policyholders 'took their money elsewhere'. Following a successful marketing campaign offering a free toaster to all applicants, the number of policyholders aged 40 last birthday rose from 325 at 1 June 2018 to 613 at 1 September 2018.

Calculate an approximate value for the central exposed to risk at age 40 last birthday for the calendar year 2018.

Solution

Using the notation $P_{40,t}$ to denote the number of policyholders at time t aged 40 last birthday, and measuring time in years from 1 January 2018, the central exposed to risk is:

$$E_{40}^c = \int_0^1 P_{40,t}\,dt$$

We know the numbers of policyholders on 1 January, 1 June, 1 September and 31 December, *ie* at times $0, {}^5\!/_{12}, {}^8\!/_{12}, 1$. Splitting the integral at these times, we have:

$$E_{40}^c = \int_0^{5/12} P_{40,t}\,dt + \int_{5/12}^{8/12} P_{40,t}\,dt + \int_{8/12}^{1} P_{40,t}\,dt$$

Now assuming that $P_{40,t}$ is linear between the census dates:

$$\begin{aligned} E_{40}^c &= \frac{5}{12}\times\frac{1}{2}\left[P_{40,0}+P_{40,{}^5\!/_{12}}\right]+\frac{3}{12}\times\frac{1}{2}\left[P_{40,{}^5\!/_{12}}+P_{40,{}^8\!/_{12}}\right]+\frac{4}{12}\times\frac{1}{2}\left[P_{40,{}^8\!/_{12}}+P_{40,1}\right] \\ &= \frac{5}{24}[547+325]+\frac{3}{24}[325+613]+\frac{4}{24}[613+547] \\ &= 492.25 \end{aligned}$$

6 Deaths classified using different definitions of age

In Section 5, we used a definition of age 'x last birthday', which identifies the year of age $[x, x+1]$.

All this is saying is that if someone is aged x last birthday, then their actual age is somewhere between x and $x+1$.

Other definitions could be used, for example:

$d_x^{(2)}$ = total number of deaths age x *nearest* birthday during calendar years $K, K+1, \ldots, K+N$

$d_x^{(3)}$ = total number of deaths age x *next* birthday during calendar years $K, K+1, \ldots, K+N$

Each of these identifies a different year of age, called the *rate interval*.

Rate interval

A *rate interval* is a period of one year during which a life's recorded age remains the same, *eg* the period during which an individual is 'aged 36 last birthday' or 'aged 42 nearest birthday'.

The key concept is that lives carry the same age label throughout a rate interval. Given this, it follows that the rate interval starts on the day when a life's age label changes.

For example, if the age label is 'age nearest birthday', a life will go from 'age 42 nearest birthday' to 'age 43 nearest birthday' 6 months before the life's 43rd birthday, *ie* at exact age 42½. In this case, the age label changes halfway between birthdays.

Consequently, estimates of μ, or q based on μ, obtained from these data ($d_x^{(2)}$ and $d_x^{(3)}$) will not be estimates of $\mu_{x+½}$ or q_x, but will be estimates of μ and q at other ages.

μ measures the average instantaneous rate of mortality that we observe over the rate interval and the μ-type rate applies to the age in the middle of the rate interval. In contrast q measures the probability of death over the next year of age or, more generally, over the next rate interval. So the q-type rate applies to the age at the start of the rate interval.

We summarise the possibilities as follows:

Definition of x	Rate interval	$\hat{\mu}$ estimates	$\hat{q}$ estimates
Age last birthday	$[x, x+1]$	$\mu_{x+½}$	q_x
Age nearest birthday	$[x-½, x+½]$	μ_x	$q_{x-½}$
Age next birthday	$[x-1, x]$	$\mu_{x-½}$	q_{x-1}

Age at which estimate applies

Once the rate interval has been identified (from the age definition used in d_x) the rule is that:

- **the crude $\hat{\mu}$ estimates μ in the middle of the rate interval, and**
- **the crude $\hat{q}$ estimates q at the start of the rate interval.**

For example, suppose we have details of the number of deaths aged 40 nearest birthday in a recent mortality investigation. In order for a death to contribute to this total, it must have occurred between ages 39½ and 40½, *ie* in the rate interval $[39½, 40½]$.

We can use the data to estimate the force of mortality at the age at the midpoint of the rate interval, *ie* to estimate μ_{40}. We can then estimate the probability of dying during the rate interval, $q_{39½}$, as $1-e^{-\hat{\mu}}$, where $\hat{\mu}$ is the estimated value of μ_{40}.

Question

Suppose that an investigation into mortality covers the period 1 January 2017 to 1 January 2018. Time is measured in years from 1 January 2017 and $P_x(t)$ denotes the number of lives at time t aged x next birthday. The following data have been recorded for each x:

$d_x =$ number of deaths aged x next birthday

$P_x(0)$ and $P_x(1)$

Explain which values of μ can be estimated using these data values, and how this can be done.

Solution

Since x is defined to be the age next birthday, we have a rate interval that starts at exact age $x-1$ and ends at exact age x. So the exact age in the middle of the rate interval is $x-½$. So $\mu_{x-½}$ is estimated by $\dfrac{d_x}{E_x^c}$, where:

$$E_x^c = \int_0^1 P_x(t)\,dt = \frac{1}{2}\left[P_x(0)+P_x(1)\right]$$

This assumes that $P_x(t)$ is linear over calendar year 2017.

6.1 Consistency between census data and death data

Whatever the definition of age used to classify the lives, we can calculate the exact exposed to risk if we have full information about the dates of entry to and exit from observation. In practice, the information will not be complete but will take the form of census data.

We must ensure that the census data are consistent with the death data. We invoke the principle of correspondence; we must check the following:

The census data $P_{x,t}$ are consistent with the death data d_x if and only if, were any of the lives counted in $P_{x,t}$ to die *on the census date*, he or she would be included in d_x.

The definition of census data corresponding to the rate interval $[x-½, x+½]$ (*ie* corresponding to an age definition of age *nearest* birthday) **is:**

$P^{(2)}_{x,t}$ = **Number of lives under observation, age x nearest birthday at time t, where $t = 1$ January in calendar year $K, K+1, ..., K+N, K+N+1$**

and the definition of census data corresponding to the rate interval $[x-1, x]$ (*ie* corresponding to an age definition of age *next* birthday) **is:**

$P^{(3)}_{x,t}$ = **Number of lives under observation, age x next birthday at time t, where $t =$ 1 January in calendar year $K, K+1, ..., K+N, K+N+1$**

When different age labels are used for the death data and the census data

In the event that the death data and the census data use different definitions of age, we must adjust the census data. Unless it is unavoidable, we *never* adjust the death data, since that 'carries most information' when rates of mortality are small. Hence it is always the death data that determine what rate interval to use.

For example, the CMI uses the definition 'age *nearest* birthday' in its work; that is, death data as in $d^{(2)}_x$. However, some life offices contribute census data classified by 'age *last* birthday', because that is what is available from their records. The latter must be adjusted in some way.

For example, if we define:

$$P'_{x,t} = ½(P_{x-1,t} + P_{x,t})$$

we can see that $P'_{x,t}$ approximates $P^{(2)}_{x,t}$.

This is because $P^{(2)}_{x,t}$ represents the number of lives under observation, aged x *nearest* birthday at time t. This group comprises all lives between ages $x-½$ and $x+½$. Those between the ages of $x-½$ and x are aged $x-1$ last birthday. Those between ages x and $x+½$ are aged x last birthday. Assuming that birthdays are uniformly distributed over the calendar year, then ½ of the $P^{(2)}_{x,t}$ lives will be aged $x-1$ last birthday and ½ of the $P^{(2)}_{x,t}$ lives will be aged x last birthday.

So $P_{x,t}^{(2)}$ can be approximated by taking the average of:

- the number of lives aged between $x-1$ and x, *ie* the number of policyholders aged $x-1$ *last* birthday at time t given by $P_{x-1,t}$, and
- the number of lives aged between x and $x+1$, *ie* the number of policyholders aged x *last* birthday at time t given by $P_{x,t}$.

This approximation assumes that the birthdays of the individuals involved are spread uniformly over the calendar year, which is usually approximately true.

Question

A mortality investigation covered the period 1 January 2017 to 1 January 2018. Time is measured in years from 1 January 2017 and $P_x(t)$ denotes the number of lives at time t aged x last birthday. The following data were recorded for each x:

$d_x =$ number of deaths aged x next birthday

$P_x(0)$ and $P_x(1)$

(i) Obtain an expression for the central exposed to risk in terms of the available census data that may be used to estimate the force of mortality μ_{x+f}, stating your assumptions.

(ii) Determine the value of f.

Solution

(i) ***Central exposed to risk***

Since the death data and the census data don't match, we define a new census function $P'_x(t)$ that does match the death data, *ie*:

$P'_x(t) =$ number of lives at time t aged x next birthday

Then the central exposed to risk at age x next birthday is:

$$E_x^c = \int_0^1 P'_x(t)\,dt$$

and, assuming that $P'_x(t)$ is linear between the census dates:

$$E_x^c = \frac{1}{2}\left[P'_x(0) + P'_x(1)\right]$$

Now:

$$P'_x(0) = \text{number of lives at time 0 aged } x \text{ next birthday}$$

$$= \text{number of lives at time 0 aged } x-1 \text{ last birthday}$$

$$= P_{x-1}(0)$$

Similarly:

$$P'_x(1) = P_{x-1}(1)$$

So, in terms of the recorded census data:

$$E_x^c = \frac{1}{2}\left[P_{x-1}(0) + P_{x-1}(1)\right]$$

(ii) ***Value of f***

Since x is defined to be the age next birthday, we have a rate interval that ends on the xth birthday. So the exact age at the end of the rate interval is x, and the exact age at the midpoint of the rate interval is $x - ½$. Hence $\frac{d_x}{E_x^c}$ estimates $\mu_{x-½}$. So $f = -½$.

The chapter summary starts on the next page so that you can keep the chapter summaries together for revision purposes.

Chapter 9 Summary

In order to reduce heterogeneity amongst the lives observed in a mortality investigation, we should divide our data into homogeneous subgroups according to characteristics known to have a significant effect on mortality. Factors typically used are:

- sex
- age
- type of policy
- smoker status
- level of underwriting
- duration in force
- sales channel
- policy size
- occupation of policyholder
- known impairments
- postcode/geographical location
- marital status.

This will only be possible if the appropriate information is available and we have sufficient data to make such detailed analysis possible.

The principle of correspondence states that the death data and the exposed to risk must be defined consistently, *ie* the numerator (d_x) and denominator (E_x^c) must correspond.

The exposed to risk can be calculated exactly if we have complete information for every life. In practice we may have only limited information relating to the size of the population at certain dates known as census dates. We can use this information to approximate the exposed to risk.

The data will be classified in terms of a rate interval. A rate interval is a period of one year during which a life has a particular age label.

Central exposed to risk

Suppose that $P_x(t)$ is the number of lives in the investigation at time t with age label x, and the same age classification has been used in both the census data (*ie* the $P_x(t)$ function) and the death data. If we know the values of $P_x(t)$ for $t = K, K+1, K+2, ..., K+N+1$, then:

$$E_x^c = \int_K^{K+N+1} P_x(t)\,dt = \sum_{t=K}^{K+N} \frac{1}{2}\big(P_x(t) + P_x(t+1)\big)$$

assuming $P_x(t)$ varies linearly between the census dates.

We often define the start of the mortality investigation to be time 0, so that $K = 0$.

If the census data and the death data do not match, an adjustment has to be made to the formula above to reflect the difference. In this case, we:

- define a new population function, $P_x'(t)$ say, that uses the same age classification as the death data
- write a formula for the central exposed to risk in terms of $P_x'(t)$
- then work out how to express $P_x'(t)$ in terms of the available census data, *ie* the given values of the $P_x(t)$ function.

Any assumptions should be clearly stated.

Chapter 9 Practice Questions

9.1 Explain the importance of dividing the data for a mortality investigation into homogeneous classes.

9.2 You have been given the following census counts for a population (covering all ages):

$$P_{2016} = \text{Number in population on 1 January 2016} = 20{,}000$$
$$P_{2017} = \text{Number in population on 1 January 2017} = 40{,}000$$
$$P_{2018} = \text{Number in population on 1 January 2018} = 30{,}000$$

Estimate the central exposed to risk (all ages) for this population over each of the following periods, given only the census counts P_{2016}, P_{2017} and P_{2018}. In each case, state any assumptions you have made.

(i) Period: 1 January 2016 to 31 December 2017

(ii) Period: 1 July 2016 to 30 June 2017

(iii) Period: 1 January 2018 to 31 December 2018

(iv) Period: 1 April 2017 to 31 March 2018

9.3 Exam style

A mortality investigation was held between 1 January 2016 and 1 January 2018. The following information was collected. The figures in the table below are the numbers of lives on each census date with the specified age labels.

	Date		
Age last birthday	1.1.16	1.1.17	1.1.18
48	3,486	3,384	3,420
49	3,450	3,507	3,435
50	3,510	3,595	3,540

During the investigation there were 42 deaths at age 49 nearest birthday. Estimate μ_{49} stating any assumptions that you make. [7]

9.4 Exam style

(i) List the data required for the exact calculation of the central exposed to risk of lives aged *x* last birthday in a mortality investigation over the two-year period from 1 January 2016 to 1 January 2018. [2]

(ii) In an investigation of mortality during the period 1 January 2016 to 1 January 2018, data are available on the number of lives under observation, aged x last birthday, on 1 January 2016, 1 July 2016 and 1 January 2018.

Derive an approximation for the central exposed to risk at age x last birthday over the period in terms of the populations recorded on each of these three dates. [3]

[Total 5]

9.5 Exam style

A researcher is studying the mortality rates of older males in a certain population over the calendar years 2016 and 2017. The researcher has obtained the following data:

- the number of males in the population at each age, classified by age next birthday, on 1 April in 2015, 2016, 2017 and 2018
- the number of deaths at each age, classified by age next birthday at the time of death.

You are given the following extract from the data:

Number of males in population

Age next birthday	At 1/4/15	At 1/4/16	At 1/4/17	At 1/4/18
81	6,010	5,980	6,130	6,200
82	5,320	5,310	5,480	5,520
83	5,680	5,800	5,750	6,030
84	5,150	5,230	5,250	5,150

Number of deaths

Age next birthday	In 2016	In 2017
81	354	348
82	375	391
83	430	432
84	442	437

Estimate $\mu_{81.5}$ using these data values. [8]

9.6 Exam style A national mortality investigation is carried out over the calendar years 2015, 2016 and 2017. Data are collected from a number of insurance companies.

Deaths during the period of the investigation, θ_x , are classified by age nearest birthday at death.

Each insurance company provides details of the number of in-force policies on 1 January 2015, 2016, 2017 and 2018, where policyholders are classified by age nearest birthday, $P_x(t)$.

(i) Describe the rate interval being used in this investigation, stating the ages of the lives at the start of the rate interval. [1]

(ii) Derive an expression for the exposed to risk, in terms of $P_x(t)$, which may be used to estimate the force of mortality at each age. State any assumptions you make. [3]

(iii) Describe how your answer to (ii) would change if the census information provided by some companies was $P_x^*(t)$, the number of in-force policies on 1 January each year, where policyholders are classified by age last birthday. [3]

[Total 7]

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 9 Solutions

9.1 If the groups are not homogeneous, any rates derived will be a weighted average of the underlying rates for the different individuals in the group. The weightings may change with time, which will make it very difficult to establish what patterns are emerging.

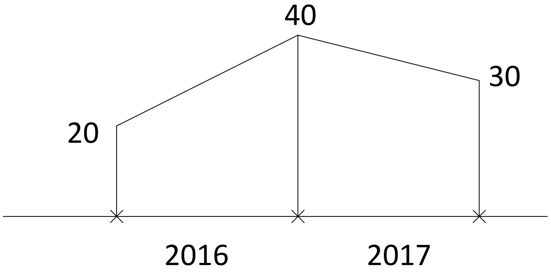
If premiums are calculated based on mortality rates derived from heterogeneous groups, then anti-selection may occur, with the more healthy lives choosing to insure themselves with an office where they will not be charged a premium based on others with an inherent higher level of risk.

9.2 For periods that fall between the census dates, the approach to use is to apply the trapezium rule ($base \times average\, height$). For periods that fall outside the census dates, the simplest approach is to assume that the population size has remained constant.

(i) ***Period: 1 January 2016 to 31 December 2017***

Here we would assume that the population has varied linearly over each calendar year, and we would use the approximation:

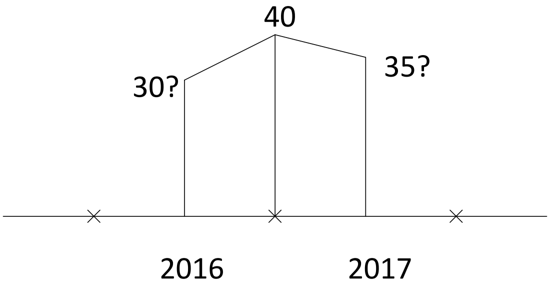
$$\tfrac{1}{2}(P_{2016}+P_{2017})+\tfrac{1}{2}(P_{2017}+P_{2018})=30,000+35,000=65,000$$



(ii) ***Period: 1 July 2016 to 30 June 2017***

Again we would assume that the population has varied linearly over each calendar year. This means that the estimated population sizes in the middle of 2016 and 2017 would be 30,000 and 35,000. Here, the widths of each section are ½ year, so we would use the approximation:

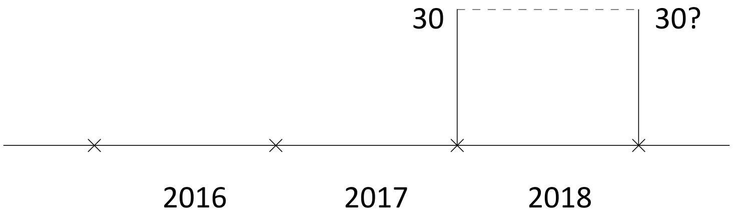
$$\tfrac{1}{2}(30,000+P_{2017})\times\tfrac{1}{2}+\tfrac{1}{2}(P_{2017}+35,000)\times\tfrac{1}{2}=17,500+18,750=36,250$$



(iii) ***Period: 1 January 2018 to 31 December 2018***

In the absence of any additional information about the population size after 1 January 2018, we would have to assume that it remained constant. So we would use the approximation:

$$P_{2018} = 30,000$$

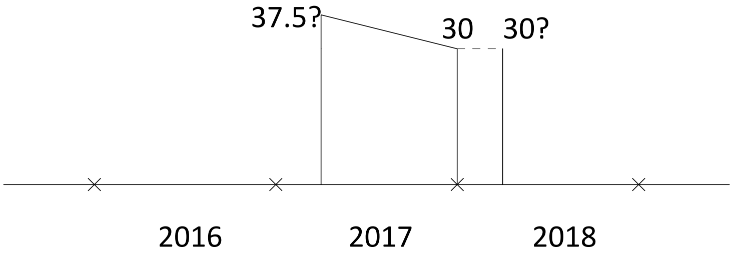


This approach would be unreliable and hence not very satisfactory. For example, the population numbers could continue on a downward 'trend' to 20,000, leading to a central exposed to risk of 25,000, which is a significantly different figure.

(iv) ***Period: 1 April 2017 to 31 March 2018***

Here we would assume that the population has varied linearly over the 2017 calendar year and has then remained constant. This means that the estimated population size on 1 April 2017 would be 37,500. So we would use the approximation:

$$\tfrac{1}{2}(37,500 + P_{2018}) \times \tfrac{3}{4} + P_{2018} \times \tfrac{1}{4} = 25,312.5 + 7,500 = 32,812.5$$



If we had reason to believe that the downward trend seen during 2017 would continue into 2018, we would estimate the population at 31 March 2018 to be 27,500, leading to an estimate of 32,500.

9.3 The deaths are classified according to age nearest birthday. Lives aged 49 nearest birthday are between the exact ages of 48½ and 49½. The age in the middle of this rate interval is 49. So $\dfrac{d_{49}}{E^c_{49}}$ estimates μ_{49} assuming that the force of mortality is constant over the rate interval. [1]

From the investigation, we have:

$$d_{49} = 42$$

If we define $P_x(t)$ to be the number of lives at time t (measured in years from 1 January 2016) aged x last birthday, then we know the values of $P_x(t)$ for $x = 48, 49, 50$ and $t = 0, 1, 2$.

However, the death data and the census data do not match. So we define a function $P'_x(t)$ that does match with the death data.

Let $P'_x(t)$ denote the number of lives at time t aged x nearest birthday. [½]

Then:

$$E^c_{49} = \int_0^2 P'_{49}(t)\,dt \qquad [½]$$

and assuming that $P'_{49}(t)$ varies linearly between the census dates … [½]

$$E^c_{49} = \frac{1}{2}\left[P'_{49}(0) + P'_{49}(1)\right] + \frac{1}{2}\left[P'_{49}(1) + P'_{49}(2)\right]$$

$$= \frac{1}{2}P'_{49}(0) + P'_{49}(1) + \frac{1}{2}P'_{49}(2) \qquad [1]$$

Now we need to determine the numerical values of the terms in the expression above. We have:

$P'_{49}(0) =$ the number of lives at time 0 aged 49 nearest birthday

Lives aged 49 nearest birthday are between the exact ages of 48½ and 49½. So their age last birthday is either 48 or 49. [½]

Assuming that birthdays are uniformly distributed over the calendar year … [½]

$$P'_{49}(0) = \frac{1}{2}\left[P_{48}(0) + P_{49}(0)\right] = \frac{1}{2}\left[3,486 + 3,450\right] = 3,468 \qquad [½]$$

Similarly:

$$P'_{49}(1) = \frac{1}{2}\left[P_{48}(1) + P_{49}(1)\right] = \frac{1}{2}\left[3,384 + 3,507\right] = 3,445.5 \qquad [½]$$

and:

$$P'_{49}(2) = \frac{1}{2}\left[P_{48}(2) + P_{49}(2)\right] = \frac{1}{2}\left[3,420 + 3,435\right] = 3,427.5 \qquad [½]$$

So:

$$E^c_{49} = \frac{1}{2} \times 3,468 + 3,445.5 + \frac{1}{2} \times 3,427.5 = 6,893.25 \qquad [½]$$

and:

$$\hat{\mu}_{49} = \frac{42}{6,893.25} = 0.006093 \qquad [½]$$

9.4 *This is Subject 104, September 2003, Question 3 (with the dates changed).*

(i) ***Data required for exact calculation of central exposed to risk***

For each life observed during the investigation period we need:

- date of birth
- date of joining the investigation (if after 1 January 2016)
- date of leaving the investigation (if before 1 January 2018). [2]

(ii) ***Census formula***

Let $P_x(t)$ denote the number of lives under observation at time t years aged x last birthday, and suppose that time is measured in years from 1 January 2016.

We have recorded the values of $P_x(0), P_x(½)$ and $P_x(2)$, but not $P_x(1)$.

The central exposed to risk for lives aged x last birthday over the two-year investigation period is:

$$E_x^c = \int_0^2 P_x(t)\,dt = \int_0^{½} P_x(t)\,dt + \int_{½}^2 P_x(t)\,dt \qquad [1]$$

Assuming $P_x(t)$ varies linearly between the census dates … [1]

… we have:

$$E_x^c = \frac{1}{2} \times \frac{1}{2}\left[P_x(0) + P_x(½)\right] + \frac{3}{2} \times \frac{1}{2}\left[P_x(½) + P_x(2)\right] = \frac{1}{4}P_x(0) + P_x(½) + \frac{3}{4}P_x(2) \qquad [1]$$

9.5 The rate interval 'age x next birthday' starts at exact age $x-1$ and ends at exact age x. So the exact age in the middle of this rate interval is $x-½$.

If d_x is the number of deaths aged x next birthday during the investigation and E_x^c is the corresponding central exposed to risk, then $\dfrac{d_x}{E_x^c}$ estimates $\mu_{x-½}$. So $\mu_{81.5}$ is estimated by $\dfrac{d_{82}}{E_{82}^c}$ assuming that the force of mortality is constant over the rate interval. [½]

From the given data values, we have:

$$d_{82} = 375 + 391 = 766 \qquad [½]$$

We have to estimate the value of E_{82}^c using the census method. Suppose that time is measured in years from 1/4/15 and let $P_{82}(t)$ be the number of males aged 82 next birthday in the population at time t. The investigation covers the years 2016 and 2017. So:

$$E_{82}^c = \int_{0.75}^{2.75} P_{82}(t)\,dt \qquad [1]$$

Time 0.75 corresponds to 1/1/16, the start of the investigation, and time 2.75 corresponds to 1/1/18, the end of the investigation.

Splitting the integral at those times when census data are available, *ie* time 1 (which corresponds to 1/4/16) and time 2 (which corresponds to 1/4/17), we have:

$$E_{82}^{c} = \int_{0.75}^{1} P_{82}(t)\,dt + \int_{1}^{2} P_{82}(t)\,dt + \int_{2}^{2.75} P_{82}(t)\,dt$$

Assuming that $P_{82}(t)$ is linear between the census dates:

$$E_{82}^{c} = 0.25 \times \frac{1}{2}\left[P_{82}(0.75) + P_{82}(1)\right] + \frac{1}{2}\left[P_{82}(1) + P_{82}(2)\right] + 0.75 \times \frac{1}{2}\left[P_{82}(2) + P_{82}(2.75)\right]$$

$$= 0.125P_{82}(0.75) + 0.625P_{82}(1) + 0.875P_{82}(2) + 0.375P_{82}(2.75) \qquad [2]$$

We know that:

$$P_{82}(1) = 5,310$$

and:

$$P_{82}(2) = 5,480$$

However, we must estimate the values of $P_{82}(0.75)$ and $P_{82}(2.75)$. Assuming that $P_{82}(t)$ is linear between time 0 and time 1:

$$P_{82}(0.75) = 0.25P_{82}(0) + 0.75P_{82}(1) = 0.25 \times 5,320 + 0.75 \times 5,310 = 5,312.5 \qquad [1]$$

Similarly, assuming that $P_{82}(t)$ is linear between time 2 and time 3:

$$P_{82}(2.75) = 0.25P_{82}(2) + 0.75P_{82}(3) = 0.25 \times 5,480 + 0.75 \times 5,520 = 5,510 \qquad [1]$$

So:

$$E_{82}^{c} = 0.125 \times 5,312.5 + 0.625 \times 5,310 + 0.875 \times 5,480 + 0.375 \times 5,510 = 10,844.0625 \qquad [1]$$

and hence the estimated value of $\mu_{81.5}$ is:

$$\frac{766}{10,844.0625} = 0.07064 \qquad [1]$$

9.6 (i) ***Rate interval***

The rate interval is 'x nearest birthday'. This is the year of age starting at exact age $x - ½$ and ending at exact age $x + ½$. [1]

(ii) ***Exposed to risk formula***

Looking at the definition of the exposed to risk, we see that the age definition for $P_x(t)$ is exactly the same as the age definition for the deaths. So we have correspondence between the two definitions.

Suppose that time is measured in years from 1 January 2015 (so that $t=0$ corresponds to 1 January 2015 and $t=3$ corresponds to 1 January 2018). Then the central exposed to risk at age x nearest birthday is:

$$E_x^c = \int_0^3 P_x(t)\,dt \qquad [1]$$

and assuming that $P_x(t)$ is linear between the census dates … [1]

$$\begin{aligned} E_x^c &= ½\big[P_x(0)+P_x(1)\big]+½\big[P_x(1)+P_x(2)\big]+½\big[P_x(2)+P_x(3)\big] \\ &= ½P_x(0)+P_x(1)+P_x(2)+½P_x(3) \qquad [1] \end{aligned}$$

(iii) ***New exposed to risk formula***

For those companies using the census information $P_x^*(t)$, there is no longer correspondence between the age definition of the deaths and the age in the exposed to risk. So we need to adapt the formula. For these companies we actually want $P_x(t)$, the number of lives aged x nearest birthday at time t. How would these lives be classified, using their age last birthday?

Consider a group of lives aged x nearest birthday. Their true age lies between $x-½$ and $x+½$. Some of these lives will have a true age which lies between $x-½$ and x. These people are all currently aged $x-1$ last birthday. The rest will have a true age which lies between x and $x+½$. These are all currently aged x last birthday. So $P_x(t)$ is actually a mixture of $P_x^*(t)$ and $P_{x-1}^*(t)$.

Assuming that birthdays are uniformly distributed over the calendar year … [1]

$$P_x(t) = ½\left[P_x^*(t)+P_{x-1}^*(t)\right] \qquad [1]$$

So the exposed to risk formula is now:

$$E_x^c = ¼\left[P_x^*(0)+P_{x-1}^*(0)\right]+½\left[P_x^*(1)+P_{x-1}^*(1)+P_x^*(2)+P_{x-1}^*(2)\right]+¼\left[P_x^*(3)+P_{x-1}^*(3)\right] \qquad [1]$$

10

Graduation and statistical tests

Syllabus objectives

4.5 Graduation and graduation tests

4.5.1 Describe and apply statistical tests of the comparison of crude estimates with a standard mortality table testing for the overall fit, the presence of consistent bias, the presence of individual ages where the fit is poor, the consistency of the 'shape' of the crude estimates and the standard table.

For each test describe the formulation of the hypothesis, the test statistic, the distribution of the test statistic using approximations where appropriate, the application of the test statistic.

4.5.2 Describe the reasons for graduating crude estimates of transition intensities or probabilities, and state the desirable properties of a set of graduated estimates.

4.5.3 Describe a test for smoothness of a set of graduated estimates.

4.5.5 Describe how the tests in 4.5.1 should be amended to compare crude and graduated sets of estimates.

4.5.7 Carry out a comparison of a set of crude estimates and a standard table, or of a set of crude estimates and a set of graduated estimates.

0 Introduction

0.1 Graduation of observed mortality rates

In previous chapters we have introduced models for mortality over a single year of age, x to $x+1$, and we have seen how to use these models to estimate $\mu_{x+½}$. In practice, an investigation will include a considerable range of ages. For example a national life table will include all ages from 0 to over 100.

We will now consider an investigation where lives are classified according to their age nearest birthday. This will enable us to calculate a crude estimate of μ_x for each age x.

The crude mortality rates derived from a mortality investigation will not be the final rates that are published for use in actuarial calculations. The rates will have to pass through a further process called *graduation*.

Graduation refers to the process of using statistical techniques to improve the estimates provided by the crude rates. The aims of graduation are to produce a smooth set of rates that are suitable for a particular purpose, to remove random sampling errors (as far as possible) and to use the information available from adjacent ages to improve the reliability of the estimates. Graduation results in a 'smoothing' of the crude rates.

The graduation process itself is covered in Chapter 11, *Methods of graduation*. In this chapter we look at the aims of graduation and the statistical tests that are used to check the reasonableness of the graduated rates.

0.2 The underlying assumptions

We now suppose that we have data for all ages from the lowest, denoted x_1, to the highest, x_m, depending on the investigation.

Using the Poisson or multiple-state model, for $x = x_1, x_2, ..., x_m$, we have:

Number of deaths at age x nearest birthday	D_x
Central exposed to risk at age x nearest birthday	E_x^c
Crude estimate of the force of mortality at exact age x	$\hat{\mu}_x$
Estimator of the force of mortality at exact age x	$\tilde{\mu}_x$

and we will use the approximate asymptotic distribution:

$$D_x \sim \text{Normal}\left(E_x^c \mu_x, E_x^c \mu_x\right)$$

or:

$$\tilde{\mu}_x \sim \text{Normal}\left(\mu_x, \frac{\mu_x}{E_x^c}\right)$$

If we assume that the force of mortality is constant between exact age $x-½$ and exact age $x+½$, then the estimator of μ_x is $\tilde{\mu}_x = \dfrac{D_x}{E_x^c}$. The approximate distribution of $\tilde{\mu}_x$ was first introduced in Chapter 3.

1 Comparison with another experience

1.1 Introduction

Given the data above (the observed numbers of deaths and the exposed to risk values for each age, and our crude estimates)**, we often want to know if it is consistent with another, known experience. For example, if it is the recent experience of the policyholders of a life insurance company, we might ask:**

- **Is it consistent with the company's own past experience, or is the experience changing? This could be important for pricing life insurance contracts.**
- **Is it consistent with the published life tables? This is important if the company plans to use published tables for any financial calculations.**

It is important for an insurance company to be aware of the extent to which the mortality experienced by its policyholders differs from that of its past experience and published life tables. The difference will be reflected in the premiums charged for life assurance contracts.

Question

Describe the major problems associated with charging premiums that are:

(a) too low

(b) too high.

Solution

(a) If premiums are too low, the business will be unprofitable. The insurance company may pay out more in claims and maturities than the invested premiums can provide.

(b) If premiums are too high, the insurance company is likely to be uncompetitive and may lose business. The office may not write enough business to cover its fixed costs and may ultimately need to cease trading.

1.2 Standard tables

Published life tables based on large amounts of data are called *standard tables*. The main examples are:

- **National life tables, based on the census data and death registration data of a whole country. In the UK, these are published every 10 years; the largest are the English Life Tables (actually based on the population of England and Wales).**

  Most countries have a similar approach.

- **Tables based on data from life insurance companies. In the UK, most life insurance companies contribute data to the Continuous Mortality Investigation (CMI), which publishes extensive tables for different types of business.**

The latest are based on 1999-2002 data, and are known as the '00 series' tables, and self-administered pension scheme tables based on 2000-06 data, known as 'SAPS S1'. Most life insurance companies use these standard tables very extensively, so it is important that they check whether or not their own mortality experience is consistent with that of the tables.

The term 'consistent' covers two concepts: the *shape* of the mortality curve over the range of ages and the *level* of mortality rates.

1.3 Comparison with standard tables

We introduce the following notation. The superscript '*s*' will denote a quantity from a published standard table, *eg* μ_x^s.

In rough terms, the question is whether our estimates $\hat{\mu}_x$ are consistent with the given μ_x^s. We will formulate this more precisely, in a way that allows us to derive statistical tests.

We have:

- **the probabilistic model (multiple-state or Poisson);**
- **the data** (the observed numbers of deaths, the exposed to risk values and our crude estimates $\hat{\mu}_x$)**; and**
- **a standard table.**

The hypothesis that we wish to test is that the standard table quantities $\{\mu_x^s\}$ are the 'true' parameters of the model at each age *x*.

In other words, our null hypothesis is:

H_0 : the mortality rates being tested are consistent with those from the standard table.

We will reject this null hypothesis if we find evidence that the rates being tested are significantly different from those in the standard table.

We can derive tests of this hypothesis using the distributional assumptions of Section 0.2 under the hypothesis:

$$D_x \sim N\left(E_x^c \mu_x^s , E_x^c \mu_x^s\right) \quad \text{(approximately)}$$

Hence we can find test statistics comparing the *actual* deaths d_x (*ie* the observed value of D_x) with the *expected* deaths given by these distributions. We will describe suitable statistical tests later in this chapter.

First, however, we must discuss some general features of mortality experiences, and the extent to which we might want to adjust the crude estimates so that they reflect these features.

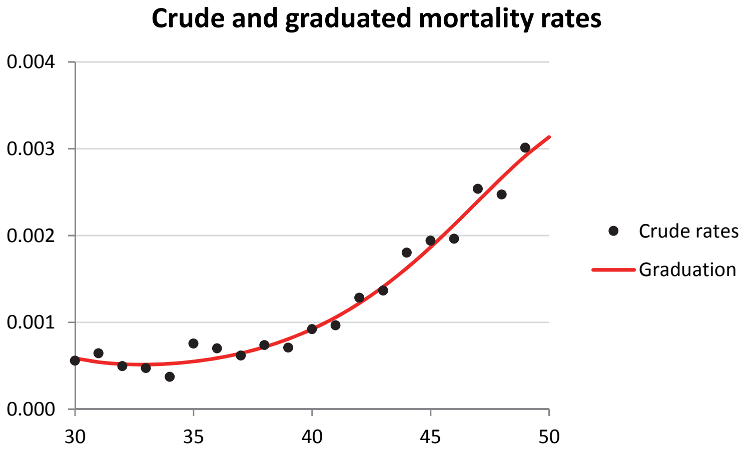
2 Graduation

The crude estimates $\{\hat{\mu}_x\}$ will progress erratically from age to age. In large part, this is because they have each been estimated independently and hence suffer independent sampling errors.

The smaller the sample size (*ie* the smaller the population studied), the less smoothly the crude estimates are likely to progress.

For several reasons (discussed below) we would prefer to work with μ_x values which are *smooth* functions of age. Therefore, we *graduate* or smooth the crude estimates, to produce a set of *graduated estimates* that do progress smoothly with age. We denote these $\{\mathring{\mu}_x\}$.

The graph below illustrates the type of relationship we would expect to see between the crude rates and the graduated rates. The true underlying rates are likely to be very close to the graduated rates.



Three questions that we must answer are:

(a) Why do we want smoothed estimates? We discuss this in Section 3.

(b) How do we carry out the graduation? (*ie* produce the $\mathring{\mu}_x$ from the $\hat{\mu}_x$). This is the subject of Chapter 11.

(c) How do we decide that a given attempt to graduate the crude estimates is satisfactory? We discuss this in Section 4, before resuming our discussion of statistical tests of a mortality experience, because statistical tests form part of the answer.

3 Reasons for graduation

3.1 The theoretical argument

At the heart of our desire to graduate is the intuitive idea that μ_x should be a smooth function of age. There is some evidence from large investigations to support this, but it is nevertheless an *assumption*.

It follows that a crude estimate of μ_x for any age x also carries information about the values of μ_{x-1}, μ_{x+1} *etc*. For example, if the force of mortality is smooth and not changing too rapidly, then our estimate of μ_x should not be too far away from estimating μ_{x-1} and μ_{x+1}, as well as being the 'best' estimate, in some sense, of μ_x. By smoothing, we can make use of the data at adjacent ages to improve the estimate at each age.

Another way of looking at this is that smoothing reduces the sampling errors at each age.

It is intuitively sensible to think that mortality is a smooth function of age. However, mortality rates may show some significant changes at certain ages. For example, there is often a marked increase in mortality amongst young males around the age when individuals start to drive cars or ride motorbikes, or start drinking alcohol. This feature is spread over a period of several years and is often referred to as the 'accident hump'.

3.2 The practical argument

A purely practical reason for smoothing mortality data is that we will use the life table to compute financial quantities, such as premiums for life insurance contracts. It is very desirable that such quantities progress smoothly with age, since irregularities (jumps or other anomalies) are hard to justify in practice.

We could calculate these quantities using our crude mortality rates, and then smooth the premium rates *etc* directly, but it is much more convenient to have smoothed mortality rates to begin with.

We would never, in any case, apply the results of a mortality experience directly to some financial problem without considering carefully its suitability. This means comparing it with other relevant experiences and tables, not just in aggregate but over age ranges of particular financial significance. It is often the case that a mortality experience must be adjusted in some way before use, in which case there is little point in maintaining the roughness of the crude estimates.

3.3 Limitations

What graduation *cannot* do is remove any bias in the data arising from faulty data collection or otherwise.

Graduation can only produce results as reliable as the original data. This principle is known as 'garbage in, garbage out'.

3.4 Summary

The crude estimates of mortality ($\hat{\mu}_x$) provide an estimate of the true underlying mortality for a particular age. However, since we believe that the underlying rates of mortality will follow a smooth curve as the age varies, we can use the additional information provided by the numbers of deaths at nearby ages to improve our estimate. This process of applying statistical techniques to improve the estimates provided by crude rates over a range of ages is called *graduation*.

The aims of graduation are:

- to produce a smooth set of rates that are suitable for a particular purpose
- to remove random sampling errors
- to use the information available from adjacent ages.

Question

Comment on the following statement:

If the data set includes the whole population, there is no need to graduate the crude rates because there will be no sampling errors.

Solution

There will still be sampling errors (*ie* the actual numbers will not be the same as the expected numbers) because the study involves a finite population and a finite time period. Also there are other reasons for graduating crude mortality rates, other than to remove sampling errors.

4 Desirable features of a graduation

We list three desirable features of a graduation:

- **smoothness;**
- **adherence to data; and**
- **suitability for the purpose to hand.**

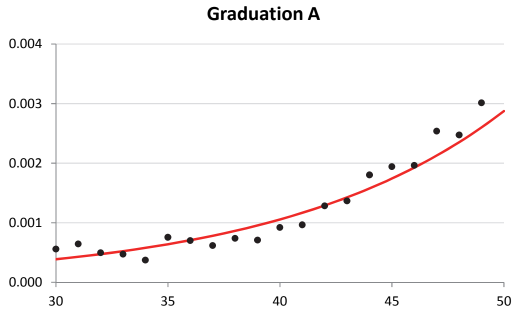
4.1 Smoothness versus adherence to data

The reasons for desiring smoothness were discussed above. At one extreme, we could easily smooth the crude estimates by ignoring the data altogether; we want to avoid such extremes since we want the graduation to be representative of the experience. We say that we require *adherence to data* or *goodness of fit*.

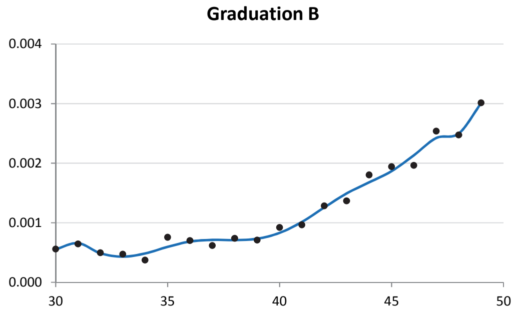
Smoothness and adherence to data are usually conflicting requirements. Perfect smoothness (extreme example: a straight line) pays little or no attention to the data, while perfect adherence to the data means no smoothing at all.

If the graduation process results in rates that are smooth but show little adherence to the data, then we say that the rates are *overgraduated*. The graph in Graduation A (see below) is very smooth, but it tends to overestimate the crude rates at the younger ages and underestimates them at the older ages.

Overgraduation has an opposite, referred to as *undergraduation*. This refers to the case where insufficient smoothing has been carried out. This will tend to produce a curve of inadequate smoothness, but better adherence to data. In this case, the graduated rates will follow the crude rates very closely, but will show an irregular progression over the range of ages. The graph in Graduation B (which uses the same data as Graduation A) adheres very closely to the crude rates, but it twists and turns erratically.



Graduation A - overgraduated



Graduation B - undergraduated

4.2 Testing smoothness and adherence to data

The 'art' of graduation lies in finding a satisfactory compromise.

The compromise is between smoothness and adherence to data. The balance between the two will be addressed when selecting the graduation method and carrying out the graduation itself (see Chapter 11).

To assist in this task, we have a battery of tests of smoothness and of adherence to data. We describe the usual test of smoothness in Section 5. The tests of adherence to data have much in common with the statistical tests of an experience against a standard table, and we will consider these together later in this chapter. They rely on the assumption that the 'true' parameters of the underlying probability model are the graduated estimates $\{\mathring{\mu}_x\}$.

When comparing observed data against a standard table, our null hypothesis is:

H_0 : the mortality rates being tested are consistent with those from the standard table, *ie* there are no significant differences between the two sets of rates.

When testing the adherence of a graduation to the observed data, our null hypothesis becomes:

H_0 : the true underlying mortality rates for the experience are the graduated rates.

That is, we replace the assumption:

$$D_x \sim N\left(E_x^c \mu_x^s , E_x^c \mu_x^s\right)$$

with the assumption:

$$D_x \sim N\left(E_x^c \mathring{\mu}_x , E_x^c \mathring{\mu}_x\right)$$

and then proceed to test the statistical hypothesis (almost) as before.

4.3 Suitability for the purpose in hand

The suitability of a graduation for practical work depends very much on what that work is, and can only be assessed in particular cases. However, two very important observations are:

(a) **In life insurance work, losses result from premature deaths (benefits are paid sooner than expected) so we must not *underestimate* mortality.**

In the case of term assurance policies, the insurance company will pay a benefit in respect of policyholders who die within the specified term. If we were to underestimate the mortality rates when calculating the premiums to charge, the insurance company would make a loss – the premiums would be insufficient to cover the benefits paid.

(b) **In pensions or annuity work, losses result from delayed deaths (benefits are paid for longer than expected) so we must not *overestimate* mortality.**

When an individual buys an annuity, the insurance company agrees to provide a regular income for the remaining lifetime of that individual. The company will provide a higher income for an individual with a lower expected lifetime, *ie* a higher rate of mortality. To limit potential losses, the company should err on the low side when determining the appropriate rates of mortality to use.

4.4 Two examples of graduation

The examples in this chapter are based on Graduation A and Graduation B, which were shown graphically in Section 4.1. The data for these graduations are shown in the tables that follow.

The left-hand part of each table shows the crude data (E_x^c , d_x and $\hat{\mu}_x$). The centre column shows the graduated rates ($\overset{\circ}{\mu}_x$). The right-hand columns show some quantities we will use in the graduation tests, which we will cover in detail later.

The details of each graduation will be explained in Chapter 11.

Graduation A

x	E_x^c	d_x	$\hat{\mu}_x$	$\mathring{\mu}_x$	$E_x^c\,\mathring{\mu}_x$	$z_x = \dfrac{d_x - E_x^c\mathring{\mu}_x}{\sqrt{E_x^c\mathring{\mu}_x}}$	z_x^2
30	70,000	39	0.000557	0.000388	27.16	2.27	5.16
31	66,672	43	0.000645	0.000429	28.60	2.69	7.25
32	68,375	34	0.000497	0.000474	32.41	0.28	0.08
33	65,420	31	0.000474	0.000524	34.28	−0.56	0.31
34	61,779	23	0.000372	0.000579	35.77	−2.14	4.56
35	66,091	50	0.000757	0.000640	42.30	1.18	1.40
36	68,514	48	0.000701	0.000708	48.51	−0.07	0.01
37	69,560	43	0.000618	0.000782	54.40	−1.55	2.39
38	65,000	48	0.000738	0.000865	56.23	−1.10	1.20
39	66,279	47	0.000709	0.000956	63.36	−2.06	4.23
40	67,300	62	0.000921	0.001056	71.07	−1.08	1.16
41	65,368	63	0.000964	0.001168	76.35	−1.53	2.33
42	65,391	84	0.001285	0.001291	84.42	−0.05	0.00
43	62,917	86	0.001367	0.001427	89.78	−0.40	0.16
44	66,537	120	0.001804	0.001577	104.93	1.47	2.16
45	62,302	121	0.001942	0.001743	108.59	1.19	1.42
46	62,145	122	0.001963	0.001926	119.69	0.21	0.04
47	63,856	162	0.002537	0.002129	135.95	2.23	4.99
48	61,097	151	0.002471	0.002353	143.76	0.60	0.36
49	61,110	184	0.003011	0.002601	158.95	1.99	3.95
Total		1,561			1,516.50		43.17

Graduation A assumed that $\ln(e^{\mu_x} - 1)$ *could be modelled as* $\alpha + \beta x$ *(2 parameters), which was fitted using the method of least squares.*

Graduation B

x	E_x^c	d_x	$\hat{\mu}_x$	$\overset{\circ}{\mu}_x$	$E_x^c\,\overset{\circ}{\mu}_x$	$z_x = \dfrac{d_x - E_x^c \overset{\circ}{\mu}_x}{\sqrt{E_x^c \overset{\circ}{\mu}_x}}$	z_x^2
30	70,000	39	0.000557	0.000555	38.85	0.02	0.00
31	66,672	43	0.000645	0.000658	43.87	−0.13	0.02
32	68,375	34	0.000497	0.000488	33.37	0.11	0.01
33	65,420	31	0.000474	0.000432	28.26	0.52	0.27
34	61,779	23	0.000372	0.000486	30.02	−1.28	1.64
35	66,091	50	0.000757	0.000596	39.39	1.69	2.86
36	68,514	48	0.000701	0.000685	46.93	0.16	0.02
37	69,560	43	0.000618	0.000713	49.60	−0.94	0.88
38	65,000	48	0.000738	0.000709	46.09	0.28	0.08
39	66,279	47	0.000709	0.000733	48.58	−0.23	0.05
40	67,300	62	0.000921	0.000831	55.93	0.81	0.66
41	65,368	63	0.000964	0.001015	66.35	−0.41	0.17
42	65,391	84	0.001285	0.001259	82.33	0.18	0.03
43	62,917	86	0.001367	0.001494	94.00	−0.82	0.68
44	66,537	120	0.001804	0.001679	111.72	0.78	0.61
45	62,302	121	0.001942	0.001866	116.26	0.44	0.19
46	62,145	122	0.001963	0.002134	132.62	−0.92	0.85
47	63,856	162	0.002537	0.002423	154.72	0.59	0.34
48	61,097	151	0.002471	0.002498	152.62	−0.13	0.02
49	61,110	184	0.003011	0.003008	183.82	0.01	0.00
Total		1,561			1,555.31		9.39

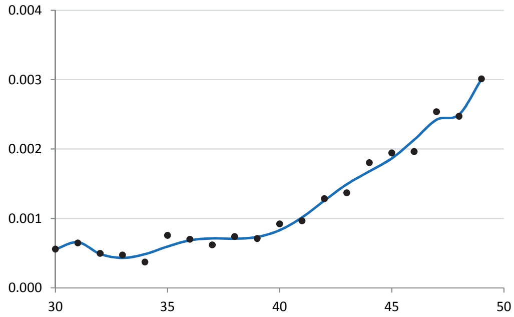
Graduation B assumed that $\ln(e^{\mu_x} - 1)$ *could be modelled as a polynomial of degree 10 (11 parameters), which was fitted using the method of least squares.*

5 Testing the smoothness of a graduation

5.1 What is a smooth graduation?

Mathematical smoothness is usually defined in terms of differentiability, but this is of little use in graduation work because many functions that misbehave wildly between integer ages are nevertheless differentiable many times.

The test for smoothness will be used as a check for undergraduation. It is possible to fit a high-order polynomial to any set of observed data, as in the graph below. The fitted polynomial is smooth in the mathematical sense, *ie* it is differentiable many times, but it does not progress smoothly from age to age.



Instead, we seek a more rough-and-ready measure of smoothness having regard to the scale with which we work (usually the year of age).

To test smoothness, we need to calculate the *third differences* of the graduated quantities $\{\mathring{\mu}_x\}$. For example:

The first difference $\Delta\mathring{\mu}_x = \mathring{\mu}_{x+1} - \mathring{\mu}_x$.

The second difference $\Delta^2\mathring{\mu}_x = \Delta\mathring{\mu}_{x+1} - \Delta\mathring{\mu}_x$.

The third difference $\Delta^3\mathring{\mu}_x = \Delta^2\mathring{\mu}_{x+1} - \Delta^2\mathring{\mu}_x$.

The third differences measure the change in curvature.

The criterion of smoothness usually used is that the third differences of the graduated quantities $\{\mathring{\mu}_x\}$ should:

- **be small in magnitude compared with the quantities themselves; and**
- **progress regularly.**

How to judge if this criterion is met takes some practice. However, since most methods of graduation now in use automatically give smooth results, this is not of great importance.

5.2 Smoothness test

In this section we describe how to carry out the smoothness test.

Question

Compare the smoothness of the rates in Graduations A and B over the age range 30 to 35.

Solution

The third differences are shown in the tables below.

Graduation A

x	$\overset{\circ}{\mu}_x$	$\Delta\overset{\circ}{\mu}_x$	$\Delta^2\overset{\circ}{\mu}_x$	$\Delta^3\overset{\circ}{\mu}_x$
30	0.000388	0.000041	0.000004	0.000001
31	0.000429	0.000045	0.000005	0.000000
32	0.000474	0.000050	0.000005	0.000001
33	0.000524	0.000055	0.000006	
34	0.000579	0.000061		
35	0.000640			

For this graduation, the third differences are very small, which indicates that the graduated rates are very smooth.

Graduation B

x	$\overset{\circ}{\mu}_x$	$\Delta\overset{\circ}{\mu}_x$	$\Delta^2\overset{\circ}{\mu}_x$	$\Delta^3\overset{\circ}{\mu}_x$
30	0.000555	0.000103	–0.000273	0.000387
31	0.000658	–0.000170	0.000114	–0.000004
32	0.000488	–0.000056	0.000110	–0.000054
33	0.000432	0.000054	0.000056	
34	0.000486	0.000110		
35	0.000596			

The third differences are larger for Graduation B than for Graduation A (especially when $x = 30$) and they progress in a less regular manner. This indicates that Graduation B is not as smooth as Graduation A.

6 Statistics refresher

In Section 7 we will look at the statistical tests to assess the reasonableness of a graduation. This section gives a brief review of the background to statistical tests.

6.1 Statistical tests

Hypotheses

Statistical tests assess the plausibility of a particular *null hypothesis* in relation to an *alternative hypothesis*. The null hypothesis (denoted by H_0) corresponds to a neutral conclusion. In graduation tests, the null hypothesis will correspond to a statement that some aspect of a proposed graduation is 'OK'. The alternative hypothesis (denoted by H_1) corresponds to a definite conclusion. In graduation tests, the alternative hypothesis will correspond to a statement that some aspect of a proposed graduation is 'no good'.

Test process

In statistical tests, we start by first assuming that the null hypothesis is correct. A *test statistic* is then calculated from the data, on that assumption. Using statistical theory, the distribution of the values that might be obtained from the test statistic, assuming that the null hypothesis is correct, can be determined. If it turns out that the value actually obtained for the test statistic is one that would be very unlikely if the null hypothesis were correct, then we conclude that the null hypothesis is not plausible, and we reject it in favour of the alternative hypothesis.

Probability value

In order to decide what can be considered as 'very unlikely', a *significance level* must be selected at the beginning of the test. The significance level usually used is 5%, which means that if H_0 were true, the value of the test statistic would only be this extreme by chance 1 time in 20. Using a significance level of 1% would give a stricter test in the sense that we would require a more extreme result to indicate that the null hypothesis is not valid.

The *probability value* (or *p-value*) is the probability, calculated assuming H_0 is true, of obtaining a value of the test statistic as extreme as the actual value obtained. If the probability value is smaller than the significance level chosen, then we reject the null hypothesis. A smaller probability value indicates a more definite ('significant') result.

A statistical test may be *one-tailed* or *two-tailed*, depending on the nature of the test and the feature we are interested in.

One-tailed tests

In a one-tailed (or one-sided) test, we will be suspicious about a test value that is unusually extreme in one direction only. For instance, a very high test value might worry us whereas a very low value would not.

For example, if our hypotheses were:

$$H_0 : P\text{ (heads on a coin)} = 0.5$$

$$H_1 : P\text{ (heads on a coin)} > 0.5$$

then we would use a one-tailed test because only a high number of heads would lead us to reject the null hypothesis in favour of the alternative hypothesis.

Two-tailed tests

In a two-tailed (or two-sided) test, we will be suspicious about a test value that is unusually extreme in either direction, *ie* either very high or very low.

For example, if our hypotheses were:

$$H_0 : P\text{ (heads on a coin)} = 0.5$$

$$H_1 : P\text{ (heads on a coin)} \neq 0.5$$

then we would use a two-tailed test because a low or high number of heads would cast doubt on the validity of the null hypothesis.

Conclusions

The test will result in one of two outcomes:

1. The probability value (*eg* 2%) is lower than the significance level (*eg* 5%). In this case, we conclude that the test provides sufficient evidence for us to reject the null hypothesis.

2. The probability value (*eg* 12%) is not lower than the significance level (*eg* 5%). In this case, we conclude that the test did not provide sufficient evidence for us to reject the null hypothesis.

Question

Consider the test:

H_0 : Smoking has no effect on mortality, *versus*

H_1 : Smoking increases mortality

(a) State whether this test is one-sided or two-sided.

(b) Describe the possible conclusions of this test.

Solution

(a) This is a one-sided test, since we are only concerned about an *increase* in mortality. The hypotheses for the corresponding two-sided test are:

H_0 : Smoking has no effect on mortality *vs* H_1 : Smoking affects mortality

Here H_1 includes the possibility that smoking could also *reduce* mortality.

(b) If the probability value of the calculated test statistic is smaller than the significance level, the conclusion is: 'The test provides sufficient evidence to reject the null hypothesis and conclude that smoking increases mortality.'

(This conclusion may not actually be correct. If we were using a significance level of 5%, we would arrive at this conclusion 5% of the time, even if smoking didn't increase mortality. However, if the probability value is very small *eg* 0.1%, this is so unlikely that nobody would doubt the result.)

If the probability value of the calculated test statistic is greater than the significance level, the conclusion is: 'The test does not provide sufficient evidence to reject the null hypothesis that smoking has no effect on mortality.'

(Again, this conclusion may not actually be correct. It may be that our study wasn't big enough or the test we used wasn't powerful enough to give a convincing result.)

6.2 Continuity correction

We often use a continuous distribution as an approximation to a discrete distribution, *eg* the normal distribution as an approximation to the binomial. When we do so, we must be careful to take into account the fact that the discrete distribution can only take certain (usually integer) values, whilst the continuous distribution can take any value.

To ensure that the approximation is acceptable, we estimate the probability of observing a particular value under the discrete distribution (*ie* $X = x$) by calculating the probability that the continuous distribution takes a value between $x - \tfrac{1}{2} \times \text{step size}$ and $x + \tfrac{1}{2} \times \text{step size}$. If X can take any integer value, then the step size between consecutive values is 1, and $P(X = x)$ becomes $P(x - \tfrac{1}{2} < X < x + \tfrac{1}{2})$. This adjustment is called a *continuity correction*.

For example, if we toss a fair coin 20 times, the number of heads has a $Binomial(20, \tfrac{1}{2})$ distribution. Under the Central Limit Theorem, the number of heads (X) will have an approximate normal distribution with mean $20 \times \tfrac{1}{2} = 10$ and variance $20 \times \tfrac{1}{2} \times \tfrac{1}{2} = 5$. Incorporating a continuity correction:

$$P(10\text{ heads}) = P(9.5 \le X \le 10.5) \simeq P\left(-\frac{0.5}{\sqrt{5}} \le Z \le \frac{0.5}{\sqrt{5}}\right) = \Phi(0.2236) - \Phi(-0.2236) = 0.177$$

and:

$$P(16\text{ heads or more}) = P(X \ge 15.5) \simeq P\left(Z \ge \frac{5.5}{\sqrt{5}}\right) = 1 - \Phi(2.4597) = 0.007$$

If, however, X can only take the values 0, 100, 200, …, then the step size between consecutive values is 100 and, using a continuity correction, $P(X \ge 200)$ becomes $P(X > 150)$.

6.3 Chi-squared tests

Purpose

A chi-squared test can be used to assess whether the observed numbers of individuals who fall into specified categories are consistent with a model that predicts the expected numbers in each category. It is a test for overall *goodness of fit*.

For example, the following categories might be used:

Dead/alive at each age: This would enable us to test whether the observed numbers of deaths and survivors at each age are consistent with the numbers predicted by the forces of mortality $\overset{\circ}{\mu}_x$ for a particular graduation.

Cause of death: If the deaths within a population have been classified by cause of death, this would enable us to test whether the numbers dying from each cause are consistent with the numbers predicted from an assumed set of proportions.

Rationale

The formula for the test statistic is $\sum \frac{(O-E)^2}{E}$, where:

- O is the observed number in a particular category
- E is the corresponding expected number predicted by the assumed probabilities
- the sum is over all possible categories.

Each term in the sum represents the square of the discrepancy between the actual and expected values for one group (with an appropriate weighting factor applied). A high value for the total indicates that the overall discrepancy is quite large and would lead us to reject the model. A low value indicates that the model is a good fit to the data.

In some cases, the model assumed in the null hypothesis doesn't specify the precise probabilities, but just gives a general formula or a family of distributions. In such cases, it will be necessary to estimate any unknown parameters to calculate the *E*'s.

Chi-squared distribution

The theory of multinomial distributions tells us that, in situations where a large number of individuals can be allocated to different categories based on fixed (but unknown) probabilities, this statistic has a *chi-squared distribution* (approximately), which is tabulated in the statistics section of the *Tables*.

Degrees of freedom

The chi-squared distribution has one parameter, called the number of *degrees of freedom* (DF), which can take the values 1,2,3,…. This parameter reflects the number of independent pieces of information present in the data. The correct number of degrees of freedom to use in a chi-squared test depends on the number of constraints that restrict the way individuals can be allocated to the different categories.

To determine the number of degrees of freedom:

1. Start with the number of groups. (Each combined group counts as one group. See below.)
2. If the groups form a set of mutually exclusive and exhaustive categories (so that their probabilities must add up to 1) or the expected numbers for each category are determined based on the total number for all groups, then subtract 1.
3. Subtract a further 1 for each parameter that has been estimated.

Small groups

The chi-squared distribution provides a good approximation provided the numbers in each group are not too small. If the expected number in a group is small (less than 5 say), a difference of just one observation can make a big difference to the value of the test statistic and the approximation becomes unreliable. This problem can be overcome by combining the expected and actual numbers in small groups.

Question

A study of causes of death in elderly men in the 1970s showed the proportions given in the table below. Carry out a chi-squared test to assess whether these percentages can still be considered to provide an accurate description of causes of death in 2015.

Cause of death	Proportion of deaths in 1975	Number of deaths in 2015
Cancer	8%	286
Heart disease	22%	805
Other circulatory disease	40%	1,548
Respiratory diseases	19%	755
Other causes	11%	464

Solution

The total number of deaths is $286+805+1,548+755+464=3,858$.

We can calculate the expected numbers of deaths from each cause by applying the proportions to this total. For example, the expected number of cancer deaths is $0.08\times 3,858=308.64$.

The figures are given in the table below.

Cause of death	**Observed frequency, *O***	**Expected frequency, *E***	$\frac{(O-E)^2}{E}$
Cancer	286	308.64	1.661
Heart disease	805	848.76	2.256
Other circulatory disease	1,548	1,543.20	0.015
Respiratory diseases	755	733.02	0.659
Other causes	464	424.38	3.699
Total	**3,858**	**3,858**	**8.290**

Here, we have 5 categories. We haven't estimated any parameters, but we have calculated the expected numbers by assuming that the total is the same as for the actual numbers. So the number of degrees of freedom to use is $5-1=4$.

From page 169 of the *Tables*, the upper 5% point of the χ_4^2 distribution is 9.488. The observed value of our test statistic is 8.290, which is less than 9.488. So there is insufficient evidence to conclude that there has been a change in the pattern of causes of death.

7 Statistical tests of a mortality experience

Here we describe some statistical tests based on the hypothesis that:

(a) the numbers of deaths at different ages are independent; and

(b) $D_x \sim N\left(E_x^c \mu_x^s, E_x^c \mu_x^s\right)$ in the case where we are comparing the experience with a standard table; or

(c) $D_x \sim N\left(E_x^c \overset{\circ}{\mu}_x, E_x^c \overset{\circ}{\mu}_x\right)$ in the case where we are testing the adherence to data of a graduation.

In parts (b) and (c) of the hypothesis, the normal distribution is used as an approximation to the Poisson distributions with small intensity and large exposure.

Many of the tests that we will describe can be based on the standardised deviations, which we now define.

Deviations and standardised deviations

The *deviation* at age x is defined to be:

Actual deaths – Expected deaths

$$= \quad D_x - E_x^c \mu_x^s \quad \text{or} \quad D_x - E_x^c \overset{\circ}{\mu}_x$$

and the *standardised deviation*, denoted z_x is:

$$z_x = \frac{D_x - E_x^c \mu_x^s}{\sqrt{E_x^c \mu_x^s}} \quad \text{or} \quad \frac{D_x - E_x^c \overset{\circ}{\mu}_x}{\sqrt{E_x^c \overset{\circ}{\mu}_x}}$$

The z_x's are often referred to as *individual* standardised deviations to distinguish them from *cumulative* deviations, which we will meet shortly. When calculating a numerical value for z_x, we replace the random variable D_x with the observed value d_x.

Under the assumption that there is a sufficient number of (independent) lives at each age x, we can replace our hypotheses with the following, by virtue of the Central Limit Theorem:

- $z_x \sim N(0,1) \qquad x = x_1, x_2, \ldots, x_m$
- **The z_x's at different ages are mutually independent.**

Question

Verify the figure shown in the table for the standardised deviation at age 30 in Graduation B.

Solution

$$z_{30}=\frac{d_{30}-E_{30}^{c}\overset{\circ}{\mu}_{30}}{\sqrt{E_{30}^{c}\overset{\circ}{\mu}_{30}}}=\frac{39-70,000\times 0.000555}{\sqrt{70,000\times 0.000555}}=\frac{39-38.85}{\sqrt{38.85}}=0.0241$$

7.1 Chi-squared (χ^2) test

The first test we describe is the χ^2-test. Unfortunately, this is the one test where we must pay attention to whether we are comparing an experience with a standard table, or testing the adherence to data of a graduation.

In either case, the test statistic is the same.

As we shall see, the difference relates to the appropriate number of degrees of freedom to use.

Purpose

To test whether the observed numbers of deaths at each age are consistent with a particular set of graduated mortality rates or a particular standard table. The chi-squared test will indicate overall goodness of fit.

Rationale

The chi-squared test can be applied to the numbers of deaths at each age (or in age groups).

A high value of the chi-squared statistic indicates that the discrepancies between the observed numbers and those predicted by the graduated rates or standard table are large, *ie* the fit is not very good. This may be because of overgraduation.

Assumptions

1. There is no heterogeneity of mortality (*ie* no variation in the mortality rates) within each age group and lives are independent.

2. The expected numbers of deaths are high enough (usually at least 5 in each cell) for the chi-squared approximation to be valid.

Method

Step 1

Combine any small groups by pooling the actual and expected deaths, so that the expected number of deaths is never less than 5.

Step 2

The test statistic for the chi-squared goodness-of-fit test is:

$$\sum \frac{(O-E)^2}{E}$$

Here O represents the observed number of deaths and E represents the expected number of deaths predicted by the graduation or standard table. The sum is taken over all age groups.

So:

$$\sum \frac{(O-E)^2}{E} = \sum z_x^2$$

and Step 2 is to calculate the observed value of:

$$\boldsymbol{X} = \sum_{\substack{\textbf{all ages}\\ \boldsymbol{x}}} \boldsymbol{z_x^2}$$

X is called the χ^2-statistic.

Step 3

Determine the appropriate number of degrees of freedom and compare the observed value of the test statistic with the appropriate percentage point of the chi-squared distribution given on page 169 of the *Tables*.

- **If we are comparing an experience with a standard table, then X can be assumed to have a χ^2 distribution with m degrees of freedom. (m is just the number of age groups in our notation.) Large values of X indicate excessive deviations, so we will test X against the upper 5% percentage point of the χ_m^2 distribution, and say that the test fails if $X > \chi_{m;0.95}^2$.**

  For example, suppose we are comparing the mortality of a population that is divided into 30 age groups with a standard table. Then the critical value for the chi-squared test is the upper 5% point of χ_{30}^2, *ie* 43.77.

- **If we are testing the adherence to data of a graduation, X can be assumed to have a χ^2 distribution, but with fewer than m degrees of freedom. How many fewer depends on the method of graduation, so we defer further comment to Chapter 11.**

Conclusion

If the value of the test statistic exceeds the upper 5% point of the relevant χ^2 distribution, this indicates a poor fit or overgraduation. The contributions to the test statistic from each term in the sum can be used to identify the ages where the fit is worst.

If the observed value of the chi-squared statistic is very low, this may indicate undergraduation. However, if we suspect undergraduation is a problem we will usually test for it in other ways (*eg* by considering the standardised deviations or applying the smoothness test described in Section 5).

R can carry out a χ^2 goodness-of-fit test using `chisq.test()`. See Subject CS1 for more details.

Details are also given in the R part of this course.

Strengths and weaknesses

The chi-squared test is a good test for overall goodness of fit. It can, however, miss certain important features.

Deficiencies of the chi-squared test

The χ^2-test will fail to detect several defects that could be of considerable financial importance. (These comments apply particularly when we are testing a graduation, and for ease of exposition we will write as if that were the case.)

(a) There could be a few large deviations offset by a lot of very small deviations. In other words, the χ^2-test could be satisfied although the data do not satisfy the distributional assumptions that underlie it. This is, in essence, because the χ^2-statistic summarises a lot of information in a single figure.

(b) The graduation might be biased above or below the data by a small amount. The χ^2-statistic can often fail to detect consistent bias if it is small, but we should still wish to avoid it for the reasons given in Section 4.3.

(c) Even if the graduation is not biased as a whole, there could be significant groups of consecutive ages (called *runs* or *clumps*) over which it is biased up or down. This is still to be avoided.

It should be noted that because the χ^2-test is based on squared deviations, it tells us nothing about the direction of any bias or the nature of any lack of adherence to data of a graduation, even if the bias is large or the lack of adherence manifest. To ascertain this there is no substitute for an inspection of the experience.

Accordingly, we devise tests that will detect these defects (at least, will do so better than does the χ^2-test).

These tests are described in the following sections. We will present these tests in terms of testing the goodness of fit of a set of graduated rates to the crude data, but they can equally be applied to comparing a set of mortality rates with a standard table.

Question

Apply the chi-squared test to Graduation A.

Solution

From the table of values for Graduation A given on page 12, we see that:

$$\sum z_x^2 = 43.17$$

In this example, it is not difficult to work out how many degrees of freedom to use. There are 20 ages. We have not constrained the totals. The graduated rates have been calculated by estimating 2 parameters. So, the number of degrees of freedom is $20-2=18$.

From the *Tables*, the upper 5% point for the χ_{18}^2 distribution is 28.87. The observed value of the test statistic exceeds this, so we reject the null hypothesis. (In fact, the test statistic also exceeds 42.31, the upper 0.1% point.)

So, we conclude that the mortality experience does not conform to a formula of the type assumed in the graduation.

7.2 Standardised deviations test

Purpose

We can use the *standardised deviations test* to look for the first defect of the chi-squared test.

The defect to which the Core Reading is referring, is the failure to detect a number of excessively large deviations. This test can detect overall goodness of fit. Where the test reveals a problem this might be due to under/overgraduation or the presence of duplicates. (The problem of duplicates is discussed briefly in Chapter 11.)

Rationale

The test looks at the distribution of the values of the standardised deviations.

Under the null hypothesis, the z_x's comprise a random sample of size m from the $N(0,1)$ distribution. This test just tests for that normality.

If the graduated rates are not a good fit, the distribution will not be 'centred correctly'. If we have undergraduation, then we would expect the standardised deviations to be tightly bunched. Conversely, if we have overgraduation, we would expect the standardised deviations to be too spread out.

Assumptions

The normal approximation provides a good approximation at all ages.

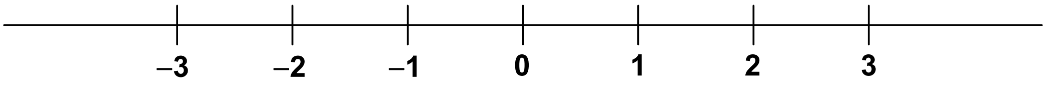
Method

Step 1

Calculate the standardised deviation z_x for each age or age group.

Step 2

Divide the real (number) **line into any convenient intervals (the more age groups, the more intervals it might be reasonable to use). For example:**



where the intervals at either end are (–∞, –3] and [+3, +∞).

Plot or count the number of standardised deviations falling into each of the ranges.

Step 3

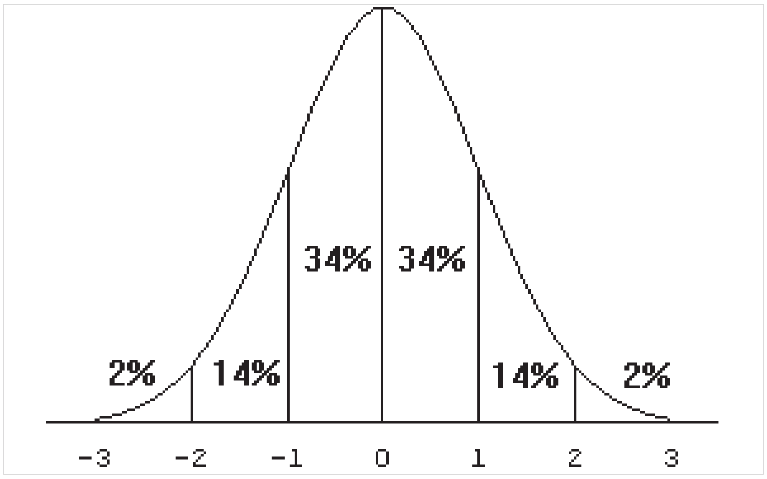
We can then compare:

- **the observed number of the z_x that fall in each interval; and**
- **the expected number of the z_x that should fall in each interval, under the hypothesis.**

The hypothesis states that the z_x values are realisations of a standard normal random variable.

In this example, the expected numbers are:

Interval	**(–∞,–3)**	**(–3,–2)**	**(–2,–1)**	**(–1,0)**	**(0,1)**	**(1,2)**	**(2,3)**	**(3,∞)**
Expected number	**0**	**0.02*m***	**0.14*m***	**0.34*m***	**0.34*m***	**0.14*m***	**0.02*m***	**0**



To formalise the comparison, we can form a χ^2-statistic (nothing to do with the use of the χ^2-test mentioned previously):

$$X = \sum_{\substack{\text{all}\\ \text{intervals}}} \frac{(\text{Actual} - \text{Expected})^2}{\text{Expected}}$$

which here should have a χ^2-distribution with 7 degrees of freedom (since we have used 8 intervals).

Note how this differs from the way we previously applied the χ^2-test, which was to test whether the observed numbers of deaths were consistent with a given set of graduated rates. Here, we are testing whether the observed pattern of the individual standardised deviations (*ie* the numbers falling in each interval) is consistent with a standard normal distribution.

If the number of age groups is small, we should use a smaller number of intervals, ensuring that the expected number of standardised deviations in each interval is not less than five (as a rule of thumb), and we then reduce the number of degrees of freedom in the χ^2-test appropriately.

If there are only a few age groups, a test must be carried out 'by eye', by considering the following features of the normal distribution.

Overall shape

The number of values in each of the ranges should conform broadly with the percentages for the normal distribution.

Absolute deviations

A related test derives from the fact that, if the z_x's are a random sample from the $N(0,1)$ distribution, half of them should lie in the interval $(-2/3, 2/3)$ – that is, half of them should have an absolute value greater than 2/3. Thus if we have m ages, the number of standardised deviations exceeding 2/3 in absolute value is a binomial random variable with parameters m and 0.5.

If there are a lot of values in the tails (*ie* the absolute deviations are too big), this indicates overgraduation or the existence of duplicates.

In this case a one-tailed test is appropriate, as we usually wish only to identify instances where the number of absolute deviations exceeding 2/3 is large. We reject the null hypothesis (of no difference between the standard table and the mortality underlying the experience, or of no difference between the graduated rates and the mortality underlying the experience) if this number falls in the upper 5% tail of the *Binomial*$(m, 0.5)$ distribution. If $m > 20$ a normal approximation to the binomial can be used.

Outliers

In addition to these two tests, we should also look at the values of the individual standardised deviations. If the z_x 's are $N(0,1)$, individual z_x 's with absolute values greater than 1.96 should form at most 1 in 20 of the whole set, and there should be only 1 in 100 with an absolute value greater than 2.57. In practice the number of ages we work with is often quite small, so these rules should be applied with some flexibility. As a guideline, we can say that with fewer than 20 ages we should be suspicious about any individual standardised deviation with a value greater than about 2.0, and regardless of the number of ages we should be concerned about any z_x with an absolute value greater than about 2.5.

Symmetry

There should be roughly equal numbers of positive and negative standardised deviations (since the normal distribution is symmetrical). An excess of positive values indicates that the graduated rates are too low. An excess of negative values indicates that the graduated rates are too high.

Conclusion

If the standardised deviations do not appear to conform to a standard normal distribution, this indicates that the observed mortality rates do not conform to the model with the rates assumed in the graduation. The features considered above will indicate the nature of the discrepancy.

Strengths and weaknesses

Looking at the distribution of the standardised deviations is a good all round test that detects most of the problems that might be present in a graduation.

Question

Analyse the distribution of the standardised deviations for Graduation A.

Solution

The observed and expected numbers in each range are shown in the table below.

Interval	$(-\infty,-3)$	$(-3,-2)$	$(-2,-1)$	$(-1,0)$	$(0,1)$	$(1,2)$	$(2,3)$	$(3,\infty)$
Observed	0	2	4	4	3	4	3	0
Expected	0.0	0.4	2.8	6.8	6.8	2.8	0.4	0

There are only 7 values in the range $(-2/3, 2/3)$. So, there appear to be too few values in the centre of the distribution and too many in the tails. This might indicate overgraduation (an inappropriate graduation formula) or the presence of duplicates.

The distribution of the standardised deviations is fairly symmetrical, with 10 positive and 10 negative values. So there is no evidence of bias in the graduated rates.

If we combine the small groups by pooling the values in the ranges $(-\infty,-1)$ and $(1,\infty)$, we can apply a chi-squared test to the resulting 4 groups. The value of the test statistic is:

$$\frac{(6-3.2)^2}{3.2}+\frac{(4-6.8)^2}{6.8}+\frac{(3-6.8)^2}{6.8}+\frac{(7-3.2)^2}{3.2}=10.24$$

This exceeds 7.815, the upper 5% point of the chi-squared distribution with 3 degrees of freedom, which confirms that the deviations do not conform to a standard normal distribution.

(Strictly speaking, the χ^2 -test should not be used even with this broad grouping since we have $E=3.2<5$ in the intervals $(-\infty,-1)$ and $(1,\infty)$.)

7.3 Signs test

Purpose

The signs test is a simple test for overall bias.

In other words, this test checks whether the graduated rates are too high or too low.

It is designed to identify the second deficiency of the chi-squared test, *ie* failure to detect where there is an imbalance between positive and negative deviations.

Rationale

If the graduated rates do not tend to be higher or lower than the crude rates on average, we would expect roughly half the graduated values to be above the crude rates and half below. So, if there are m age groups, the number above (or below) should have a $Binomial(m,½)$ distribution. An excessively high number of positive or negative deviations will indicate that the rates are biased.

This is normally applied as a two-tailed test, *ie* we are looking for both positive and negative bias.

Assumptions

None.

Method

Step 1

Count how many of the graduated rates lie above/below the crude rates. We will do this by looking at the signs of the individual standardised deviations. (This can also be done by comparing the crude rates with a graduated mortality curve plotted on a graph or by comparing the numerical values of $\hat{\mu}_x$ and $\overset{\circ}{\mu}_x$.)

Step 2

Calculate the probability value for the test by finding the probability of obtaining a split of positive/negative values as extreme as observed.

Define the test statistic:

$\boldsymbol{P}$ **= Number of** $\boldsymbol{z_x}$ **that are positive**

Alternatively, we could base the test on the number of negative z_x's.

Under the hypothesis, $\boldsymbol{P \sim Binomial(m,½)}$**.**

So the probability function of P is:

$$P(P=x)=\binom{m}{x}\left(\frac{1}{2}\right)^m \;, \; x=0,1,\ldots,m$$

An excess of either negative or positive deviations is a defect, so we apply a two-tailed test. We cannot do so exactly, since the binomial distribution is discrete. We could find the *smallest* value of $\boldsymbol{k}$ **for which:**

$$\boldsymbol{\sum_{j=0}^{k}\binom{m}{j}\left(\frac{1}{2}\right)^m \geq 0.025}$$

ie the smallest value of k for which $P(P<k)<0.025$. By the symmetry of the $Binomial(m,½)$ distribution, $P(P<k)=P(P>m-k)$. So we want the smallest value of k for which $P(P>m-k)<0.025$.

The test would be satisfied (at the 5% level) if $\boldsymbol{k \leq P \leq m-k}$**. Or, (perhaps more satisfactorily) we could just find the *p*-value corresponding to** $\boldsymbol{P}$**.**

If possible, we should calculate the p-value of the test using the probabilities for the binomial distribution given on pages 186-188 of the *Tables*. However, we can only use the *Tables* if the sample size is one of those listed ($n=2,3,\ldots,10$, 12 or 20).

Question

A graduation covers 20 age groups and has resulted in 6 positive and 14 negative deviations. Carry out a signs test on these data values.

Solution

Under the null hypothesis, $P \sim Binomial(20,½)$. Here we have fewer positive deviations than expected. The p-value of the test is:

$$2P\left(P \leq 6\right)=2\times 0.0577=0.1154$$

The probability of 0.0577 can be found on page 188 of the Tables by looking up $n=20$, $p=0.5$, $x=6$. We multiply this probability by 2 because this is a two-tailed test.

Since the p -value is greater than 5%, there is insufficient evidence to reject the null hypothesis at the 5% significance level. So we can conclude that the rates are not biased.

If the number of age groups is large, we can use the approximation:

$$P \sim N(\tfrac{1}{2}m, \tfrac{1}{4}m)$$

If we use a normal approximation, then we should use a continuity correction since we are approximating a discrete (binomial) distribution with a continuous (normal) distribution.

For a two-tailed test, the probability value must be based on the total probability for both tails of the distribution. For example, we can only reject H_0 at the 5% level if the observed number of positives or negatives is greater than the upper 2.5% point or less than the lower 2.5% point of the binomial distribution.

Conclusion

If the test shows that the number of positive values is very high or very low, this indicates that the rates are on average too low or too high (respectively). An examination of the pattern of the signs will indicate the range of ages where the bias is worst.

Strengths and weaknesses

Just looking at the signs of the deviations provides no indication of the extent of the discrepancy. This test is qualitative rather than quantitative.

Question

State the conclusion that can be drawn from an examination of the signs of the deviations for Graduation B.

Solution

There are 12 positive and 8 negative values for Graduation B. Here we have more positive deviations than expected. So the p-value is:

$$2P(P \geq 12) = 2\left[1 - P(P \leq 11)\right] = 2\left[1 - 0.7483\right] = 0.5034$$

The value of 0.7483 comes from page 188 of the Tables.

Since this is (much) greater than 5%, there is very little evidence of bias in the graduated rates.

7.4 Cumulative deviations

Purpose

The cumulative deviations test checks whether the overall number of deaths conforms to the model with the mortality rates assumed in the graduation. This test can detect overall goodness of fit. It addresses the problem of the inability of the chi-squared test to detect a large positive or negative cumulative deviation over part (or the whole) of the age range.

The *cumulative deviations* test detects overall bias or long runs of deviations of the same sign.

Rationale

Consider the hypothesis:

$$D_x \sim N(E_x^c \mathring{\mu}_x, E_x^c \mathring{\mu}_x)$$

Here, the deviation has (approximate) distribution:

$$D_x - E_x^c \mathring{\mu}_x \sim N(0, E_x^c \mathring{\mu}_x)$$

So the accumulated deviation, over the whole age range, has distribution:

$$\sum_{\substack{\text{all}\\\text{ages}}} (D_x - E_x^c \mathring{\mu}_x) \sim N(0, \sum_{\substack{\text{all}\\\text{ages}}} E_x^c \mathring{\mu}_x)$$

and, upon standardising,

$$\frac{\sum_{\substack{\text{all}\\\text{ages}}} (D_x - E_x^c \mathring{\mu}_x)}{\sqrt{\sum_{\substack{\text{all}\\\text{ages}}} E_x^c \mathring{\mu}_x}} \sim N(0,1)$$

This can be tested in the usual way, using a two-tailed test, since either positive or negative deviations are of concern. As well as applying this test to the whole age range, we can apply it to parts of the age range of possible financial significance, *provided* we choose which sub-ranges to test without reference to the data.

Assumptions

The normal approximation is reasonable at all ages.

Method

Step 1

Decide which range of ages to test. This might be the whole table or just the range of ages that has the most financial importance. The test is not valid if the age range is selected *after* looking at the pattern of the data. For example, if we spot a blip in one part of the table and decide to apply the cumulative deviation test to that part of the table only, the results would be meaningless.

Step 2

Calculate $\sum d_x$ (the total observed deaths) and $\sum E_x^c \overset{\circ}{\mu}_x$ (the total expected deaths), where the sum is over the selected age range.

Step 3

Calculate the test statistic $\dfrac{\sum d_x - \sum E_x^c \overset{\circ}{\mu}_x}{\sqrt{\sum E_x^c \overset{\circ}{\mu}_x}}$ and use this to determine the *p*-value using the tables for the standard normal distribution.

Since the total number of deaths must be a whole number, we should, in theory, apply a continuity correction when carrying out this test. However, since the denominator is usually quite big (as we are summing over a range of ages), the continuity correction does not often make much difference to the value of the test statistic. As a result, it is usually omitted.

Conclusion

If the magnitude (*ie* the absolute value) of the calculated test statistic is high, this indicates that either:

- the graduated rates are biased (too low if the test statistic is positive, too high if the test statistic is negative), or
- the variance is higher than predicted by the model for the range of ages considered. This could be as a result of duplicate policies, which we will discuss in Section 6 of Chapter 11.

The cumulative deviations test can only detect features that are present over the whole age range considered. An excess of positive deviations over one age range may 'cancel out' an excess of negatives over another range.

A word of warning: many methods of graduation result in a cumulative deviation of zero as part of the fitting process, in which case this test cannot be applied.

Question

State the conclusion that can be drawn from applying the cumulative deviations test to the whole age range of Graduation A.

Solution

From the table given on page 12, the value of the test statistic is:

$$\frac{\sum d_x - \sum E_x^c \overset{\circ}{\mu}_x}{\sqrt{\sum E_x^c \overset{\circ}{\mu}_x}} = \frac{1,561 - 1,516.50}{\sqrt{1,516.50}} = 1.143$$

This is a two-tailed test, so we compare the value of the test statistic with the upper and lower 2.5% points of $N(0,1)$, *ie* ± 1.96. As $-1.96 < 1.143 < 1.96$, there is insufficient evidence to reject the null hypothesis.

So, the cumulative deviations test does not provide evidence that the graduated rates are biased.

7.5 Grouping of signs test

Purpose

To test for overgraduation.

The grouping of signs test (also called Stevens' test) detects 'clumping' of deviations of the same sign. It relies on some simple combinatorics.

Rationale

The test looks at the number of groups (or *runs*) of deviations of the same sign and compares this with the number that would be expected if the positive and negative signs were arranged in random order.

If the graduated rates are overgraduated, the standardised deviations will not swap from positive to negative very often and there will be fewer runs than expected. If the rates are undergraduated, the standardised deviations will swap from positive to negative very often and there will be more runs than expected. However, we do not usually use this test to look for undergraduation. This is because any problems resulting from undergraduation will usually already have been picked up by either the smoothness test (if adhering too closely to the crude rates has led to an erratic pattern of rates) or by the chi-squared test (if adhering too closely to the crude rates over one part of the age range has led to large discrepancies in another part). So the grouping of signs test is a one-sided test as we are worried about a *low* number of groups.

Define the test statistic:

$G =$ Number of *groups* of positive z_x's

Also, suppose that of the m deviations, n_1 are positive and n_2 are negative.

The hypothesis is that the given n_1 positive deviations and n_2 negative deviations are in random order. We, therefore, compute the probability that the number of positive groups will be at least G given n_1 and n_2. Let $t \le G$.

(a) There are $\binom{n_2+1}{t}$ ways to arrange t positive *groups* among n_2 negative *signs*.

There are (n_2+1) places in which the t positive groups can be located: before the first negative sign, after the last negative sign or in any of the (n_2-1) gaps between the signs.

(b) There are $\binom{n_1-1}{t-1}$ ways to arrange n_1 positive *signs* into t positive *groups*.

Question

Explain why there are $\binom{n_1-1}{t-1}$ ways to arrange n_1 positive *signs* into t positive *groups*, assuming that the positive signs occur in a given order.

Solution

We can define a 'separator' as a boundary marking the end of one group and the start of another group. The problem of splitting the n_1 signs into t groups is equivalent to the problem of placing $t-1$ separators in the n_1-1 gaps between the n_1 signs. (No separator can come before the first or after the last sign as a group must contain at least one sign.)

There are $\binom{n_1-1}{t-1}$ ways of placing the $t-1$ separators in the n_1-1 gaps, and so this is the number of ways of splitting the n_1 signs into t groups.

(c) There are $\binom{m}{n_1}$ ways to arrange n_1 positive and n_2 negative *signs*,

since, by definition, $m = n_1 + n_2$.

Hence, the probability of exactly t positive groups is $\dfrac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{m}{n_1}}$.

This formula is given on page 34 of the *Tables*.

Assumptions

None.

Method

Step 1

Determine the sign of the deviation at each age.

Step 2

Count the number of groups of positive signs ($=G$).

Step 3

Since every pair of positive groups must be separated by a negative group, the numbers of positive and negative groups will be small or large alike, so a one-tailed test is appropriate. We should find the smallest k such that:

$$\sum_{t=1}^{k} \frac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{m}{n_1}} \geq 0.05$$

and say that the test has been failed (at the 5% level) if $G < k$.

Alternatively, we could look up the critical value of the test on page 189 of the *Tables.* If the number of groups of positive deviations is less than or equal to the critical value given in the *Tables*, we reject the null hypothesis.

Another alternative is to determine the *p*-value for the test by calculating the probability that the number of groups of positive deviations, G , takes a value less than or equal to the value we have observed.

Question

A graduation covers 20 age groups. The number of positive deviations is 6, and the number of groups of positive deviations is 2. Carry out a grouping of signs test using these data values.

Solution

From page 189 of the *Tables*, we see that the critical value is 2 when $n_1 = 6$ and $n_2 = 14$. Since we have observed 2 groups of positive deviations, we reject the null hypothesis at the 5% significance level and conclude that there is evidence of grouping of deviations of the same sign.

However, if *m* is large enough ($m \geq 20$ or so), we can use a normal approximation as follows:

$$G \sim N\left(\frac{n_1(n_2+1)}{n_1+n_2}, \frac{(n_1 n_2)^2}{(n_1+n_2)^3}\right)$$

This result is also given on page 34 of the *Tables*. Because the test statistic can only take integer values, a continuity correction should be applied when using a normal approximation.

Conclusion

If there are too few runs, this indicates that the rates are overgraduated. The rates do not adhere closely enough to the crude data and may be consistently too high or too low over certain parts of the table.

Strengths and weaknesses

When applying this test, we arbitrarily chose to count the *positive* groups, rather than the *negative* groups. This test can, in some cases, lead to different conclusions depending on whether positive or negative groups are considered.

Question

Test Graduation A for overgraduation using the grouping of signs test.

Solution

Here there are 20 age groups, with 10 positive and 10 negative deviations. From page 189 of the *Tables*, we see that the critical value is 3 when $n_1 = 10$ and $n_2 = 10$. Looking at the column of z_x values on page 12, we see that there are 3 runs of positive deviations. So we reject the null hypothesis at the 5% significance level and conclude that there is evidence of grouping of deviations of the same sign.

Alternatively, we could carry out the test using a normal approximation. The expected number of positive runs is:

$$\frac{10(10+1)}{(10+10)} = 5.5$$

and the variance is:

$$\frac{(10 \times 10)^2}{(10+10)^3} = 1.25$$

The *p*-value is $P(G \le 3)$. Using $N(5.5, 1.25)$ as an approximation to the distribution of G and incorporating a continuity correction:

$$P(G \le 3) \simeq P\left(N(0,1) \le \frac{3.5 - 5.5}{\sqrt{1.25}} \right) = \Phi(-1.78885) = 1 - \Phi(1.78885) = 3.7\%$$

Since the *p*-value is less than 5%, we reject the null hypothesis at the 5% level and conclude that there is evidence of grouping of deviations of the same sign.

7.6 Serial correlations test

A final test, detecting grouping of signs of deviations, is the serial correlations test.

Purpose

To test for clumping of deviations of the same sign (an issue that will not be picked up by a chi-squared test). If clumping is present, then the graduation has the wrong shape.

Rationale

If the graduated rates are neither overgraduated nor undergraduated, we would expect the individual standardised deviations at consecutive ages to behave as if they were independent.

However, if the graduated rates are overgraduated, the graduated mortality curve will tend to stay the same side of the crude rates for relatively long periods and, although there will be random variations in the numbers of deaths, we would expect the values of consecutive deviations to have similar values, *ie* they will be positively correlated.

Conversely, if the rates are undergraduated, the graduated curve will cross the crude rates quite frequently and the values of consecutive deviations will tend to oscillate, *ie* they will be negatively correlated. However, we will use this test as a one-sided test to test for overgraduation since undergraduation is usually tested for using the smoothness test.

If correlations are present, we would expect the effect to be strongest at adjacent ages or at ages separated by 2 or 3 years.

Under the null hypothesis, the two sequences (of length $m-1$):

$$z_1, z_2, \ldots, z_{m-2}, z_{m-1}$$

and:

$$z_2, z_3, \ldots, z_{m-1}, z_m$$

should be uncorrelated.

So should the sequences (of length $m-2$):

$$z_1, z_2, \ldots, z_{m-3}, z_{m-2}$$

and:

$$z_3, z_4, \ldots, z_{m-1}, z_m$$

We call these the lagged sequences, with lag 1 and lag 2 respectively, and we define sequences with longer lags in the obvious way.

The correlation coefficient of the j th lagged sequences is:

$$r_j = \frac{\sum_{i=1}^{m-j}(z_i - \bar{z}^{(1)})(z_{i+j} - \bar{z}^{(2)})}{\sqrt{\sum_{i=1}^{m-j}(z_i - \bar{z}^{(1)})^2 \sum_{i=1}^{m-j}(z_{i+j} - \bar{z}^{(2)})^2}}$$

where $\bar{z}^{(1)} = \frac{1}{m-j}\sum_{i=1}^{m-j} z_i$ **and** $\bar{z}^{(2)} = \frac{1}{m-j}\sum_{i=1}^{m-j} z_{i+j}$ **.**

This ratio gives the *serial correlation coefficients* r_j, which can take values in the range $-1 \le r_j \le 1$. A positive value indicates that nearby values of z_x tend to have similar values, whereas a negative value indicates that they tend to have opposite values.

If m is large enough, we can approximate $\bar{z}^{(1)}$ and $\bar{z}^{(2)}$ by $\bar{z} = \frac{1}{m}\sum_{i=1}^{m} z_i$, and simplify the above, to obtain:

$$r_j \cong \frac{\sum_{i=1}^{m-j}(z_i - \bar{z})(z_{i+j} - \bar{z})}{\frac{m-j}{m}\sum_{i=1}^{m}(z_i - \bar{z})^2} \qquad \textbf{(10.1)}$$

or:

$$r_j \cong \frac{\frac{1}{m-j}\sum_{i=1}^{m-j}(z_i - \bar{z})(z_{i+j} - \bar{z})}{\frac{1}{m}\sum_{i=1}^{m}(z_i - \bar{z})^2}$$

The second form shows that this is just the ratio of two averages. This formula for the approximation is given on page 34 of the *Tables*.

The difference between this approximation and the exact formula above is negligible for large m. However, the approximate formula may be inappropriate for small values of m. It is generally acceptable in practice, but its limitations should be borne in mind and, in particular, if the test is to be carried out on *real* data where m is small (*eg* less than 20) then the exact formula should be used.

It is known that $r_j \sim N(0,1/m)$, under the null hypothesis.

So multiplying r_j by $\sqrt{m}$ should give a value that comes from a standard normal distribution.

Hence, $r_j\sqrt{m}$ can be tested against the $N(0,1)$ distribution. Too high a value indicates a tendency for deviations of the same sign to cluster.

Assumptions

None.

Method

Step 1

Calculate the standardised deviations z_x for each age or age group.

Step 2

Calculate the serial correlation coefficients using the formula:

$$r_j = \frac{\dfrac{1}{m-j}\displaystyle\sum_{i=1}^{m-j}(z_i - \bar{z})(z_{i+j} - \bar{z})}{\dfrac{1}{m}\displaystyle\sum_{i=1}^{m}(z_i - \bar{z})^2}$$

where $\bar{z} = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} z_i$ is the overall average of z_x for the m ages (or age groups).

We can use the fact that r_j must take values in the range $-1 \le r_j \le 1$ to check the calculations for reasonableness.

Step 3

Multiply by $\sqrt{m}$ to obtain the value of the test statistic and compare this with the percentage points of the standard normal distribution.

Conclusion

If the test statistic is 'too positive', this indicates that the rates are overgraduated. The rates do not adhere closely enough to the crude data and may be consistently too high or too low over certain parts of the table.

Strengths and weaknesses

The serial correlation test takes into account actual numerical values of the standardised deviations (whereas the grouping of signs test ignores the magnitude of the deviations). As a result, it is possible for correlations in one part of the age range to be cancelled out by opposite correlations in another part. This means that the grouping of signs test is usually more powerful, *ie* it is more likely to detect overgraduation if this is present.

Question

Carry out the serial correlation test at lag 1 for Graduation A.

Solution

The mean of the individual standardised deviations is:

$$\bar{z} = \frac{1}{20}\sum_{x=30}^{49} z_x = \frac{1}{20}(2.27 + 2.69 + \ldots + 1.99) = 0.18$$

The denominator of r_1 is:

$$\frac{1}{20}\sum_{x=30}^{49} (z_x - \bar{z})^2 = 2.13$$

The numerator of r_1 is:

$$\frac{1}{19}\sum_{x=30}^{48} (z_x - \bar{z})(z_{x+1} - \bar{z}) = 0.94$$

So:

$$r_1 = \frac{0.94}{2.13} = 0.44$$

and the value of the test statistic is

$$\sqrt{20} \times 0.44 = 1.97$$

This is more than 1.6449, the upper 5% point of the standard normal distribution. So there is evidence of grouping of deviations of the same sign.

The serial correlations test can be carried out in R by considering the series of z_x's as if they were a time series and computing the first-order autocorrelation.

Alternatively, to compute r_1 based on formula (10.1) in R from a set of 50 deviations z_x, use:

```
z1x <- zx[1:49]
z2x <- zx[2:50]
cor(z1x, z2x)
```

7.7 Testing actual versus expected rates

We have mainly looked at the tests in the context of testing graduations. However, the tests above can also be used where we wish to test a set of observed rates against an existing table to which we think the rates conform.

The tests will be carried out as previously. In this case though there is no equivalent to undergraduation. We are looking only for goodness of fit.

The statistical tests that assess the significance of differences between observed and expected values operate on the basis that the true underlying mortality rates at each age are those specified by the expected mortality basis.

The null hypothesis in this case is:

H_0 : The mortality rates being tested are consistent with the rates from the standard table.

The purpose of each of the various tests in this case is summarised below.

Chi-squared test

This will be a one-sided test for goodness of fit. The null hypothesis will be rejected if the test statistic exceeds the upper 5% level. In this case the number of degrees of freedom will normally just be the number of age groups being considered.

Distribution of ISDs

This test will again be used to examine goodness of fit. In particular it will identify any excessively large deviations.

Signs test

This test is used to identify any imbalance between positive and negative deviations, *ie* to ensure that the observed rates are not consistently above or below the expected rates.

Cumulative deviations test

This test will detect a large positive or negative cumulative deviation, as previously.

Grouping of signs test

This test detects excessive clumping of deviations of the same sign. A small number of groups of positive deviations indicates that the shape of the true rates underlying the observed rates is significantly different from the expected mortality rates, at least over part of the range.

Serial correlation test

This is another one-sided test to detect clumping of deviations of the same sign. A large positive value of the test statistic indicates that the shape of the true rates underlying the observed rates is significantly different from the expected mortality rates, at least over part of the range.

The chapter summary starts on the next page so that you can
keep all the chapter summaries together for revision purposes.

Chapter 10 Summary

Purpose of graduation

The crude mortality rates derived from a mortality investigation are graduated to make them acceptable for use in actuarial calculations. Graduation refers to the process of using statistical techniques to improve the estimates provided by the crude rates. The aims of graduation are:

- to produce a smooth set of rates that are suitable for a particular purpose
- to remove random sampling errors (as far as possible)
- to use the information available from adjacent ages to improve the reliability of the estimates.

Graduated rates should move smoothly between adjacent years of age. This is based on the theoretical assumption that underlying mortality rates progress smoothly from year to year and on the practical desire to perform financial calculations (*eg* to calculate premiums) that are consistent from one age to the next.

The process of graduation involves a trade-off between smoothness and goodness of fit. The suitability of a graduation can be assessed using statistical tests.

Null hypothesis

The null hypothesis for each of these tests is that the graduated rates are equal to the true underlying mortality rates.

Testing smoothness

Smoothness (undergraduation) can be judged by examining the third differences of the graduated rates.

$$\Delta \mathring{\mu}_x = \mathring{\mu}_{x+1} - \mathring{\mu}_x \qquad \Delta^2 \mathring{\mu}_x = \Delta \mathring{\mu}_{x+1} - \Delta \mathring{\mu}_x \qquad \Delta^3 \mathring{\mu}_x = \Delta^2 \mathring{\mu}_{x+1} - \Delta^2 \mathring{\mu}_x$$

Testing goodness of fit

Many of the tests of goodness of fit are based on the values of the individual standardised deviations, which provide information about the individual ages or age groups. The individual standardised deviations are given by:

$$z_x = \frac{D_x - E_x^c \mathring{\mu}_x}{\sqrt{E_x^c \mathring{\mu}_x}}$$

Under the null hypothesis, these random variables follow the standard normal distribution.

Chi-squared test

The test statistic is:

$$\sum_{\substack{\text{all ages}\\ x}} z_x^2$$

Under the null hypothesis, this has a chi-squared distribution. The number of degrees of freedom depends on the number of ages and the method of graduation.

The chi-squared test gives an overall assessment of the goodness of fit, but can miss features such as:

- bias over the whole age range
- clumping of the signs of deviations (which may indicate overgraduation)
- outliers balanced by small deviations.

Cumulative deviations test

The test statistic is:

$$\frac{\sum_{\substack{\text{all}\\ \text{ages}}} (D_x - E_x^c \overset{\circ}{\mu}_x)}{\sqrt{\sum_{\substack{\text{all}\\ \text{ages}}} E_x^c \overset{\circ}{\mu}_x}}$$

Under the null hypothesis, this should come from the standard normal distribution.

The cumulative deviations test looks at the overall deviation over a range of ages.

Signs test

If there are m age groups in total, then under the null hypothesis:

Number of positive deviations $\sim Binomial(m, ½)$

The signs test looks at the distribution of positive and negative deviations, but it ignores the magnitude of the deviations. The test can be carried out by calculating the p-value.

Cumulative probabilities are listed on pages 186-188 of the *Tables* for certain values of m. Alternatively, if m is large, we can use a normal approximation with a continuity correction.

Grouping of signs test

This is a one-tailed test that checks for excessive clumping of deviations of the same sign. If n_1 = number of positive signs and n_2 = number of negative signs, then:

$$\text{P}(t \text{ positive groups}) = \frac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{n_1+n_2}{n_1}}, \quad n_1 \geq 1, t \geq 1$$

$$\text{Number of positive groups } \sim N\left(\frac{n_1(n_2+1)}{n_1+n_2}, \frac{(n_1 n_2)^2}{(n_1+n_2)^3}\right) \text{ approximately}$$

These formulae are given on page 34 of the *Tables.*

The test can also be carried out by comparing the number of groups of positive deviations with the critical value. Critical values are given on page 189 of the *Tables*.

Serial correlations test

The correlation coefficient at lag 1 is:

$$r_1 = \frac{\frac{1}{m-1}\sum_{i=1}^{m-1}(z_i - \overline{z})(z_{i+1} - \overline{z})}{\frac{1}{m}\sum_{i=1}^{m}(z_i - \overline{z})^2} \qquad \text{where} \quad \overline{z} = \frac{1}{m}\sum_{i=1}^{m} z_i$$

and the test statistic is:

$$r_1\sqrt{m}$$

These formulae are given on page 34 of the *Tables*. Under the null hypothesis, the test statistic follows the standard normal distribution. This is a one-tailed test, which tests for positive correlation.

Each of these tests concentrates on different features of the graduation. Several tests must be applied before deciding that a set of graduated rates is acceptable. However, an unfavourable result on just one test may be sufficient to reject a set of rates.

We can also use these tests to check whether a set of observed mortality rates conforms to an existing standard table. In this case, the null hypothesis is that the rates being tested are consistent with those from the standard table.

The practice questions start on the next page so that you can keep all the chapter summaries together for revision purposes.

Chapter 10 Practice Questions

10.1 Explain why it is necessary to graduate crude rates of mortality for practical use.

10.2 State the characteristics of a good graduation.

10.3 Two graduations have been carried out and you have obtained the following results:

Graduation 1 covers 40 age groups and has produced 13 positive and 27 negative standardised deviations.

Graduation 2 covers 7 age groups and has produced 5 positive and 2 negative standardised deviations.

Carry out the signs test on these graduations, clearly stating the probability value and your conclusion in each case.

10.4 Apply the chi-squared test to Graduation B, given on page 13 of this chapter.

10.5 Analyse the distribution of the standardised deviations for Graduation B.

10.6 Apply the grouping of signs test to Graduation B.

10.7 Carry out the serial correlation test at lag 1 on Graduation B.

For Graduation B you are given that:

$$\sum(z_x - \bar{z})(z_{x+1} - \bar{z}) = -5.10 \quad \text{and} \quad \sum(z_x - \bar{z})^2 = 9.36$$

10.8 Carry out the smoothness test on the following set of graduated rates:

x	$\mathring{\mu}_x$
55	0.00429
56	0.00478
57	0.00535
58	0.00596
59	0.00667
60	0.00754
61	0.00867

10.9 A graduation of a set of mortality rates from age 25 to age 64 has 15 positive individual standardised deviations, which occurred in 8 groups.

Exam style

Carry out two tests to check the suitability of this graduation. [6]

10.10 Exam style

(i) Explain why graduated rates, rather than crude estimates of mortality rates, are used in the construction of standard mortality tables. [3]

(ii) A graduation of the mortality experience of the male population of a region of the United Kingdom has been carried out using a formula with 3 parameters. The following is an extract from the results.

Age nearest birthday	Actual number of deaths	Graduated mortality rate	Central exposed to risk	
x	θ_x	$\overset{\circ}{\mu}_x$	E_x^c	$E_x^c\overset{\circ}{\mu}_x$
14	3	0.00038	12,800	4.864
15	8	0.00043	15,300	6.579
16	5	0.00048	12,500	6.000
17	14	0.00053	15,000	7.950
18	17	0.00059	16,500	9.735
19	9	0.00066	10,100	6.666
20	15	0.00074	12,800	9.472
21	10	0.00083	13,700	11.371
22	10	0.00093	11,900	11.067
Total	91			73.704

Use the chi-squared test to test the adherence of the graduated rates to the data. State clearly the null hypothesis you are testing and comment on the result. [4]

(iii) Perform two other tests that detect different aspects of the adherence of the graduation to the data. For each test state clearly the features of the graduation that the test is able to detect, and comment on your results. [8]

[Total 15]

10.11 An insurance company is concerned that the ratio between the mortality of its female and male pensioners is unlike the corresponding ratio among insured pensioners in general. It conducts an investigation and estimates the mortality of male and female pensioners, $\hat{\mu}_x^m$ and $\hat{\mu}_x^f$. It then uses the $\hat{\mu}_x^m$ to calculate what the expected mortality of its female pensioners would be if the ratio between male and female mortality rates reflected the corresponding ratio in the PMA92 and PFA92 tables, S_x, using the formula:

Exam style

$$\tilde{\mu}_x^f = \hat{\mu}_x^m \, S_x$$

The table below shows, for a range of ages, the numbers of female deaths actually observed in the investigation and the number which would be expected from the $\tilde{\mu}_x^f$.

Age, *x*	Actual deaths $E_x^c \hat{\mu}_x^f$	Expected deaths $E_x^c \tilde{\mu}_x^f$
65	30	28.4
66	20	30.1
67	25	31.2
68	40	33.5
69	45	34.1
70	50	41.8
71	50	46.5
72	45	44.5

(i) Describe and carry out an overall test of the hypothesis that the ratios between male and female death rates among the company's pensioners are the same as those of insured pensioners in general. Clearly state your conclusion. [5]

(ii) Investigate further the possible existence of unusual ratios between male and female death rates among the company's pensioners, using two other appropriate statistical tests. [6]

[Total 11]

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 10 Solutions

10.1 It is necessary to graduate crude rates for practical use for the following reasons:

- to make them fit for the purpose for which they are intended
- to remove random sampling errors, thus better estimating the true underlying mortality rates
- to allow the rate at a particular age to be set with reference to the rates at adjacent ages
- to produce a set of mortality rates that progresses smoothly from age to age, which allows a practical smooth set of premium rates to be produced.

10.2 A set of graduated rates should adhere to the data and be smooth enough for the purposes that it will be used for.

10.3 For Graduation 1, there are enough values to use a normal approximation. There are fewer positives than expected and:

$$\begin{aligned}P(\leq 13 \text{ positives}) &= P[Binomial(40,½) \leq 13]\\ &\approx P[N(20,10) < 13.5]\\ &= P\left[N(0,1) < \frac{13.5-20}{\sqrt{10}}\right]\\ &= \Phi(-2.0555)\\ &= 1-\Phi(2.0555)\\ &= 0.0199\end{aligned}$$

Since this is a two-tailed test, the *p*-value is approximately $2 \times 0.0199 = 0.0398$. This is less than 0.05, so there is evidence at the 5% significance level to conclude that there is bias in the graduated rates.

For Graduation 2 there are only 7 values, so an exact calculation must be used. There are more positives than expected and:

$$\begin{aligned}P(\geq 5 \text{ positives}) &= P[Binomial(7,½)=5]+P[Binomial(7,½)=6]+P[Binomial(7,½)=7]\\ &= \binom{7}{5}\left(\frac{1}{2}\right)^7+\binom{7}{6}\left(\frac{1}{2}\right)^7+\binom{7}{7}\left(\frac{1}{2}\right)^7\\ &= \frac{21+7+1}{2^7}\\ &= 0.2266\end{aligned}$$

Alternatively:

$$P(\geq 5 \text{ positives}) = 1-P(\leq 4 \text{ positives}) = 1-P(Binomial(7,½) \leq 4)$$

Page 186 of the Tables gives:

$$P(Binomial(7,½) \le 4) = 0.7734$$

Hence:

$$P(\ge 5 \text{ positives}) = 1 - 0.7734 = 0.2266$$

as before.

So, the *p*-value is $2 \times 0.2266 = 0.453$, which is not significant. So, in this case, we conclude that the signs test doesn't provide enough evidence to conclude that the graduation is biased.

10.4 From the given table of values, we see that:

$$\sum z_x^2 = 9.39$$

There are 20 ages. We have not constrained the totals. The graduated rates have been calculated by estimating 11 parameters. So, the number of degrees of freedom is $20 - 11 = 9$.

From page 169 of the *Tables*, the upper 5% point of χ_9^2 distribution is 16.92. The observed value of the test statistic is less than this, so we conclude that the graduated rates are a good fit to the data.

10.5 The observed and expected frequencies are as follows:

Interval	$(-\infty,-3)$	$(-3,-2)$	$(-2,-1)$	$(-1,0)$	$(0,1)$	$(1,2)$	$(2,3)$	$(3,\infty)$
Observed	0	0	1	7	11	1	0	0
Expected	0.0	0.4	2.8	6.8	6.8	2.8	0.4	0

There are no z_x values greater than 2 in magnitude, where we would expect less than 1. So this seems fine. We would expect around 50% of the z_x values, *ie* 10, to fall in the range $(-2/3, 2/3)$, but we actually have 13. This indicates that the deviations are a little smaller in size than expected, but this is not a sufficiently extreme result to cast doubt on the null hypothesis. The distribution of the z_x values is fairly symmetrical (12 positive and 8 negative deviations).

If we combine the small groups by pooling the values in the ranges $(-\infty,-1)$ and $(1,\infty)$, we can apply a chi-squared test to the resulting 4 groups. The value of the test statistic is:

$$\frac{(1-3.2)^2}{3.2} + \frac{(7-6.8)^2}{6.8} + \frac{(11-6.8)^2}{6.8} + \frac{(1-3.2)^2}{3.2} = 5.625$$

This is less than 7.815, the upper 5% point of χ_3^2, so there is insufficient evidence to reject the hypothesis that the graduated rates are the true mortality rates underlying the data. We conclude from this test that the graduated rates are a good fit to the data.

The chi-squared test is not strictly valid here as two of the expected frequencies are less than 5.

10.6 The grouping of signs test is a one-tailed test and we are testing to see if there are too few positive groups. Here we have $n_1 = 12$ and $n_2 = 8$. From page 189 of the *Tables*, we find that the critical value of the test is 3. The observed number of runs of positive deviations is 9. This is greater than the critical value, which indicates that there is no problem with clumping of deviations of the same sign.

10.7 The serial correlation coefficient is:

$$r_1 = \frac{\frac{1}{m-1}\sum(z_x - \bar{z})(z_{x+1} - \bar{z})}{\frac{1}{m}\sum(z_x - \bar{z})^2} = \frac{-5.10/19}{9.36/20} = -0.57$$

and the value of the test statistic is:

$$r_1\sqrt{m} = -2.56$$

This serial correlation test is a one-tailed test, and a negative test statistic does not lead us to reject the null hypothesis that the graduated rates are the true mortality rates underlying the data.

10.8 The first, second and third differences in the graduated rates are shown in the table below:

x	$\overset{\circ}{\mu}_x$	$\Delta\overset{\circ}{\mu}_x$	$\Delta^2\overset{\circ}{\mu}_x$	$\Delta^3\overset{\circ}{\mu}_x$
55	0.00429	0.00049	0.00008	−0.00004
56	0.00478	0.00057	0.00004	0.00006
57	0.00535	0.00061	0.00010	0.00006
58	0.00596	0.00071	0.00016	0.00010
59	0.00667	0.00087	0.00026	
60	0.00754	0.00113		
61	0.00867			

The third differences are small in magnitude compared to the graduated rates, and the numbers in the third difference column change fairly smoothly. So we conclude that the graduated rates are smooth.

10.9 *The statistical information given allows us to carry out a signs test and a grouping of signs test only.*

The null hypothesis is that the graduated rates are the true mortality rates for the population. [½]

Signs test

The observed value of the test statistic, P, is 15. If the null hypothesis is true, then the sampling distribution of the test statistic will be $Binomial(40, 0.5)$. [1]

The number of age groups is large enough so that we can use a normal approximation. We are using a continuous distribution to approximate a discrete distribution, so we should use a continuity correction.

Using a normal approximation with continuity correction:

$$\begin{aligned} P(Binomial(40,0.5)\leq 15) &\approx P\left(N(0,1)<\frac{15.5-40\times 0.5}{\sqrt{40\times 0.5\times 0.5}}\right) \\ &=\Phi(-1.4230) \\ &=1-\Phi(1.4230) \\ &=1-0.9226 \\ &=0.0774 \end{aligned}$$ [1]

Since this is a two-tailed test, the *p*-value is approximately $2\times 0.0774 = 0.155$. Since this is greater than 5%, the result is not significant at the 5% level. [1]

Grouping of signs test

The observed value of the test statistic, G, is 8. [½]

This is a one-tailed test and small values of the test statistic are significant. The critical values of the test statistic are given on page 189 of the *Tables*. [½]

Here n_1, the number of positive deviations is 15 and n_2, the number of negative deviations is 25. The critical value of the test statistic at the 5% level of significance is 6. [1]

So the observed value of the test statistic does not lie in the critical region and the data support the null hypothesis. [½]

10.10 (i) ***Reasons for graduation***

We expect the true rates to progress smoothly, with no irregularities. [1]

Graduation reduces sampling errors at each age by using information from adjacent ages. [½]

Standard tables are used for premium and reserve calculations, where it is important to have unbiased estimates of the true underlying rates. [½]

Premiums should vary smoothly with age (as policyholders would expect). [1]

(ii) ***Chi-squared test***

The null hypothesis is:

H_0 : the graduated rates are the true underlying mortality rates for the population [½]

We calculate the individual standardised deviations at each age using the formula:

$$z_x = \frac{\theta_x - E_x^c \mathring{\mu}_x}{\sqrt{E_x^c \mathring{\mu}_x}}$$

The ISDs are:

$-0.845,\ 0.554,\ -0.408,\ 2.146,\ 2.328,\ 0.904,\ 1.796,\ -0.407,\ -0.321$ [1]

The test statistic for the chi-squared test (based on unrounded z_x values) is:

$\sum z_x^2 = 15.53$ [1]

We now compare this with a χ^2 distribution. We were given data from 9 ages. Since the graduation was carried out using a formula with 3 parameters, we lose 3 degrees of freedom. So we are left with 6 degrees of freedom. [½]

From the *Tables*, we see that the upper 5% point of χ_6^2 is 12.59. As the value of the test statistic is greater than this, we reject the null hypothesis and conclude that the graduated rates do not provide a good fit to the data. In particular, it looks like the graduated rates are too low for ages 17 to 20. [1]

(iii) ***Two other tests***

You can take your pick here from the individual standardised deviations test, the signs test, the cumulative deviations test, the grouping of signs test and the serial correlation test.

The null hypothesis for all the tests is:

H_0 : the graduated rates are the true underlying mortality rates for the population [½]

ISD Test

This is a good all round test that detects most of the problems that might be present in a graduation including any outliers. [½]

For this test we compare the z_x values with a standard normal distribution:

	$(-\infty,-3)$	$(-3,-2)$	$(-2,-1)$	$(-1,0)$	$(0,1)$	$(1,2)$	$(2,3)$	$(3,\infty)$
Obs	0	0	0	4	2	1	2	0
Exp	0	0.18	1.26	3.06	3.06	1.26	0.18	0

[½]

There are 4 things to consider here:

- Outliers – there are no ISDs greater than 3 in absolute value, which is good; however, with fewer than 20 age groups, we should be suspicious about any ISD greater than 2 in magnitude, and here we have 2 ISDs greater than 2. [½]
- The balance of positive and negative deviations – this is OK. [½]

- Symmetry – the distribution is a bit positively skewed, which is not so good. [½]
- Proportion of ISDs lying in the range $(-\tfrac{2}{3},\tfrac{2}{3})$ should be $\tfrac{1}{2}$ – it is $\tfrac{4}{9}$ here, which is OK. [½]

The graduated rates fail this test as the ISDs do not appear to be normally distributed. In particular, the graduated rates appear to be too low at ages 17 and 18. [½]

Signs test

This is a simple two-tailed test for overall bias. [½]

There should be roughly equal numbers of positive and negative ISDs. Under the null hypothesis, the number of positive deviations has a $Binomial(9,0.5)$ distribution. [1]

We have 5 positives and 4 negatives, which is fine. [1]

So we do not reject the null hypothesis and we conclude that there is no evidence of overall bias in the graduated rates. [1]

Cumulative deviations test

This is a two-tailed test for overall bias. [½]

The observed value of the test statistic is:

$$\frac{\sum \theta_x - \sum E_x^c \overset{\circ}{\mu}_x}{\sqrt{\sum E_x^c \overset{\circ}{\mu}_x}} = \frac{91-73.704}{\sqrt{73.704}} = 2.015 \qquad [2]$$

For a test at the 5% significance level, we compare the value of the test statistic with the lower and upper 2.5% points of $N(0,1)$, *ie* with ±1.96. Since 2.015 is greater than 1.96, we reject the null hypothesis and conclude that the graduated rates are too low overall. [1]

Make sure that you don't do both of the signs test and the cumulative deviations test as they both test for the same thing.

Grouping of signs test

This is a one-tailed test that detects clumping of deviations of the same sign. [½]

The observed number of positive deviations is 5, and the observed number of negative deviations is 4. [1]

From the *Tables,* we find that the critical value is 1 and we reject the null hypothesis if the observed number of positive runs is less than or equal to this. [1]

The observed number of positive runs is 2, so we do not reject the null hypothesis in this case, and we conclude that there is no evidence of grouping of signs. [1]

Serial correlation test

This is an alternative test for grouping of signs, but it takes much longer to carry out this test so we don't recommend that you do it unless you absolutely have to. Make sure that you don't carry out both the grouping of signs test and the lag-1 serial correlation test since they test for the same thing.

This is a one-tailed test that detects clumping of deviations of the same sign. [½]

The serial correlation coefficient at lag 1 is:

$$r_1 = \frac{\frac{1}{8}\sum_{x=14}^{21}(z_x - \bar{z})(z_{x+1} - \bar{z})}{\frac{1}{9}\sum_{x=14}^{22}(z_x - \bar{z})^2} = \frac{0.2165}{1.3172} = 0.1643 \qquad [2]$$

and the value of the test statistic is:

$$r_1\sqrt{m} = 0.1643 \times 3 = 0.493 \qquad [½]$$

As we are only testing for positive correlation, we compare the value of the test statistic with 1.6449, the upper 5% point of $N(0,1)$. We find that there is insufficient evidence to reject the null hypothesis or, in other words, there is no evidence of grouping of signs. [½]

Comment

The graduation has not fully taken into account the accident hump, *ie* the increase in mortality around the late teens and early twenties. [½]

10.11 (i) ***Goodness-of-fit test***

The null hypothesis is:

H_0 : the ratios between male and female death rates among the company's pensioners are the same as those of insured pensioners in general. [1]

An overall test of this hypothesis can be done using a χ^2 goodness-of-fit test. The test statistic for this test is:

$$\sum_x \frac{(O-E)^2}{E}$$

where O denotes the observed number of deaths and E denotes the expected number of deaths.

Using the given data, the value of the test statistic is:

$$\frac{(30-28.4)^2}{28.4}+\frac{(20-30.1)^2}{30.1}+\frac{(25-31.2)^2}{31.2}+\frac{(40-33.5)^2}{33.5}$$

$$+\frac{(45-34.1)^2}{34.1}+\frac{(50-41.8)^2}{41.8}+\frac{(50-46.5)^2}{46.5}+\frac{(45-44.5)^2}{44.5}$$

$$=0.0901+3.3890+1.2321+1.2612+3.4842+1.6086+0.2634+0.0056$$

$$=11.334$$ [2]

This is a one-tailed test. When comparing an experience with a standard table, the number of degrees of freedom in the chi-squared test is the number of age groups considered. So in this case we compare the value of the test statistic with the χ_8^2 distribution. [1]

The upper 5% point of χ_8^2 is 15.51. Since the value of the test statistic is less than 15.51, there is insufficient evidence to reject the null hypothesis. So we conclude that the ratios between male and female death rates among the company's pensioners are the same as those of insured pensioners in general. [1]

(ii) ***Other appropriate tests***

The null hypothesis for all the tests is:

H_0 : the ratios between male and female death rates among the company's pensioners are the same as those of insured pensioners in general. [½]

We could also carry out the signs test on these data. This test checks for overall bias, *ie* whether the actual deaths are systematically higher or lower than the expected deaths. [½]

The differences (actual – expected) at each age are:

1.6, –10.1, –6.2, 6.5, 10.9, 8.2, 3.5, 0.5 [½]

6 of these are positive and 2 are negative.

Let N denote the number of positive differences. Under the null hypothesis:

$$N \sim Binomial(8,½)$$ [½]

We have observed 6 positives, which is more than the expected number of 4. From page 187 of the *Tables*:

$$P(N\geq 6)=1-P(N\leq 5)=1-0.8555=0.1445$$ [½]

Since this is a two-tailed test, its p-value is:

$$2P(N\geq 6)=2\times 0.1445=0.289$$ [½]

This is (much) greater than 0.05, so there is insufficient evidence to reject the null hypothesis. [½]

Another test that checks for overall bias is the cumulative deviations test. So you could do this one instead of the signs test. Don't do both the signs test and the cumulative deviations test as they test for the same thing (overall bias).

The test statistic for the cumulative deviations test is:

$$\frac{\sum O - \sum E}{\sqrt{\sum E}}$$

Under the null hypothesis, this has a standard normal distribution. From the given data:

$$\sum O = 305 \qquad \sum E = 290.1$$

So the value of the test statistic is:

$$\frac{305 - 290.1}{\sqrt{290.1}} = 0.875$$

This is a two-tailed test. Using a 5% significance level, the critical values are ± 1.96. *Since the value of the test statistic lies between these two critical values, there is insufficient evidence to reject the null hypothesis (at the 5% significance level).*

Another test you could do is the standardised deviations test, which tests for outliers. [½]

We calculate the individual standard deviation (ISD) at each age using the formula $\frac{O-E}{\sqrt{E}}$. The ISDs are:

0.3002, –1.8409, –1.1100, 1.1230, 1.8666, 1.2683, 0.5133, 0.0750 [1]

Under the null hypothesis, these should be random values from the standard normal distribution. None of the ISDs are greater than 1.96 in absolute value, *ie* there are no outliers. There are 3 values in the range $(-\frac{2}{3}, \frac{2}{3})$, which is slightly lower than the expected value of 4, but this is OK. [½]

The balance of positive/negative deviations has already been investigated by the signs test and has been found to be satisfactory.

So there is no evidence to reject the null hypothesis. [½]

You could also carry out the serial correlation test here. However, the serial correlation test involves a great deal of calculation, and is to be avoided if at all possible.

You may also have considered carrying out a grouping of signs test. This test checks for bias over parts of the age range.

Using the notation that is consistent with that used in the Tables, we have:

n_1 = the number of positive differences $= 6$

n_2 = the number of negative differences $= 2$

G = the number of positive groups $= 2$

From page 189 of the Tables, we see that no value of G *would be significant for the given values of* n_1 *and* n_2.

11

Methods of graduation

Syllabus objectives

4.5 Graduation and graduation tests

4.5.4 Describe the process of graduation by the following methods, and state the advantages and disadvantages of each:

- parametric formula
- standard table
- spline functions.

(The student will not be required to carry out a graduation.)

4.5.5 Describe how the tests in 4.5.1 (for the comparison of crude estimates with a standard table) should be amended to compare crude and graduated sets of estimates.

4.5.6 Describe how the tests in 4.5.1 (for the comparison of crude estimates with a standard table) should be amended to allow for the presence of duplicate policies.

4.5.7 Carry out a comparison of a set of crude estimates and a standard table, or of a set of crude estimates and a set of graduated estimates.

0 Introduction

In this chapter, we will look at three methods of carrying out a graduation:

- graduation by parametric formula
- graduation by reference to a standard table
- graduation using spline functions.

The most appropriate method of graduation to use will depend on the quality of the data available and the purpose for which the graduated rates will be used.

The general methodology of graduation is essentially the same under each method. Once we have decided on the appropriate method, we will choose a model to represent the underlying force of mortality, fit the model to the crude observed rates and test the graduation for adherence to data and (if necessary) smoothness. Each method can produce many possible graduations. The graduation chosen will be the one whose adherence and smoothness best meet the requirements for which the rates are intended.

Graduation is a compromise between adherence to data (goodness of fit) and smoothness. The balance that we want between these two conflicting objectives is a subjective choice and will depend on how the graduated rates will be used. For example:

- If we are constructing a standard table of national population mortality, we will be interested in maximising the accuracy. We will put more emphasis on adherence and less emphasis on smoothness.
- If the rates are to be used to calculate premiums and reserves for a life insurance company, we will want to ensure that the rates (and hence the premiums and reserves) progress smoothly from age to age to avoid sudden changes and inconsistencies. We will put more emphasis on smoothness and less emphasis on adherence. The mortality rates at ages around the accident hump will be less important in this situation as few policyholders are likely to be in the age range 18-22.

Recall that the precise form of some of the statistical tests that we described in Chapter 10 depends on the method of graduation used. It's a good idea to re-read the relevant sections of Chapter 10 after completing this chapter.

1 Graduation by parametric formula

1.1 Overview

The method of graduation most often used for reasonably large experiences is to fit a parametric formula to the crude estimates.

The underlying assumption is that μ_x can be modelled using an appropriate mathematical formula with unknown parameters. The parameters are typically calculated automatically by a computer using numerical methods.

If the formula used does not include enough parameters, it will not be flexible enough to follow the crude rates closely, which may result in overgraduation. If too many parameters are included, sudden bends may appear in the graduated curve, which may result in undergraduation.

For different values of the parameters, we can assess the smoothness and adherence to data of the fitted model. (In practice we will not need to check smoothness if the number of parameters is sufficiently small.)

We will choose the values of the parameters that provide the most appropriate model, according to some pre-defined criterion in respect of goodness of fit.

1.2 Choosing and fitting parametric formulae

Two simple (but useful) formulae are:

Gompertz (1825) $\mu_x = Bc^x$

Makeham (1860) $\mu_x = A + Bc^x$

We described these simple laws of mortality in Chapter 6.

In practice, it is usually found that μ_x follows an exponential curve quite closely over middle and older ages (in human populations) so most successful formulae include a Gompertz term. Makeham's formula is interpreted as the addition of accidental deaths, independent of age, to a Gompertz term representing senescent deaths.

The most recent standard tables produced for use by UK life insurance companies used formulae of the form

$$\mu_x = \textbf{polynomial}_1 + \textbf{exp}(\textbf{polynomial}_2)$$

which includes Gompertz and Makeham as special cases.

Question

Define $polynomial_1$ and $polynomial_2$ for the special case of Makeham's formula.

Solution

In Makeham's formula:

$$\text{polynomial}_1 = A$$

$$\text{polynomial}_2 = \log B + x \log c$$

We then have:

$$\mu_x = A + \exp(\log B + x \log c) = A + Bc^x$$

A wide range of techniques is available to choose and to fit a curve to a set of crude estimates. Here we just describe how the fitting was carried out in the case of some recent UK life insurance standard tables.

1.3 A practical example

The available data were deaths and central exposed to risk, and the Poisson model was used. The data were collected by life insurance companies during 1999-2002, and were analysed by the Continuous Mortality Investigation (CMI). Different tables were prepared for males and females, and for different classes of insurance and pension business; collectively they are known as the '00 series' tables.

'Classes' of insurance refers to the different types of policy, *eg* whole life annuity or term assurance.

The formulae were of the form:

$$\mathring{\mu}_x = f(\alpha_1, \alpha_2, \ldots, \alpha_r, \alpha_{r+1}, \alpha_{r+2}, \ldots, \alpha_{r+s}, x)$$

where

$$\textbf{polynomial(1)} = \alpha_1 + \alpha_2 x + \alpha_3 x^2 + \ldots + \alpha_r x^{r-1}$$

$$\textbf{polynomial(2)} = \alpha_{r+1} + \alpha_{r+2} x + \alpha_{r+3} x^2 + \ldots + \alpha_{r+s} x^{s-1}$$

In other words, a formula with $(r+s)$ parameters of the form:

$$\mu_x = \text{polynomial}_1 + \exp(\text{polynomial}_2)$$

was fitted for each table .

Under the Poisson model (covered in Chapter 3), the probability of observing d_x deaths from a central exposed to risk E_x^c where x denotes age nearest birthday is given by:

$$P\left(D_x = d_x\right) = \frac{(E_x^c \mu_x)^{d_x} \exp(-E_x^c \mu_x)}{d_x!}$$

Therefore, in respect of the age interval $[x-½, x+½]$, the likelihood in the Poisson model is:

$$(\mu_x)^{d_x}.\exp(-\mu_x E_x^c) \times \text{constants}$$
$$= f(\alpha_1,\ldots,\alpha_{r+s},x)^{d_x}.\exp(-f(\alpha_1,\ldots,\alpha_{r+s},x)E_x^c) \times \text{constants}$$

So the total likelihood, ignoring constants, is:

$$\prod_{\substack{\text{all ages}\\ x}} f(\alpha_1,\ldots,\alpha_{r+s},x)^{d_x}.\exp(-f(\alpha_1,\ldots,\alpha_{r+s},x)E_x^c)$$

This likelihood was maximised numerically to obtain maximum likelihood estimates of the parameters $\alpha_1,\alpha_2,\ldots,\alpha_{r+s}$, and hence $\mathring{\mu}_x$.

Other ways to fit a parametric formula include:

- minimising the χ^2-statistic $\sum\frac{(O-E)^2}{E}$ where, for each age group, O denotes the observed number of deaths and E denotes the expected number of deaths (calculated using the fitted model); and
- minimising the value of the weighted least squares, *ie* $\sum w_x(\hat{\mu}_x-\mathring{\mu}_x)^2$, the sum of the squares of the differences between the crude and fitted values of μ_x with a weighting w_x based on the exposed to risk at each age.

1.4 Other considerations

Using additional information from other investigations

For practical use, it is not sufficient to choose and fit a formula using statistical methods alone. It is always necessary to inspect the results in the light of previous knowledge of mortality experiences, especially at very young and very old ages where the data may be scarce.

Question

Describe the experiences that might be available to help us check our results.

Solution

We might be able to check our results against:

- previous investigations of the same population
- recent investigations of a different population with similar characteristics
- changes to mortality observed in recent investigations in other countries.

The graduated estimates should also be compared with other experiences to see if they behave as we would expect. Examples of the checks that would be applied are:

- **The mortality of males is higher than the mortality of females.**
- **The mortality of persons with life insurance policies is lower than that of the population as a whole.**
- **The mortality of persons who have recently taken out life insurance is lower than that of persons who took out life insurance a long time ago (because they have to be in good health to obtain life insurance).**

It might be necessary to adjust the graduation to obtain a satisfactory final result.

Our observations may suggest a different pattern of mortality to that experienced in previous investigations. In these circumstances, we must decide on the relative levels of reliance that we can place on each source of information. If our study is small, we may be more confident that the previous investigations reflect the true position. If our study is large, we may be more confident that the pattern of underlying mortality is genuinely different from that of previous investigations.

Question

Explain why the mortality of people with life insurance policies might be lower than that of the population as a whole.

Solution

The mortality of people with life insurance policies might be lower than that of the population as a whole for several reasons. These include the facts that individuals in poor health may be refused life insurance and that policyholders may generally belong to a higher socio-economic group with a lower rate of mortality.

Financial risks

We should always consider where the financial risks lie:

- **A life insurance contract pays out on death, so the insurance company will charge inadequate premiums if it *underestimates* mortality.**
- **A pension or annuity contract pays out on continued survival, so the insurance company will charge inadequate premiums if it *overestimates* mortality.**

So, if an insurance company wishes to protect itself against the risk of charging inadequate premiums, it will try to overestimate mortality for life insurance contracts and underestimate mortality for pension or annuity contracts. At the same time, the insurance company will need to ensure that its margins are not so large as to make the premiums uncompetitive.

Changes in mortality

Since insurance companies will use graduated mortality tables to estimate *future* mortality (under insurance contracts yet to be sold) but investigations must be of *past* mortality, the trend of mortality is important. In most countries mortality has been falling for a long time, which means that past mortality is likely to be on the safe side for insurance business but not adequate for pension or annuity business. In respect of the latter it is necessary to make some projection of future improvements in mortality.

Mortality projection is discussed in Chapter 12.

1.5 The graduation process

The curve-fitting process described above is only one of several stages that must be carried out, often repeatedly, before a satisfactory result is obtained.

Step 1 – select a graduation formula

A particular parametric family of curves must be chosen. For example, the first few (useful) families of the general type used by the CMI are:

$\alpha_1 \exp(\alpha_2 x)$ **(Gompertz)**

$\alpha_1 + \alpha_2 \exp(\alpha_3 x)$ **(Makeham)**

$\alpha_1 + \alpha_2 \exp(\alpha_3 x + \alpha_4 x^2)$

and so on.

Step 2 – determine parameter values

Given a family of curves, the best-fitting values of the parameters must be found. The CMI used maximum likelihood, but there are many other suitable procedures.

Other procedures include weighted least squares estimation. Parameter estimation is usually performed on a computer using a statistics package.

Step 3 – calculate graduated rates

Calculate the graduated rates at each age using the fitted parametric formula. This can be done on a computer using a spreadsheet.

Step 4 – test

Given the best-fitting curve of a given family, the graduated rates must be compared with the original data to see if they are acceptably close, according to some test procedures (see Chapter 10).

Usually this process will be carried out for several families of curves, and the final choice will be influenced by the 'goodness of fit'. However, many other factors influence the outcome, and it is not always the best-fitting graduation (in the statistical sense) that gives the most suitable result for practical use.

2 Graduation by reference to a standard table

2.1 Overview

The underlying rationale of the method is as follows: if the class of lives involved in the graduation is sufficiently similar to the class of lives on which the standard table is based, then the true underlying mortality rates of the lives involved in the graduation should be quite similar to those of the standard table. Even if overall levels of mortality differ between the two, it would still be expected that the overall progression of rates from age to age would be similar.

A 'standard table' means a published life table based upon sufficient data to be regarded as reliable (for appropriate applications). Examples include national life tables based on a country's entire population (*eg* the English Life Tables) and insured lives tables based on large numbers of insured lives (*eg* the '00 series' tables).

A standard table will always be based on a well-defined class of lives, although this does not mean that that class of lives will be perfectly homogeneous. If we are given the mortality experience of a similar group of lives, we might reasonably suppose that it should share some of the characteristics of the experience underlying the standard table, such as its overall shape. This is useful if we do not have much data from the experience in which we are interested.

So, we'll tend to use this method of graduation when we do not have a large amount of data *and* there is a standard table that we think is appropriate. We make use of the valuable information provided by the standard table relating to the general shape of mortality. An appropriate simple equation, involving unknown parameters, is selected to reflect the relationship between the mortality rates for the new experience and the rates from the standard table. For example, if we think that the true underlying forces of mortality are a linear function of those from the standard table, we might try an equation of the form $\mathring{\mu}_x = a + b\mu_x^s$. The graduated rates are then a combination of the shape provided by the standard table and the level of mortality observed in our investigation.

Only if the standard table is sufficiently similar to the underlying experience will a satisfactory fit to the crude estimates be possible.

2.2 The graduation process

Step 1 – select standard table

We select an existing standard mortality table that is believed to have a similar pattern of mortality rates over the age range of interest. The appropriateness of a particular standard table will be assessed by comparing the characteristics of the lives on which it was based and those in the current investigation, *eg* sex, geographical area, period of investigation.

Step 2 – find simple link to the standard table

Let μ_x^s be the forces of mortality of the standard table. Then we try to exploit the assumed similarity of the experiences by seeking a reasonably simple function $f()$ such that

$$\mathring{\mu}_x = f(\mu_x^s).$$

Examples include:

$$\mathring{\mu}_x = a + b\mu_x^s$$
$$\mathring{\mu}_x = (a + bx)\mu_x^s$$
$$\mathring{\mu}_x = \mu_x^s + k$$
$$\mathring{\mu}_x = \mu_{x+k}^s$$

where a, b and k are suitable constants.

The search for a suitable function $f()$ can be aided by making some simple plots, for example a plot of $\hat{\mu}_x$ against μ_x^s might indicate a linear relationship. If it is not possible to find a simple relationship, then the supposition that the experiences have similar characteristics should perhaps be reconsidered. It should be remembered that if data are scarce, too close a fit to any suggested relationship is not to be expected, especially at extreme ages.

Step 3 – determine parameter values

Once a possible relationship has been identified, the best-fitting parameters must be found. Any suitable method might be used, for example:

(a) maximum likelihood: the underlying model is that:

$$\mu_x = f(\alpha_1,\ldots,\alpha_n,\mu_x^s)$$

where the $\alpha_1,\ldots,\alpha_n$ are unknown parameters. The MLEs are then found by maximising the likelihoods as in Section 1.3.

(b) least squares: the parameter values are found that minimise:

$$\sum_{\substack{\text{all ages}\\ x}} w_x(\hat{\mu}_x - \mathring{\mu}_x)^2$$

where the $\{w_x\}$ are suitable weights. Natural weights would be the exposures to risk (E_x^c) at each age, or the inverse of the estimated variance of $\tilde{\mu}_x$.

Question

Explain why weights based on the *inverse* of the estimated variance are a sensible choice.

Solution

Weights that are based on the inverse (*ie* the reciprocal) of the variance give more weighting to the ages where the estimated variance is low (*ie* where we are more confident about the true rate) and less weighting to the ages where the estimated variance is high (*ie* where we are less confident about the true rate).

For example, suppose that a set of crude forces of mortality ($\hat{\mu}_x$) is to be graduated by reference to a standard table using the relationship $\overset{\circ}{\mu}_x = a + b\mu_x^s$. The parameters are to be determined using the method of weighted least squares. So we need to minimise:

$$S = \sum_x w_x (\hat{\mu}_x - \overset{\circ}{\mu}_x)^2 = \sum_x w_x [\hat{\mu}_x - (a + b\mu_x^s)]^2$$

For each age x, the weight w_x should be proportional to the reciprocal of $\text{var}(\tilde{\mu}_x)$. So here we could use:

$$w_x = \frac{E_x^c}{\hat{\mu}_x}$$

By differentiating S with respect to a and b, we can determine the values of a and b that minimise S.

Step 4 – calculate graduated rates

The assumed relationship with the standard table is then used to calculate the graduated rates using the formula $\overset{\circ}{\mu}_x = f(\mu_x^s)$.

Step 5 – test

The resulting graduation would be subjected to tests of goodness-of-fit to the data (see Chapter 10) and, if there is more than one candidate function $f()$ (or more than one suitable standard table), **the goodness-of-fit may be used to assist in the final choice.**

The remarks in Section 1.4 apply also to experiences graduated by this method.

3 Graduation using spline functions

3.1 Overview

An alternative approach to graduation is to use spline functions. These are polynomials of a specified degree which are defined on a piecewise basis across the age range. The pieces join together at *knots*, where certain continuity conditions for the functions themselves and their derivatives are required.

So we use a different function over each section of the age range, whilst ensuring that, over the whole age range, the function is continuous. We also require the function to have continuous first and second derivatives at the knots.

The method may be illustrated using cubic splines, or polynomials of degree 3, which are commonly used. They were used, for example, to graduate English Life Table 14 (1980-82). Similar, but more complex methods, have been used to graduate more recent English Life Tables. For example, English Life Table 17 (2010-2012) used a linear combination of basis splines and a penalisation term to control the smoothness of the resulting fit (see Chapter 17 for more discussion of this).

Suppose we wish to fit a spline through a set of mortality rates μ_x for ages x with knots at ages $x_1, x_2, ..., x_n$, where $x_1 < x_2 < \cdots < x_n$. **The smoothest interpolating spline is the *natural cubic spline*. This is linear at ages below x_1 and above x_n. This can be written as:**

$$\mu_x = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n} \beta_j \phi_j(x)$$

where:

$$\phi_j(x) = \begin{cases} 0 & x < x_j \\ (x - x_j)^3 & x \geq x_j \end{cases}$$

With this cubic function, we have for $j = 1, 2, ..., n$:

$$\phi_j(x_j) = 0$$

$$\phi_j'(x) = \begin{cases} 0 & \text{if } x < x_j \\ 3(x - x_j)^2 & \text{if } x \geq x_j \end{cases} \quad \Rightarrow \quad \phi_j'(x_j) = 0$$

$$\phi_j''(x) = \begin{cases} 0 & \text{if } x < x_j \\ 6(x - x_j) & \text{if } x \geq x_j \end{cases} \quad \Rightarrow \quad \phi_j''(x_j) = 0$$

which ensures that the function is smooth at the knots.

Given the definition of $\phi_j(x)$, we see that for $x < x_1$:

$$\phi_1(x) = \phi_2(x) = \cdots = \phi_n(x) = 0$$

So, for $x < x_1$:

$$\mu_x = \alpha_0 + \alpha_1 x$$

This is indeed a linear function of x.

For $x_1 \le x < x_2$:

$$\phi_1(x) = (x - x_1)^3$$

and: $\phi_2(x) = \cdots = \phi_n(x) = 0$

So, for $x_1 \le x < x_2$:

$$\mu_x = \alpha_0 + \alpha_1 x + \beta_1 (x - x_1)^3$$

For $x_2 \le x < x_3$:

$$\phi_1(x) = (x - x_1)^3$$

$$\phi_2(x) = (x - x_2)^3$$

and: $\phi_3(x) = \cdots = \phi_n(x) = 0$

So, for $x_2 \le x < x_3$:

$$\mu_x = \alpha_0 + \alpha_1 x + \beta_1 (x - x_1)^3 + \beta_2 (x - x_2)^3$$

More generally, for $x_k \le x < x_{k+1}$, where $k = 1, 2, \ldots, n-1$, we have:

$$\mu_x = \alpha_0 + \alpha_1 x + \beta_1 (x - x_1)^3 + \cdots + \beta_k (x - x_k)^3$$

and for $x > x_n$:

$$\begin{aligned}\mu_x &= \alpha_0 + \alpha_1 x + \beta_1 (x - x_1)^3 + \cdots + \beta_n (x - x_n)^3 \\ &= \alpha_0 + \alpha_1 x + \beta_1 (x^3 - 3x^2 x_1 + 3x x_1^2 - x_1^3) + \cdots + \beta_n (x^3 - 3x^2 x_n + 3x x_n^2 - x_n^3)\end{aligned}$$

For this to be a linear function of x, the coefficients of x^2 and x^3 must be 0. The coefficient of x^2 is:

$$-3\beta_1 x_1 - 3\beta_2 x_2 - \cdots - 3\beta_n x_n$$

and the coefficient of x^3 is:

$$\beta_1 + \beta_2 + \cdots + \beta_n$$

So, for linearity, we must have:

$$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = 0 \tag{1}$$

and: $$\beta_1 + \beta_2 + \cdots + \beta_n = 0 \tag{2}$$

The definition of $\phi_j(x)$ leads to the following form for the natural cubic spline over the whole age range:

$$\boldsymbol{\mu_x = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \Phi_j(x)}$$

where:

$$\boldsymbol{\Phi_j(x) = \phi_j(x) - \left[\frac{x_n - x_j}{x_n - x_{n-1}}\right]\phi_{n-1}(x) + \left[\frac{x_{n-1} - x_j}{x_n - x_{n-1}}\right]\phi_n(x)}$$

To see that this is equivalent to the formula for μ_x given earlier, which states that $\mu_x = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n} \beta_j \phi_j(x)$, we will make use of equations (1) and (2).

Multiplying (2) by x_n gives:

$$\beta_1 x_n + \beta_2 x_n + \cdots + \beta_n x_n = 0 \tag{3}$$

Then subtracting (1) from (3):

$$\beta_1(x_n - x_1) + \beta_2(x_n - x_2) + \cdots + \beta_{n-1}(x_n - x_{n-1}) = 0$$

$$\Rightarrow \beta_1(x_n - x_1) + \beta_2(x_n - x_2) + \cdots + \beta_{n-2}(x_n - x_{n-2}) = -\beta_{n-1}(x_n - x_{n-1})$$

$$\Rightarrow \beta_{n-1} = -\left[\beta_1\left(\frac{x_n - x_1}{x_n - x_{n-1}}\right) + \cdots + \beta_{n-2}\left(\frac{x_n - x_{n-2}}{x_n - x_{n-1}}\right)\right] = -\sum_{j=1}^{n-2} \beta_j \left(\frac{x_n - x_j}{x_n - x_{n-1}}\right) \tag{4}$$

In addition, rearranging (2) gives:

$$\beta_n = -(\beta_1 + \beta_2 + \cdots + \beta_{n-1})$$

and using (4):

$$\beta_n = -(\beta_1 + \beta_2 + \cdots + \beta_{n-2}) + \beta_1\left(\frac{x_n - x_1}{x_n - x_{n-1}}\right) + \cdots + \beta_{n-2}\left(\frac{x_n - x_{n-2}}{x_n - x_{n-1}}\right)$$

$$= \beta_1\left(\frac{x_n - x_1}{x_n - x_{n-1}} - 1\right) + \cdots + \beta_{n-2}\left(\frac{x_n - x_{n-2}}{x_n - x_{n-1}} - 1\right)$$

$$= \beta_1\left(\frac{x_{n-1} - x_1}{x_n - x_{n-1}}\right) + \cdots + \beta_{n-2}\left(\frac{x_{n-1} - x_{n-2}}{x_n - x_{n-1}}\right) = \sum_{j=1}^{n-2} \beta_j \left(\frac{x_{n-1} - x_j}{x_n - x_{n-1}}\right)$$

Alternatively, this expression for β_n can be obtained by multiplying (2) by x_{n-1} and then subtracting (1).

So:

$$\begin{aligned}
\mu_x &= \alpha_0 + \alpha_1 x + \sum_{j=1}^{n} \beta_j \phi_j(x) \\
&= \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \phi_j(x) + \beta_{n-1}\phi_{n-1}(x) + \beta_n \phi_n(x) \\
&= \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \phi_j(x) - \sum_{j=1}^{n-2} \beta_j \left(\frac{x_n - x_j}{x_n - x_{n-1}} \right) \phi_{n-1}(x) + \sum_{j=1}^{n-2} \beta_j \left(\frac{x_{n-1} - x_j}{x_n - x_{n-1}} \right) \phi_n(x) \\
&= \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \left[\phi_j(x) - \left(\frac{x_n - x_j}{x_n - x_{n-1}} \right) \phi_{n-1}(x) + \left(\frac{x_{n-1} - x_j}{x_n - x_{n-1}} \right) \phi_n(x) \right] \\
&= \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \Phi_j(x)
\end{aligned}$$

3.2 The graduation process

The stages in spline graduation are, therefore:

Step 1 – make decisions about knots

Choose the number and value of the knots.

For each knot, the position of x (*ie* the age) is specified, but the value of μ_x is not. It is not necessary for the knots to be equally spaced.

Step 2 – preliminary calculations

Calculate the $\Phi_j(x)$.

Step 3 – estimate the parameter values

Fit the equation $\mathring{\mu}_x = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \Phi_j(x)$ using weighted least squares, where the weights are proportional to the inverse of the estimated variance of $\tilde{\mu}_x$.

That is, we determine the values of α_0, α_1, β_1, β_2, …, β_{n-2} that minimise the expression:

$$S = \sum_x w_x (\hat{\mu}_x - \mathring{\mu}_x)^2 = \sum_x w_x \left[\hat{\mu}_x - \left(\alpha_0 + \alpha_1 x + \beta_1 \Phi_1(x) + \beta_2 \Phi_2(x) + \cdots + \beta_{n-2} \Phi_{n-2}(x) \right) \right]^2$$

The value of w_x should be proportional to $\dfrac{E_x^c}{d_x}$.

Step 4 – calculate the graduated rates

Calculate the graduated rates using the estimated values of α_0, α_1, β_1, β_2, ..., β_{n-2} from Step 3.

Step 5 – test

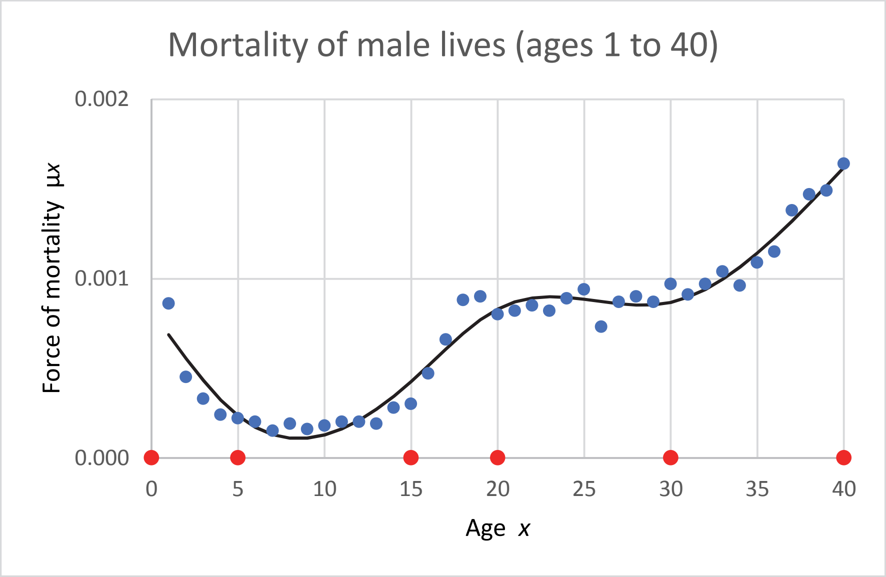
The greater the number of knots, the more closely will the graduated rates adhere to the crude rates, but the less smooth the graduation will be.

The resulting graduation would be subjected to tests of goodness-of-fit to the data (see Chapter 10) which may assist in finding the optimal number of knots.

The remarks in Section 1.4 apply also to experiences graduated by this method.

3.3 Examples of graduations using spline functions

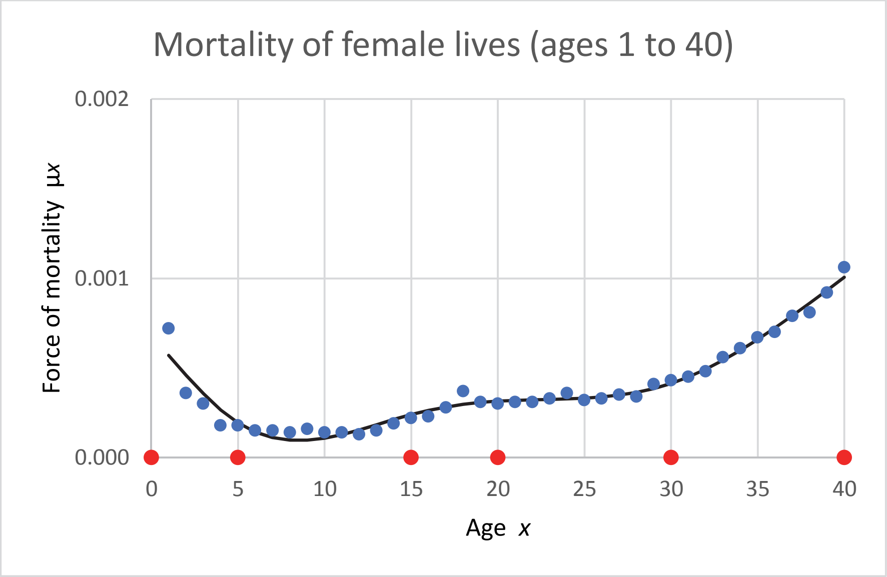
The graphs below show two graduations (one for males and one for females) for a large population of lives similar to those used in English Life Tables No 15 (ELT15). The circles show the crude estimates of the force of mortality at each age (calculated by dividing the number of deaths aged x nearest birthday by the corresponding central exposed to risk). The solid line shows the graduated values, which have been calculated by fitting a cubic spline function with 6 knots, positioned at ages 0, 5, 15, 20, 30 and 40 (marked by circles on the x-axis).



The mortality for the male lives over the age range shown includes several contrasting features:

- relatively high infant mortality at the start (see age 1), which then decreases
- a flat period of very low mortality between ages 5 and 15
- an 'accident hump' in the late teenage years (see ages 18 and 19)
- another flat period between ages 20 and 30
- increasing mortality rates from around age 30.

As can be seen, the spline function is able to follow these twists and turns, producing a set of graduated rates that progress smoothly but also adhere closely to the crude rates.



The mortality of the female lives has the same features, although these are less pronounced. Again, the spline function produces a set of graduated rates that progress smoothly but also adhere closely to the crude rates.

Question

(i) Write down a formula for calculating the values of μ_x in the spline graduations illustrated in the graphs above in terms of the fitted parameters, the age x and the functions $\Phi_j(x)$.

(ii) Let $\phi_j(x) = \max\left\{(x - x_j)^3, 0\right\}$. Write down the formula used to calculate $\Phi_1(x)$ in terms of the functions $\phi_j(x)$.

Solution

(i) These graduations each use $n = 6$ knots. So the equation for μ_x will have the form:

$$\mu_x = \alpha_0 + \alpha_1 x + \beta_1 \Phi_1(x) + \beta_2 \Phi_2(x) + \beta_3 \Phi_3(x) + \beta_4 \Phi_4(x)$$

(ii) The formula for calculating $\Phi_1(x)$ is:

$$\Phi_1(x) = \phi_1(x) - \left(\frac{x_6 - x_1}{x_6 - x_5}\right)\phi_5(x) + \left(\frac{x_5 - x_1}{x_6 - x_5}\right)\phi_6(x)$$

Here, the first knot is at $x_1 = 0$ and the last two are at $x_5 = 30$ and $x_6 = 40$, so this is:

$$\Phi_1(x) = \phi_1(x) - 4\phi_5(x) + 3\phi_6(x)$$

4 Comparison of different methods

First, we note that the three methods of graduation described above by no means cover all possible methods. We take parametric formula graduation to be an example of approaches used with reasonably large data sets, and the other two to be examples of methods used with smaller data sets.

4.1 Graduation by parametric formula

The mathematical formula method produces a smooth set of rates and can be a fully automated process. The ability to optimise statistically eliminates any subjectivity from the fitting process. The only subjectivity remaining is in the ultimate choice of formula to use. These properties, along with the independence of such a graduation from any other experience, make this method ideally suited for use in constructing standard tables.

The main problem with the method is that a phenomenon such as mortality, which is subject to a great number of different influences to different extents over an individual's lifetime, may be impossible to represent adequately by a single mathematical formula. As a result no single function may produce a satisfactory fit, at least over the whole age range. Heterogeneous data (*ie* data covering a mixture of people with different patterns of mortality – males and females, for example) can also make it more difficult to find a function that produces an adequate fit.

Some specific points about parametric formula graduation are:

- **It is a natural extension of the simple probabilistic models for single years of age, parameterised by μ_x. It is straightforward to extend the statistical theory of estimation from one parameter to several, including estimation of standard errors and so on, and very often computer software is available to carry out the necessary optimisations.**
- **The graduation will inherit its smoothness from the smoothness of the function in the parametric formula. In general, formulae with a small number of parameters will produce an acceptably smooth graduation.**
- **Sometimes, when comparing several experiences, it is useful to fit the same parametric formula to all of them. Differences between the fitted parameters, given their standard errors, then give insight into the differences between the experiences.**

  For example, the difference between parameters may help us to identify trends in mortality over time.
- **The approach is very well-suited to the production of standard tables from large amounts of data.**

  It is not possible to use the method successfully where data are scanty over a wide age range.
- **It can, however, be very difficult to find a suitable curve that fits an experience well at all ages. Partly this is because of the different features that predominate at different ages (*eg* infant mortality, the accident hump and exponential mortality after middle age). Partly it may be because cross-sectional studies mix up different generations at different ages. A very likely reason is that there is still a good deal of heterogeneity in all mortality studies, even if we classify the data by age, sex, policy type, calendar period and so on.**

For example, if we tried to remove heterogeneity by limiting our investigation to 40-year old female policyholders, we might still find large differences in the underlying rates because of differences in diet, smoking habits, levels of exercise *etc*.

- **Care is required when extrapolating. Most methods of curve fitting will result in a good fit where there is most data, which in graduation usually means at middle ages. The form of the curve at the extreme ages is therefore sometimes determined by the best-fitting parameters at other ages, which means that the curve is, to a large degree, extrapolated from the middle ages. The results at extreme ages can, therefore, be quite poor, and might require adjustment. The same warning applies if the graduation is extrapolated beyond the ages for which there are data.**

  Given this problem, we may decide to abandon the formula at extreme ages and use an adjusted standard table instead.

- The optimisation procedures can make quantitative allowance for our relative confidence in each observed rate by reflecting the amount of data available at each age via the weighting factors.

4.2 Graduation by reference to a standard table

This method is a very simple way of obtaining a workable set of graduated rates in many practical situations. The process will often follow naturally from an experience investigation involving a comparison against a standard table, as described earlier. Because of the fact that a great deal of the form of the function $f(x)$ is provided by the standard table, the graduation formula (and hence the process itself) is greatly simplified compared with other methods. As only a few parameters may need to be fitted from the data, the amount of information that the data need to provide is also less than other methods.

This dependence on the form of the standard table often leads to a very significant difficulty. The features of the actual experience (and hence the probable progression of the true underlying rates with age) may differ significantly from the features displayed by any standard table. Where this is the case, it would never be possible to obtain satisfactory adherence to the data using this method, at least without further adjustment to the graduated rates. It also makes the method unsuitable for the purpose of *producing* standard tables, which need to be fully representative of the data. In this case it would be inappropriate for the graduation to be influenced in this way by another experience.

Some points about graduation by reference to a standard table are:

- **It can be used to fit relatively small data sets where a suitable standard table exists.**

  Results based on small data sets will be correspondingly less reliable.

- **Provided a simple function is chosen** (*eg* a polynomial or exponential function of low order)**, and the standard table is smooth to begin with, a smooth graduation should result.**

- **The collateral information obtained from the standard table can be particularly useful in deciding the shape of the graduation at the extreme ages, where there might be little or no data.**

- **The method is not suitable for the preparation of standard tables based on large amounts of data.**
- **The choice of standard table is important; choosing an inappropriate table could impart the wrong shape to the entire graduation.**

  Features exhibited by the standard table will also be exhibited by the graduated rates. These may not be desirable (or representative of the data) for the graduation being performed.
- **It is not always easy to choose an appropriate standard table.**
- The simple form of the function means that the fitting process (*ie* estimation of the parameters) is usually easy to carry out.

4.3 Graduation using spline functions

Some points about graduation using a spline function are:

- **Provided the number of knots is small, the graduation will usually be smooth.**
- **Alternative graduations can be tried by varying the number and position of the knots.**
- **The method is suitable for quite small experiences as well as very large experiences (such as national populations). It can also be used to produce standard tables. It is, however, not suitable for very small experiences with scanty data at many ages.**
- **It is not easy to choose the knots, and experiments with different numbers and locations are likely to be needed. However, past attempts to graduate similar data using splines can guide the choice. Generally, sections of the age range over which mortality rates are changing rapidly require more frequent knots than do sections of the age range where mortality progresses regularly. Thus, for human mortality, more frequent knots are usually needed at younger ages than, say at ages between 30 and 90 years.**
- **Care is required at extreme ages where data can be scanty. With a natural cubic spline, the form of the curve above the highest knot is linear. Therefore the fit at the oldest ages can be quite poor and need adjustment.**
- Splines are particularly useful when the pattern of mortality rates shows significant changes in shape. For example, the spline method would be appropriate if we were graduating human mortality rates up to the age of about 30. At the youngest ages, we expect to see high infant mortality. Mortality is then low through childhood and increases in the late teens, where we expect to see the start of the accident hump. Mortality rates decrease for a few years towards the end of the accident hump, before starting to increase exponentially after the late 20s.

The spline method is a special case of graduation using a parametric formula.

Question

An actuary has conducted investigations into the mortality of the following classes of lives:

(a) the male members of a medium-sized pension scheme

(b) the female population of a large developed country.

The actuary wishes to graduate the crude rates. State, with reasons, an appropriate method of graduation for both of these classes of lives.

Solution

(a) ***Medium-sized pension scheme***

The population is unlikely to be large enough for us to be able to fit a parametric formula directly. However, the experience is likely to be similar to one of the published tables, which can be adjusted to match this experience.

So graduation by reference to a standard table would be appropriate here.

Alternatively, the spline method could be used.

(b) ***Large developed country***

The population is likely to be large enough for us to be able to fit a parametric formula directly. Flexible formulae, such as the Gompertz-Makeham family, are available. We may want to use the data to produce a standard table, and this should be independent of other standard tables. In addition, we will probably be interested in comparing this experience with other experiences (*eg* the same population 10 years ago), so we don't want to base the rates on an existing table.

So graduation by parametric formula or spline function would be appropriate here.

5 Statistical tests of a graduation

We discussed statistical tests of graduations in Chapter 10. The precise form of some tests depends on the method of graduation. In this section we will return to the subject of statistical tests and cover the outstanding issues, now that we have covered these graduation methods.

In this chapter we will be looking at just two of the tests – the chi-squared (χ^2) test and the cumulative deviations test – as these are the only ones that we need to modify according to the method of graduation employed.

5.1 Comparing one experience with another

In Chapter 10, we introduced statistical tests of the hypothesis that one experience was the same as another. Often, the question is whether or not an experience for which we have data and crude estimates is consistent with a given standard table.

The tests depended on comparison of the actual deaths observed at each age x in one experience, d_x , with the number expected on the basis of the other experience.

For example, if the second experience was represented by a standard table $\{\mu_x^s\}$, we devised tests based on the deviations $D_x - E_x^c \mathring{\mu}_x$.

5.2 Testing a graduation

The same tests can be used to test the hypothesis that the graduation adheres to the data, by substituting the graduated estimates for the standard table quantities above, and using the deviations $D_x - E_x^c \mathring{\mu}_x$.

In effect, we are asking whether or not the observed numbers of deaths are consistent with the numbers expected if the graduated estimates are 'correct'.

There are two problems. The first is with the χ^2 -test. The second relates to the cumulative deviations test.

Chi-squared test

Given m age groups $(x_1,\ldots,x_m)$, the χ^2 -statistic:

$$\sum_{x=x_1}^{x_m} \frac{(D_x - E_x^c \mu_x^s)^2}{E_x^c \mu_x^s}$$

had a χ^2 distribution with m degrees of freedom. (The superscript '*s*' denotes the standard table as usual.) This led to a simple statistical test.

A crucial point in the above reasoning is that the two experiences in question should not be the same. In other words, the data upon which the *observed* deaths are based should *not* be the same as the data upon which the *expected* deaths are based.

This clearly does not hold when we are testing the adherence of a graduation to the observed data – we compute the expected deaths using the graduated quantities $\{\overset{\circ}{\mu}_x\}$, which are themselves based on the observed deaths.

It is still legitimate to use the χ^2-test in these circumstances. The χ^2-statistic is unchanged (except that mortality according to the standard table has been replaced by mortality according to our graduation):

$$\sum_{x=x_1}^{x_m} \frac{(D_x - E_x^c \overset{\circ}{\mu}_x)^2}{E_x^c \overset{\circ}{\mu}_x}$$

but we must reduce the number of degrees of freedom.

- **If we used parametric formula graduation, we lose one degree of freedom for each parameter fitted.**
- **If we used standard table graduation, we lose one degree of freedom for each parameter fitted, and we lose some further (indeterminate) number of degrees of freedom because of the constraints imposed by the choice of standard table. Rather than suggest how many, it is more important to be aware of the problem when it comes to interpreting the result of the test.**
- **If we used spline function graduation, one degree of freedom should be subtracted for each fitted constant (*ie* parameter) estimated from the data. With a natural cubic spline, this will involve subtracting one degree of freedom for each knot used.**

  As we have already seen, if we use a natural cubic spline with a total of n knots, then we have n parameters to estimate (α_0, α_1, β_1, β_2, ..., β_{n-2}). So we lose n degrees of freedom.

  If the knots are chosen after inspecting the crude rates, additional degrees of freedom might be subtracted.

Question

For a given set of data, crude estimates $\{\hat{\mu}_x\}$ for ages $x = 30, 31, ..., 79$ have been calculated. These rates have been graduated assuming that the underlying force of mortality follows Makeham's law $\mu_x = A + Bc^x$. The graduated rates are now being tested for adherence to data.

Give a formula for the test statistic and explain how the test is carried out.

Solution

The formula for the test statistic is $\displaystyle\sum_{x=30}^{79} \frac{(D_x - E_x^c \overset{\circ}{\mu}_x)^2}{E_x^c \overset{\circ}{\mu}_x}$.

We have 50 age groups (ages $x = 30, 31, ..., 79$). In fitting the model, we will have estimated three parameters (*A*, *B* and *c*).

We compare the value of the test statistic against the χ^2 distribution with 47 (*ie* 50 – 3) degrees of freedom. We reject the graduation if the observed value of the test statistic exceeds the upper 5% point of χ^2_{47} .

Cumulative deviations test

The second problem when testing a set of graduated rates is that the cumulative deviations test cannot be used if the cumulative deviation is automatically zero because of the graduation procedure.

Some methods of graduation may force the cumulative deviation to be close to zero over the age range being graduated. The test is invalidated for such graduations due to the imposition of this extra constraint when the curve has been fitted to the data.

6 The effect of duplicate policies

The investigations of life office mortality carried out in the UK by the CMI have one particular feature that affects the statistical properties of the resulting estimates: they are based on policies and not lives.

That is, instead of observing persons and recording:

E_x^c = Number of person-years observed

d_x = Number of deaths

the CMI observe policies, and record:

E_x^c = Number of policy-years observed

d_x = Number of policies becoming claims by death.

So, if a female policyholder born on 1 January 1975 were to own 3 separate life assurance policies, she would contribute a maximum of 3 years to the value of E_{45}^c in the year 2020. If she were to die in 2021, she would contribute 3 to the value of d_{46} in that year.

The reasons for observing policies rather than lives are that life office record-keeping is based on policies, and that it can be very difficult to establish when two policies are, in fact, owned by the same person, especially if many life offices pool their data (as in the CMI investigations).

It can be particularly difficult to establish if two policies are owned by the same person if the policies were bought from different insurance companies.

The outcome is that we can no longer be sure that we are observing a collection of independent claims; it is quite possible that two distinct death claims are the result of the death of the same life. This is called the problem of *duplicate policies*.

The effect of duplicate policies is to increase the variance of the number of claims. This may be seen intuitively by noting that if a person has several policies, then the death of that person will cause a greater increase in the number of claims than would the death of a person having only one policy. The ratio by which the variance is increased depends on the extent to which people own duplicate policies: it may also vary with age x.

If the ratios by which the variance was increased were known, we could make allowance for the increased variances in tests of a graduation and so on. Usually they are not known for any particular investigation, but the CMI has carried out special investigations from time to time to match up duplicate policies in force and hence derive estimates of the ratios suitable for use.

Chapter 11 Summary

Methods of graduation

Three of the most common methods of graduation are:

- graduation by parametric formula – we assume that mortality can be modelled using a mathematical formula
- graduation by reference to a standard table – we assume that there is a simple relationship between the observed mortality and an appropriate standard table
- graduation using spline functions – we divide the age range into sections and assume that the mortality rates in each section can be modelled using a polynomial of a certain degree.

The strengths and weaknesses of these methods can be assessed in terms of the following criteria:

- smoothness
- adherence to crude rates (goodness of fit)
- ease of use
- amount of data required
- flexibility in allowing for special features
- validity of the method given the problems.

We must take care when using some of the statistical tests to assess the adherence of a graduation to the observed crude estimates. Since the actual number of deaths (those observed) and the expected number (based on the graduated rates) are based on the same set of data, we must reduce the number of degrees of freedom for the χ^2 -test.

Duplicate policies

Duplicate policies (*ie* lives with more than one policy) can distort the results of an investigation. Allowance can be made for the increase in the variance of the number of claims observed due to the existence of duplicate policies.

The practice questions start on the next page so that you can keep all the chapter summaries together for revision purposes.

Chapter 11 Practice Questions

11.1 State whether each of the following statements is true or false.

I When graduating by reference to a standard table you always need to test for smoothness.

II If the mortality of a whole population was recorded, there would be no need to graduate since random sampling errors would not have occurred.

III Graduating by reference to a standard table can produce good results even with scanty data.

11.2 Describe the circumstances under which it would be appropriate to graduate the rates in a mortality investigation using a parametric formula.

11.3 An investigation has been carried out into the mortality rates of males under the age of 30 in a deprived area of the UK. The rates are to be graduated using the spline method. Explain why this method is appropriate in this situation.

11.4 Exam style

(i) (a) Describe the general form of the polynomial formula used to graduate recent standard tables produced for use by UK life insurance companies.

(b) Show how the Gompertz and Makeham formulae arise as special cases of this formula. [3]

(ii) An investigation was undertaken of the mortality of persons aged between 40 and 75 years who are known to be suffering from a degenerative disease. It is suggested that the crude estimates be graduated using the formula:

$$\overset{\circ}{\mu}_x = \exp\left[b_0 + b_1 x + b_2 x^2\right]$$

(a) Explain why this might be a sensible formula to choose for this class of lives.

(b) Suggest two techniques that can be used to estimate the parameter values. [3]

[Total 6]

11.5 Exam style

An insurance company is investigating the mortality of its annuity policyholders. The crude mortality rates are to be graduated for use in future premium calculations.

(i) (a) Suggest, giving reasons, a suitable method of graduation in this case.

(b) Describe how you would graduate the crude rates. [5]

(ii) Comment on any further considerations that the company should take into account before using the graduated rates in its premium calculations. [2]

[Total 7]

The solutions start on the next page so that you can
separate the questions and solutions.

Chapter 11 Solutions

11.1 Most standard tables are already smooth. Hence a simple transformation of it will leave smoothness undisturbed, so I is false.

Even if the investigation covers the whole population, the numbers of deaths observed will still be random and we will still need to graduate the crude rates, so II is false. (If we use the whole population the random variations will probably be quite small, but we would need an infinite population to remove them completely!)

III is true.

11.2 It would be appropriate to graduate the results of a mortality investigation using a parametric formula if:

- a suitable mathematical formula can be found that can describe mortality rates adequately over the entire age range of interest
- the expected number of deaths is great enough at all ages to give reliable answers
- the data values can be considered to be complete and accurate, and they are adequately subdivided with respect to age, sex and other relevant categories
- an analytic method or computer software that can determine the optimal parameter values is available.

An example of such a situation would be an investigation of the mortality of a large group of life office policyholders or a national mortality investigation.

11.3 The pattern of mortality rates shows significant changes in shape over the age range 0 to 25 – mortality rates are high at the earliest ages, decrease rapidly in the first few months of age, are low through the childhood years, increase in the late teens (a feature known as the accident hump), and fall away again through the early to mid-20s. By selecting appropriate knots, we can produce a spline function that captures these features.

The spline method works well with fairly small data sets as well as larger samples. We are not told the population size here, but that is not an issue when using the spline method (as long as the population is not very small).

11.4 *This question is based on Subject CT4, September 2006, Question B6, parts (i) and (ii).*

(i)(a) *Form of the polynomial formula*

The general form of the polynomial formula is:

$$\overset{\circ}{\mu}_x = poly(1) + \exp\left[poly(2)\right]$$

where $poly(1)$ and $poly(2)$ are polynomials in x. [1]

(i)(b) ***Gompertz and Makeham formulae***

We can generate Gompertz' formula by setting $poly(1)$ equal to zero, and setting $poly(2)$ equal to a linear function of x, say $\alpha_1 + \alpha_2 x$. We can now write:

$$\mathring{\mu}_x = \exp(\alpha_1 + \alpha_2 x) = e^{\alpha_1}(e^{\alpha_2})^x \qquad [½]$$

This follows Gompertz' law, *ie* $\mathring{\mu}_x$ is of the form Bc^x with $B = e^{\alpha_1}$ and $c = e^{\alpha_2}$. [1]

We can obtain Makeham's law in exactly the same way, but by setting $poly(1)$ equal to a constant A instead of to zero. [½]

(ii)(a) ***Explanation***

Mortality is likely to be an exponentially increasing function over the age range from 40 to 75, so this might be a sensible formula since it behaves in the correct way. [½]

In addition, the x^2 term might allow for mortality worsening significantly with age for lives suffering from degenerative conditions. [½]

(ii)(b) ***Estimating the parameter values***

Two possible methods for estimating the parameter values are maximum likelihood estimation and least squares estimation. [2]

11.5 *This question is based on Subject CT4, April 2007, Question 2.*

(i)(a) ***A suitable method of graduation***

Graduation by reference to a standard table would be appropriate when graduating crude mortality rates for use in future premium calculations. [1]

This method would be appropriate for the following reasons:

- There is likely to be a suitable standard table available for graduating the mortality rates of annuitants. [½]
- There are likely to be ages at which the data are scarce, which means that graduation using a parametric formula or a spline function would not be appropriate. [½]
- Graduation by reference to a standard table will produce smooth graduated rates and will use information about the pattern of rates by age from the standard table. Smoothness is a desirable feature of graduated rates. [½]

(i)(b) ***Description of the method***

1. Select a standard table that is believed to have a similar pattern of mortality rates over the age range of interest, in this case the latest available table for annuitants (divided by gender). [½]

2. Look for a simple link between the crude mortality rates and the standard table rates. We would make some simple plots to help us here. For example, a plot of $\hat{\mu}_x$ against μ_x^s might indicate a linear relationship between the crude rates and the standard table rates. If this is the case, we could model the graduated rates using a function of the form $\overset{\circ}{\mu}_x = a + b\mu_x^s$. [½]

3. Estimate the parameter values using a mathematical technique, *eg* the method of maximum likelihood or least squares estimation. [½]

4. Calculate the graduated rates using the model with the parameters replaced by their estimated values. [½]

5. Test the graduated rates for goodness of fit. We don't need to test for smoothness. This is because the standard table should be smooth already and, provided we have used a simple function of the standard table rates to calculate the graduated rates, the graduated rates will automatically be smooth. If there is more than one possible simple link between the crude rates and the standard table rates, or more than one suitable standard table, the goodness-of-fit test may be used to help us make our final choice. [½]

(ii) ***Other considerations***

Since these rates are to be used to calculate premiums for annuitants, the company must be careful not to overestimate mortality. If it used mortality rates that were too high, the premiums charged to policyholders would not be sufficient to cover their annuity benefits. [1]

On the other hand, if the company uses mortality rates that are lower than its competitors, its policies will be expensive and it is likely to lose business. [½]

The company should also take account of expected future changes in mortality. [½]

12

Mortality projection

Syllabus objectives

4.6 Mortality projection

4.6.1 Describe the approaches to the forecasting of future mortality rates based on extrapolation, explanation and expectation, and their advantages and disadvantages.

4.6.2 Describe the Lee-Carter, age-period-cohort, and *p*-spline regression models for forecasting mortality.

4.6.3 Use an appropriate computer package to apply the models in 4.6.2 to a suitable mortality dataset.

4.6.4 List the main sources of error in mortality forecasts.

Objective 4.6.3 is covered in more detail in the R part of Subject CS2.

0 Introduction

The projection of future mortality rates is vital for many actuarial purposes, most obviously life insurance and pensions. If its estimates of future mortality are too low, a life insurance company may run into financial difficulties (premiums charged may be too low to allow it to meet its financial obligations). On the other hand, a company offering pensions and annuities could become uncompetitive. If estimates of future mortality are too high, on the other hand, the opposite problems may arise: the financial commitments of a pension scheme may outweigh its resources, but a life insurance company may offer uncompetitive rates and lose business to rival companies whose mortality projections are more accurate.

Governments, too, need accurate forecasts of future mortality rates. The future population of a country, especially the older age population, depends very much on future mortality trends. In ageing populations, this will have major implications for the provision of state pensions (including both the amounts offered and decisions about the pensionable age), and the funding of health and social care.

In the past, mortality projections generally relied on an *expectation* that previous trends in mortality improvements would continue, accompanied by consultation with demographers and other experts. For much of the twentieth century this proved reasonably satisfactory. However, more recently these methods have generally underestimated improvements in mortality, particularly in the period 1990-2010. An alternative approach along these lines is to suppose that, by some future date, mortality will have reached some target level, and to examine different ways in which the schedule of age-specific mortality rates might move from their present pattern in order to achieve that target.

Expectation-based methods are described in Section 1.

Since the early 1990s, interest has moved towards methods which involve the *extrapolation* of past mortality trends based on fitting statistical models to past data. Some use time series methods like those described in Chapters 13 and 14. Others are extensions to the graduation methods described in Chapter 11.

These methods are described in Section 2.

Standard epidemiological analyses presume that changes in mortality are associated with changes in mortality rates from specific causes of death. This is the rationale behind the cause-deleted life table approach of examining what would happen if mortality from a certain cause of death were eliminated. This leads to a third set of *explanatory* methods of forecasting which attempt to capture the reasons why mortality rates might vary in the future.

A life table is a computational tool that captures a certain mortality experience (*eg* based on a set of graduated mortality rates) which can be used to calculate probabilities of mortality and survival that reflect that experience. The construction and use of life tables is described in Subject CM1.

Explanatory methods of mortality forecasting are covered in Section 3.

The Core Reading in this chapter draws on two very good discussions of mortality projection methods: H. Booth and L. Tickle (2008) *Mortality modelling and forecasting: a review of methods* (Australian Demographic and Social Research Institute Working Paper no. 3, Canberra, Australian National University); and Chapters 12 and 13 of A.S. Macdonald, S.J Richards and I.D. Currie, *Modelling Mortality with Actuarial Applications* (Cambridge, Cambridge University Press).

Anyone considering using the methods described in this chapter should consult sources such as those mentioned in the above Core Reading. The details of the above reading, however, will not be required for the CS2 exam.

1 Methods based on expectation

Official statistical agencies have traditionally based their mortality projections on simple expectations (for example a continued exponential decline in age-specific mortality rates). The parameters of the future evolution of mortality have been set either by fitting deterministic functions to recent mortality trends, or by inviting experts to indicate how far and fast they anticipate mortality to fall, and to set 'targets'.

One approach involves the use of reduction factors, $R_{x,t}$, which measure the proportion by which the mortality rate at age x, q_x, is expected to be reduced by future year t. We can write:

$$R_{x,t} = \alpha_x + (1-\alpha_x)(1-f_{n,x})^{t/n}$$

where α_x is the ultimate reduction factor, and $f_{n,x}$ represents the proportion of the total decline expected to occur in n years. Expert opinion is used to set the targets α_x and $f_{n,x}$.

Question

Identify the value of the reduction factor when (a) $t=0$, (b) $t=n$, and (c) $t\to\infty$, and so describe the amount of reduction in future mortality that is predicted by this formula in each case.

Solution

(a) $\boldsymbol{t=0}$

The reduction factor becomes:

$$R_{x,0} = \alpha_x + (1-\alpha_x)(1-f_{n,x})^0 = \alpha_x + 1 - \alpha_x = 1$$

So the formula predicts the mortality rate for projection year 0 to be the same as the base mortality rate, as it should be.

(b) $\boldsymbol{t=n}$

Now the reduction factor becomes:

$$R_{x,n} = \alpha_x + (1-\alpha_x)(1-f_{n,x})^1 = \alpha_x + 1 - \alpha_x - (1-\alpha_x)f_{n,x} = 1-(1-\alpha_x)f_{n,x}$$

As α_x is the lowest possible value for the reduction factor, it follows that $(1-\alpha_x)$ is the maximum amount of reduction in future mortality that is assumed to be possible at this age. So, at time n, the formula predicts that a proportion $f_{n,x}$ of the maximum possible reduction will have been achieved.

(c) $\boldsymbol{t \to \infty}$

In this case the reduction factor becomes:

$$\lim_{t\to\infty} R_{x,t} = \lim_{t\to\infty}\left[\alpha_x + (1-\alpha_x)(1-f_{n,x})^{t/n}\right] = \alpha_x$$

because $0 < f_{n,x} < 1$. So the formula predicts that the reduction factor at age x can never be lower than α_x, regardless of how far we project mortality rates into the future.

An example of the use of this formula is shown in the *Tables* on page 109. This was used for projecting the future mortality of insurance company pensioner lives in the construction of the PMA92C20 and PFA92C20 tables. The calendar year 1992 is the base year for the projection, and the mortality rate for lives aged x in calendar year $1992+t$ is:

$$q_{x,t} = q_{x,0} \times R_{x,t}$$

where $q_{x,0}$ is the (graduated) mortality rate for calendar year 1992. The value of n was chosen to be 20 and examples of the factors used are shown below:

Age x	α_x	$f_{20,x}$
60	0.13	0.55
70	0.304	0.498
80	0.478	0.446
90	0.652	0.394
100	0.826	0.342
110	1	0.29

So, for example, at the time this projection method was proposed, it was expected that mortality rates for people aged 70 in any future year could never reduce to below around 30% of the current (1992) mortality rate for people aged 70 – *ie* a reduction of 70%. The choice of 0.498 for $f_{20,70}$ meant that about 50% of the total reduction was assumed to occur over the first 20 years (*ie* by calendar year 2012). So, according to this projection model, the mortality rate for people aged 70 in 2012 was projected to be:

$$q_{70,20} = q_{70,0} \times R_{70,20} = q_{70,0} \times \left[0.304 + 0.696 \times \left(1-0.498\right)\right] = q_{70,0} \times 0.653$$

ie a reduction of around 35% from 1992 values.

Question

(i) Using the values of α and f given above, calculate the reduction factors and hence the projected mortality rates for the shaded entries in the table below:

Age	Year					
	1992		1997		2002	
	R	q	R	q	R	q
60	1	0.005914	0.843	0.004983		
80	1	0.075464			0.867	0.065392

(ii) Use the values in the completed table to describe the view that was taken, about the expected future progression of mortality rates, when these projections were made.

Solution

(i) ***Reduction factors and projected mortality rates***

For age 60 in calendar year 2002, the reduction factor is:

$$R_{60,10} = 0.13 + 0.87 \times (1 - 0.55)^{10/20} = 0.713614$$

Hence the projected mortality rate in 2002 is:

$$q_{60,10} = q_{60,0} \times R_{60,10} = 0.005914 \times 0.713614 = 0.004220$$

Similarly, for age 80 in calendar year 2002, we have:

$$R_{80,5} = 0.478 + 0.522 \times (1 - 0.446)^{5/20} = 0.928348$$

and so:

$$q_{80,5} = q_{80,0} \times R_{80,5} = 0.075464 \times 0.928348 = 0.070057$$

So the completed table is:

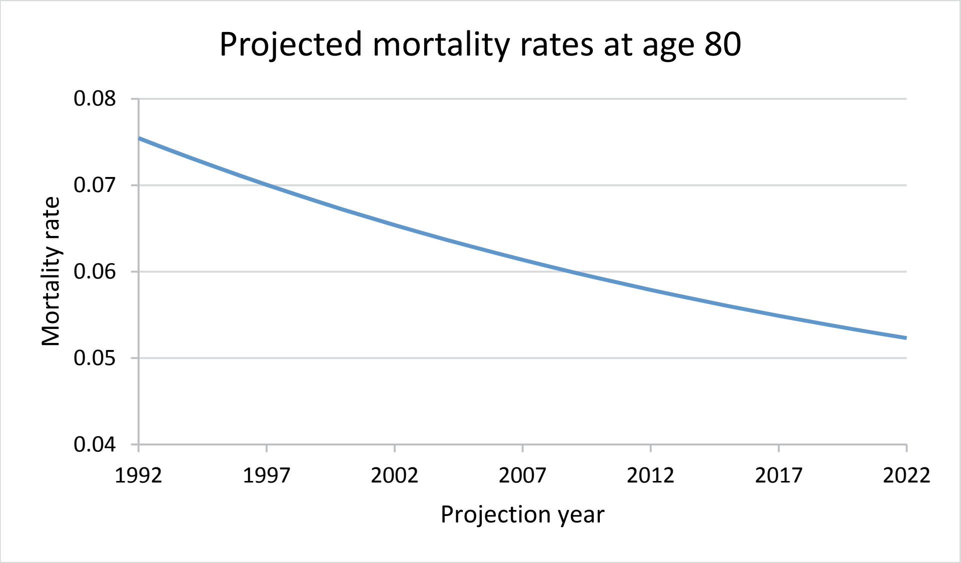
Age	Year					
	1992		1997		2002	
	R	q	R	q	R	q
60	1	0.005914	0.843	0.004983	0.714	0.004220
80	1	0.075464	0.928	0.070057	0.867	0.065392

(ii) ***Expected progression of mortality rates***

The projection reflects the expectation that:

- mortality will reduce more quickly at younger ages, and will ultimately reduce by a greater percentage than at older ages
- mortality will reduce more quickly in the short term, the speed of reduction reducing gradually the longer the projection period.

The second bullet in the above solution can be more clearly appreciated in the following plot of the projected mortality rates at age 80 over the first 30 projection years.



Methods based on expectation have been widely used in the past, and have the advantage of being straightforward and easy to implement. However, in recent decades they have tended to underestimate improvements, especially for male mortality. One reason for this is that progress on reducing male mortality was slow during the 1950s, 1960s and 1970s because of lifestyle factors such as cigarette smoking. The proportion of men who smoked was at a maximum among the cohorts born in the late nineteenth and early twentieth centuries, and in the post-War decades these cohorts were reaching the age when smoking-related deaths peak. Since the 1980s, changes in lifestyles, and developments in the prevention and treatment of leading causes of death such as heart disease and strokes, have meant rapid improvements in age-specific mortality for older men which most experts failed to see coming.

There are theoretical problems, too, with targeting. The setting of the 'target' is a forecast, which implies that the method is circular; and setting a target leads to an underestimation of the true level of uncertainty around the forecast.

Question

List and briefly describe the various sources of uncertainty involved in the mortality rates projected in the previous question.

Solution

Errors in the estimation of the graduated (base) mortality rates

- random error in the observed experience data (on which the base-year graduated rates were based)
- errors in fitting the model for the mortality rates to the experience data in the graduation process

Errors in the projection of the base mortality rates into the future

- Model error: the formula for $R_{x,t}$ will not have exactly the right functional form to produce a realistic projection of future mortality rates, regardless of the parameter values used.
- Parameter error: the parameter values will be incorrect, *ie* the values chosen for α_x and $f_{n,x}$ for each x will be subject to error.
- Random error: the actual mortality rates will differ from the projected rates due to random fluctuations in the mortality experience.
- Changes to the class of lives involved between past and future, *eg* if we are projecting pensioner mortality, the type of people becoming pensioners in future may differ from those who were pensioners when the base mortality rates were calculated. This may be because of changing types of pension arrangements or alternative options available, changes in the standard of living amongst pensioners, changes in the mix of pensioner lives by nationality, *etc*.

An alternative approach to generating expectations is to ask the population as a whole about its health status and general well-being. Self-reported health status has been found to be a good way of identifying 'healthy' and 'unhealthy' individuals, but it is not clear that it can provide useful information on future longevity at the population level.

2 Methods based on extrapolation

2.1 Stochastic models

Deterministic approaches based on expectation have been largely superseded, other than for short-term forecasting. More advanced approaches use stochastic forecasting models.

In this section we describe some commonly used models, but we start with a discussion of the factors that apply in forecasting mortality.

2.2 Age, period and cohort factors

In general, when forecasting mortality, the problem is to produce estimates of $m_{x,t}$, the central rate of mortality at age x at time t, for some future time period, based on data for $m_{x,t}$ over some past time period.

Recall from Chapter 6 that the central rate of mortality at age x is defined as:

$$m_x = \frac{q_x}{\int_{t=0}^{1} {}_tp_x \, dt} = \frac{\int_{t=0}^{1} {}_tp_x \, \mu_{x+t} \, dt}{\int_{t=0}^{1} {}_tp_x \, dt}$$

m_x is therefore a weighted average of the force of mortality over the year of age $[x, x+1]$. So:

$$m_x \approx \mu_{x+½}$$

and, if the force of mortality is assumed to be a constant ($\mu_{\overline{x}}$) over the year of age, then:

$$m_x = \mu_{\overline{x}}$$

So we can see that m_x is just a slightly different way of representing the force of mortality over a given year of age. As a result, the maximum likelihood estimate of the force of mortality is also an estimate of m_x, *ie*:

$$\hat{m}_x = \frac{\theta_x}{E_x^c}$$

where θ_x is the number of observed deaths over the year of age, and E_x^c is the corresponding central exposed to risk.

When considering the projection of mortality, we need to consider the future year in which the mortality rate is expected to apply, as well as the age of the person in that year. We include the additional time argument t for this purpose.

The $m_{x,t}$ are defined on the basis of two factors: age x and time (period) t.

If we define t to be the projection year and x to be the age reached during that projection year, then, for example, $m_{60,2030}$ is the expected mortality rate of those people who reach the age of 60 during the year 2030.

Age and period can be combined to produce a third factor, the *cohort*, defined, say, on the basis of date of birth. Because a person aged x at time t will have been born at time $t-x$, age, period and cohort are not independent.

So the mortality rates $m_{60,2030}$, $m_{61,2031}$, $m_{62,2032}$, *etc* all relate to the same cohort of lives born in 1970.

Forecasting models can be classified according to the number of factors taken into account in the forecasting processes, as follows:

One-factor models	**Age**
Two-factor models	**Age, period OR Age, cohort**
Three-factor models	**Age, period, cohort**

The mortality rates for these three models could be written as:

- One factor: m_x (1)
- Two factor (age and period) $m_{x,t}$ (2)
- Two factor (age and cohort) $m_{x,c}$ (3)
- Three factor (age, period and cohort) $m_{x,t,c}$ (4)

So, using these conventions for our example:

$$(2)=m_{60,2030};\quad (3)=m_{60,\,1970};\quad (4)=m_{60,2030,1970}$$

In two-factor models, it has been usual to work with age and period. It is possible to work with age and cohort, but the cohort approach makes heavy data demands and there is the largely insoluble problem that recent cohorts have their histories truncated by the present day.

Question

Describe and compare the data requirements if we wish to produce a table of projected mortality rates for pensioners aged 60-100 in 12 years' time, using the two alternative forms of the two-factor model.

Solution

Age-period model

Here we need a long enough period of past data to enable us to assess how mortality has been changing with calendar time, so that we can construct the period component of the model. The data will also need to relate to persons attaining all the required ages (60-100) over this historical period.

Age-cohort model

Here we need sufficient past data to identify differences in mortality according to cohort (year of birth). The historical data will also need to include experience at the required ages (60-100) so that the age component of the model can be assessed.

Comparison

For an age-period model, the main issue will be having a long enough investigation period to assess the effect of time period accurately enough for our projection purposes. However, as we are only projecting over a relatively short (12-year) period, then this problem should be surmountable.

For an age-cohort model, we need enough years of past data to enable the separate cohort effects to be accurately assessed. For example, persons aged 80 in 12 years' time will be 68 now, and it is probable that we will have past data on this cohort of lives (*eg* for lives aged 67 last year, 66 in the year before that, and so on.) However, for lives aged 60 in 12 years' time, it will be very unlikely that we will have any historical data for this cohort (as they will now be aged 48 and the past data relating to pensioners at these ages will be very sparse or even entirely absent).

Moreover, even for cohorts with *some* relevant past data, the data may not be sufficient to assess the mortality effects accurately enough. So while we might be able to assess the period component of the age-period model on 10-20 years of past data (say), we might need 30 or 40 years of past data per cohort for the cohort component of the age-cohort model.

For these reasons the age-cohort model is generally harder to use than the age-period model.

In general, most research has found that cohort effects are smaller than period effects, though cohort effects are non-negligible. Examples include the 'smoking cohort' mentioned above, of males born between 1890 and 1910, which had higher mortality than expected in middle-age, and the 'golden generation' born between 1925 and 1945 which had lower mortality than expected.

Three-factor models also have the logical problem that each factor is linearly dependent on the other two. Various approaches have been developed to overcome this problem, though none are entirely satisfactory because the problem is a logical one not an empirical one.

2.3 The Lee-Carter model

One of the most widely used models is that developed by Lee and Carter in the early 1990s. The Lee-Carter model has two factors, age and period, and may be written as follows:

Lee-Carter model

$$\log_e m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t}$$

where:

- $m_{x,t}$ **is the central mortality rate at age** x **in year** t
- a_x **describes the general shape of mortality at age** x **(more exactly it is the mean of the time-averaged logarithms of the central mortality rate at age** x**)**
- b_x **measures the change in the rates in response to an underlying time trend in the level of mortality** k
- k_t reflects the effect of the time trend on mortality at time t, and
- $\varepsilon_{x,t}$ **are independently distributed random variables with means of zero and some variance to be estimated.**

So, ignoring the error term, the mortality rate at age x in projection year t is:

$$m_{x,t} = \exp\left(a_x + b_x k_t\right)$$

As written above, the Lee-Carter model is not 'identifiable'. To obtain unique estimates of the parameters a_x**,** b_x**, and** k_t **constraints need to be imposed.**

Constraints imposed in the Lee-Carter model

The usual constraints are that $\sum_x b_x = 1$ **and** $\sum_t k_t = 0$ **.**

For example, without these constraints, there would be an infinite number of combinations of b_x and k_t values that could be fitted to the same set of observed data values.

To understand how all the factors contribute to the overall mortality rate in this model, let's look at a hypothetical example.

If we assume that mortality is improving with time (something that cannot, however, be taken as a given), then the time trend factor k_t would be reducing with increasing time t. So let's assume for simplicity that k_t decreases linearly over time.

b_x is the relative impact that the time trend has on mortality rates for a given age x. Evidence generally suggests that the time trend is less apparent at older ages, so we would therefore assume that the absolute value of b_x reduces with increasing age x.

Question

In a Lee-Carter model the function k_t is decreasing over time. At ages below 68 the b_x values are positive, and at older ages the b_x values are negative (while satisfying the overall constraint that $\sum_x b_x = 1$.)

Explain how the time trend in mortality rates in the model differs between the two age ranges.

Solution

As k_t is a decreasing function of t, a positive value of b_x means that $b_x k_t$ is also a decreasing function of t. So the model predicts that mortality rates will reduce over time at ages below 68.

Conversely, a negative value of b_x means that $b_x k_t$ is an increasing function of t. So the model predicts that mortality rates will increase over time at ages 68 and above.

---

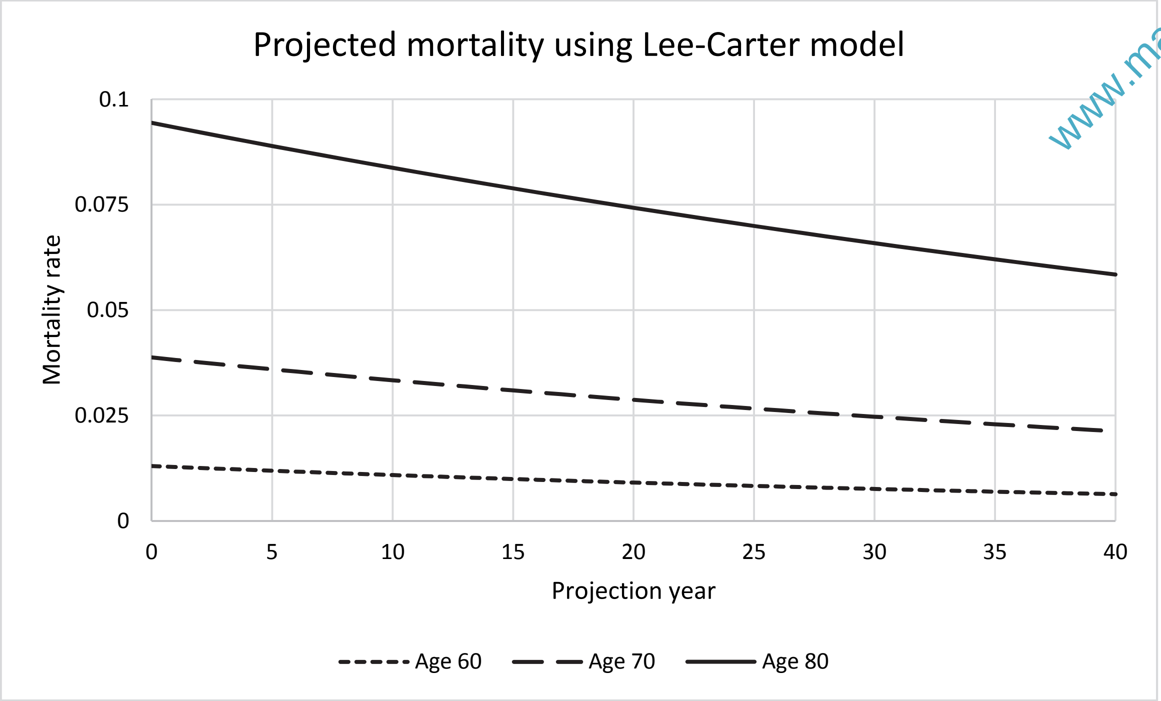
a_x is the overall 'age-effect' on mortality in the model. We can show that this is equal to the average value of $\ln\left(m_{x,t}\right)$ over the time period for a given age x. For a time period of n years and again ignoring the error terms:

$$\frac{1}{n}\sum_{t=1}^{n}\ln\left(m_{x,t}\right)=\frac{1}{n}\left(\sum_{t=1}^{n}a_x+b_x\sum_{t=1}^{n}k_t\right)=\frac{1}{n}\left(na_x+b_x\times 0\right)=a_x$$

The constraint $\sum k_t = 0$ was deliberately chosen to make this result hold.

We can also think of $\exp\left(a_x\right)$ as the projected mortality rate at age x at the median point in the projection period (*ie* at the time point where $k_t = 0$).

The following is a Lee-Carter projection model with a 40-year projection period based on the above assumptions. For simplicity, we have only produced projections for ages 60, 65, 70, 75 and 80.



In this example we have assumed that the time trend factor k_t decreases linearly from 1.5 at time 0 down to –1.5 at time 40. So $k_t = 1.5 - 0.075t$. The values used for a_x and b_x are:

x	a_x	$\exp(a_x) = m_x$	b_x
60	–4.70	0.00910	0.24
65	–4.10	0.01657	0.22
70	–3.55	0.02872	0.20
75	–3.05	0.04736	0.18
80	–2.60	0.07427	0.16

So, for example:

$$m_{60,t} = \exp(a_{60} + b_{60}\,k_t) = \exp\left[-4.70 + 0.24(1.5 - 0.075t)\right] = e^{-4.34} e^{-0.018t}$$

and hence:

$$m_{60,0} = e^{-4.34} = 0.01304$$

$$m_{60,10} = e^{-4.34} e^{-0.018\times 10} = 0.01089$$

$$m_{60,20} = e^{-4.34} e^{-0.018\times 20} = 0.00910$$

$$m_{60,30} = e^{-4.34} e^{-0.018\times 30} = 0.00760$$

$$m_{60,40} = e^{-4.34} e^{-0.018\times 40} = 0.00635$$

From the above, we see that:

- for every year that we project into the future, the mortality rate $m_{60,t}$ is multiplied by a factor of $e^{-0.018}$ (or 0.98216), *ie* it decreases by approximately 1.8% *pa*
- the projected mortality rate in 20 years' time is equal to the value of m_{60} .

Similarly:

$$m_{70,t} = \exp(a_{70} + b_{70}\,k_t) = \exp\left[-3.55 + 0.20(1.5 - 0.075t)\right] = e^{-3.25}e^{-0.015t}$$

which gives:

$$m_{70,0} = e^{-3.25} = 0.03877$$

$$m_{70,10} = e^{-3.25}e^{-0.015\times 10} = 0.03337$$

$$m_{70,20} = e^{-3.25}e^{-0.015\times 20} = 0.02872$$

$$m_{70,30} = e^{-3.25}e^{-0.015\times 30} = 0.02472$$

$$m_{70,40} = e^{-3.25}e^{-0.015\times 40} = 0.02128$$

So this time we multiply by a factor of $e^{-0.015}$ (or 0.98511) for each year that we project into the future and hence the values of $m_{70,t}$ decrease by approximately 1.5% *pa*. The percentage reduction is smaller than that for $m_{60,t}$ since $b_{70} < b_{60}$. In fact, we can see from the formula:

$$m_{x,t} = \exp(a_x + b_x\,k_t)$$

that the larger the value of b_x , the greater the effect of the time trend factor k_t .

Estimation of the Lee-Carter model

There are several approaches to estimating the Lee-Carter model.

- **The original approach of Lee and Carter first estimated the a_x as the time-averaged logarithms of mortality at each age x .**

  For example, if $\hat{m}_{x,t}$ is the maximum likelihood estimate (MLE) of $m_{x,t}$, then:

  $$\hat{a}_x = \frac{1}{n}\sum_{t=1}^{n}\ln\left(\hat{m}_{x,t}\right)$$

  is an MLE of:

  $$\frac{1}{n}\sum_{t=1}^{n}\ln\left(m_{x,t}\right) = a_x$$

They then used singular value decomposition of the matrix of centred age profiles of mortality $\log_e m_{x,t} - \hat{a}_x$ to estimate the b_x and k_t. (Singular value decomposition (SVD) is a way of decomposing a matrix into three component matrices, two of which are orthogonal and one diagonal.)

The details of this are not needed for the CS2 exam, but an outline of how it works goes as follows. The values of the residual errors $\log_e m_{x,t} - \hat{a}_x$ for different values of x and t form a matrix $(M,$ say$)$. The SVD method then writes this in the form $M = UDV$ where D is a diagonal matrix (*ie* a matrix with zeros everywhere apart from down the main diagonal), and U and V are matrices chosen to give the correct entries for the matrix M. If we now ignore the smaller entries along this diagonal and call this new matrix D^*, the values of $M^* = UD^*V$ still provide a good approximation to the original values but they can now be calculated using a formula with just a few parameters (one for each of the non-zero diagonal entries). We can then use these values in place of the actual residuals to obtain a set of smoothed mortality rates.

From the formula for the Lee-Carter model:

$$\ln\left(m_{x,t}\right) - a_x = b_x\, k_t + \varepsilon_{x,t}$$

So the idea of the above process is to find the 'best' (*eg* maximum likelihood) combination of b_x and k_t values that fit the observed values:

$$\ln\left(\hat{m}_{x,t}\right) - \hat{a}_x = \ln\left(\hat{m}_{x,t}\right) - \frac{1}{n}\sum_{t=1}^{n} \ln\left(\hat{m}_{x,t}\right)$$

In selecting these estimates, the required constraints $\sum_x b_x = 1$ and $\sum_t k_t = 0$ would be imposed.

- **Macdonald *et al* propose an alternative method which makes use of the `gnm` package in R. In this approach the `gnm` function is used to obtain estimates of the a_x, b_x, and k_t. These estimates will not satisfy the constraints that $\sum_x b_x = 1$ and $\sum_t k_t = 0$. However, a simple adjustment of the estimates produced by the `gnm` function will recover estimates of a_x, b_x, and k_t which do satisfy these constraints.**

In R `gnm` (generalised non-linear models) is a separate package which needs downloading and installing independently of the basic R system.

Forecasting with the Lee-Carter model

The Lee-Carter model has three sets of parameters, a_x, b_x, and k_t. Two of these relate to the age pattern of mortality, whereas the third, k_t, measures how mortality evolves over time. Forecasting with the model in practice involves forecasting the k_t while holding the a_x and b_x constant.

One obvious approach is to use time series methods such as those described in Chapters 13 and 14 to forecast the k_t. Lee and Carter originally used a random walk on the differenced k_t series, but other auto-regressive and moving average models described in Chapter 13 could be used.

The random walk model may be written as:

$$k_t - k_{t-1} = \Delta k_t = \mu + \varepsilon_t$$

where μ measures the average change in the k_t and the ε_t are independent normally distributed error terms with variance σ^2.

Because the random walk model assumes that the average increase in k_t is a constant μ per unit time, we are essentially assuming k_t varies linearly over time. This is what we assumed in the example Lee-Carter model shown earlier.

Suppose that t_0 is the latest year for which we have data. Having estimated μ using data for $t < t_0$, forecasting can be achieved for the first future period, as:

$$\hat{k}_{t_0+1} = k_{t_0} + \hat{\mu}$$

and, in general, for l years ahead:

$$\hat{k}_{t_0+l} = k_{t_0} + l\hat{\mu}$$

Question

Suppose the last year for which we have data is 2017. The estimated value of k_{2017} from the model is 0.93, and the estimate of μ is -0.007.

Calculate the predicted value of k_{2025}.

Solution

The predicted value of k_{2025} is:

$$\hat{k}_{2025} = k_{2017} + \hat{\mu} \times (2025 - 2017) = 0.93 - 0.007 \times 8 = 0.874$$

While the predicted values for this random walk model appear identical in form to those obtained using a standard linear regression model, the error terms are modelled differently.

Question

A linear regression model for k_t might be written as:

$$k_t = \alpha + \beta t + \varepsilon_t$$

where:

$$\varepsilon_t \sim N(0, \sigma^2)$$

for all values of t. Explain how this differs from the random walk model of k_t defined in the Core Reading.

Solution

In the linear regression model, ε_t is the error between the actual value of k_t and its predicted value $\alpha + \beta t$.

In the random walk model, ε_t is the error between the actual value of the *increment* (or increase) in the value of k_t from its previous value k_{t-1}.

So, when we make forecasts using the random walk model, the errors accumulate over time, meaning that we are increasingly uncertain about our forecasts further into the future. In a linear regression model, the error (between actual and predicted values) is assumed to be stationary, *ie* it has the same distribution (with constant variance) over time.

$\hat{\mu}$ is calculated from observed historical data, and so is itself also subject to uncertainty. If we denote the estimator of μ as $\tilde{\mu}$, then:

$$\text{var}(\tilde{\mu}) = \frac{\sigma^2}{t_n - 1}$$

where t_n is the number of past years' data used to estimate μ.

So:

$$\text{var}\left(k_{t_0} + l\tilde{\mu}\right) = l^2 \text{var}(\tilde{\mu}) = l^2 \left(\frac{\sigma^2}{t_n - 1}\right)$$

The standard error of the forecast is therefore given by:

$$\boldsymbol{SE(k_{t_0+l}) = l\left(\frac{\sigma}{\sqrt{t_n - 1}}\right)}$$

The standard error can be used to draw confidence intervals around the predicted values of k_t. However, these will underestimate the true forecast error, as they only take into account error in the estimation of μ. The overall error in the forecast is made up of two components:

- **error in the estimation of μ**
- **the fact that the actual future observations will vary randomly according to the value of σ^2.**

So the actual future value of the k parameter in year $t_0 + l$ will be given by:

$$\hat{k}_{t_0+l} = k_{t_0} + l\hat{\mu} + \sum_{j=1}^{l} \varepsilon_{t_0+j}$$

This formula shows the accumulation of all the (incremental) future error terms we mentioned earlier.

See Macdonald *et al* for a fuller discussion of this point.

Question

Consider again the previous calculation question, where the estimated value of k_{2017} is 0.93, and the estimate of μ is -0.007. From this we calculated the predicted value of k_{2025} to be 0.874.

Assume the error terms ε_t are normally distributed with zero mean and common standard deviation $\sigma = 0.0055$, and that the estimates of the model parameters were based on 37 years of historical data. Calculate a 95% confidence interval for the value of k_{2025}.

Solution

A 95% confidence interval for k_{2025} is:

$$\hat{k}_{2025} \pm 1.96 \times l\left(\frac{\sigma}{\sqrt{t_n - 1}}\right)$$

where the values ± 1.96 are the upper and lower 2.5% points of the standard normal distribution. Putting in the given values, the confidence interval is:

$$0.874 \pm 1.96 \times 8 \times \left(\frac{0.0055}{\sqrt{37-1}}\right) = 0.874 \pm 0.01437 = (0.860, 0.888)$$

Don't confuse the t_n notation used here with the Student's t distribution, which features in some similar formulae where we have to estimate the value of σ when the variance is unknown.

Predicted future mortality rates in year $t_0 + l$ are then obtained as:

$$\log_e \hat{m}_{x,t_0+l} = \hat{a}_x + \hat{b}_x \hat{k}_{t_0+l}$$

Advantages and disadvantages of the Lee-Carter model

The Lee-Carter model has the advantage that, once the parameters have been estimated, forecasting is straightforward and can proceed using standard time-series methods, the statistical properties of which are well known. The degree of uncertainty in parameter estimates, and hence the extent of random error in mortality forecasts, can be assessed. The Lee-Carter model can also be extended and adapted to suit particular contexts, for example by smoothing the age patterns of mortality using penalised regression splines (see Section 2.5 below).

Disadvantages of the Lee-Carter model include:

(1) Future estimates of mortality at different ages are heavily dependent on the original estimates of the parameters a_x and b_x. The forecasting assumes that these remain constant into the future. The parameters a_x (the general shape of mortality at different ages), and b_x (the change in the rates in response to an underlying time trend in the level of mortality) are estimated from past data, and will incorporate any roughness in the past data. They may be distorted by past period events which affected different ages to different degrees. If the estimated b_x values show variability from age to age, it is possible for the forecast age-specific mortality rates to 'cross over' (such that, for example, $\hat{m}_{65,t+j} < \hat{m}_{66,t+j}$ but $\hat{m}_{65,t+j+1} > \hat{m}_{66,t+j+1}$).

So in some future years the forecast mortality rate is higher at age 65 than it is at age 66.

This can be avoided by smoothing the estimates of a_x and b_x.

(2) There is a tendency for Lee-Carter forecasts to become increasingly rough over time.

(3) The model assumes that the underlying rates of mortality change are constant over time across all ages, when there is empirical evidence that this is not so.

(4) The Lee-Carter model does not include a cohort term, whereas there is evidence from UK population mortality experience that certain cohorts exhibit higher improvements than others.

(5) Unless observed rates are used for the forecasting, it can produce 'jump-off' effects (*ie* an implausible jump between the most recent observed data and the forecast for the first future period).

Extensions to the Lee-Carter model

Modifications and extensions to the Lee-Carter approach have been suggested to try to overcome the disadvantages listed above. There is evidence that most of these modifications and alternatives to the original Lee-Carter model give lower forecast errors than the original in respect of age-specific mortality rates, though there is little difference when the expectation of life is considered (For details of this, see H. Booth, R.J. Hyndman, L. Tickle and P. de Jong (2006) 'Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions', *Demographic Research* 15, pp. 289-310).

2.4 The age-period-cohort model

The use of cohort effects in forecasting

The Lee-Carter model is a two-factor model, taking into account age and period. It does not account for variations in mortality across cohorts.

Generally, cohort effects are smaller than period effects, but they have been observed for certain countries, including the United Kingdom, the United States, France and Japan. Models that take into account cohort effects as well as age and period effects have, in some circumstances, proved superior to two-factor models in forecasting.

Age-period-cohort models have substantial disadvantages:

- **The identification problem: any one factor is linearly dependent on the other two. Various solutions to this problem have been proposed, including the use of three-way classification in data collection, and the imposition of constraints on model parameters.**
- **Models incorporating cohort effects impose heavy data demands. To observe the mortality of a cohort over all ages requires around 100 years of data. Cohort data truncated at the current age of the cohort can be used, but a model will be needed to estimate the experience of the cohort at older ages.**

Adding cohort effects to the Lee-Carter model

An age-period-cohort extension of the Lee-Carter model may be written:

$$\log_e m_{x,t} = a_x + b_x^1 k_t + b_x^2 h_{t-x} + \varepsilon_{x,t}$$

where h_{t-x} is the overall level of mortality for persons born in year $t-x$. See A. E. Renshaw and S. Haberman (2006) 'A cohort-based extension of the Lee-Carter model for mortality reduction factors', *Insurance, Mathematics and Economics* 38, pp. 556-70.

So in this model we now have two 'b_x' parameters for each age x:

- b_x^1, which is the extent to which the time trend affects mortality rates at age x, and
- b_x^2, which is the extent to which the cohort affects mortality rates at age x.

The superscripts on the symbols are *not* powers, but indicate the two different parameter values at the given age. b_x^1 is exactly equivalent to b_x in the standard Lee-Carter model. If we use c to represent the cohort year (as we did in Section 2.2 above), then we could alternatively write the age-period-cohort Lee-Carter model as:

$$\log_e m_{x,t,c} = a_x + b_x^1 k_t + b_x^2 h_c + \varepsilon_{x,t}$$

which shows more clearly how the three factors (age x, period t and cohort c) all influence the projected mortality rate.

In this case the h_{t-x} can be estimated from past data and the forecasting achieved using time series methods similar to those described for the k_t parameter in Section 2.3 above.

2.5 Forecasting using *p*-splines

In Chapter 11, the idea of graduation using splines was introduced. In this section, we extend this idea to forecasting.

Splines

We first recap how splines are used for modelling a set of observed mortality rates that are assumed to vary only according to age x, *ie* using a single factor model.

Recall from Chapter 11 that a spline is a polynomial of a specified degree defined on a piecewise basis. The pieces join together at knots, where certain continuity conditions are fulfilled to ensure smoothness. In Chapter 11, the splines were fitted to an age schedule of mortality, so the knots were specified in terms of ages. Typically, the polynomials used in splines in mortality forecasting are of degree 3 (*ie* cubic).

In Chapter 11 we saw that the (natural) cubic spline function, with n knots at values $x_1, x_2, \dots x_n$, is defined as:

$$f(x)=\alpha_0+\alpha_1 x+\sum_{j=1}^{n}\beta_j\,\phi_j(x)$$

where:

$$\phi_j(x)=\begin{cases}0 & x<x_j\\ (x-x_j)^3 & x\geq x_j\end{cases}$$

Question

State the continuity conditions that are incorporated in this function, which ensure the smoothness of the joins between the successive cubic functions at each knot.

Solution

The continuity conditions are, for each knot (*ie* at values $x_1, x_2, \dots x_n$):

- the value of the cubic leading into the knot and leading out of the knot are equal at the knot
- the first derivative of the of the cubic leading into the knot and leading out of the knot are equal at the knot
- the second derivative of the cubic leading into the knot and leading out of the knot are equal at the knot.

We also impose the condition that before the first knot, and after the last knot, the spline function is linear.

So, we can find the 'best' cubic spline function that fits the observed mortality rates across the observed age range, choosing a function that gives the desired compromise between smoothness from age to age and adherence to the observed data values at each age.

To construct the model, we choose the number of knots (and hence the number of splines to use), and the degree of the polynomials in each spline. We can then use the splines in a regression model, such as the Gompertz model.

To illustrate, the Gompertz model can be written as:

$$\log_e[E(D_x)] = \log_e E_x^c + \alpha + \beta x \qquad (1)$$

where $E(D_x)$ is the expected deaths at age x, E_x^c is the central exposed to risk at age x, and α and β are parameters to be estimated.

Rearranging (1):

$$\ln[E(D_x)] - \ln E_x^c = \alpha + \beta x$$

which is equivalent to:

$$\ln\left[\frac{E(D_x)}{E_x^c}\right] = \alpha + \beta x \qquad (2)$$

where $\dfrac{E(D_x)}{E_x^c}$ is the true force of mortality μ for lives labelled as aged x (since $E(D_x) = \mu E_x^c$).

If the age definition for deaths and exposed to risk is 'age nearest birthday', then this will be the force of mortality at *exact* age x, *ie* μ_x.

Question

Show how the Gompertz model just defined relates to the Gompertz law (as shown on page 32 of the *Tables*).

Solution

With deaths and exposed to risk aged x nearest birthday, from (2) the Gompertz model can be written:

$$\ln\left[\frac{E(D_x)}{E_x^c}\right] = \ln \mu_x = \alpha + \beta x \qquad (3)$$

The Gompertz law for the force of mortality at exact age x is usually written as:

$$\mu_x = Bc^x$$

This gives:

$$\ln \mu_x = \ln B + x \ln c$$

So (3) represents the Gompertz law with $B = e^{\alpha}$ and $c = e^{\beta}$.

However the Gompertz law is a deterministic formula, whereas the Gompertz *model* is a stochastic model of the number of deaths occurring in a specified age group. So instead of writing the model as in (3) we could alternatively define it in stochastic form by replacing $E(D_x)$ with D_x:

$$\ln\left[\frac{D_x}{E_x^c}\right] = \alpha + \beta x + e_x$$

where e_x is a random error term with mean zero (and for which a probability distribution would need to be assumed).

If in (1) we replace the term $\alpha + \beta x$ by a smooth function defined using splines, we have:

$$\log_e[E(D_x)] = \log_e E_x^c + \sum_{j=1}^{s} \theta_j B_j(x) \qquad (4)$$

where $B_j(x)$ are the set of splines, and θ_j are the parameters to be estimated. The number of splines is s (see Macdonald *et al*).

So all we are doing here is replacing the (relatively inflexible) function $\alpha + \beta x$ with the (much more flexible) cubic spline function. We should be able to get a much better fit to the observed data than if we just tried to fit the Gompertz law itself.

In this equation the θ_j and $B_j(x)$ correspond to the β_j and $\phi_j(x)$ that we used earlier in this chapter (and in Chapter 11) for splines.

A difference between (4) and the spline graduations we described in Chapter 11 is that, in Chapter 11 we fitted the spline directly to observed values of $\hat{\mu}_x$, whereas here we are fitting the spline to observed values of $\ln \hat{\mu}_x$, *ie* using log-transformed data.

To see that this is the case, consider:

$$\frac{E(D_x)}{E_x^c} = \mu_x \;\Rightarrow\; \ln\left[\frac{E(D_x)}{E_x^c}\right] = \ln \mu_x \;\Rightarrow\; E(D_x) = \ln E_x^c + \ln \mu_x$$

Comparing this with Equation (4) we can see that the spline function is being used to model $\ln \mu_x$.

Question

Give one reason for using the log-transformed data in this way.

Solution

Mortality for the middle and older ages generally varies approximately exponentially with increasing age, so the logarithm of the mortality rate would be expected to follow an approximately linear progression. By transforming the data in this way, we can expect to be able to use a simple polynomial to fit the data.

p-splines

Spline models will adhere more closely to past data if the number of knots is large and the degree of the polynomial in the spline function is high. For forecasting, we ideally want a model which takes account of important trends in the past data but is not influenced by short-term 'one-off' variations. This is because it is likely that the short-term variations in the past will not be repeated in the future, so that taking account of them may distort the model in a way which is unhelpful for forecasting. On the other hand, we do want the model to capture trends in the past data which are likely to be continued into the future.

Models which adhere too closely to the past data tend to be 'rough' in the sense that the coefficients for adjacent years do not follow a smooth sequence.

By this we mean that the sequence of estimated parameter values $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s$ forms an uneven progression. In practice, it is found that roughness in these parameters also leads to a corresponding roughness in the mortality rates that are predicted by the fitted model. So if we can reduce the roughness in the fitted parameters, we should consequently produce a model with a smoother progression of mortality rates from age to age.

The method of *p*-splines attempts to find the optimal model by introducing a penalty for models which have excessive 'roughness'. The method may be implemented as follows:

- **Specify the knot spacing and degree of the polynomials in each spline.**
- **Define a *roughness penalty*, $P(\theta)$, which increases with the variability of adjacent coefficients. This, in effect, measures the amount of roughness in the fitted model.**

  $P(\theta)$ is a function of the fitted parameter values $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s$ such that, the more irregular the progression of $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s$ is, the higher $P(\theta)$ will be.
- **Define a smoothing parameter, λ, such that if $\lambda = 0$, there is no penalty for increased roughness, but as λ increases, roughness is penalised more and more.**
- **Estimate the parameters of the model, including the number of splines, by maximising the penalised log-likelihood**

$$l_p(\theta) = l(\theta) - \frac{1}{2}\lambda P(\theta)$$

  where $l(\theta)$ is the log-likelihood from model (4).

So when we wish to estimate the parameters $\theta_1, \theta_2, ..., \theta_s$ (and also the value of the number of knots s), we first define a likelihood function (in terms of those parameters) that is proportionate to the probability of the observed mortality rates occurring. The 'log-likelihood' is the natural log of this function.

The penalised log-likelihood is effectively trying to balance smoothness and adherence to the data.

As the 'rougher' values of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$ would cause the mortality rates to adhere more closely to the data, so the normal (unpenalised) maximum likelihood estimates would tend to result in parameters (and hence fitted mortality rates) that do not progress as smoothly as desired. The penalty factor means that the estimation process automatically compensates for this feature. We can exercise some control over the balance between smoothness and adherence both through the choice of penalty function and by changing the value of λ.

Question

Explain why the following might be a suitable choice of penalty function:

$$P(\theta) = (\theta_1 - 2\theta_2 + \theta_3)^2 + (\theta_2 - 2\theta_3 + \theta_4)^2 + \cdots + (\theta_{s-2} - 2\theta_{s-1} + \theta_s)^2$$

Solution

The function can be alternatively written as:

$$P(\theta) = ([\theta_3 - \theta_2] - [\theta_2 - \theta_1])^2 + ([\theta_4 - \theta_3] - [\theta_3 - \theta_2])^2 + \cdots + ([\theta_s - \theta_{s-1}] - [\theta_{s-1} - \theta_{s-2}])^2$$

This is the same as:

$$\begin{aligned} P(\theta) &= (\Delta\theta_2 - \Delta\theta_1)^2 + (\Delta\theta_3 - \Delta\theta_2)^2 + \cdots + (\Delta\theta_{s-1} - \Delta\theta_{s-2})^2 \\ &= \left(\Delta^2\theta_1\right)^2 + \left(\Delta^2\theta_2\right)^2 + \cdots + \left(\Delta^2\theta_{s-2}\right)^2 \end{aligned}$$

where Δ^r indicates the value of the r th order difference. So, minimising $P(\theta)$ will attempt to select values of θ_j that minimise the sum of the 2nd differences, in a similar way to the smoothness test in graduation, where we want the 3rd differences to be small.

For example, let's define:

$$f(a,b,c) = (c-b) - (b-a)$$

and compare the values of this function for the sequence $(4,5,7)$ *and the sequence* $(4,7,5)$. *The function yields:*

$$f(4,5,7) = (7-5) - (5-4) = 2 - 1 = 1$$

$$f(4,7,5) = (5-7) - (7-4) = -2 - 3 = -5$$

So the 'rougher' progression yields the larger absolute value of the second difference.

Squaring the function ensures that the higher absolute values are the most penalised (by making all the values positive), and also places a proportionately greater penalty on the larger absolute differences. This should ultimately result in a smoother final outcome.

Forecasting

We now turn to forecasting future mortality rates by age x and time period t, *ie* using a two-factor model. The basic process is to use a (spline) function to model values of $\ln m_{x,t}$ by time period (or year) t, using a different spline function for each age (or group of ages) identified by x. So now we are fitting the function by time period, rather than by age.

So we have:

$$\log_e[E(D_{x,t})] = \log_e E^c_{x,t} + \sum_{j=1}^{s} \theta_j B_j(t)$$

or:

$$\ln\left[m_{x,t}\right] = \sum_{j=1}^{s} \theta_j B_j(t) \qquad (5)$$

Recall that if x is an integer and we are grouping data by age nearest birthday, then the mortality functions μ and m are essentially interchangeable.

Forecasting using *p*-splines is effected at the same time as the fitting of the model to past data. The past data used will consist of deaths and exposures at ages x for a range of past years t.

Forecasting may be carried out for each age separately, or for many ages simultaneously. In the case of a single age x, we wish to construct a model of the time series of mortality rates $m_{x,t}$ for age x over a period of years, so the knots are specified in terms of years.

Having decided upon the forecasting period (the number of years into the future we wish to forecast mortality), we add to the data set dummy deaths and exposures for each year in the forecast period. These are given a weight of 0 when estimating the model, whereas the existing data are given a weight of 1. This means that the dummy data have no impact on $l(\theta)$. We then choose the regression coefficients in the model (5) so that the penalty $P(\theta)$ is unchanged.

***p*-spline forcasting for single years of age or for many ages simultaneously can be carried out using the `MortalitySmooth` package in R.**

Advantages and disadvantages of the *p*-spline approach

The *p*-spline approach has the advantages that it is a natural extension of methods of graduation and smoothing, and it is relatively straightforward to implement in R.

It has the following disadvantages:

- **When applied to ages separately, mortality at different ages is forecast independently. So there is a danger that there will be roughness between adjacent ages. This can be overcome by fitting the model and forecasting in two dimensions (age and time) simultaneously.**
- **There is no explanatory element to the projection (in the way that time-series methods use a structure for mortality and an identifiable time series for projection).**

  So, when we fit the model we will obtain a set of numbers for the spline coefficients, but these don't have any natural interpretation, and so it is not easy to compare different models in terms of the differences in the parameter values obtained.

- ***p*-splines tend to be over-responsive to an extra year of data (though this can be ameliorated by increasing the knot spacing).**

  This means that if we fit the model to n years of data, and fit it again to $n+1$ years of data (*eg* because we've just gathered another complete year of observed experience), the model changes more dramatically than we would normally expect (*eg* compared to the case where we are fitting a standard mathematical formula like the Gompertz model).

Extensions and combinations with the Lee-Carter model

Several variations on the *p*-spline approach have been proposed.

R.J. Hyndman and M.S. Ullah (2007) 'Robust forecasting of mortality and fertility rates: a functional data approach', *Computational Statistics and Data Analysis* 51, pp. 4,942-4,956, proposed a model which combines the *p*-spline method with the Lee-Carter model. This can be written as follows:

$$\log_e m_{x,t} = a_x + \sum_{j=1}^{J} k_{t,j}\, b_j(x) + \text{error term}$$

Here, the a_x represent the average pattern of mortality across years, but smoothed using *p*-splines. The term $\sum_{j=1}^{J} k_{t,j}\, b_j(x)$ replaces the $b_x k_t$ in the Lee-Carter model: $b_j(x)$ are basis functions and $k_{t,j}$ are time series coefficients. For each age x there is a set of J such functions and coefficients, estimated using principal components decomposition.

The second term is an attempt to allow the model to be structured according to the factors that the historical data *indicate* are significant for the mortality projection. For example, the data may indicate that different time trends are operating for smokers (S) and non-smokers (NS). So, we might then construct the model with the following parameters (showing two ages for example):

Category j	$k_{t,j}$	$b_j(x)$	
		$x=40$	$x=60$
Non-smoker	$k_{t,NS}=1.6-0.08t$	$b_{NS}(40)=0.24$	$b_{NS}(60)=0.18$
Smoker	$k_{t,S}=0.8-0.04t$	$b_S(40)=0.32$	$b_S(60)=0.15$

So, according to this model, the time trend has a greater effect:

- for non-smokers than smokers (because k_t has a steeper negative gradient with t for non-smokers than smokers)
- for younger ages than older ages (because in all cases $b(40)$ is greater than $b(60)$), there being a bigger age effect for smokers than non-smokers (because the difference between $b(40)$ and $b(60)$ is bigger for smokers than for non-smokers).

3 Methods based on explanation

In this section, we will consider models that take into account the different causes of mortality, such as cancer and heart disease.

The previous approaches take no, or only limited, cognisance of the causal factors underlying mortality. Since causal factors are quite well understood, at least at a general level, it might be thought sensible to use this knowledge in forecasting. For example, if cancer is a leading cause of death in a country, and if it seems likely that a significant breakthrough in the treatment of cancer is likely, this could be explicitly taken account of in mortality projections for that country.

Question

Suppose the current mortality rate at age 70 for males in a particular country is 0.01 *pa*. An analysis of the causes of mortality at this age reveals that 35% of deaths are due to heart disease, 40% from cancer, and 25% from all other causes.

Due to the introduction of a revolutionary new treatment, next year it is predicted that the mortality rate of males aged 70 due to cancer will be 90% of the current rate. At the same time, it is predicted that the mortality rate due to heart disease will increase by 2% and that due to other causes will increase by 1% of their current rates.

(i) Give a possible reason why the deaths from causes other than cancer might have increased.

(ii) Calculate the expected population mortality rate for male lives aged 70 in one year's time.

Solution

(i) ***Why mortality rates from other causes have increased***

Those people who survive the risk of dying from cancer are nevertheless still exposed to the risk of dying from the other possible causes. As the 'exposed to risk' of dying from other causes has increased, all else being equal we would expect the total number of deaths due to these other causes to rise slightly.

(ii) ***Projecting the overall mortality rate***

The current rates of mortality of males aged 70 are:

Heart disease	0.0035
Cancer	0.0040
Other	0.0025
Total	0.01

The projected changes lead to the following expected mortality rates by cause of death in one year's time:

Heart disease	0.0035×1.02	=	0.003570
Cancer	0.0040×0.9	=	0.003600
Other	0.0025×1.01	=	0.002525
Total			0.009695

So it is predicted that overall the mortality rate for this age group will reduce by around 3% over the next year.

Cause-deleted life table approach

Consider the following extract from the ELT15 (Males) mortality table (shown on page 69 of the *Tables*):

Age x	l_x	d_x	q_x
65	79,293	1,940	0.02447
66	77,353	2,097	0.02711
67	75,256		

In this table:

- l_x is the expected number of people surviving to exact age x
- d_x is the expected number of people dying in the year between exact age x and exact age $x+1$
- q_x is the probability of a person who is currently aged exactly x, dying between exact age x and exact age $x+1$.

Question

Write down a formula for q_x in terms of l_x, d_x, or both.

Solution

The probability of dying during the year can be calculated as the proportion of lives, currently aged x, who are expected to die during the year. That is:

$$q_x = \frac{\text{expected deaths between } x \text{ and } x+1}{\text{number of survivors at exact age } x} = \frac{d_x}{l_x}$$

Alternatively, because the number dying during the year is the difference between the numbers of survivors at the start and end of the year, we can write:

$$q_x = \frac{l_x - l_{x+1}}{l_x}$$

or:

$$q_x = 1 - \frac{l_{x+1}}{l_x}$$

Now let's suppose that 10% of the deaths at each of the ages shown in this extract are caused by a particular type of cancer. A new cure for this illness has been introduced that is expected to halve the deaths from this cause in the coming year and to eliminate them altogether in all years thereafter.

Question

Consider a cohort of people currently aged exactly 65, who in the absence of the new cure were originally expected to follow the mortality of the ELT15 (Males) table over their next two years of age. Recalculate the entries for d_x, l_x, and q_x in the above life table, allowing for the changes in death rates from this particular cancer that have been predicted for the next two years.

Solution

In the first year, 10% of the expected deaths at age 65 were due to this cancer. This is:

$$0.1 \times 1{,}940 = 194$$

Halving the expected number of deaths from this cause means that the revised value of d_{65} is:

$$d'_{65} = 1{,}940 - 0.5 \times 194 = 1{,}843$$

(Alternatively this can be calculated as $1{,}940 \times 0.95$.)

The revised mortality probability is therefore:

$$q'_{65} = \frac{1{,}843}{79{,}293} = 0.02324$$

The revised number of expected survivors at age 66 is:

$$l'_{66} = l_{65} - d'_{65} = 79{,}293 - 1{,}843 = 77{,}450$$

The mortality rate at age 66 will now reduce to 90% of its original value (as we have eliminated all 10% of the deaths that were previously expected in this year of age). This mortality rate is therefore:

$$q'_{66} = 0.02711 \times 0.9 = 0.02440$$

So the revised expected number of deaths at age 66 is:

$$l'_{66} \times q'_{66} = 77,450 \times 0.02440 = 1,890$$

The number of survivors expected at age 67 is:

$$l'_{67} = l'_{66} - d'_{66} = 77,450 - 1,890 = 75,560$$

So the revised life table looks like this:

Age x	l'_x	d'_x	q'_x
65	79,293	1,843	0.02324
66	77,450	1,890	0.02440
67	75,560		

The above projection is not quite accurate as we have not allowed for the increased mortality that would be expected from other causes.

Multiple state modelling

Greater sophistication in this approach might be achieved using multiple state Markov jump process models (as described elsewhere in this course). The states in the model might include:

- Healthy
- Disease state 1
- Disease state 2
- …
- Disease state n
- Dead.

The model could incorporate transitions:

- from healthy to each disease state
- from healthy to dead (without visiting any disease state)
- from each disease state to dead
- from each disease state back to healthy (as appropriate)
- between the different disease states (as appropriate).

Such an approach is obviously complex and can involve a huge number of parameters (transition rates). The reliability with which this modelling can be carried out is also doubtful at the present time. Some reasons for this are described below.

It might seem surprising that progress in using the explanatory approach has been limited. However, the methods require either that long lags are specified between changes in the risk factors and changes in mortality, or that the risk factors themselves be forecasted, and forecasting developments in the risk factors is likely to be almost as difficult as forecasting mortality.

Decomposition of mortality by cause of death is an integral part of the explanatory approach. However, in practice it is difficult to achieve successfully. The reasons include:

- **cause of death reporting is unreliable, especially at older ages (where an increasing proportion of deaths occur)**
- **causes of death often act synergistically, so it is not realistic to posit a single cause of death**
- **elimination of one cause of death might 'unmask' another cause that would not have been identified previously**
- **the time series of data are often rather short.**

4 Sources of error in mortality forecasts

Mortality forecasts are always wrong. It is of interest to know how wrong they are likely to be, and what the main sources of error are. The latter is important not so much because it will help to eliminate errors (this is not possible) but so effort can be focused on the areas most likely to cause the forecasts to be at variance with reality, or on the elements of the process to which the sensitivity of the outcome is greatest.

Alho, J.M. (1990) 'Stochastic methods in population forecasting', *International Journal of Forecasting* 6, pp. 521-30, classified sources of error as follows:

1. **Model mis-specification. We might have the wrong parameterisation function, or the wrong model.**

2. **Uncertainty in parameter estimates.**

3. **Incorrect judgement or prior knowledge. The data set we use as the basis may not accurately reflect the mortality we wish to model.**

4. **Random variability, including the randomness in the process generating the mortality, short term variation due to severe winters or hot summers.**

5. **Errors in data (for example age-misstatement).**

In addition, in actuarial applications, especially with small portfolios, the financial risk may be concentrated (for example in a small number of high net worth policyholders).

See S.J. Richards and I.D. Currie (2009) 'Longevity risk and annuity pricing with the Lee-Carter model', *British Actuarial Journal* 15(65), pp. 317-65 for a further discussion of the risks inherent in mortality forecasting.

There is a tendency to focus on uncertainty in parameter estimates. This focus may be misdirected. Stochastic models are good at calculating the level of uncertainty in parameter estimates (for example through bootstrapping) and random variability. But they cannot, of themselves, help with model mis-specification or incorrect judgement.

Bootstrapping is where our existing sample of data (of say n observations) is used as a population from which random smaller samples (of say $m<n$ observations) are repeatedly taken and the parameter value estimated each time. The extent of error observed in these parameter estimates can give a good indication of the uncertainty inherent in parameters estimated from sampling in general.

Forecast errors are often correlated, either across age or time. If errors are positively correlated, they will tend to reinforce one another to widen the prediction interval (the interval within which we believe future mortality to lie). If they are negatively correlated, they will tend to cancel out. Normally, positive correlation is to be expected in mortality forecasting. This is because we believe age patterns of mortality to be smooth, so the mortality rate at any age contains information about the mortality rate at adjacent ages; and because any period-based fluctuations in mortality are likely to affect mortality at a range of ages.

For example, if a large (unforecast) reduction in the mortality rate at age 70 occurs as the result of a new medical treatment, this is likely to affect the mortality rates at neighbouring ages as well. So the forecast error will be more similar at these ages than if the errors were independent (*ie* the errors will tend to be positively correlated).

The chapter summary starts on the next page so that you can keep all the chapter summaries together for revision purposes.

Chapter 12 Summary

Mortality projection

Projections of mortality can be made using two-factor models (age x and time period t, or age x and cohort c), or three-factor models (age, time period and cohort). In three-factor models the factors are linked by $x = t - c$.

It is important to know the advantages and disadvantages of using the various two and three-factor models.

The three projection approaches are based on expectation, extrapolation, and explanation.

Methods based on expectation

These use simple deterministic models (*eg* reduction factors), based on expectations of target future mortality rates based on expert opinion and/or on recent historical trends.

Methods based on extrapolation

Lee-Carter model

Two-factor stochastic model (age and period):

$$\log_e m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t}$$

where:

a_x is the mean value of $\ln\left[m_{x,t}\right]$ averaged over all periods t

k_t is the effect of time t on mortality (with $\sum_t k_t = 0$)

b_x is the extent to which mortality is affected by the time trend at age x (with $\sum_x b_x = 1$)

$b_x k_t$ is the effect of time t on mortality at age x

$\varepsilon_{x,t}$ is the error term (independently and identically distributed with zero mean and common variance).

Time series methods are used to model k_t.

The parameters are estimated by:

$\frac{1}{n}\sum_{t=1}^{n}\ln\left(\hat{m}_{x,t}\right)$ is used to estimate a_x

$\ln\left(\hat{m}_{x,t}\right)-\hat{a}_x$ is used to estimate b_x and k_t

Age-period-cohort version of the Lee-Carter model:

$$\log_e m_{x,t} = a_x + b_x^1 k_t + b_x^2 h_{x-t} + \varepsilon_{x,t}$$

where:

a_x can be fitted (and therefore smoothed) using *p*-splines

h_c is the effect of cohort year c on mortality

b_x^2 is the extent to which mortality is influenced by the cohort effect at age x.

Multi-factor extension of the Lee-Carter model:

$$\log_e m_{x,t} = a_x + \sum_{j=1}^{J} k_{t,j}\, b_j(x) + \varepsilon_{x,t}$$

where:

$k_{t,j}$ is the effect of the time trend on mortality at time t for group j

$b_j(x)$ is the extent to which mortality is influenced by time for group j at age x

It is important to know the advantages and disadvantages of the Lee-Carter model.

Splines

Spline functions can be used for modelling mortality rates by age using:

$$\log_e[E(D_x)] = \log_e E_x^c + \sum_{j=1}^{s}\theta_j B_j(x) \;\Rightarrow\; \ln\left(m_x\right) = \sum_{j=1}^{s}\theta_j B_j(x)$$

For a cubic spline with s knots at ages $x_1, x_2, \ldots x_s$:

$$B_j(x) = \begin{cases} 0 & x < x_j \\ (x-x_j)^3 & x \ge x_j \end{cases}$$

To use splines for modelling mortality rates by time period, rather than age, the $B_j(x)$ would be replaced by $B_j(t)$ with respect to knots at times $t_1, t_2, \ldots t_s$.

p-splines

When estimating the parameters $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_s$ using maximum likelihood techniques, we maximise the penalised log-likelihood:

$$L_p(\theta) = L(\theta) - \frac{1}{2}\lambda P(\theta)$$

where $P(\theta)$ is a roughness penalty that increases with the degree of irregularity in the progression of $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_s$. This is designed to produce a smoother progression of fitted rates with age and/or duration.

Methods based on explanation

Projections are made separately by cause of death and combined.

Possible methods include:

- cause-deleted life table approach
- multiple state (Markov) modelling

Difficulties of the approach include:

- forecasting future changes in the risk factors / disease states
- allowing for the lag between changes in the risk factors and their effect on mortality
- difficulties in identifying and categorising the cause of death

Sources of error in mortality forecasting

The main sources of error are:

- model mis-specification
- parameter uncertainty
- incorrect judgement or prior knowledge
- random variation, including seasonal effects
- data errors.

The practice questions start on the next page so that you can keep the chapter summaries together for revision purposes.

Chapter 12 Practice Questions

12.1 (i) Explain the notation and meaning of the parameters α_x and $f_{n,x}$ in the following reduction factor formula:

$$R_{x,t} = \alpha_x + (1-\alpha_x)\left(1-f_{n,x}\right)^{t/n}$$

(ii) State briefly how the values of these parameters are usually determined.

(iii) The mortality rate for the base year of a mortality projection has been estimated to be:

$$m_{60,0} = 0.006$$

It is believed that the minimum possible mortality rate for lives aged 60 is 0.0012. It is also believed that 30% of the maximum possible reduction in mortality at this age will have occurred by ten years' time.

Using an appropriate reduction factor, calculate the projected mortality rate for lives aged 60 in 20 years' time.

(iv) Describe the advantages and disadvantages of using an expectation-based approach to mortality projections.

12.2 (i) Discuss a major difficulty that is present in a three-factor age-period-cohort mortality projection model that is not found in either an age-period or age-cohort model. [1]

Exam style

(ii) The following Lee-Carter model has been fitted to mortality data covering two age groups (centred on ages 60 and 70), and a 41-year time period from 1990 to 2030 inclusive:

$$\ln\left(m_{x,t}\right) = a_x + b_x k_t + \varepsilon_{x,t}$$

(a) Define in words the symbols a_x, b_x, k_t and $\varepsilon_{x,t}$.

(b) State the constraints that are normally imposed on b_x and k_t in order for the model to be uniquely specified.

(c) In this model k_t has been set to cover a 41-year time period from 1990 to 2030 inclusive, such that for projection (calendar) year t:

$$k_{t+1} = k_t - 0.02 + e_t$$

where e_t is a normally distributed random variable with zero mean and common variance.

Identify the numerical values of k_t ($t = 1990, 1991, \ldots 2029, 2030$), ignoring error terms. *Hint: they need to satisfy the constraint for k_t that you specified in part (b).* [5]

(iii) Mortality has been improving over time for both ages included in the model in part (ii). You have been given the following further information about the model:

$$b_{60} = 3b_{70}$$

$$\hat{m}_{60,2010} = 0.00176$$

$$\hat{m}_{70,2010} = 0.01328$$

where is the predicted mortality rate at age in calendar year calculated from the fitted model (*ie* ignoring error terms).

(a) State what the above information indicates about the impact of the time trend on mortality at the two ages.

(b) Use the above information to complete the specification of the model.

(c) Use the model to calculate the projected values of and . [6]

(iv) Describe the main disadvantages of the Lee-Carter model. [3]

[Total 15]

12.3 You have fitted a model to mortality data that are subdivided by age and time period , with a view to using the model to project future mortality rates. For a particular age , the model is defined as:

$$\ln\left[E(D_{x,t})\right] = \ln E^c_{x,t} + a + bt + ct^2$$

where $D_{x,t}$ is the random number of deaths, and $E^c_{x,t}$ is the central exposed to risk for age group x in time period t ($t=0$ is the year 1975).

(i) If $m_{x,t}$ is the central rate of mortality for exact age x in time period t, show that the above model is equivalent to:

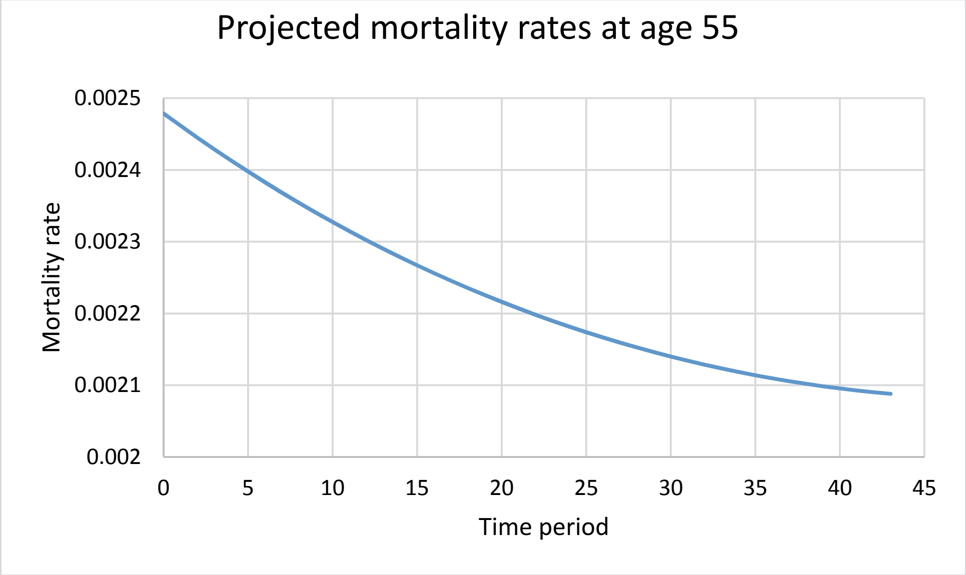
$$m_{x,t} = A\,B^t C^{t^2}$$

stating the values of the parameters A, B and C.

(ii) The model had been fitted to existing data covering the years 1975 to 2017 inclusive. At age 55 the maximum likelihood estimates of the parameters are:

$$\hat{a} = -6, \quad \hat{b} = -0.007, \quad \hat{c} = 0.00007$$

and a plot of the predicted values of $m_{55,t}$ is shown in the graph below:



A colleague has commented that this model is not an adequate fit to the observed data and suggests replacing the quadratic function with a cubic spline function, again fitting a different function for each age.

(a) Set out the revised mortality projection model that uses a cubic spline function as suggested by your colleague, defining all the symbols used.

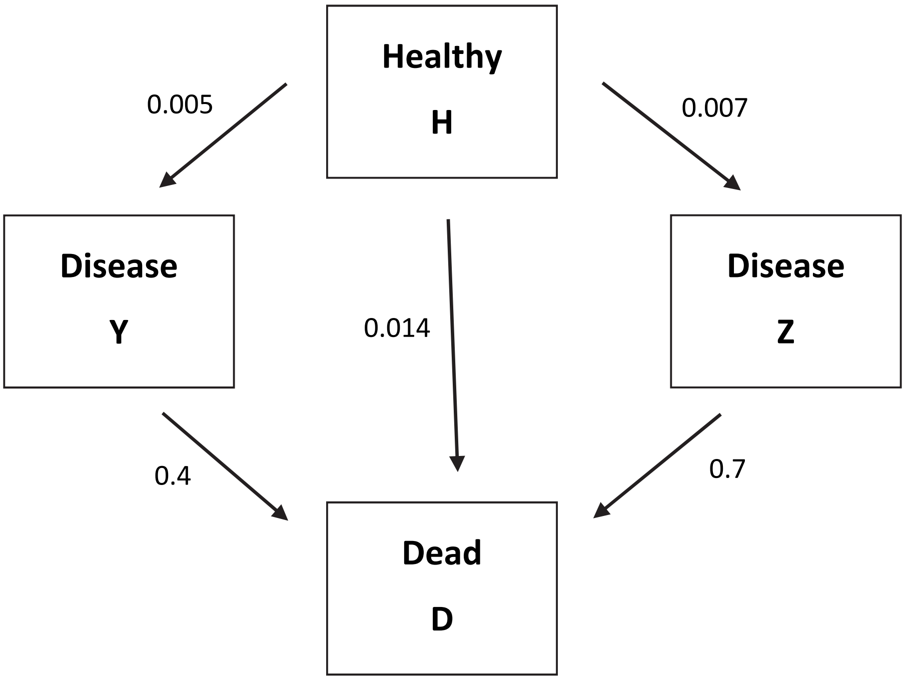
(b) Give a possible reason for the inadequate fit of the original model and explain how the use of the cubic spline function could improve the model as suggested.

(c) A second colleague has challenged the use of cubic splines for this purpose, arguing that the resulting fitted model tends to be too 'rough'.

Explain what is meant by 'rough' in this context, and describe how the method of *p*-splines could be used to help address this difficulty.

(iii) Describe the disadvantages of using the *p*-spline approach.

12.4 In a particular country, Y and Z are important terminal diseases that are significant causes of death for men at older ages. The following represents a Markov jump model of the process, for male lives aged 70, showing annual constant transition rates:



(i) Calculate the probability that a healthy male life aged exactly 70 is dead by the end of the coming year.

(ii) An early diagnosis of Disease Z can prevent the disease from entering the terminal phase and can lead to a full recovery.

A national screening programme has been planned that will increase the rates of early diagnosis of Disease Z, and this is expected to reduce the rate of contracting the terminal phase of the illness by 70% of the current rate (*ie* the transition rate from H to Z in the above Markov model should reduce by 70%). All other transition rates are expected to remain the same as before.

Calculate the revised probability of dying over the year, and hence the percentage reduction in the overall probability of mortality achieved.

(iii) Without performing any more calculations, explain whether a similar screening programme for Disease Y (which would reduce the transition rate from H to Y by 70%) would result in a greater or lower percentage reduction in the overall 1-year probability of mortality.

Chapter 12 Solutions

12.1 (i) ***Interpretation of the reduction factor parameters***

α_x is the lowest level, expressed as a proportion of the current mortality rate at age x, to which the mortality rate at age x can reduce at any time in the future.

$f_{n,x}$ is the proportion of the maximum possible reduction (of $(1-\alpha_x)$) that is expected to have occurred by n years' time.

(ii) ***How the parameters are determined***

Both parameters could be set by expert opinion, perhaps assisted by some analysis of relevant recent observed mortality trends.

(iii) ***Projected mortality rate at age 60 in 20 years' time***

We can first calculate α_{60} as:

$$\alpha_{60} = \frac{0.0012}{0.006} = 0.2$$

We are also given that $f_{10,60} = 0.3$, so we need:

$$R_{60,20} = \alpha_{60} + (1-\alpha_{60})(1-f_{10,60})^{20/10} = 0.2 + 0.8 \times (1-0.3)^2 = 0.592$$

Hence the projected mortality rate at age 60 in 20 years' time is:

$$m_{60,20} = m_{60,0}\, R_{60,20} = 0.006 \times 0.592 = 0.003552$$

(iv) ***Advantages and disadvantages of using an expectation approach***

Advantages

- The method is easy to implement.

Disadvantages

- The effect of such factors as lifestyle changes and prevention of hitherto major causes of death are difficult to predict, as they have not occurred before, and experts may fail to judge the extent of the impact of these on future mortality adequately.
- Because the parameters are themselves target forecasts, there is a circularity in the theoretical basis of the projection model (because forecasts are being used to construct a model whose purpose should be to produce those forecasts).
- Setting the target levels leads to an under-estimation of the true level of uncertainty around the forecasts.

12.2 (i) ***Difficulty of age-period-cohort models***

Three-factor models have the logical problem that each factor is linearly dependent on the other two. So we need to ensure that the three arguments of the function work together in a consistent way in the formulae. [1]

(ii)(a) ***Definitions***

In the Lee-Carter model:

- a_x is the mean value of $\ln\left(m_{x,t}\right)$ averaged over all periods t [½]
- is the effect of time on mortality [½]
- b_x is the extent to which mortality is affected by the time trend at age x [½]
- $\varepsilon_{x,t}$ is the error term (independently and identically distributed with zero mean and common variance). [½]

(b) ***Constraints***

The constraints are:

- $\sum\limits_t k_t = 0$ [½]
- $\sum\limits_x b_x = 1$ [½]

(c) ***Numerical values of k_t***

k_t is a linear function of calendar year t, whose values must sum to zero over the 41-year time period. So the function needs to pass through zero when t takes its central value (2010). Hence:

$$k_t = -0.02 \times (t - 2{,}010)$$

which gives:

$$k_t = 0.4, 0.38, ... -0.38, -0.4 \text{ for } t = 1990, 1991, ... 2029, 2030 \text{ respectively.}$$ [2]

(iii)(a) ***Effect of time trend at different ages***

Mortality rates at age 60 are assumed to be improving at three times the rate at which they are improving at age 70. [1]

(b) ***Complete the specification of the model***

We need values of a_x and b_x at both ages.

As mortality rates are improving at both ages, the values of b_{60} and b_{70} are both positive. Using $b_{60} + b_{70} = 1$ we have:

$$3b_{70} + b_{70} = 1 \Rightarrow b_{70} = 0.25, b_{60} = 0.75$$ [1]

a_x is the value of $\ln\left(\hat{m}_{x,t}\right)$ when $k_t = 0$, which is when $t = 2010$. [½]

So:

$$a_{60} = \ln\left(\hat{m}_{60,2010}\right) = \ln(0.00176) = -6.34244$$ [½]

$$a_{70} = \ln\left(\hat{m}_{70,2010}\right) = \ln(0.01328) = -4.32150$$ [½]

(c) ***Projected values***

We need:

$$k_{2025} = -0.02 \times (2,025 - 2,010) = -0.3$$ [½]

At age 60, we have:

$$\ln\left(\hat{m}_{60,2025}\right) = a_{60} + b_{60}\, k_{2025} = -6.34244 + 0.75 \times (-0.3) = -6.56744$$

$$\Rightarrow \quad \hat{m}_{60,2025} = e^{-6.56744} = 0.00141$$ [1]

At age 70:

$$\ln\left(\hat{m}_{70,2025}\right) = a_{70} + b_{70}\, k_{2025} = -4.32150 + 0.25 \times (-0.3) = -4.39650$$

$$\Rightarrow \quad \hat{m}_{70,2025} = e^{-4.39650} = 0.01232$$ [1]

(iv) ***Disadvantages of Lee-Carter model***

Future estimates of mortality at different ages are heavily dependent on the original estimates of the parameters a_x and b_x, which are assumed to remain constant into the future. These parameters are estimated from past data, and will incorporate any roughness contained in the data. In particular, they may be distorted by past period events which might affect different ages to different degrees. [1½]

If the estimated b_x values show variability from age to age, it is possible for the forecast age-specific mortality rates to 'cross over' (such that, for example, projected rates may increase with age at one duration, but decrease with age at the next). [½]

There is a tendency for Lee-Carter forecasts to become increasingly rough over time. [½]

The model assumes that the underlying rates of mortality change are constant over time across all ages, when there is empirical evidence that this is not so. [½]

The Lee-Carter model does not include a cohort term, whereas there is evidence from some countries that certain cohorts exhibit higher mortality improvements than others. [½]

Unless observed rates are used for the forecasting, it can produce 'jump-off' effects (*ie* an implausible jump between the most recent observed mortality rate and the forecast for the first future period). [½]

[Maximum 3 for (iv)]

12.3 (i) ***Formula for*** $m_{x,t}$

Rearranging the original model we have:

$$\ln\left[E\left(D_{x,t}\right)\right]-\ln E_{x,t}^c=a+bt+ct^2$$

$$\Rightarrow \ln\left[E\left(\frac{D_{x,t}}{E_{x,t}^c}\right)\right]=a+bt+ct^2$$

$\dfrac{D_{x,t}}{E_{x,t}^c}$ is an unbiased estimator of $m_{x,t}$ so:

$$\ln\left[m_{x,t}\right]=a+bt+ct^2$$

$$\Rightarrow m_{x,t}=\exp\left[a+bt+ct^2\right]$$

$$=e^a e^{bt} e^{ct^2}$$

$$=A\,B^t\,C^{t^2}$$

where:

$$A=e^a$$

$$B=e^b$$

$$C=e^c$$

(ii)(a) ***Revised projection model using cubic spline function***

The mortality projection model would now be:

$$\ln\left[E\left(D_{x,t}\right)\right]=\ln E_{x,t}^c+\sum_{j=1}^{J}\theta_j B_j(t)$$

where there are J knots positioned at values $t_1, t_2, \ldots, t_J$, θ_j are parameters to be fitted from the data, and:

$$B_j(t)=\begin{cases}0 & t<t_j\\ (t-t_j)^3 & t\geq t_j\end{cases}$$

(b) ***Reasons for inadequate fit and how it could be improved by cubic spline function***

The trend in mortality over time is unlikely to follow a quadratic function, even after it has been log-transformed, as in this model, because the progression of predicted values is likely to be too smooth.

There may be significant variations in the trends in the past data that may be relevant to future projections and which we would therefore like the model to take into account.

Spline functions are very flexible models in terms of the shape of the function being fitted.

Adherence to data can be improved both by increasing the number of knots used, and by placing the knots in locations where the greatest changes in curvature of the trend line occur.

However, some smoothing is still a requirement, and using cubic splines generally produces the smoothest result (compared to using splines of higher orders).

(c) *Use of p-splines*

The problem with splines is that they can be *too* flexible, and may cause the model to include historical trend variations that are either short-term or past-specific, and which are not expected to recur in future.

To include these features in the model may then be inappropriate or unhelpful when we attempt to use the model for forecasting purposes.

One symptom of this over-adherence, or roughness, in the model, is that the sequence of estimated parameters $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_J$ may form an uneven progression, and smoothing this progression can help reduce the roughness in the predicted values from the model.

The method of *p*-splines attempts to find an optimal model by introducing a penalty for models which have excessive 'roughness'.

The method may be implemented as follows:

- Specify the knot spacing and degree of the polynomials in each spline.
- Define a *roughness penalty*, $P(\theta)$, which increases with the variability of adjacent coefficients. This, in effect, measures the amount of roughness in the fitted model.
- Define a smoothing parameter, λ, such that if $\lambda = 0$, there is no penalty for increased roughness, but as λ increases, roughness is increasingly penalised.
- Estimate the parameters of the model, including the number of splines, by maximising the penalised log-likelihood:

  $$l_p(\theta) = l(\theta) - \frac{1}{2}\lambda P(\theta)$$

  where $l(\theta)$ would be the usual log-likelihood for the model.
- The penalised log-likelihood is effectively trying to balance smoothness and adherence to the data.

(iii) *Disadvantages of using p-splines*

When applied to ages separately, mortality at different ages is forecast independently so there is a danger that there will be roughness between adjacent ages.

There is no explanatory element to the projection (in the way that time series methods use a structure for mortality and an identifiable time series for projection).

p-splines tend to be over-responsive to adding an extra year of data.

12.4 (i) ***Calculating the 1-year probability of dying q_{70}***

A healthy person aged 70 can die over one year by following any one of the following three paths:

(1) transition directly from H to D within one year (HD)

(2) transition from H to Y followed by transition from Y to D in one year (HYD)

(3) transition from H to Z followed by transition from Z to D in one year (HZD)

We need the sum of the probabilities of following each path.

The one-year probabilities are denoted by P_{HD}, P_{HYD} and P_{HZD} respectively.

Healthy directly to dead

We need:

$$P_{HD}=\int_{t=0}^{1} p_{HH}(t)\,\mu_{HD}\,dt$$

which is the probability of someone staying healthy until time t, then dying from healthy at that point, integrated over all possible times t within the one year $(0<t<1)$.

Now:

$$p_{HH}(t)=p_{\overline{HH}}(t)=e^{-(\mu_{HD}+\mu_{HY}+\mu_{HZ})t}=e^{-0.026t}$$

So:

$$P_{HD}=\int_{t=0}^{1} e^{-0.026t}\times 0.014\,dt=0.014\times\left[\frac{e^{-0.026t}}{-0.026}\right]_0^1=\frac{0.014}{0.026}\left(1-e^{-0.026}\right)=0.013820$$

Healthy to dead via Y

Now we need:

$$P_{HYD}=\int_{t=0}^{1} p_{HH}(t)\,\mu_{HY}\,p_{YD}(1-t)\,dt=\mu_{HY}\int_{t=0}^{1} p_{HH}(t)\,p_{YD}(1-t)\,dt$$

This is the probability of staying healthy until time t, contracting disease Y at that point, and then dying (from Disease Y) at some point in the remaining $(1-t)$ of the year.

First we see that:

$$p_{YD}(1-t) = 1 - p_{YY}(1-t)$$

This is because there are only two outcomes over the $(1-t)$ period for lives starting in state Y – if they are not dead by the end of the period, they must still be in state Y. So:

$$p_{YD}(1-t) = 1 - e^{-0.4(1-t)} = 1 - e^{-0.4}e^{0.4t}$$

So:

$$\begin{aligned}
P_{HYD} &= 0.005 \int_{t=0}^{1} e^{-0.026t}\left(1 - e^{-0.4}e^{0.4t}\right)dt \\
&= 0.005\left(\int_{t=0}^{1} e^{-0.026t}dt - e^{-0.4}\int_{t=0}^{1} e^{0.374t}dt\right) \\
&= 0.005\left(\frac{1-e^{-0.026}}{0.026} - e^{-0.4}\left[\frac{e^{0.374}-1}{0.374}\right]\right) \\
&= 0.005 \times (0.987112 - 0.812874) \\
&= 0.005 \times 0.174237 \\
&= 0.000871
\end{aligned}$$

Healthy to dead via Z

Similarly:

$$P_{HZD} = \int_{t=0}^{1} p_{HH}(t)\,\mu_{HZ}\,p_{ZD}(1-t)\,dt = \mu_{HD}\int_{t=0}^{1} p_{HH}(t)\,p_{ZD}(1-t)\,dt$$

where:

$$p_{ZD}(1-t) = 1 - p_{ZZ}(1-t) = 1 - e^{-0.7(1-t)} = 1 - e^{-0.7}e^{0.7t}$$

So:

$$\begin{aligned}
P_{HZD} &= 0.007\int_{t=0}^{1} e^{-0.026t}\left(1 - e^{-0.7}e^{0.7t}\right)dt = 0.007\left(\int_{t=0}^{1} e^{-0.026t}dt - e^{-0.7}\int_{t=0}^{1} e^{0.674t}dt\right) \\
&= 0.007\left(0.987112 - e^{-0.7}\left[\frac{e^{0.674}-1}{0.674}\right]\right) = 0.007 \times 0.278284 = 0.001948
\end{aligned}$$

So the total probability of dying is:

$$q_{70} = P_{HD} + P_{HYD} + P_{HZD} = 0.013820 + 0.000871 + 0.001948 = 0.016639$$

(ii) ***Effect of the screening programme for disease Z***

The transition rate from H to Z has reduced to:

$$\mu'_{HZ} = 0.3 \times 0.007 = 0.0021$$

The individual probabilities change as follows.

We need:

$$P'_{HD} = \int_{t=0}^{1} p'_{HH}(t)\,\mu_{HD}\,dt$$

where:

$$p'_{HH}(t) = p'_{\overline{HH}}(t) = e^{-(\mu_{HD}+\mu_{HY}+\mu'_{HZ})t} = e^{-(0.005+0.014+0.0021)t} = e^{-0.0211t}$$

So:

$$P'_{HD} = \int_{t=0}^{1} e^{-0.0211t} \times 0.014\,dt = 0.014 \times \left(\frac{1-e^{-0.0211}}{0.0211}\right) = 0.014 \times 0.989524 = 0.013853$$

$$P'_{HYD} = \int_{t=0}^{1} p'_{HH}(t)\,\mu_{HY}\,p_{YD}(1-t)\,dt = \int_{t=0}^{1} e^{-0.0211t} \times 0.005 \times \left(1-e^{-0.4}e^{0.4t}\right)dt$$

$$= 0.005\left(\int_{t=0}^{1} e^{-0.0211t}\,dt - e^{-0.4}\int_{t=0}^{1} e^{0.3789t}\,dt\right)$$

$$= 0.005\left(\frac{1-e^{-0.0211}}{0.0211} - e^{-0.4}\left[\frac{e^{0.3789}-1}{0.3789}\right]\right)$$

$$= 0.005 \times \left(0.989524 - 0.814993\right) = 0.000873$$

$$P'_{HZD} = \int_{t=0}^{1} p'_{HH}(t)\,\mu'_{HZ}\,p_{ZD}(1-t)\,dt = \int_{t=0}^{1} e^{-0.0211t} \times 0.0021 \times \left(1-e^{-0.7}e^{0.7t}\right)dt$$

$$= 0.0021\left(\int_{t=0}^{1} e^{-0.0211t}\,dt - e^{-0.7}\int_{t=0}^{1} e^{0.6789t}\,dt\right)$$

$$= 0.0021\left(\frac{1-e^{-0.0211}}{0.0211} - e^{-0.7}\left[\frac{e^{0.6789}-1}{0.6789}\right]\right)$$

$$= 0.0021 \times \left(0.989524 - 0.710761\right) = 0.000585$$

So the revised total probability of dying is:

$$q_{70} = P'_{HD} + P'_{HYD} + P'_{HZD} = 0.013853 + 0.000873 + 0.000585 = 0.015311$$

which is 0.001328 lower than the previous value of 0.016639. This is a reduction of 8.0%.

(iii) ***Effect of the screening programme for Disease Y***

The reduction in mortality rate would be less, for two reasons:

(1) People with Disease Y live for longer on average than those with disease Z.

(Recall that the expected survival time in state s is:

$$E\left[T_s\right] = \frac{1}{\lambda_s}$$

So, with $\lambda_Y < \lambda_Z$, we must have $E\left[T_Y\right] > E\left[T_Z\right]$.)

So, cutting the number of people contracting Disease Y will have a proportionately lower impact on the total number dying during the year compared to Disease Z (*ie* Z is a more serious disease than Y, so reducing the incidence of Z should have the bigger impact on mortality rates).

(2) The transition rate from H to Y is lower than from H to Z. So reducing this rate to 30% of its current level will cause a smaller reduction in the number of people contracting Disease Y over the year. So, even if the mortality rates for the two diseases were the same, the impact on the number of people dying would be less (*ie* Disease Z is commoner than Disease Y, so there are fewer deaths from Disease Y that can be prevented).

End of Part 3

What next?

1. Briefly **review** the key areas of Part 3 and/or re-read the **summaries** at the end of Chapters 8 to 12.

2. Ensure you have attempted some of the **Practice Questions** at the end of each chapter in Part 3. If you don't have time to do them all, you could save the remainder for use as part of your revision.

3. Attempt **Assignment X3**.

13

Time series 1

Syllabus objectives

2.1 Concepts underlying time series models

2.1.1 Explain the concept and general properties of stationary, $I(0)$, and integrated, $I(1)$, univariate time series.

2.1.2 Explain the concept of a stationary random series.

2.1.4 Know the notation for backwards shift operator, backwards difference operator, and the concept of roots of the characteristic equation of time series.

2.1.5 Explain the concepts and basic properties of autoregressive (AR), moving average (MA), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) time series.

2.1.6 Explain the concepts and properties of discrete random walks and random walks with normally distributed increments, both with and without drift.

2.1.9 Show that certain univariate time series models have the Markov property and describe how to rearrange a univariate time series model as a multivariate Markov model.

2.2 Applications of time series models

2.2.3 Describe simple applications of a time series model, including random walk, autoregressive and cointegrated models as applied to security prices and other economic variables.

0 Introduction

The time series material for Subject CS2 has been split into two parts.

This chapter gives the basic definitions and simple examples which we will outline in more detail below.

The sections correspond roughly to the syllabus objectives but not necessarily in the same order.

Firstly, the idea of a time series is introduced. This is just a sequence of observations that we record at regular time intervals. Financial time series would therefore include the closing price of the FTSE 100 index on successive days, the retail prices index in successive months and so on.

This chapter is mainly concerned with a class of processes that is commonly used to model time series data, the so-called *ARIMA* class.

It also deals with some of the necessary background theory and definitions that are required. In particular, stationary processes play a major role. Indeed, it turns out that we can only model time series data efficiently if that data is a realisation of a stationary process. We therefore spend some time discussing the stationarity of various ARIMA processes, and if they are not stationary, how they can be transformed into stationary ones for the process of modelling. The only technique looked at here for transforming a non-stationary process into a stationary one will be differencing the data.

The next chapter concentrates on some of the practical issues surrounding time series analysis:

- is the data set a realisation of a stationary process, and if not, how can the data be transformed?
- fitting a model to the (transformed) data
- forecasting future values.

Chapter 14 also introduces multivariate time series.

1 Properties of a univariate time series

A *univariate time series* is a sequence of observations of a single process taken at a sequence of different times. Such a series can in general be written as:

$$x(t_1), x(t_2), \dots, x(t_n)$$

***ie* as:**

$$\{x(t_i) : i = 1,2,3,\dots,n\}$$

Most applications involve observations taken at equally-spaced times. In this case the series is written as:

$$x_1, x_2, \dots, x_n$$

***ie* as:**

$$\{x_t : t = 1,2,3,\dots,n\}$$

For instance, a sequence of daily closing prices of a given share constitutes a time series, as does a sequence of monthly inflation figures.

The fact that the observations occur in time order is of prime importance in any attempt to describe, analyse and model time series data. The observations are related to one another and cannot be regarded as observations of independent random variables. It is this very dependence amongst the members of the underlying sequence of variables which any analysis must recognise and exploit.

For example, a list of returns of the stocks in the FTSE 100 index on a particular day is not a time series, and the order of records in the list is irrelevant. At the same time, a list of values of the FTSE 100 index taken at one-minute intervals on a particular day is a time series, and the order of records in the list is of paramount importance.

Note that the observations x_t can arise in different situations. For example:

- **the time scale may be inherently discrete (as in the case of a series of 'closing' share prices)**
- **the series may arise as a sample from a series observable continuously through time (as in the case of hourly readings of atmospheric temperature)**
- **each observation may represent the results of aggregating a quantity over a period of time (as in the case of a company's total premium income on new business each month).**

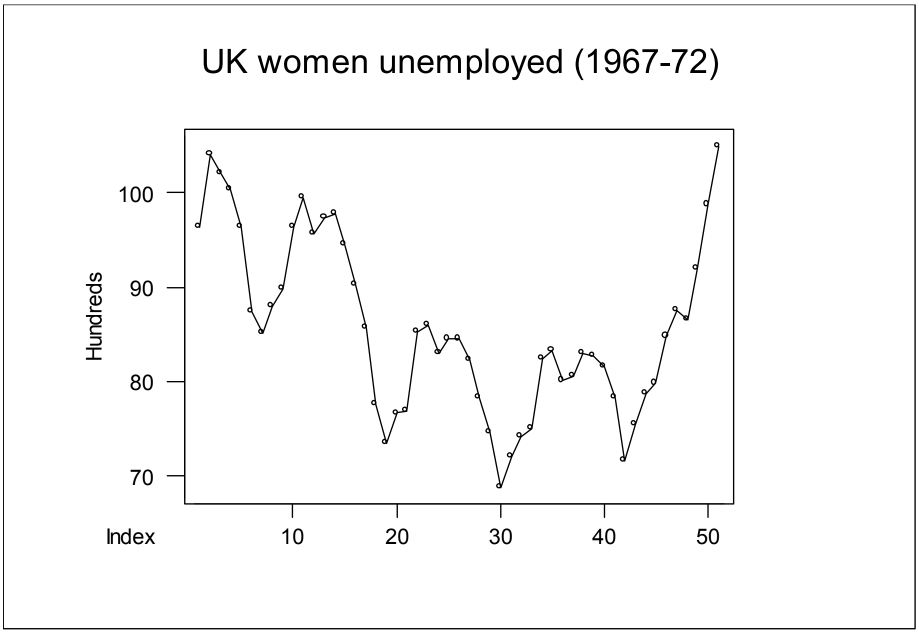


Figure 13.1: a time series

The purposes of a practical time series analysis may be summarised as:

- **description of the data**
- **construction of a model which fits the data**
- **forecasting future values of the process**
- **deciding whether the process is out of control, requiring action**
- **for vector time series, investigating connections between two or more observed processes with the aim of using values of some of the processes to predict those of the others.**

These five key aims will be discussed in more detail throughout this chapter.

A univariate time series is modelled as a realisation of a sequence of random variables:

$$\{X_t : t = 1,2,3,\ldots,n\}$$

called a *time series process*. (Note, however, that in the modern literature the term 'time series' is often used to mean both the data and the process of which it is a realisation.) A time series process is a stochastic process indexed in discrete time with a continuous state space.

It is important to appreciate the difference between x_t, which is just a number, and X_t, which is a random variable. The latter will be used to model the former; the procedure for finding a suitable model is the second of our key objectives given above.

The sequence $\{X_t : t = 1,2,3,\ldots,n\}$ may be regarded as a sub-sequence of a doubly infinite collection $\{X_t : t = \ldots,-2,-1,0,1,2,\ldots\}$. This interpretation will be found to be helpful in investigating notions such as convergence to equilibrium.

We will often use the shorthand notation X instead of $\{X_t : t = \ldots, -2, -1, 0, 1, 2, \ldots\}$ or $\left\{X_t : t = 1, 2, 3, \ldots, n\right\}$.

The phrase, 'convergence to equilibrium' may require some explanation. We will see shortly that a *stationary* process is basically in a (statistical) equilibrium, *ie* the statistical properties of the process remain unchanged as time passes. If a process is currently non-stationary, then it is a natural question to ask whether or not that process will ever settle down and reach (converge to) equilibrium. In this case we can think about what will happen as *t* gets very large.

Alternatively, we might think of the process as having started some time ago in the past, perhaps indexed by negative *t*, so that it has already had time to settle down.

This will be made clearer later, when stationarity is discussed in more detail.

2 Stationary random series

2.1 Stationary time series processes

The concept of stationarity was introduced in Chapter 1, along with the ideas of strict and weak stationarity.

Question

Explain what it means for a process X to be:

(i) strictly stationary

(ii) weakly stationary.

Solution

(i) A process X is strictly stationary if the joint distributions of $X_{t_1}, X_{t_2}, \dots, X_{t_n}$ and $X_{k+t_1}, X_{k+t_2}, \dots, X_{k+t_n}$ are identical for all $t_1, t_2, \dots, t_n$ and $k+t_1, k+t_2, \dots, k+t_n$ in J and all integers n. This means that the statistical properties of the process remain unchanged as time elapses.

(ii) A process X is weakly stationary if $E(X_t)$ is independent of t and $\text{cov}(X_t, X_{t-s})$ depends only on the lag, s.

A weakly stationary process has constant variance since, for such a process, $\text{var}(X_t) = \text{cov}(X_t, X_t)$ is independent of t.

In the study of time series it is a convention that the word 'stationary' on its own is a shorthand notation for 'weakly stationary', though in the case of a multivariate normal process, strict and weak stationarity are equivalent.

This is because a distribution of a multivariate normal random variable is completely determined by its mean vector and covariance matrix. We will consider this further in Chapter 14.

But we do need to be careful in our definition, as there are some processes which we wish to exclude from consideration but which satisfy the definition of weak stationarity.

Purely indeterministic processes

A process X is called *purely indeterministic* if knowledge of the values of $X_1, \dots, X_n$ is progressively less useful at predicting the value of X_N as $N \to \infty$. When we talk of a 'stationary time series process' we shall mean a weakly stationary purely indeterministic process.

Question

Let Y_t be a sequence of independent standard normal random variables. Determine which of the following processes are stationary time series (given the definition above).

(i) $X_t = \sin(\omega t + U)$, where U is uniformly distributed on the interval $[0, 2\pi]$

(ii) $X_t = \sin(\omega t + Y_t)$

(iii) $X_t = X_{t-1} + Y_t$

(iv) $X_t = Y_{t-1} + Y_t$

(v) $X_t = 2 + 3t + 0.5X_{t-1} + Y_t + 0.3Y_{t-1}$

Solution

(i) For this process, $X_0 = \sin U$, $X_1 = \sin(\omega + U)$, $X_2 = \sin(2\omega + U)$, …. Given the value of X_0, future values of the process are fully determined. So this is not purely indeterministic, and is not therefore a stationary time series in the sense defined in the Core Reading.

(ii) For this process, $X_0 = \sin Y_0$, $X_1 = \sin(\omega + Y_1)$, $X_2 = \sin(2\omega + Y_2)$, …. Since $E(X_t)$ varies over time, this process is not stationary.

(iii) Here we have:

$$E(X_t) = E(X_{t-1} + Y_t) = E(X_{t-1}) + E(Y_t) = E(X_{t-1})$$

So the process has a constant mean. However:

$$\text{var}(X_t) = \text{var}(X_{t-1} + Y_t) = \text{var}(X_{t-1}) + \text{var}(Y_t) = \text{var}(X_{t-1}) + 1$$

Here we are using the fact that Y_t is a sequence of independent standard normal random variables. Since the variance is not constant, the process is not stationary.

(iv) This process is weakly stationary:

$$E(X_t) = E(Y_{t-1}) + E(Y_t) = 0$$

and:

$$\text{cov}(X_t, X_{t+k}) = \begin{cases} 2 & k = 0 \\ 1 & |k| = 1 \\ 0 & |k| \geq 2 \end{cases}$$

For example:

$$\begin{aligned}\text{cov}(X_t, X_t) &= \text{cov}(Y_{t-1} + Y_t, Y_{t-1} + Y_t) \\ &= \text{cov}(Y_{t-1}, Y_{t-1}) + 2\,\text{cov}(Y_t, Y_{t-1}) + \text{cov}(Y_t, Y_t) \\ &= \text{var}(Y_{t-1}) + \text{var}(Y_t) = 2\end{aligned}$$

In addition, it is purely indeterministic. From the defining equation, we have:

$$X_1 = Y_0 + Y_1$$

So $Y_1 = X_1 - Y_0$ and:

$$X_2 = Y_1 + Y_2 = (X_1 - Y_0) + Y_2$$

Rearranging gives:

$$Y_2 = X_2 - X_1 + Y_0$$

and hence:

$$X_3 = Y_2 + Y_3 = (X_2 - X_1 + Y_0) + Y_3$$

Continuing in this way, we see that:

$$X_n = X_{n-1} - X_{n-2} + X_{n-3} + \cdots + (-1)^{n-2} X_1 + (-1)^{n-1} Y_0 + Y_n$$

From this formula, we see that knowledge of the values of X_1, X_2 and X_3, say, becomes progressively less useful in predicting the value of X_n as $n \to \infty$.

(v) This process has a deterministic trend via the '3*t*' term, *ie* its mean varies over time. So it is not stationary.

A particular form of notation is used for time series: X is said to be $I(0)$ (read 'integrated of order 0') if it is a stationary time series process, X is $I(1)$ if X itself is not stationary but the increments $Y_t = X_t - X_{t-1}$ form a stationary process, X is $I(2)$ if it is non-stationary but the process Y is $I(1)$, and so on.

We will see plenty of examples of integrated processes when we study the ARIMA class of processes in Section 3.8.

The theory of stationary random processes plays an important role in the theory of time series because the calibration of time series models (that is, estimation of the values of the model's parameters using historical data) can be performed efficiently only in the case of stationary random processes. A non-stationary random process has to be transformed into a stationary one before the calibration can be performed. (See Chapter 14.)

Question

Suppose that we have a sample set of data that looks to be a realisation of an integrated process of order 2. Explain what can we do to the data set in order to model it.

Solution

We can difference the data twice, *ie* look at the increments of the increments.

2.2 Autocovariance function

The mean function (or trend) of the process is $\mu_t = E[X_t]$, the covariance function $\text{cov}(X_s, X_t) = E[X_s X_t] - E[X_s]E[X_t]$. Both of these functions take a simpler form in the case where X is stationary:

- **The mean of a stationary time series process is constant, *ie* $\mu_t \equiv \mu$ for all *t*.**
- **The covariance of any pair of elements X_r and X_s of a stationary sequence X depends only on the difference *r* – *s*.**

Autocovariance function

We can therefore define the autocovariance function $\{\gamma_k : k \in Z\}$ of a stationary time series process *X* as follows:

$$\gamma_k \equiv \text{cov}\left(X_t, X_{t+k}\right) = E\left[X_t X_{t+k}\right] - E\left[X_t\right]E\left[X_{t+k}\right]$$

The common variance of the elements of a stationary process is given by:

$$\gamma_0 = \text{var}(X_t)$$

If a process is not stationary, then the autocovariance function depends on two variables, namely the time *t* and the lag *k*. This could be denoted, for example, $\gamma(t,k) = \text{cov}\left(X_t, X_{t+k}\right)$. However, one of the main uses of the autocovariance function is to determine the type of process that will be used to model a given set of data. Since this will be done only for stationary series, as mentioned above, it is the autocovariance function for stationary series that is most important.

Because of the importance of the autocovariance function, we will have to calculate it for various processes. This naturally involves calculating covariances and so we need to be familiar with all of the properties of the covariance of two random variables. The following question is included as a revision exercise.

Question

Let X and Y denote any random variables.

(i) Express $\text{cov}(X,Y)$ in terms of $E(X)$, $E(Y)$ and $E(XY)$.

(ii) Express each of the following in terms of $\text{cov}(X,Y)$:

(a) $\text{cov}(Y,X)$

(b) $\text{cov}(X,c)$, where c is a constant

(c) $\text{cov}(2X,3Y)$

(iii) Give an equivalent expression for $\text{cov}(X,X)$.

(iv) Prove, using your formula in (i), that:

$$\text{cov}(X+Y,W)=\text{cov}(X,W)+\text{cov}(Y,W)$$

(v) Simplify each of the following expressions assuming that $\{X_t\}$ denotes a stationary time series defined at integer times and $\{Z_t\}$ are independent $N(0,\sigma^2)$ random variables.

(a) $\text{cov}(Z_2,Z_3)$

(b) $\text{cov}(Z_3,Z_3)$

(c) $\text{cov}(X_2,Z_3)$

(d) $\text{cov}(X_2,X_3)$

(e) $\text{cov}(X_2,X_2)$

Solution

(i) $\text{cov}(X,Y)=E(XY)-E(X)E(Y)$

(ii) (a) $\text{cov}(Y,X)=\text{cov}(X,Y)$

(b) $\text{cov}(X,c)=0$

(c) $\text{cov}(2X,3Y)=6\,\text{cov}(X,Y)$

(iii) $\text{cov}(X,X)=E(X^2)-[E(X)]^2=\text{var}(X)$

(iv)
$$\begin{aligned}\text{cov}(X+Y,W)&=E[(X+Y)W]-E(X+Y)E(W)\\&=E(XW+YW)-[E(X)+E(Y)]E(W)\\&=E(XW)+E(YW)-E(X)E(W)-E(Y)E(W)\\&=\text{cov}(X,W)+\text{cov}(Y,W)\end{aligned}$$

(v) (a) $\text{cov}(Z_2, Z_3) = 0$, since they are independent.

(b) $\text{cov}(Z_3, Z_3) = \text{var}(Z_3) = \sigma^2$

(c) $\text{cov}(X_2, Z_3) = 0$

(d) and (e) will depend on the actual process. If it is stationary, then $\text{cov}(X_2, X_3) = \gamma_1$, and $\text{cov}(X_2, X_2) = \gamma_0$.

2.3 Autocorrelation function

The autocovariance function is measured in squared units, so that the values obtained depend on the absolute size of the measurements. We can make this quantity independent of the absolute sizes of X_n by defining a dimensionless quantity, the *autocorrelation function*.

Autocorrelation function

The *autocorrelation function* (ACF) of a stationary process is defined by:

$$\boldsymbol{\rho_k = \text{corr}(X_t, X_{t+k}) = \frac{\gamma_k}{\gamma_0}}$$

The ACF of a purely indeterministic process satisfies $\boldsymbol{\rho_k \to 0}$ as $\boldsymbol{k \to \infty}$.

This statement is intuitive. We do not expect two values of a (purely indeterministic) time series to be correlated if they are a long way apart.

Question

Write down the formula for the correlation coefficient between the random variables X and Y.

Hence deduce the formula for the autocorrelation function given above.

Solution

The formula for the correlation coefficient is:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

So:

$$\rho_k = \frac{\text{cov}(X_t, X_{t+k})}{\sqrt{\text{var}(X_t)\text{var}(X_{t+k})}} = \frac{\gamma_k}{\sqrt{\gamma_0 \gamma_0}} = \frac{\gamma_k}{\gamma_0}$$

For a non-stationary process we could define an autocorrelation function by:

$$\rho(t,k) = \frac{\text{cov}(X_t, X_{t+k})}{\sqrt{\text{var}(X_t)}\sqrt{\text{var}(X_{t+k})}} = \frac{\gamma(t,k)}{\sqrt{\gamma(t,0)}\sqrt{\gamma(t+k,0)}}$$

However, as with the autocovariance function, it is the stationary case that is of most use in practice.

A simple class of weakly stationary random processes is the white noise processes. A random process $\{e_t : t \in Z\}$ is a *white noise* if $E[e_t] = 0$ for any t, and:

$$\gamma_k = \text{cov}(e_t, e_{t+k}) = \begin{cases} \sigma^2 & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases}$$

An important representative of the white noise processes is a sequence of independent normal random variables with common mean 0 and variance σ^2.

Strictly speaking a white noise process only has to be a sequence of *uncorrelated* random variables, *ie* not necessarily a sequence of *independent* random variables. We can also have white noise processes without zero mean.

Result 13.1

The autocovariance function γ and autocorrelation function ρ of a stationary random process are even functions of k, that is, $\gamma_k = \gamma_{-k}$ and $\rho_k = \rho_{-k}$.

Proof

Since the autocovariance function $\gamma_k = \text{cov}(X_t, X_{t+k})$ does not depend on *t*, we have:

$$\gamma_k = \text{cov}(X_{t-k}, X_{t-k+k}) = \text{cov}(X_{t-k}, X_t) = \text{cov}(X_t, X_{t-k}) = \gamma_{-k}$$

Thus γ is an even function, which in turn implies that ρ is even.

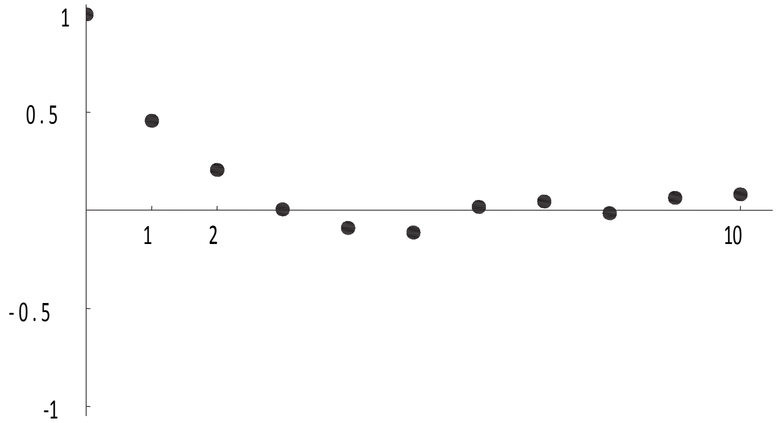
This result allows us to concentrate on positive lags when finding the autocorrelation functions of stationary processes.

2.4 Correlograms

Autocorrelation functions are the most commonly used statistic in time series analysis. A lot of information about a time series can be deduced from a plot of the sample autocorrelation function (as a function of the lag). Such a plot is called a *correlogram.*

Typical stationary series

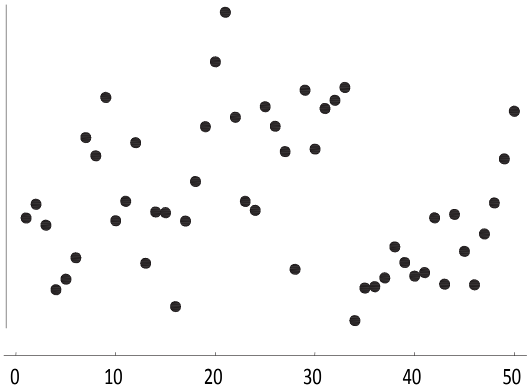
A typical sample autocorrelation function for a stationary series looks like the one shown below. The lag is shown on the horizontal axis, and the autocorrelation on the vertical.



At lag 0 the autocorrelation function takes the value 1, since $\rho_0 = \frac{\gamma_0}{\gamma_0} = 1$. Often the function starts out at 1 but decays fairly quickly, which is indicative of the time series being stationary. The above correlation function tells us that at lags 0, 1 and 2 there is some positive correlation so that a value on one side of the mean will tend to have a couple of values following that are on the same side of the mean. However, beyond lag 2 there is little correlation.

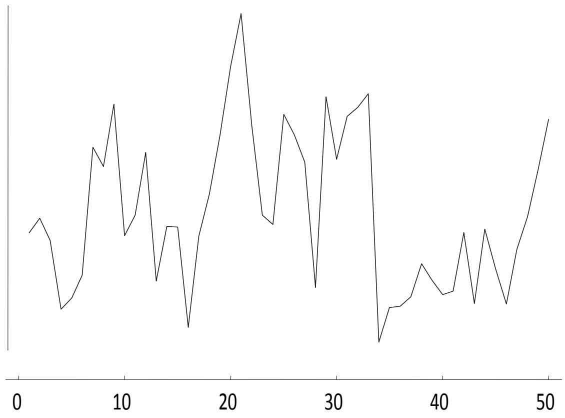
In fact, the above function comes from a sample path of a stationary $AR(1)$ process, namely $X_n = 0.5X_{n-1} + e_n$. (We look in more detail at such processes in the next section.)

The data used for the first 50 values is plotted below. (The actual data used to produce the autocorrelation function used the first 1,000 values.)



The 'gap' in the axes here is deliberate; the vertical axis does not start at zero. The horizontal axis on this and the next graph shows time, and the vertical axis shows the value of the time series X.

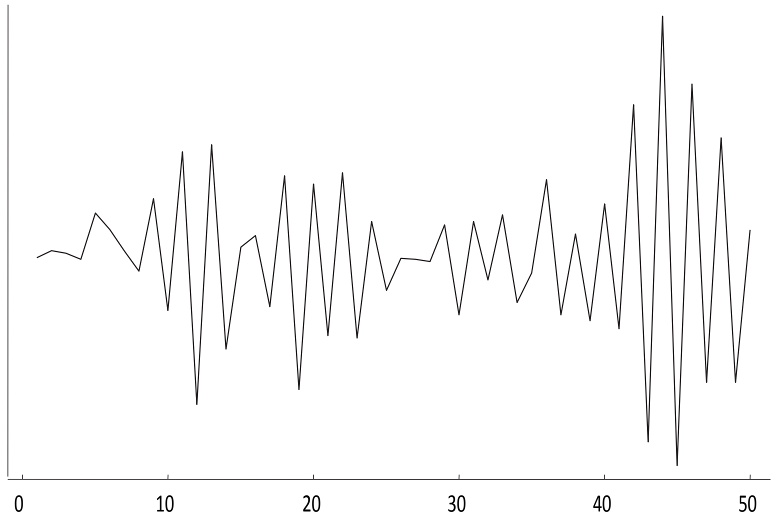
This form of presentation is difficult to interpret. It's easier to see if we 'join the dots'.



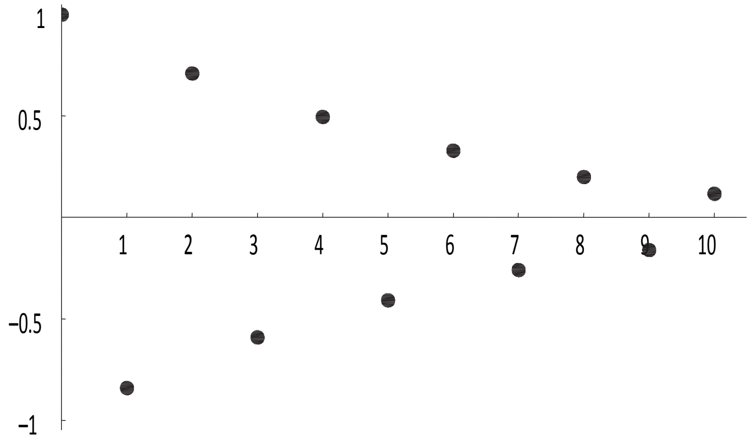
By inspection of this graph we can indeed see that one value tends to be followed by another similar value. This is also true at lag 2, though slightly less clear. Once the lag is 3 or more, there is little correlation.

The previous data set is in stark contrast to the following one.

Alternating series



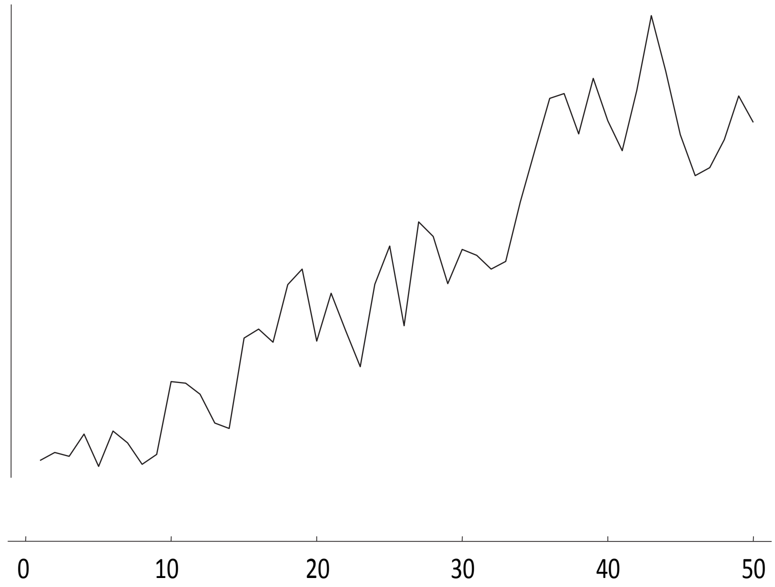
The average of this data is obviously roughly in the middle of the extreme values. Given a particular value, the following one tends to be on the other side of the mean. The series is *alternating*. This is reflected in the autocorrelation function shown below. At lag 1 there is a negative correlation. Conversely, at lag 2, the two points will generally be on the same side of the mean and therefore will have positive correlation, and so on. The autocorrelation therefore also alternates as shown.



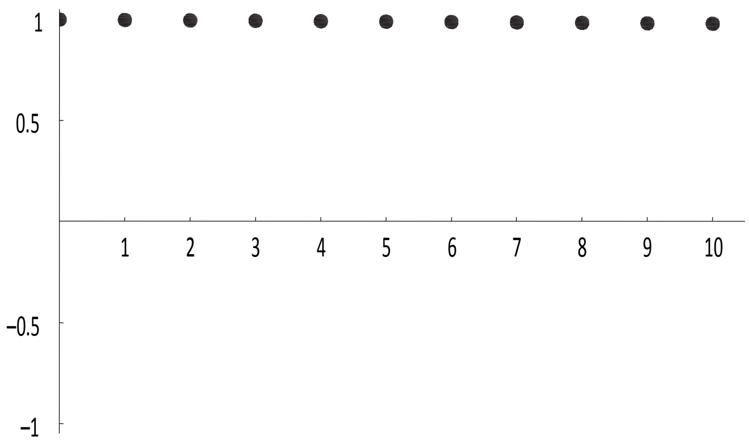
The data in this case actually came from a stationary autoregressive process, this time $X_n = -0.85X_{n-1} + e_n$. This is stationary, but because the coefficient of X_{n-1} is larger in magnitude, *ie* 0.85 *vs* 0.5, the decay of the autocorrelation function is slower. This is because the X_{n-1} term is not swamped by the random factors e_n as quickly. It is the fact that the coefficient is negative that makes the series alternate.

Series with a trend

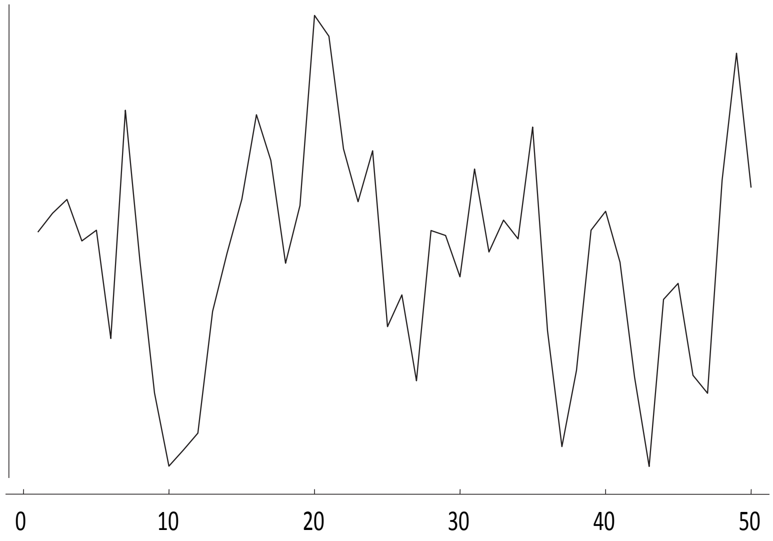
A final example comes from the following data generated from $X_n = 0.1n + 0.5X_{n-1} + e_n$.



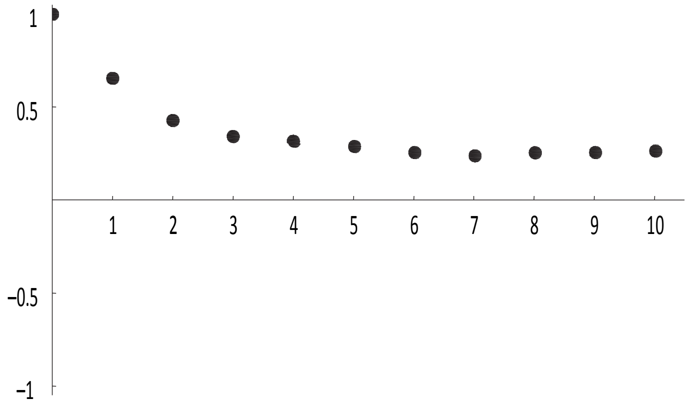
In this time series, a strong trend is clearly visible. The effect of this is that any given value is followed, in general, by terms that are greater. This gives positive correlation at all lags. The decay of the autocorrelation function will be very slow, if it occurs at all.



If the trend is weaker, for example $X_n = 0.001n + 0.5X_{n-1} + e_n$, then there may be some decay at first as the trend is swamped by the other factors, but there will still be some residual correlation at larger lags.



The trend is difficult to see from this small sample of the data but shows up in the autocorrelation function as the residual correlation at higher lags.



Question

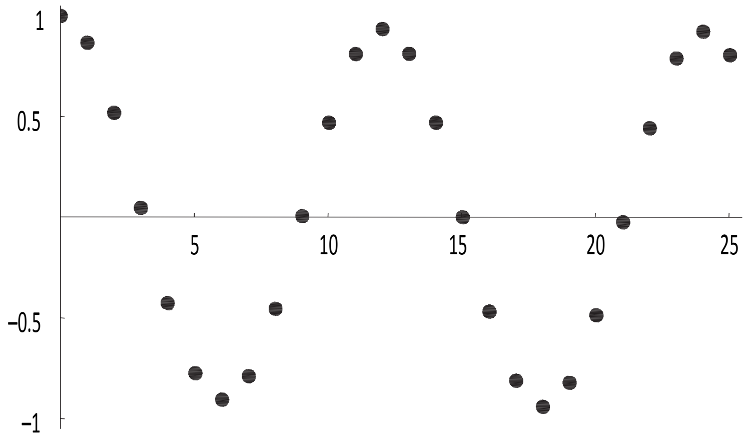
Describe the associations we would expect to find in a time series representing the average daytime temperature in successive months in a particular town, and hence sketch a diagram of the autocorrelation function of this series.

Solution

We expect the temperature in different years to be roughly the same at the same time of year, and hence there should be very strong positive correlation at lags of 12 months, 24 months and so on.

Within each year we would also expect a positive correlation between nearby times, for example with lags of 1 or 2 months, with decreasing correlation as the lag increases. On the other hand, once we reach a lag of 6 months there should be strong *negative* correlation since one temperature will be above the mean, the other below it. For example comparing June with December.

The autocorrelation function will therefore oscillate with period of 12 months.



2.5 Partial autocorrelation function

Another important characteristic of a stationary random process is the *partial autocorrelation function* (PACF), $\{\phi_k : k = 1,2,\ldots\}$, defined as the conditional correlation of X_{t+k} with X_t given $X_{t+1},\ldots,X_{t+k-1}$.

Unlike the autocovariance and autocorrelation functions, the PACF is defined for positive lags only.

This may be derived as the coefficient $\phi_{k,k}$ in the problem to minimise:

$$E\left[\left(X_t - \phi_{k,1}X_{t-1} - \phi_{k,2}X_{t-2} - \cdots - \phi_{k,k}X_{t-k}\right)^2\right]$$

We can explain the last expression as follows. Suppose that at time $t-1$ we are trying to estimate X_t, but we are going to limit our choice of estimator to linear functions of the k previous values $X_{t-k},\dots,X_{t-1}$. The most general linear estimator will be of the form:

$$\phi_{k,1}X_{t-1}+\phi_{k,2}X_{t-2}+\dots+\phi_{k,k}X_{t-k}$$

where $\phi_{k,i}$ are constants. We can choose the coefficients to minimise the mean square error, which is the expression given above in Core Reading. The partial autocorrelation for lag *k* is then the weight that we assign to the X_{t-k} term.

Question

Consider the process $X_t=0.5X_{t-2}+e_t$, where e_t forms a white noise process.

Determine the partial autocorrelation function for this process.

Solution

For $k=1$ we just have the correlation itself. However, in this case it is clear that the X_t for even values of *t* are independent of those for odd values. It follows that the correlation at lag 1 is 0.

For $k=2$ the partial autocorrelation is the coefficient of X_{t-2} in the best linear estimator:

$$\phi_{2,1}X_{t-1}+\phi_{2,2}X_{t-2}$$

Comparing this to the defining equation suggests that $\phi_2=0.5$.

Similarly, the defining equation suggests that the best linear estimator will not involve $X_{t-3},X_{t-4},\dots$. It follows that for $k\geq 3$, we have $\phi_k=0$.

For the time series in the previous question, we have $\phi_4=0$. This is in contrast to the actual correlation at lag four, since X_t depends on X_{t-2}, which in turn depends on X_{t-4}. X_t and X_{t-4} will therefore be correlated. The partial autocorrelation is zero, however, because it effectively removes the impact of the correlation at smaller lags.

In general it is difficult to calculate the PACF by hand.

The formula for calculating ϕ_k involves a ratio of determinants of large matrices whose entries are determined by $\rho_1,\dots,\rho_k$; it may be found in standard works on time series analysis, and is readily available in common computer packages like R.

The diagrams below show the autocorrelation function and partial autocorrelation of an *ARMA*(1,1) series. ARMA processes are discussed in detail in Section 3.7.

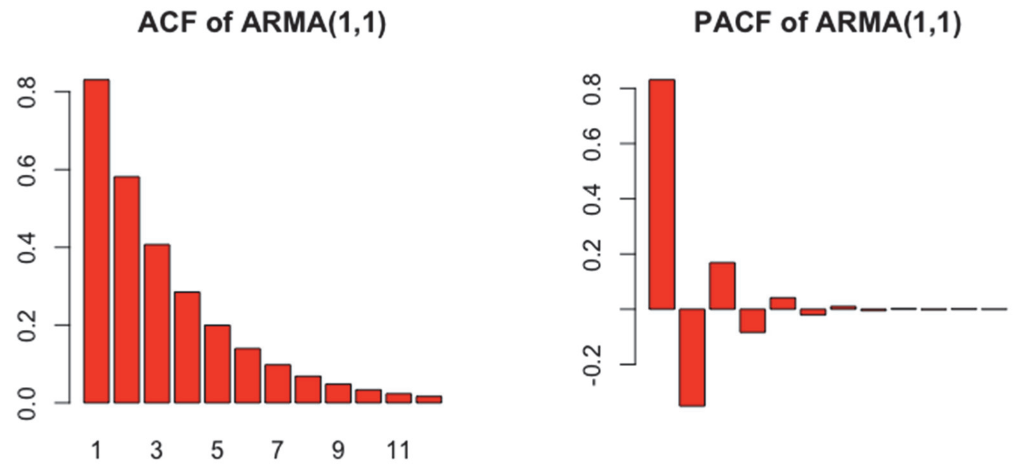


Figure 13.1: ACF and PACF values of some stationary time series model.

In particular the formulae for ϕ_1 and ϕ_2 are as follows:

Partial autocorrelation function at lags 1 and 2

$$\phi_1 = \rho_1, \qquad \phi_2 = \frac{\det\begin{pmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{pmatrix}}{\det\begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

Note that for each k, ϕ_k depends on only $\rho_1, \rho_2, \ldots, \rho_k$.

These formulae can be found on page 40 of the *Tables.* Their derivations are not required.

It is important to realise that the PACF is determined by the ACF, as the above expressions suggest. The PACF does not therefore contain any extra information; it just gives an alternative presentation of the same information. However, as we will see, this can be used to identify certain types of process.

3 Main linear models of time series

3.1 Introduction

The main linear models used for modelling stationary time series are:

- **Autoregressive process (*AR*)**
- **Moving average process (*MA*)**
- **Autoregressive moving average process (*ARMA*).**

The definitions of each of these processes, presented below, involve the standard zero-mean white noise process $\{e_t : t = 1,2,\ldots\}$ defined in Section 2.3.

In practice we often wish to model processes which are not $I(0)$ (stationary) but $I(1)$. For this purpose a further model is considered:

- **Autoregressive integrated moving average (*ARIMA*).**

ARMA and *ARIMA* are pronounced as single words (similar to 'armour' and 'areema').

Autoregressive

An *autoregressive process* of order p (the notation $AR(p)$ is commonly used) is a sequence of random variables $\{X_t\}$ defined consecutively by the rule:

$$X_t = \mu + \alpha_1\left(X_{t-1}-\mu\right) + \alpha_2\left(X_{t-2}-\mu\right) + \cdots + \alpha_p\left(X_{t-p}-\mu\right) + e_t$$

Thus the autoregressive model attempts to explain the current value of X as a linear combination of past values with some additional externally generated random variation. The similarity to the procedure of linear regression is clear, and explains the origin of the name 'autoregression'.

Moving average

A *moving average process* of order q, denoted $MA(q)$, is a sequence $\{X_t\}$ defined by the rule:

$$X_t = \mu + e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q}$$

The moving average model explains the relationship between the X_t as an indirect effect, arising from the fact that the current value of the process results from the recently passed random error terms as well as the current one. In this sense, X_t is 'smoothed noise'.

Autoregressive moving average

The two basic processes (AR and MA) can be combined to give an autoregressive moving average, or ARMA, process. The defining equation of an $ARMA(p,q)$ process is:

$$X_t = \mu + \alpha_1\left(X_{t-1}-\mu\right) + \cdots + \alpha_p\left(X_{t-p}-\mu\right) + e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q}$$

Note: $ARMA(p,0)$ is $AR(p)$; $ARMA(0,q)$ is $MA(q)$.

Autoregressive integrated moving average

The definition of an $ARIMA(p,d,q)$ process is given in Section 3.8.

3.2 The backwards shift operator, *B*, and the difference operator, ∇

Further discussion of the various models will be helped by the use of two operators which operate on the whole time series process *X*.

The *backwards shift operator*, B, acts on the process X to give a process BX such that:

$$(BX)_t = X_{t-1}$$

If we apply the backwards shift operator to a constant, then it doesn't change it:

$$B\mu = \mu$$

The *difference operator*, ∇, is defined as $\nabla = 1 - B$, or in other words:

$$(\nabla X)_t = X_t - X_{t-1}$$

Both operators can be applied repeatedly. For example:

$$(B^2 X)_t = (B(BX))_t = (BX)_{t-1} = X_{t-2}$$

$$(\nabla^2 X)_t = (\nabla X)_t - (\nabla X)_{t-1} = X_t - 2X_{t-1} + X_{t-2}$$

and can be combined as, for example:

$$(B\nabla X)_t = (B(1-B)X)_t = (BX)_t - (B^2 X)_t = X_{t-1} - X_{t-2}$$

The usefulness of both of these operators will become apparent in later sections.

We could also work out $\nabla^2 X_t$ as follows:

$$\nabla^2 X_t = (1-B)^2 X_t = (1-2B+B^2)X_t = X_t - 2X_{t-1} + X_{t-2}$$

Similarly:

$$\nabla^3 X_t = \left(1-B\right)^3 X_t = (1-3B+3B^2-B^3)X_t = X_t - 3X_{t-1} + 3X_{t-2} - X_{t-3}$$

In addition, we use the difference operator to write $X_t - 5X_{t-1} + 7X_{t-2} - 3X_{t-3}$ as follows:

$$\begin{aligned} X_t - 5X_{t-1} + 7X_{t-2} - 3X_{t-3} &= (X_t - X_{t-1}) - 4(X_{t-1} - X_{t-2}) + 3(X_{t-2} - X_{t-3}) \\ &= \nabla X_t - 4\nabla X_{t-1} + 3\nabla X_{t-2} \\ &= (\nabla X_t - \nabla X_{t-1}) - 3(\nabla X_{t-1} - \nabla X_{t-2}) \\ &= \nabla^2 X_t - 3\nabla^2 X_{t-1} \end{aligned}$$

Question

Suppose that $w_n = \nabla x_n$. Give a formula for x_n in terms of x_0 and the differences:

$$w_n = x_n - x_{n-1}, \quad w_{n-1} = x_{n-1} - x_{n-2}, \ldots, \quad w_1 = x_1 - x_0$$

Solution

We have:

$$\begin{aligned} x_n &= w_n + x_{n-1} \\ &= w_n + w_{n-1} + x_{n-2} \\ &= \cdots \\ &= w_n + w_{n-1} + \cdots + w_1 + x_0 \\ &= x_0 + \sum_{i=1}^{n} w_i \end{aligned}$$

The R commands for generating the differenced values of some time series x are:

```
diff(x,lag=1,differences=1)
```

for ordinary difference ∇.

```
diff(x,lag=1,differences=3)
```

for differencing three times ∇^3, and:

```
diff(x,lag=12,differences=1)
```

for a simple seasonal difference with period 12, ∇_{12} (see Section 1.4 in Chapter 14).

3.3 The first-order autoregressive model, *AR*(1)

The simplest autoregressive process is the $AR(1)$, given by:

$$X_t = \mu + \alpha(X_{t-1} - \mu) + e_t \qquad \textbf{(13.1)}$$

A process satisfying this recursive definition can be represented as:

$$X_t = \mu + \alpha^t (X_0 - \mu) + \sum_{j=0}^{t-1} \alpha^j e_{t-j} \qquad \textbf{(13.2)}$$

This representation can be obtained by substituting in for X_{t-1}, then for X_{t-2}, and so on, until we reach X_0. It is important to realise that X_0 itself will be a random variable in general – it is not necessarily a given constant, although it might be.

It follows that the mean function μ_t is given by:

$$\mu_t = \mu + \alpha^t \left(\mu_0 - \mu\right)$$

Here the notation μ_t is being used in place of $E(X_t)$. This result follows by taking expectations of both sides of Equation (13.2), and noting that the white noise terms have zero mean.

The white noise terms are also uncorrelated with each other, and with X_0. It follows that the variance of X_t can be found by summing the variances of the terms on the right-hand side.

The same representation (13.2) gives the variance:

$$\text{var}(X_t) = \sigma^2 \frac{1-\alpha^{2t}}{1-\alpha^2} + \alpha^{2t}\,\text{var}(X_0)$$

where, as before, σ^2 denotes the common variance of the white noise terms $\{e_t\}$.

Question

Derive this expression.

Solution

From Equation (13.2), we have:

$$\begin{aligned}
\text{var}(X_t) &= \text{var}\left(\mu + \alpha^t (X_0 - \mu) + \sum_{j=0}^{t-1} \alpha^j e_{t-j}\right) \\
&= \alpha^{2t}\,\text{var}(X_0 - \mu) + \sum_{j=0}^{t-1} \alpha^{2j}\,\text{var}(e_{t-j}) \\
&= \alpha^{2t}\,\text{var}(X_0) + \sigma^2 \sum_{j=0}^{t-1} \alpha^{2j}
\end{aligned}$$

Now using the formula $a + ar + ar^2 + \cdots + ar^{n-1} = \dfrac{a(1-r^n)}{1-r}$ for summing the first n terms of a geometric progression, we see that:

$$\text{var}(X_t) = \alpha^{2t}\,\text{var}(X_0) + \sigma^2\left(\frac{1-\alpha^{2t}}{1-\alpha^2}\right)$$

For the process X to be stationary, its mean and variance must both be constant. A quick look at the expressions above is enough to see that this will not be the case in general. It is therefore natural to ask for conditions under which the process *is* stationary.

From this it follows that a stationary process X satisfying (13.1) can exist only if $|\alpha| < 1$. Further requirements are that $\mu_0 = \mu$ and that $\text{var}(X_0) = \sigma^2 / (1 - \alpha^2)$.

It should be clear that we require $\mu_0 = \mu$ in order to remove the t-dependence from the mean. Similarly, we require $\text{var}(X_0) = \dfrac{\sigma^2}{1-\alpha^2}$ in order to make the variance constant. (We are assuming $\alpha \neq 0$, otherwise X is a white noise process, which is certainly stationary.)

We also require $|\alpha| < 1$. One way of seeing this is to note that the variance has to be a finite non-negative number.

Notice that this implies that X can be stationary only if X_0 is random. If X_0 is a known constant, then $\text{var}(X_0) = 0$ and $\text{var}(X_t)$ is no longer independent of t, whereas if X_0 has expectation different from μ then the process X will have non-constant expectation.

We now consider the situation in which $\mu_t \neq \mu$, and/or $\text{var}(X_0) \neq \dfrac{\sigma^2}{1-\alpha^2}$. From what we've just said, the process will then be non-stationary. However, what we are about to see is that even if the process is non-stationary, as long as $|\alpha| < 1$, the process will become stationary in the long run, without any extra conditions.

It is easy to see that the difference $\mu_t - \mu$ is a multiple of α^t and that $\text{var}(X_t) - \sigma^2 / (1 - \alpha^2)$ is a multiple of α^{2t}.

This follows by writing the equations we derived above for the mean and variance in the form:

$$\mu_t - \mu = \alpha^t (\mu_0 - \mu) \quad \text{and} \quad \text{var}(X_t) - \frac{\sigma^2}{1-\alpha^2} = \alpha^{2t} \left[\text{var}(X_0) - \frac{\sigma^2}{1-\alpha^2} \right]$$

Both of these terms will decay away to zero for large t if $|\alpha| < 1$, implying that X will be virtually stationary for large t.

We can also turn this result on its head: if we assume that the process has already been running for a very long time, then the process will be stationary. In other words, any $AR(1)$ process with an infinite history and $|\alpha| < 1$ will be stationary.

In this context it is often helpful to assume that $X_1, \ldots, X_n$ is merely a sub-sequence of a process $\ldots, X_{-1}, X_0, X_1, \ldots, X_n$ which has been going on unobserved for a long time and has already reached a 'steady state' by the time of the first observation. A double-sided infinite process satisfying (13.1) can be represented as:

$$X_t = \mu + \sum_{j=0}^{\infty} \alpha^j e_{t-j} \qquad \textbf{(13.3)}$$

This is an explicit representation of the value of X_t in terms of the historic values of the process e. The infinite sum on the right-hand side only makes sense, *ie* converges, if $|\alpha|<1$. The equation can be derived in two ways – either using an iterative procedure as used above to derive Equation (13.2), or by using the backward shift operator.

For the latter method, we write the defining equation $X_t = \mu + \alpha(X_{t-1} - \mu) + e_t$ in the form:

$$(1-\alpha B)(X_t - \mu) = e_t$$

The expression $(1-\alpha B)$ will be invertible (using an expansion) if and only if $|\alpha|<1$. (In fact we could expand it for $|\alpha|>1$ using $(1-\alpha B)^{-1} = -(\alpha B)^{-1}\left(1-\alpha^{-1}B^{-1}\right)^{-1}$. However, this would give an expansion in terms of future values since B^{-1} is effectively a forward shift. This would not therefore be of much use. We will not point out this qualification in future.)

If $(1-\alpha B)$ is invertible, then we can write:

$$X_t - \mu = (1-\alpha B)^{-1} e_t = \left(1+\alpha B + \alpha^2 B^2 + \cdots\right) e_t$$
$$= e_t + \alpha e_{t-1} + \alpha^2 e_{t-2} + \cdots$$

In other words:

$$X_t = \mu + \sum_{j=0}^{\infty} \alpha^j e_{t-j}$$

This representation makes it clear that X_t has expectation μ and variance equal to:

$$\sum_{j=0}^{\infty} \alpha^{2j}\sigma^2 = \frac{\sigma^2}{1-\alpha^2}$$

if $|\alpha|<1$.

This last step uses the formula for the sum of an infinite geometric progression.

So, in this case, the process does satisfy the conditions required for stationarity given above.

We have only looked at the mean and variance so far, however. We also need to look at the autocovariance function.

In order to deduce that X is stationary we also need to calculate the autocovariance function:

$$\gamma_k = \text{cov}(X_t, X_{t+k}) = \sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\alpha^i\alpha^j \,\text{cov}(e_{t-j}, e_{t+k-i})$$

$$= \sum_{j=0}^{\infty}\sigma^2\alpha^{2j+k} = \alpha^k\sum_{j=0}^{\infty}\sigma^2\alpha^{2j} = \alpha^k\gamma_0$$

The double sum has been simplified here by noting that the covariance will be zero unless the subscripts $t-j$ and $t+k-i$ are equal, *ie* unless $i=j+k$. In this case the covariance equals σ^2. So in the sum over i, we only include the term for $i=j+k$.

We have also used the formula $\gamma_0 = \text{var}(X_t) = \sum_{j=0}^{\infty}\alpha^{2j}\sigma^2 = \dfrac{\sigma^2}{1-\alpha^2}$ from before.

This is independent of t, and thus a stationary process exists as long as $|\alpha| < 1$.

It is worth introducing here a method of more general utility for calculating autocovariance functions. From (13.1) we have, assuming that X is stationary:

$$\begin{aligned}\gamma_k &= \text{cov}(X_t, X_{t-k})\\ &= \text{cov}(\mu + \alpha(X_{t-1}-\mu) + e_t, X_{t-k})\\ &= \alpha\,\text{cov}(X_{t-1}, X_{t-k})\\ &= \alpha\gamma_{k-1}\end{aligned}$$

implying that:

$$\gamma_k = \alpha^k\gamma_0 = \alpha^k\frac{\sigma^2}{1-\alpha^2}$$

and:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \alpha^k$$

for $k \geq 0$.

The partial autocorrelation function ϕ_k is given by:

$$\phi_1 = \rho_1 = \alpha$$

$$\phi_2 = \frac{\alpha^2 - \alpha^2}{1-\alpha^2} = 0$$

Indeed, since the best linear estimator of X_t given $X_{t-1}, X_{t-2}, X_{t-3}$, is just αX_{t-1}, the definition of the PACF implies that $\phi_k = 0$ for all $k > 1$. Notice the contrast with the ACF, which decreases geometrically towards 0.

The following lines in R generate the ACF and PACF functions for an $AR(1)$ model:

```
par(mfrow=c(1,2))

barplot(ARMAacf(ar=0.7,lag.max = 12)[-1],main = "ACF of AR(1)",
col="red")

barplot(ARMAacf(ar=0.7,lag.max = 12,pacf = TRUE),main = "PACF of
AR(1)",col="red")
```

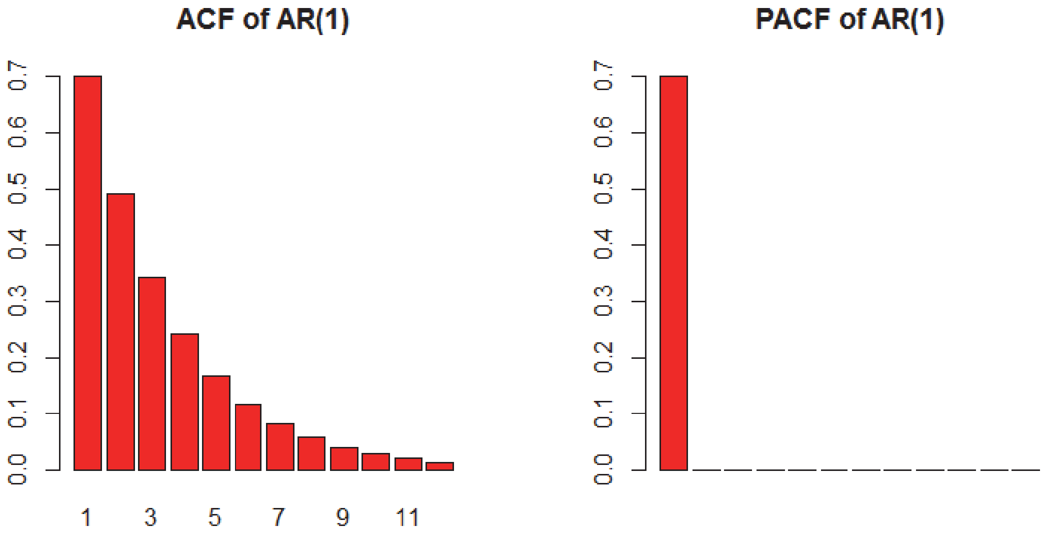


Figure 13.2: ACF and PACF of $AR(1)$ with $\alpha = 0.7$

Example

One of the well known applications of a univariate autoregressive model is the description of the evolution of the consumer price index $\{Q_t : t = 1,2,3,\dots\}$. The *force of inflation*, $r_t = \ln(Q_t / Q_{t-1})$, is assumed to follow the $AR(1)$ process:

$$r_t = \mu + \alpha(r_{t-1} - \mu) + e_t$$

One initial condition, the value for r_0, is required for the complete specification of the model for the force of inflation r_t.

The process r_t is said to be *mean-reverting*, *ie* it has a long-run mean, and if it drifts away, then it tends to be dragged back towards it. In this case, the long-run mean is μ. The equation for r_t can be written in the form:

$$r_t - \mu = \alpha(r_{t-1} - \mu) + e_t$$

If we ignore the error term, which has a mean of zero, then this equation says that the difference between r and the long-run mean at time t, is α times the previous difference. In order to be mean-reverting, this distance must reduce, so we need $|\alpha|<1$, as for stationarity. In fact, we probably wouldn't expect the force of inflation to be dragged to the other side of the mean, so a realistic model is likely to have $0<\alpha<1$.

3.4 The autoregressive model, *AR*(*p*)

The equation of the more general $AR(p)$ process is:

$$X_t = \mu + \alpha_1(X_{t-1}-\mu) + \alpha_2(X_{t-2}-\mu) + \cdots + \alpha_p(X_{t-p}-\mu) + e_t \qquad \textbf{(13.4)}$$

or, in terms of the backwards shift operator:

$$(1-\alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p)(X_t - \mu) = e_t \qquad \textbf{(13.5)}$$

As seen for $AR(1)$, there are some restrictions on the values of the α_j which are permitted if the process is to be stationary. In particular, we have the following result.

Condition for stationarity of an *AR*(*p*) process (Result 13.2)

If the time series process X given by (13.4) is stationary then the roots of the equation:

$$1-\alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_p z^p = 0$$

are all greater than 1 in absolute value.

(The polynomial $1-\alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_p z^p$ is called the *characteristic polynomial* of the autoregression.)

The equation $1-\alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_p z^p = 0$ is known as the *characteristic equation* of the process.

For an $AR(1)$ process, $X_t = \mu + \alpha\left(X_{t-1}-\mu\right) + e_t$, the characteristic equation is:

$$1-\alpha z = 0$$

The root of this equation is $z = 1/\alpha$. So for an $AR(1)$ process to be stationary, we must have:

$$\frac{1}{|\alpha|} > 1$$

ie:

$$|\alpha| < 1$$

This is the same stationarity condition that we derived in Section 3.3.

It is important to realise what this result does *not* say. Although for an $AR(1)$ process we can look at the coefficient α and deduce stationarity if and only if $|\alpha|<1$, we cannot do this for higher order autoregressive processes. For example, an $AR(2)$ process:

$$X_t = \mu + \alpha_1 \left(X_{t-1} - \mu\right) + \alpha_2 \left(X_{t-2} - \mu\right) + e_t$$

would not necessarily be stationary, just because $|\alpha_i|<1$ for both $i=1$ and 2. We would need to look at the roots of the characteristic polynomial.

We can prove the stationarity condition as follows.

Proof of Result 13.2

If X is stationary then its autocovariance function satisfies:

$$\boldsymbol{\gamma_k = \text{cov}(X_t, X_{t-k}) = \text{cov}\left(\sum_{j=1}^{p} \alpha_j X_{t-j} + e_t, X_{t-k}\right) = \sum_{j=1}^{p} \alpha_j \gamma_{k-j}}$$

for $k \geq p$.

The μ's are constant and do not therefore contribute to the covariance.

This is a p-th order difference equation with constant coefficients; it has a solution of the form:

$$\boldsymbol{\gamma_k = \sum_{j=1}^{p} A_j z_j^{-k}}$$

for all $k \geq 0$, where $z_1, \ldots, z_p$ are the p roots of the characteristic polynomial and $A_1, \ldots, A_p$ are constants. (We will show this in a moment.) **As X is purely indeterministic, we must have $\gamma_k \to 0$, which requires that $|z_j| > 1$ for each j.**

Question

Show by substitution that $\gamma_k = \sum_{j=1}^{p} A_j z_j^{-k}$ is a solution of the given difference equation.

Solution

Substituting $\gamma_k = \sum_{j=1}^{p} A_j z_j^{-k}$ into the right-hand side of the difference equation $\gamma_k = \sum_{j=1}^{p} \alpha_j \gamma_{k-j}$ we get:

$$\sum_{j=1}^{p} \alpha_j \gamma_{k-j} = \sum_{j=1}^{p} \alpha_j \left(\sum_{i=1}^{p} A_i z_i^{-k+j} \right) = \sum_{i=1}^{p} A_i z_i^{-k} \left(\sum_{j=1}^{p} \alpha_j z_i^j \right)$$

By definition of z_i as a root of the characteristic equation, we have:

$$\sum_{j=1}^{p} \alpha_j z_i^j = 1$$

So:

$$\sum_{j=1}^{p} \alpha_j \gamma_{k-j} = \sum_{i=1}^{p} A_i z_i^{-k} = \gamma_k$$

as required.

The converse of Result 13.2 is also true (but the proof is not given here): if the roots of the characteristic polynomial are all greater than 1 in absolute value, then it is possible to construct a stationary process X satisfying (13.4). In order for an arbitrary process X satisfying (13.4) to be stationary, the variances and covariances of the initial values $X_0, X_{-1}, \ldots, X_{-p+1}$ must also be equal to the appropriate values.

Although we do not give a formal proof, we will provide another way of thinking about this result.

Recall that in the $AR(1)$ case we said that the process turned out to be stationary if and only if X_t could be written as a (convergent) sum of white noise terms. Equivalently, if we start from the equation:

$$(1-\alpha B)(X_t - \mu) = e_t$$

then the process is stationary if and only if we can invert the term $(1-\alpha B)$, since this is the case if and only if $|\alpha| < 1$.

Analogously, we can write an AR(p) process in the form:

$$\left(1-\frac{B}{z_1}\right)\left(1-\frac{B}{z_2}\right)\cdots\left(1-\frac{B}{z_p}\right)(X_t - \mu) = e_t$$

where $z_1, z_2, \ldots, z_p$ are the p (possibly complex) roots of the characteristic polynomial. In other words, the characteristic polynomial factorises as:

$$1-\alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_p z^p = \left(1-\frac{z}{z_1}\right)\left(1-\frac{z}{z_2}\right)\cdots\left(1-\frac{z}{z_p}\right)$$

It follows that in order to write X_t in terms of the process e, we need to be able to invert all p of the factors $\left(1-\frac{B}{z_i}\right)$. This will be the case if and only if $|z_i| > 1$ for all $i = 1, 2, \ldots, p$.

Question

Determine the characteristic polynomial of the process defined by the equation:

$$X_t = 5 - 2(X_{t-1} - 5) + 3(X_{t-2} - 5) + e_t$$

and calculate its roots. Hence comment on the stationarity of the process.

Solution

We first rearrange the equation for the process so that all the X terms appear on the same side. Doing this we obtain:

$$X_t + 2X_{t-1} - 3X_{t-2} = e_t$$

We now replace X_{t-s} by z^s , *ie* we replace X_t by 1, X_{t-1} by z, and X_{t-2} by z^2 . So the characteristic polynomial is $1 + 2z - 3z^2$.

This polynomial can be factorised as $(1+3z)(1-z)$, so its roots are $\frac{1}{3}$ and 1.

This shows that the process is not stationary.

There is no requirement to use the letter z (or indeed any particular letter) when writing down the characteristic polynomial. The letter λ is often used instead.

Question

Given that $\lambda = 2$ is a root of the characteristic equation of the process:

$$X_n = \tfrac{11}{6} X_{n-1} - X_{n-2} + \tfrac{1}{6} X_{n-3} + e_n$$

calculate the other roots and classify the process as $I(d)$.

Solution

The process can be written in the form:

$$X_n - \tfrac{11}{6} X_{n-1} + X_{n-2} - \tfrac{1}{6} X_{n-3} = e_n$$

So that the characteristic equation is:

$$1 - \tfrac{11}{6}\lambda + \lambda^2 - \tfrac{1}{6}\lambda^3 = 0$$

We are given that $\lambda=2$ is a root. So $(\lambda-2)$ is a factor and hence:

$$1-\tfrac{11}{6}\lambda+\lambda^2-\tfrac{1}{6}\lambda^3=(\lambda-2)(a\lambda^2+b\lambda+c)$$

where a, b and c are constants. The values of a, b and c can be determined in several ways, *eg* by comparing the coefficients on both sides of this equation, by long division of polynomials, or by synthetic division. We find that:

$$1-\tfrac{11}{6}\lambda+\lambda^2-\tfrac{1}{6}\lambda^3=(\lambda-2)\left(-\tfrac{1}{6}\lambda^2+\tfrac{2}{3}\lambda-\tfrac{1}{2}\right)=-\tfrac{1}{6}(\lambda-2)(\lambda-3)(\lambda-1)$$

So the other roots of the characteristic equation are $\lambda=3$ and $\lambda=1$.

The process is not stationary, since the characteristic equation has a root that is not strictly greater than 1 in magnitude.

It is easy to see that differencing the process once will eliminate the root of 1. The two remaining roots (*ie* 2 and 3) are both strictly greater than 1 in magnitude, so the differenced process is stationary. Hence X is $I(1)$.

Often exact values for the γ_k are required, entailing finding the values of the constants A_k. From (13.4) we have:

$$\operatorname{cov}(X_t,X_{t-k})=\alpha_1\operatorname{cov}(X_{t-1},X_{t-k})+\cdots+\alpha_p\operatorname{cov}(X_{t-p},X_{t-k})+\operatorname{cov}(e_t,X_{t-k})$$

which can be re-expressed as:

$$\gamma_k=\alpha_1\gamma_{k-1}+\alpha_2\gamma_{k-2}+\cdots+\alpha_p\gamma_{k-p}+\sigma^2 1_{\{k=0\}}$$

for $0\le k\le p$. (These are known as the *Yule-Walker equations*.) Here the notation $1_{\{k=0\}}$ denotes an indicator function, taking the value 1 if $k=0$, the value 0 otherwise.

For $p=3$ we have 4 equations:

$$\gamma_3=\alpha_1\gamma_2+\alpha_2\gamma_1+\alpha_3\gamma_0$$

$$\gamma_2=\alpha_1\gamma_1+\alpha_2\gamma_0+\alpha_3\gamma_1$$

$$\gamma_1=\alpha_1\gamma_0+\alpha_2\gamma_1+\alpha_3\gamma_2$$

$$\gamma_0=\alpha_1\gamma_1+\alpha_2\gamma_2+\alpha_3\gamma_3+\sigma^2$$

The second and third of these equations are sufficient to deduce γ_2 and γ_1 in terms of γ_0, which is all that is required to find ρ_2 and ρ_1. The first and fourth of the equations are needed when the values of the γ_k are to be found explicitly.

The PACF, $\{\phi_k : k \geq 1\}$, of the $AR(p)$ process can be calculated from the defining equations, but is not memorable. In particular, the first three equations above can be written in terms of ρ_1, ρ_2, ρ_3 and the resulting solution of α_3 as a function of ρ_1, ρ_2, ρ_3 is the expression of ϕ_3. The same idea applies to all values of k , so that ϕ_k is the solution of α_k in a system of k linear equations, including those for $\phi_1 = \rho_1$ and $\phi_2 = \dfrac{\rho_2 - \rho_1^2}{1-\rho_1^2}$ that we have seen before.

It is important to note, though, that $\phi_k = 0$ for all $k > p$.

This result is worth repeating.

Behaviour of the PACF for an $AR(p)$ process

For an $AR(p)$ process:

$$\phi_k = 0 \quad \text{for } k > p$$

This property of the PACF is characteristic of autoregressive processes and forms the basis of the most frequently used test for determining whether an $AR(p)$ model fits the data. It would be difficult to base a test on the ACF as the ACF of an autoregressive process is a sum of geometrically decreasing components. (See Section 2 in Chapter 14.)

Question

Give a derivation of the equation:

$$\gamma_0 = \alpha_1\gamma_1 + \alpha_2\gamma_2 + \alpha_3\gamma_3 + \sigma^2$$

Solution

The autocovariance at lag 0 is:

$$\gamma_0 = \text{cov}\left(X_t, X_t\right)$$

Expanding the LHS only (which will always be our approach when determining the autocovariance function of an autoregressive process), and remembering that the covariance is unaffected by the mean μ, we see that:

$$\gamma_0 = \text{cov}\left(\alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + e_t, X_t\right)$$

Now, using the properties of covariance:

$$\begin{aligned}\gamma_0 &= \alpha_1 \text{cov}(X_{t-1}, X_t) + \alpha_2 \text{cov}(X_{t-2}, X_t) + \alpha_3 \text{cov}(X_{t-3}, X_t) + \text{cov}(e_t, X_t) \\ &= \alpha_1\gamma_1 + \alpha_2\gamma_2 + \alpha_3\gamma_3 + \text{cov}\left(e_t, X_t\right)\end{aligned}$$

But:

$$\begin{aligned}\text{cov}(e_t, X_t) &= \text{cov}(e_t, \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + e_t)\\ &= \alpha_1 \text{cov}(e_t, X_{t-1}) + \alpha_2 \text{cov}(e_t, X_{t-2}) + \alpha_3 \text{cov}(e_t, X_{t-3}) + \text{cov}(e_t, e_t)\\ &= 0+0+0+\sigma^2 = \sigma^2\end{aligned}$$

This is because $X_{t-1}, X_{t-2}, \ldots$ are functions of past white noise terms, and e_t is independent of earlier values. So:

$$\gamma_0 = \alpha_1\gamma_1 + \alpha_2\gamma_2 + \alpha_3\gamma_3 + \sigma^2$$

We will now derive a formula for the ACF of an $AR(2)$ process. To do this we need to remember a little about difference equations. Some formulae relating to difference equations are given on page 4 of the *Tables*.

Question

A stationary $AR(2)$ process is defined by the equation:

$$X_t = \tfrac{5}{6} X_{t-1} - \tfrac{1}{6} X_{t-2} + e_t$$

Determine the values of ρ_k and ϕ_k for $k = 1, 2, 3, \ldots$.

Solution

We do not actually need to calculate γ_0 in order to find the ACF. This is always the case for an autoregressive process.

By definition:

$$\rho_0 = \frac{\gamma_0}{\gamma_0} = 1$$

The autocovariance at lag 1 is:

$$\begin{aligned}\gamma_1 &= \text{cov}(X_t, X_{t-1})\\ &= \text{cov}(\tfrac{5}{6} X_{t-1} - \tfrac{1}{6} X_{t-2} + e_t, X_{t-1})\\ &= \tfrac{5}{6}\text{cov}(X_{t-1}, X_{t-1}) - \tfrac{1}{6}\text{cov}(X_{t-2}, X_{t-1}) + \text{cov}(e_t, X_{t-1})\\ &= \tfrac{5}{6}\gamma_0 - \tfrac{1}{6}\gamma_1\end{aligned}$$

Rearranging gives:

$$\tfrac{7}{6}\gamma_1 = \tfrac{5}{6}\gamma_0$$

So:

$$\gamma_1 = \tfrac{5}{7}\gamma_0 \qquad \text{and} \qquad \rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{5}{7}$$

Similarly, the autocovariance at lag 2 is:

$$\begin{aligned}\gamma_2 &= \text{cov}(X_t, X_{t-2})\\ &= \text{cov}(\tfrac{5}{6}X_{t-1} - \tfrac{1}{6}X_{t-2} + e_t, X_{t-2})\\ &= \tfrac{5}{6}\text{cov}(X_{t-1}, X_{t-2}) - \tfrac{1}{6}\text{cov}(X_{t-2}, X_{t-2}) + \text{cov}(e_t, X_{t-1})\\ &= \tfrac{5}{6}\gamma_1 - \tfrac{1}{6}\gamma_0\end{aligned}$$

Using the fact that $\gamma_1 = \frac{5}{7}\gamma_0$, we have:

$$\gamma_2 = \tfrac{5}{6}\left(\tfrac{5}{7}\gamma_0\right) - \tfrac{1}{6}\gamma_0 = \tfrac{3}{7}\gamma_0 \qquad \text{and} \qquad \rho_2 = \frac{\gamma_2}{\gamma_0} = \frac{3}{7}$$

In general, for $k \geq 2$, we have:

$$\rho_k = \tfrac{5}{6}\rho_{k-1} - \tfrac{1}{6}\rho_{k-2}$$

We can solve this second order difference equation. The characteristic equation (for the difference equation, which is unfortunately slightly different to the characteristic equation for the process) is:

$$\lambda^2 - \tfrac{5}{6}\lambda + \tfrac{1}{6} = \left(\lambda - \tfrac{1}{2}\right)\left(\lambda - \tfrac{1}{3}\right) = 0$$

Using the formula on page 4 of the *Tables,* the general solution of this difference equation is of the form:

$$\rho_k = A\left(\tfrac{1}{2}\right)^k + B\left(\tfrac{1}{3}\right)^k$$

In order to find the solution we want, we need to use two boundary conditions to determine the two constants. We know that $\rho_0 = 1$ and $\rho_1 = \frac{5}{7}$. So:

$$\rho_0 = A + B = 1$$

$$\rho_1 = \tfrac{1}{2}A + \tfrac{1}{3}B = \tfrac{5}{7}$$

Solving these equations gives $A = \frac{16}{7}$ and $B = -\frac{9}{7}$.

The autocorrelation function is therefore:

$$\rho_k = \tfrac{16}{7}\left(\tfrac{1}{2}\right)^k - \tfrac{9}{7}\left(\tfrac{1}{3}\right)^k$$

We are also asked for the partial autocorrelation function. Using the formulae on page 40 of the *Tables:*

$$\phi_1 = \rho_1 = \tfrac{5}{7}$$

$$\phi_2 = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2} = -\tfrac{1}{6}$$

Also, since this is an $AR(2)$ process:

$$\phi_k = 0 \text{ for } k = 3,\ 4,\ 5,\ \ldots$$

3.5 The first-order moving average model, *MA*(1)

A first-order moving average process, denoted $MA(1)$, is a process given by:

$$\boldsymbol{X_t = \mu + e_t + \beta e_{t-1}}$$

The mean of this process is $\mu_t = \mu$.

The variance and autocovariance are:

$$\gamma_0 = \text{var}(e_t + \beta e_{t-1}) = (1+\beta^2)\sigma^2$$

$$\gamma_1 = \text{cov}(e_t + \beta e_{t-1}, e_{t-1} + \beta e_{t-2}) = \beta\sigma^2$$

$$\gamma_k = 0 \text{ for } k > 1$$

Hence the ACF of the $MA(1)$ process is:

$$\rho_0 = 1$$

$$\rho_1 = \frac{\beta}{1+\beta^2}$$

$$\rho_k = 0 \text{ for } k > 1$$

Question

Show that the moving average process $X_n = Z_n + \beta Z_{n-1}$ is weakly stationary, where Z_n is a white noise process with mean μ and variance σ^2.

Solution

The mean is constant since $E(X_n)=(1+\beta)\mu$.

For the covariance:

$$\operatorname{cov}(X_n,X_n)=\operatorname{cov}(Z_n+\beta Z_{n-1},Z_n+\beta Z_{n-1})=(1+\beta^2)\sigma^2$$

or alternatively:

$$\operatorname{var}(X_n)=\operatorname{var}(Z_n+\beta Z_{n-1})=(1+\beta^2)\sigma^2$$

and:

$$\operatorname{cov}(X_n,X_{n-1})=\operatorname{cov}(Z_n+\beta Z_{n-1},Z_{n-1}+\beta Z_{n-2})=\operatorname{cov}(\beta Z_{n-1},Z_{n-1})=\beta\sigma^2$$

$$\operatorname{cov}(X_n,X_{n-2})=\operatorname{cov}(Z_n+\beta Z_{n-1},Z_{n-2}+\beta Z_{n-3})=0$$

In fact, the covariance at higher lags remains 0 since there is no overlap between the Z's. The covariances at the corresponding negative lags are the same.

Since none of these expressions depends on *n*, it follows that the process is weakly stationary.

An $MA(1)$ process is stationary regardless of the values of its parameters. The parameters are nevertheless usually constrained by imposing the condition of *invertibility*. This may be explained as follows.

It is possible to have two distinct $MA(1)$ models with identical ACFs: consider, for example, $\beta = 0.5$ and $\beta = 2$, both of which have $\rho_1 = \dfrac{\beta}{1+\beta^2} = 0.4$.

The defining equation of the $MA(1)$ may be written in terms of the backwards shift operator:

$$X - \mu = (1+\beta B)e \qquad (13.6)$$

In many circumstances an autoregressive model is more convenient than a moving average model.

We may rewrite (13.6) as:

$$(1+\beta B)^{-1}(X-\mu)=e$$

and use the standard expansion of $(1+\beta B)^{-1}=1-\beta B+\beta^2B^2-\beta^3B^3+\cdots$ to give:

$$X_t-\mu-\beta(X_{t-1}-\mu)+\beta^2(X_{t-2}-\mu)-\beta^3(X_{t-3}-\mu)+\cdots=e_t$$

The expansion referred to here is given on page 2 of the *Tables*.

The original moving average model has therefore been transformed into an autoregression of infinite order. But this procedure is only valid if the sum on the left-hand side is convergent, in other words if $|\beta| < 1$. When this condition is satisfied the $MA(1)$ is called *invertible*. Although more than one MA process may share a given ACF, at most one of the processes will be invertible.

We might want to know the historic values of the white noise process. Although the values $\{x_0, x_1, \ldots, x_t\}$ can be observed, and are therefore known at time *t*, the values of the white noise process $\{e_0, e_1, \ldots, e_t\}$ are not. Can we obtain the unknown *e* values from the known *x* values? The answer is yes, in theory, if and only if the process is invertible, since we can then write the value e_t in terms of the *x*'s, as above. In practice, we wouldn't actually have an infinite history of *x* values, but since the coefficients of the *x*'s get smaller as we go back in time, for an invertible process, the contribution of the values before time 1, say, will be negligible. We can make this more precise, as in the following question.

Question

Show that the process $X_n = \mu + e_n + \beta e_{n-1}$ may be inverted as follows:

$$e_n = (-\beta)^n e_0 + \sum_{i=0}^{n-1} (-\beta)^i (x_{n-i} - \mu)$$

Solution

A simple algebraic rearrangement shows that $X_n = \mu + e_n + \beta e_{n-1}$ can be rewritten as $e_n = X_n - \mu - \beta e_{n-1}$. Now using an iterative procedure:

$$\begin{aligned} e_n &= X_n - \mu - \beta e_{n-1} \\ &= X_n - \mu - \beta(X_{n-1} - \mu - \beta e_{n-2}) \\ &\vdots \\ &= (X_n - \mu) - \beta(X_{n-1} - \mu) + \beta^2(X_{n-2} - \mu) + \cdots + (-\beta)^{n-1}(X_1 - \mu) + (-\beta)^n e_0 \end{aligned}$$

Notice that, as *n* gets large, the dependence of e_n on e_0 will be small provided $|\beta| < 1$.

The condition for a $MA(1)$ process to be invertible is similar to the condition that an $AR(1)$ process is stationary. An $AR(1)$ process is stationary if and only if the process *X* can be written explicitly in terms of the process *e*. The invertibility condition ensures that the white noise process *e* can be written in terms of the *X* process. This relationship generalises to $AR(p)$ and $MA(q)$ processes, as we will see shortly.

It is possible, at the cost of considerable effort, to calculate the PACF of the $MA(1)$, giving:

$$\phi_k = (-1)^{k+1} \frac{(1-\beta^2)\beta^k}{1-\beta^{2(k+1)}}$$

This formula can be found on page 41 of the *Tables.*

This decays approximately geometrically as $k \to \infty$, highlighting the way in which the ACF and PACF are complementary: the PACF of a $MA(1)$ behaves like the ACF of an $AR(1)$, and the PACF of an $AR(1)$ behaves like the ACF of a $MA(1)$.

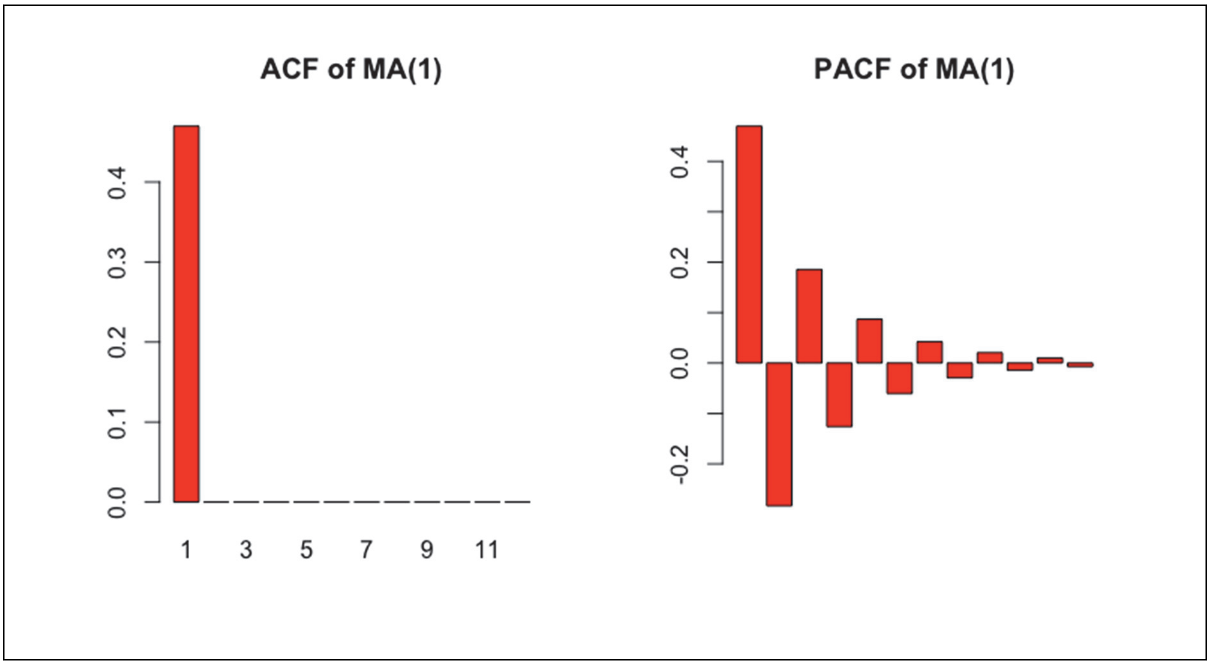


Figure 13.3: ACF and PACF of $MA(1)$ with $\beta = 0.7$

3.6 The moving average model, *MA*(*q*)

The defining equation of the general qth order moving average is, in backwards shift notation:

$$X - \mu = (1 + \beta_1 B + \beta_2 B^2 + \cdots + \beta_q B^q)e$$

In other words, it is:

$$X_t - \mu = e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \cdots + \beta_q e_{t-q}$$

Moving average processes are always stationary, as they are a linear combination of white noise, which is itself stationary.

Recall that for a stationary $AR(p)$ process, X_t can be expressed as an (infinite and convergent) sum of white noise terms. This means that any stationary autoregressive process can be considered to be a moving average of infinite order. However, by a moving average process we will usually mean one of finite order.

The autocovariance function is easier to find than in the case of $AR(p)$:

$$\gamma_k = \sum_{i=0}^{q}\sum_{j=0}^{q}\beta_i\beta_j E(e_{t-i}e_{t-j-k}) = \sigma^2\sum_{i=0}^{q-k}\beta_i\beta_{i+k}$$

provided $k \le q$. (Here β_0 denotes 1.)

Note that $\text{cov}(e_i, e_j) = E(e_ie_j) - E(e_i)E(e_j) = E(e_ie_j) = 0$ for $i \ne j$ (since the random variables e_t have 0 mean and are uncorrelated), and $\text{cov}(e_i, e_i) = \text{var}(e_i) = E(e_i^2) = \sigma^2$. So, in the double sum above, the only non-zero terms will be where the subscripts of e_{t-i} and e_{t-j-k} match, *ie* when $i = j + k$. This means that we can simplify the double sum by writing everything in terms of j. We need to get the limits right for j, which cannot go above $q - k$ because $i = j + k$ and i only goes up to q. So we get:

$$\gamma_k = \sum_{j=0}^{q-k}\beta_{j+k}\beta_j\sigma^2$$

This matches the formula above, except that i has been used in place of j.

For $k > q$ it is obvious that $\gamma_k = 0$. Just as autoregressive processes are characterised by the property that the partial ACF is equal to zero for sufficiently large k, moving average processes are characterised by the property that the ACF is equal to zero for sufficiently large k.

The importance of this observation will become apparent in Section 2 of Chapter 14. We will look at an explicit case of this result to make things clearer.

Question

Calculate γ_k, $k = 0,1,2,3,\ldots$ for the process:

$$X_n = 3 + e_n - e_{n-1} + 0.25e_{n-2}$$

where e_n is a white noise process with mean 0 and variance 1.

Solution

We have:

$$\begin{aligned}\gamma_0 &= \text{cov}(X_n, X_n) = \text{var}(X_n)\\ &= \text{var}(e_n - e_{n-1} + 0.25e_{n-2})\\ &= \text{var}(e_n) + (-1)^2\,\text{var}(e_{n-1}) + 0.25^2\,\text{var}(e_{n-2})\\ &= 1 + 1 + 0.0625\\ &= 2.0625\end{aligned}$$

$$\gamma_1 = \text{cov}\left(X_n, X_{n-1}\right)$$

$$= \text{cov}\left(e_n - e_{n-1} + 0.25e_{n-2}, e_{n-1} - e_{n-2} + 0.25e_{n-3}\right)$$

$$= \text{cov}(-e_{n-1}, e_{n-1}) + \text{cov}(0.25e_{n-2}, -e_{n-2})$$

$$= -1 - 0.25$$

$$= -1.25$$

$$\gamma_2 = \text{cov}\left(X_n, X_{n-2}\right)$$

$$= \text{cov}\left(e_n - e_{n-1} + 0.25e_{n-2}, e_{n-2} - e_{n-3} + 0.25e_{n-4}\right)$$

$$= \text{cov}(0.25e_{n-2}, e_{n-2})$$

$$= 0.25$$

$$\gamma_3 = \text{cov}\left(X_n, X_{n-3}\right)$$

$$= \text{cov}\left(e_n - e_{n-1} + 0.25e_{n-2}, e_{n-3} - e_{n-4} + 0.25e_{n-5}\right)$$

$$= 0$$

The covariance at higher lags is also 0 since there is no overlap between the *e*'s.

In the solution above, we have expanded the terms on both sides of the covariance expression. This will always be our strategy when calculating the autocovariance function for a moving average series. For all other types of series, we just expand the term on the LHS of the covariance expression.

We said above that an $MA(1)$ process $X_t = \mu + e_t + \beta e_{t-1}$ is invertible if $|\beta| < 1$, and we drew the analogy with an $AR(1)$ process being stationary. The same goes for this more general case. Recall that an $AR(p)$ process is stationary if and only if the roots of the characteristic equation are all strictly greater than 1 in magnitude.

For an $MA(q)$ process we have:

$$X_t - \mu = (1 + \beta_1 B + \beta_2 B^2 + \cdots + \beta_q B^q)e_t$$

The equation:

$$1 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_q z^q = 0$$

can be used to determine invertibility. This can be thought of as the characteristic equation of the white noise terms.

Condition for invertibility of a *MA*(*q*) process

The process X defined by the equation:

$$X_t - \mu = e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \cdots + \beta_q e_{t-q}$$

is invertible if and only if the roots of the equation:

$$1 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_q z^q = 0$$

are all strictly greater than 1 in magnitude.

This is equivalent to saying that the value e_t can be written explicitly as a (convergent) sum of X values.

Again, λ is often used instead of z in the characteristic equation.

It follows that in the same way as a stationary autoregression can be thought of as a moving average of infinite order, so a moving average can be thought of as an autoregression of infinite order.

Question

Determine whether the process $X_t = 2 + e_t - 5e_{t-1} + 6e_{t-2}$ is invertible.

Solution

The equation $1 - 5\lambda + 6\lambda^2 = (1-2\lambda)(1-3\lambda) = 0$ has roots $1/3$ and $1/2$, so the process is not invertible.

Although there may be many moving average processes with the same ACF, at most one of them is invertible, since no two invertible processes have the same autocorrelation function. Moving average models fitted to data by statistical packages will always be invertible.

3.7 The autoregressive moving average process, *ARMA*(*p*,*q*)

A combination of the moving average and autoregressive models, an ARMA model includes direct dependence of X_t on both past values of X and present and past values of e .

The defining equation is:

$$X_t = \mu + \alpha_1(X_{t-1} - \mu) + \cdots + \alpha_p(X_{t-p} - \mu) + e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q}$$

or, in backwards shift operator notation:

$$(1 - \alpha_1 B - \cdots - \alpha_p B^p)(X - \mu) = (1 + \beta_1 B + \cdots + \beta_q B^q)e$$

This might also be written:

$$\phi(B)(X_n - \mu) = \theta(B)e_n$$

where $\theta(B)$ and $\phi(B)$ are polynomials of degrees *q* and *p*, respectively.

If $\theta(B)$ and $\phi(B)$ have any factors (*ie* roots) in common, then the defining relation could be simplified. For example, we may have a stationary $ARMA(1,1)$ process defined by $\left(1-\alpha B\right)X_n = \left(1-\alpha B\right)e_n$ with $|\alpha|<1$. Since the factors $\left(1-\alpha B\right)$ are invertible, we can multiply both sides of the equation by $\left(1-\alpha B\right)^{-1}$ to see that $X_n = e_n$. So this process would actually be classified as $ARMA(0,0)$.

In general, it would be assumed that the polynomials $\theta(B)$ and $\phi(B)$ have no common roots.

Autoregressive and moving average processes are special cases of ARMA processes. $AR(p)$ is the same as $ARMA(p,0)$. $MA(q)$ is the same as $ARMA(0,q)$.

To check the stationarity of an ARMA process, we just need to examine the autoregressive part. The moving average part (which involves the white noise terms) is always stationary. The test is the same as for an autoregressive process – we need to determine the roots of the characteristic equation formed by the X terms. The process is stationary if and only if all the roots are strictly greater than 1 in magnitude.

Similarly, we can check for invertibility by examining the roots of the characteristic equation that is obtained from the white noise terms. The process is invertible if and only if all the roots are strictly greater than 1 in magnitude.

Neither the ACF nor the PACF of the ARMA process eventually becomes equal to zero. This makes it more difficult to identify an ARMA model than either a pure autoregression or a pure moving average.

Theoretically, both the ACF and PACF of a stationary ARMA process will tend towards 0 for large lags, but neither will have a cut off property.

It is possible to calculate the ACF by a method similar to the method employing the Yule-Walker equations for the ACF of an autoregression.

We will show that the autocorrelation function of the stationary zero-mean $ARMA(1,1)$ process:

$$X_t = \alpha X_{t-1} + e_t + \beta e_{t-1} \qquad (13.7)$$

is given by:

$$\rho_1 = \frac{(1+\alpha\beta)(\alpha+\beta)}{(1+\beta^2+2\alpha\beta)}$$

$$\rho_k = \alpha^{k-1}\rho_1, \quad k = 2,3,\ldots$$

These results can be obtained from the formula for ρ_k given on page 40 of the *Tables.*

Figure 13.1 in Section 2.5 shows the ACF and PACF values of such a process with $\alpha = 0.7$ and $\beta = 0.5$.

Before we tackle the Yule-Walker equations, we need a couple of preliminary results.

Using equation (13.7), *ie*:

$$X_t = \alpha X_{t-1} + e_t + \beta e_{t-1}$$

we have:

$$\begin{aligned}\text{cov}(X_t, e_t) &= \alpha\,\text{cov}(X_{t-1}, e_t) + \text{cov}(e_t, e_t) + \beta\,\text{cov}(e_{t-1}, e_t)\\ &= \sigma^2\end{aligned}$$

since e_t is independent of both e_{t-1} and X_{t-1}.

Similarly:

$$\begin{aligned}\text{cov}(X_t, e_{t-1}) &= \alpha\,\text{cov}(X_{t-1}, e_{t-1}) + \text{cov}(e_t, e_{t-1}) + \beta\,\text{cov}(e_{t-1}, e_{t-1})\\ &= (\alpha+\beta)\sigma^2\end{aligned}$$

This enables us to deduce the autocovariance function of X. Again from (13.7):

$$\text{cov}(X_t, X_t) = \alpha\,\text{cov}(X_{t-1}, X_t) + \text{cov}(e_t, X_t) + \beta\,\text{cov}(e_{t-1}, X_t)$$

$$\text{cov}(X_t, X_{t-1}) = \alpha\,\text{cov}(X_{t-1}, X_{t-1}) + \text{cov}(e_t, X_{t-1}) + \beta\,\text{cov}(e_{t-1}, X_{t-1})$$

and, for $k > 1$:

$$\text{cov}(X_t, X_{t-k}) = \alpha\,\text{cov}(X_{t-1}, X_{t-k}) + \text{cov}(e_t, X_{t-k}) + \beta\,\text{cov}(e_{t-1}, X_{t-k})$$

So:

$$\gamma_0 = \alpha\gamma_1 + (1 + \alpha\beta + \beta^2)\sigma^2$$

$$\gamma_1 = \alpha\gamma_0 + \beta\sigma^2$$

$$\gamma_k = \alpha\gamma_{k-1}$$

The solution is:

$$\gamma_0 = \frac{1 + 2\alpha\beta + \beta^2}{1-\alpha^2}\sigma^2$$

$$\gamma_1 = \frac{(\alpha+\beta)(1+\alpha\beta)}{1-\alpha^2}\sigma^2$$

$$\gamma_k = \alpha^{k-1}\gamma_1 \qquad k = 2, 3, \ldots$$

assuming that the process is stationary, *ie* that $|\alpha| < 1$.

Hence:

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{(1+\alpha\beta)(\alpha+\beta)}{(1+2\alpha\beta+\beta^2)}$$

and:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\alpha^{k-1}\gamma_1}{\gamma_0} = \alpha^{k-1}\rho_1 \, , \; k=2,3,\ldots$$

Question

Show that the process $12X_t = 10X_{t-1} - 2X_{t-2} + 12e_t - 11e_{t-1} + 2e_{t-2}$ is both stationary and invertible.

Solution

We start by rewriting the process so that all the $X's$ are on the same side and all the $e's$ are on the same side:

$$12X_t - 10X_{t-1} + 2X_{t-2} = 12e_t - 11e_{t-1} + 2e_{t-2}$$

The characteristic equation of the AR part is $12 - 10\lambda + 2\lambda^2 = (2\lambda - 4)(\lambda - 3) = 0$, which has roots 2 and 3. The process is therefore stationary.

The characteristic equation of the MA part is $12 - 11\lambda + 2\lambda^2 = (2\lambda - 3)(\lambda - 4) = 0$. The roots of this equation are 1.5 and 4. The process is therefore invertible.

3.8 Modelling non-stationary processes: the ARIMA model

This is the most general class of models we will consider. They lie at the heart of the Box-Jenkins approach to modelling time series. In order to understand the rationale underlying the definition, it will be useful to give a brief introduction to the Box-Jenkins method; a more detailed discussion will be given in Chapter 14.

Suppose we are given some time series data x_n, where *n* varies over some finite range. If we want to model the data, then we would expect to take sample statistics, in particular the sample autocorrelation function, sample partial autocorrelation function and sample mean; these will be discussed in more detail in Chapter 14. The modelling process would then involve finding a stochastic process with similar characteristics. In the Box-Jenkins approach, the model is picked from the $ARMA(p,q)$ class. However, the theoretical counterparts of the autocorrelation and partial autocorrelation functions are only defined for stationary series. The upshot of this is that we can only *directly* apply these methods if the original data values are stationary.

However, we can get around this problem as follows. First transform the data into a stationary form, which we will discuss in a moment. We can then model this stationary series as suggested above. Finally, we carry out the inverse transform on our model to obtain a model for the original series. The question remains as to what we mean by 'transform' and 'inverse transform'.

The backward difference operator can turn a non-stationary series into a stationary one. For example, a random walk:

$$X_t = X_{t-1} + e_t$$

is non-stationary (as its characteristic equation is $1-\lambda = 0$, which has a root of 1). However, the difference:

$$\nabla X_t = X_t - X_{t-1} = e_t$$

is just white noise, which *is* stationary.

For the moment we will assume that it's possible to turn the data set into a stationary series by repeated differencing. We may have to difference the series several times, the specific number usually being denoted by *d*.

Now assuming we've transformed our data into stationary form by differencing, we can model this series using a *stationary* $ARMA(p,q)$ process. The final step is to reverse the differencing procedure to obtain a model of the original series. The inverse process of differencing is integration since we must sum the differences to obtain the original series.

Question

From a data set $x_0, x_1, x_2, \ldots, x_N$ the first order differences $w_i = x_i - x_{i-1}$ are calculated.

State the range of values of i and give an expression for x_j in terms of x_0 and the $w's$.

Solution

The values of i are $1,2,\ldots,N$ and:

$$x_j = x_0 + \sum_{i=1}^{j} w_i$$

In many applications the process being modelled cannot be assumed stationary, but can reasonably be fitted by a model with stationary increments, that is, if the first difference of X, $Y = \nabla X$, is itself a stationary process.

A process X is called an $ARIMA(p,1,q)$ process if X is non-stationary but the first difference of X is a stationary $ARMA(p,q)$ process.

We will now consider some examples.

Example 1

The simplest example of an ARIMA process is the random walk, $X_t = X_{t-1} + e_t$, which can be written $X_t = X_0 + \sum_{j=1}^{t} e_j$. The expectation of X_t is equal to $E\left[X_0\right]$ but the variance is $\text{var}\left(X_0\right) + t\sigma^2$, so that X is not itself stationary.

Here we are assuming (as we usually do) that the white noise process has zero mean.

The first difference, however, is given by:

$$Y_t = \nabla X_t = e_t$$

which certainly is stationary. Thus the random walk is an $ARIMA(0,1,0)$ process.

Example 2

Let Z_t denote the closing price of a share on day t. The evolution of Z is frequently described by the model:

$$Z_t = Z_{t-1}\exp\left(\mu + e_t\right)$$

By taking logarithms we see that this model is equivalent to an $I(1)$ model, since $Y_t = \ln Z_t$ satisfies the equation:

$$Y_t = \mu + Y_{t-1} + e_t$$

which is the defining equation of a random walk with drift because $Y_t = Y_0 + \mu t + \sum_{j=1}^{t} e_j$. The model is based on the assumption that the daily returns $\ln\left(Z_t / Z_{t-1}\right)$ are independent of the past prices $Z_0, Z_1, \ldots, Z_{t-1}$.

Example 3

The logarithm of the consumer price index can be described by the $ARIMA(1,1,0)$ model:

$$(1-B)\ln Q_t = \mu + \alpha\left[(1-B)\ln Q_{t-1} - \mu\right] + e_t$$

When analysing the behaviour of an $ARIMA(p,1,q)$ model, the standard technique is to look at the first difference of the process and to perform the kind of analysis which is suitable for an *ARMA* model. Once complete, this can be used to provide predictions for the original, undifferenced, process.

ARIMA(*p*,*d*,*q*) processes

In certain cases it may be considered desirable to continue beyond the first difference, if the process X is still not stationary after being differenced once. The notation extends in a natural way.

Definition of an *ARIMA* process

If X needs to be differenced at least d times in order to reduce it to stationarity and if the d th difference $Y = \nabla^d X$ is an $ARMA(p,q)$ process, then X is termed an $ARIMA(p,d,q)$ process.

In terms of the backwards shift operator, the equation of the $ARIMA(p,d,q)$ process is:

$$(1-\alpha_1 B - \cdots - \alpha_p B^p)(1-B)^d (X-\mu) = (1+\beta_1 B + \cdots + \beta_q B^q)e$$

An $ARIMA(p,d,q)$ process is $I(d)$. We can think of the classification $ARIMA(p,d,q)$ as:

$$AR(p)I(d)MA(q)$$

We now consider an example where we classify a time series as an *ARIMA* process.

To identify the values of p , d and q for which X is an $ARIMA(p,d,q)$ process, where:

$$X_t = 0.6X_{t-1} + 0.3X_{t-2} + 0.1X_{t-3} + e_t - 0.25e_{t-1}$$

we can write the equation in terms of the backwards shift operator:

$$(1-0.6B-0.3B^2-0.1B^3)X = (1-0.25B)e$$

We now check whether the polynomial on the left-hand side is divisible by 1 – *B*; if so, factorise it out. We continue to do this until the remaining polynomial is not divisible by $1-B$.

$$(1-B)(1+0.4B+0.1B^2)X = (1-0.25B)e$$

The model can now be seen to be $ARIMA(2,1,1)$.

We should also check that the roots of the characteristic polynomial of ∇X_t , *ie* $1+0.4\lambda+0.1\lambda^2$, are both strictly greater than 1 in magnitude. In fact, the roots are $-2 \pm i\sqrt{6}$. The magnitude of the complex number $a+bi$ is $\sqrt{a^2+b^2}$. So the magnitude of $-2+i\sqrt{6}$ is:

$$\sqrt{(-2)^2+\left(\sqrt{6}\right)^2} = \sqrt{4+6} = \sqrt{10}$$

The magnitude of $-2-i\sqrt{6}$ is also $\sqrt{10}$. So both roots are strictly greater than 1 in magnitude.

Differencing once removes the factor of $(1-B)$. Hence, the process ∇X_t is stationary, as required.

Alternatively, we could write the equation for the process as:

$$X_t - 0.6X_{t-1} - 0.3X_{t-2} - 0.1X_{t-3} = e_t - 0.25e_{t-1}$$

The characteristic equation of the AR part is:

$$1-0.6\lambda-0.3\lambda^2-0.1\lambda^3=0$$

There is no simple formula for solving cubic equations, so we should start by checking whether 1 is a root of this equation. Setting $\lambda=1$, we see that the left-hand side is:

$$1-0.6-0.3-0.1=0$$

So 1 is a root, and hence $(\lambda-1)$ is a factor. The characteristic polynomial can therefore be written in the form:

$$(\lambda-1)(a\lambda^2+b\lambda+c)$$

We can determine the values of a, b and c by comparing coefficients, by long division of polynomials, or by synthetic division. We find that $a=0.1$, $b=0.4$ and $c=1$, and the roots of the equation $0.1\lambda^2+0.4\lambda+1=0$ are $-2\pm i\sqrt{6}$ as stated above, and these are strictly greater than 1 in magnitude. Since the characteristic equation has one root of 1 and the other roots are strictly greater than 1 in magnitude, differencing once will give us a stationary process. So $d=1$. There are two other roots, so $p=2$. In addition, $q=1$ since the moving average part is of order 1. Hence the process is $ARIMA(2,1,1)$.

Another alternative is to write the defining equation in terms of the differences.

The equation:

$$X_t=0.6X_{t-1}+0.3X_{t-2}+0.1X_{t-3}+e_t-0.25e_{t-1}$$

can be rearranged as:

$$X_t-0.6X_{t-1}-0.3X_{t-2}-0.1X_{t-3}=e_t-0.25e_{t-1}$$

or:

$$(X_t-X_{t-1})+0.4(X_{t-1}-X_{t-2})+0.1(X_{t-2}-X_{t-3})=e_t-0.25e_{t-1}$$

or:

$$\nabla X_t+0.4\nabla X_{t-1}+0.1\nabla X_{t-2}=e_t-0.25e_{t-1}$$

The characteristic equation formed by the ∇X terms is:

$$0.1\lambda^2+0.4\lambda+1=0$$

As we have already seen, the roots of this equation are $-2\pm i\sqrt{6}$, and these both have a magnitude of $\sqrt{10}$. So ∇X is stationary $ARMA(2,1)$ and hence X is $ARIMA(2,1,1)$.

ARIMA models play a central role in the Box-Jenkins methodology, which aims to provide a consistent and unified framework for analysis and prediction using time series models. (See Section 3.1 of Chapter 14.)

3.9 The Markov property

As we saw in Chapter 1, if the future development of a process can be predicted from its present state alone, without any reference to its past history, it possesses the Markov property. Stated precisely this reads:

$$P\left[X_t \in A \mid X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n, X_s = x\right] = P\left[X_t \in A \mid X_s = x\right]$$

for all times $s_1 < s_2 < \cdots < s_n < s < t$, all states $x_1, x_2, \ldots, x_n$ and x in S and all subsets A of S.

Recall from Section 1 that a time series process is defined as having a continuous state space. The necessity to work with subsets $A \subseteq S$ (rather than just having $X_t = a \in S$) is to cover these continuous state space cases. For these the probability that X_t takes on a particular value is zero. We therefore need to work with probabilities of X_t lying in some interval of *S*, or more generally in some subset.

A first-order autoregressive process possesses the Markov property, since the conditional distribution of X_{n+1} given all previous X_t depends only on X_n. This property does not apply, however, to higher-order autoregressive models.

Suppose X is an $AR(2)$. X does not possess the Markov property, since the conditional distribution of X_{n+1} given the history of X up until time n depends on X_{n-1} as well as on X_n. But let us define a vector-valued process $\underline{Y}$ by $\underline{Y}_t = \left(X_t, X_{t-1}\right)^T$.

Vector-valued or multivariate processes will be studied in more detail in Chapter 14. The superscript '*T*' here means transpose, *ie* we are thinking of $\underline{Y}$ as a column vector.

Given the whole history of the process X up until time n, the distribution of $\underline{Y}_{n+1}$ depends only on the values of X_n and X_{n-1} – in other words, on the value of $\underline{Y}_n$. This means that $\underline{Y}$ possesses the Markov property.

We can illustrate this as follows. The $AR(2)$ process:

$$X_n = \alpha_1 X_{n-1} + \alpha_2 X_{n-2} + e_n$$

can be written in matrix form as follows:

$$\begin{pmatrix} X_n \\ X_{n-1} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X_{n-1} \\ X_{n-2} \end{pmatrix} + \begin{pmatrix} e_n \\ 0 \end{pmatrix}$$

ie:

$$\underline{Y}_n = \begin{pmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{pmatrix} \underline{Y}_{n-1} + \begin{pmatrix} e_n \\ 0 \end{pmatrix}$$

We can see from the equation immediately above that future values of the process $\underline{Y}$ depend only on the current value, and not on the past history of the process.

Similarly, the $AR(3)$ process:

$$X_n = \alpha_1 X_{n-1} + \alpha_2 X_{n-2} + \alpha_3 X_{n-3} + e_n$$

can be written in matrix form as follows:

$$\begin{pmatrix} X_n \\ X_{n-1} \\ X_{n-2} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{n-1} \\ X_{n-2} \\ X_{n-3} \end{pmatrix} + \begin{pmatrix} e_n \\ 0 \\ 0 \end{pmatrix}$$

In general an $AR(p)$ does not possess the Markov property (for $p > 1$) but we may define a vector-valued process $\underline{Y}_t = (X_t, X_{t-1}, \ldots, X_{t-p+1})^T$ which does.

You will recall from Chapter 1 that a random walk possesses the Markov property. The discussion of the Markov property for autoregressions can be extended to include some *ARIMA* processes such as the random walk, which has already been shown to be an $ARIMA(0,1,0)$ process: an $ARIMA(p,d,0)$ process does not possess the Markov property (for $p+d > 1$) but we may define a vector-valued process $\underline{Y}_t = (X_t, X_{t-1}, \ldots, X_{t-p-d+1})^T$ which does.

A moving average, or more generally an $ARIMA(p,d,q)$ process with $q > 0$, can never be Markov, since knowledge of the value of X_n, or of any finite collection $(X_n, X_{n-1}, \ldots, X_{n-q+1})^T$ will never be enough to deduce the value of e_n, on which the distribution of X_{n+1} depends. Since a moving average has been shown to be equivalent to an autoregression of infinite order, and since a pth order autoregression needs to be expressed as a p-dimensional vector in order to possess the Markov property, a moving average has no similar finite-dimensional Markov representation.

Question

Let $X_n = e_n + e_{n-1}$ be an $MA(1)$ process where $e_n \sim N(0,1)$. Determine whether this process has the Markov property.

Solution

The process is not Markov. We can see this intuitively as follows (although we could explicitly calculate some conditional probabilities to give a more formal proof).

The problem is really whether knowing the value of X_{n-2} in addition to X_{n-1} will help in predicting the value of X_n. Suppose, for example, that $X_{n-1} = 0$, *ie* $e_{n-1} + e_{n-2} = 0$. We cannot deduce the values of the *e*'s themselves.

However, the value X_{n-2} will give us extra information as to the likely value of e_{n-2}, which in turn gives us extra information about the value of e_{n-1}, which in turn gives us information about the value of X_n.

For example, X_{n-2} might be very large and positive, which would indicate that e_{n-2} is more likely to be positive, which would indicate that e_{n-1} is more likely to be negative, which would indicate that X_n is more likely to be negative.

Chapter 13 Summary

Univariate time series

A univariate time series is a sequence of observations $\{X_t\}$ recorded at regular intervals. The state space is continuous but the time set is discrete.

Such series may follow a pattern to some extent, for example possessing a trend or seasonal component, as well as having random factors. The aim is to construct a model to fit a set of past data in order to forecast future values of the series.

Stationarity

It is easier (more efficient) to construct a model if the time series is stationary.

A time series is said to be stationary, or strictly stationary, if the joint distributions of $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ and $\{X_{t_1+k}, X_{t_2+k}, \ldots, X_{t_n+k}\}$ are identical for all $t_1, t_2, \ldots, t_n$ and $t_1+k, t_2+k, \ldots, t_n+k$ in the time set J and all integers n. This means that the statistical properties of the process remain unchanged as time elapses.

For most cases of interest to us, it is enough for the time series to be weakly stationary. This is the case if the time series has a constant mean, constant variance and the covariance depends only on the lag.

We are also interested primarily in purely indeterministic processes. This means knowledge of the values $X_1, X_2, \ldots, X_n$ is progressively less useful at predicting the value of X_N as $N \to \infty$.

We redefine the term 'stationary' to mean weakly stationary and purely indeterministic.

Importantly, the time series consisting of a sequence of white noise terms is weakly stationary and purely indeterministic. White noise is defined as a sequence of uncorrelated random variables. In time series, we assume that white noise has zero mean. This series has constant mean and variance, and covariance that depends only on whether the lag is zero or non-zero. It is purely indeterministic due to its random nature.

A time series process X is stationary if we can write it as a convergent sum of white noise terms.

It can be shown that this is equivalent to saying that the roots of the characteristic polynomial of the X terms are all greater than 1 in magnitude. For example, if the time series is defined by $X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q}$ then the characteristic polynomial is $1 - \alpha_1 \lambda - \cdots - \alpha_p \lambda^p$. To find the roots, we set this equal to zero and solve.

Invertibility

A time series process X is invertible if we can write the white noise term e_t as a convergent sum of the X terms.

It can be shown that this is equivalent to saying that the roots of the characteristic polynomial of the e terms are all greater than 1 in magnitude. For example, if the time series is given by $X_t = \alpha_1 X_{t-1} + \dots + \alpha_p X_{t-p} + e_t + \beta_1 e_{t-1} + \dots + \beta_q e_{t-q}$ then the characteristic polynomial of the e terms is $1 + \beta_1 \lambda + \dots + \beta_q \lambda^q$. To find the roots, we set this equal to zero and solve.

Invertibility is a desirable characteristic since it enables us to calculate the residual terms and hence analyse the goodness of fit of a particular model.

Backward shift and difference operators

The backwards shift operator, *B*, is defined as follows:

$$BX_t = X_{t-1}$$

$B\mu = \mu$ where μ is a constant.

The backward shift operator can be applied repeatedly so that $B^k X_t = X_{t-k}$.

The difference operator, ∇, is defined as follows:

$$\nabla X_t = X_t - X_{t-1}$$

The difference operator can be applied repeatedly so that $\nabla^k X_t = \nabla^{k-1} X_t - \nabla^{k-1} X_{t-1}$.

The difference operator and backward shift operator are linked by $\nabla = 1 - B$.

Integrated processes

A time series process X is integrated of order *d*, denoted $I(d)$, if its d th difference is stationary. So X is $I(0)$ if the process X itself is stationary, and X is $I(1)$ if ∇X is stationary.

Autocovariance function

If the time series X is stationary, then its covariance depends only on the lag *k*. In this case, we define the autocovariance function as $\gamma_k = \operatorname{cov}\left(X_t, X_{t+k}\right)$.

Autocorrelation function

If the time series X is stationary, then its autocorrelation function (ACF) is:

$$\rho_k = corr\left(X_t, X_{t+k}\right) = \frac{\gamma_k}{\gamma_0}$$

For purely indeterministic time series processes X (where the past values of X become less useful the further into the future we look), $\rho_k \to 0$ as $k \to \infty$.

Partial autocorrelation function

The partial autocorrelation function (PACF), $\{\phi_k : k = 1,2,\ldots\}$, is defined as the conditional correlation of X_{t+k} with X_t given $X_{t+1},\ldots,X_{t+k-1}$. Formulae for the partial autocorrelation are given on page 40 of the *Tables*.

Moving average processes

A time series process X is said to be $MA(q)$ (or moving average of order q) if it can be written as a weighted average of the past q white noise terms (plus a new white noise term):

$$X_t = e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q} \qquad \text{(zero mean)}$$

$$X_t = \mu + e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q} \qquad \text{(mean } \mu\text{)}$$

Features of $MA(q)$ time series include:

- always stationary (as a finite sum of white noise)
- invertible if all the roots of the characteristic equation $1 + \beta_1\lambda + \cdots + \beta_q\lambda^q = 0$ are strictly greater than 1 in magnitude
- not Markov
- ρ_k cuts off for $k > q$
- ϕ_k decays geometrically as $k \to \infty$.

Autoregressive processes

A time series process X is said to be $AR(p)$ (or autoregressive of order p) if it depends on the past p terms of the series (plus a new white noise term):

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + e_t \qquad \text{(zero mean)}$$

$$X_t = \mu + \alpha_1(X_{t-1} - \mu) + \cdots + \alpha_p(X_{t-p} - \mu) + e_t \qquad \text{(mean } \mu\text{)}$$

Features of $AR(p)$ time series include:

- stationary if all the roots of the characteristic equation $1-\alpha_1\lambda-\cdots-\alpha_p\lambda^p=0$ are strictly greater than 1 in magnitude
- always invertible
- only Markov if $p=1$
- ρ_k decays geometrically as $k\to\infty$
- ϕ_k cuts off for $k>p$.

ARMA processes

A time series process X is said to be $ARMA(p,q)$ (or autoregressive moving average of order *p*, *q*) if it is the sum of an $AR(p)$ and an $MA(q)$ time series:

$$X_t=\alpha_1X_{t-1}+\cdots+\alpha_pX_{t-p}+e_t \\ +\beta_1e_{t-1}+\cdots+\beta_qe_{t-q} \qquad \text{(zero mean)}$$

$$X_t=\mu+\alpha_1(X_{t-1}-\mu)+\cdots+\alpha_p(X_{t-p}-\mu)+e_t \\ +\beta_1e_{t-1}+\cdots+\beta_qe_{t-q} \qquad \text{(mean } \mu\text{)}$$

Features of $ARMA(p,q)$ time series include:

- stationary if all the roots of the characteristic equation of the AR part, *ie* $1-\alpha_1\lambda-\cdots-\alpha_p\lambda^p=0$, are strictly greater than 1 in magnitude
- invertible if all the roots of the characteristic equation of the MA part, *ie* $1+\beta_1\lambda+\cdots+\beta_q\lambda^q=0$, are strictly greater than 1 in magnitude
- only Markov if $p=1$, $q=0$
- ρ_k decays as $k\to\infty$
- ϕ_k decays as $k\to\infty$.

ARIMA processes

A time series process X is said to be $ARIMA(p,d,q)$ (or autoregressive integrated moving average of order *p*, *d*, *q*) if the d th difference $Y_t=\nabla^dX_t$ is a *stationary* $ARMA(p,q)$ time series process.

Chapter 13 Practice Questions

13.1 Give an expression for $2X_t - 5X_{t-1} + 4X_{t-2} - X_{t-3}$ in terms of second order differences.

13.2 A time series is defined by the relationship $X_t = X_{t-1} + Z_t$, where the Z_t are IID $N(0,\sigma^2)$ random variables.

Determine the relationship between $\text{var}(X_t)$ and $\text{var}(X_{t-1})$, and hence comment on the stationarity of this series.

13.3 Determine whether the process $X_n = X_{n-1} - \frac{1}{2}X_{n-2} + e_n$ is stationary.

13.4 An autoregressive stationary time series W_t is defined by the relationship:

$$W_t = 0.6W_{t-1} + 0.4W_{t-2} - 0.1W_{t-3} + Z_t$$

for integer times *t*, where $\{Z_t\}$ represents a set of uncorrelated random variables with mean 0 and variance σ^2.

(i) Explain why $\text{cov}(W_t, Z_{t+1}) = 0$ and $\text{cov}(W_{t-1}, W_t) = \text{cov}(W_{t-1}, W_{t-2})$.

(ii) By considering $\text{cov}(W_t, W_{t-k})$ when $k = 0,1,2,3$, write down a set of four equations relating the values of the autocovariance function γ_k at lags $k = 0,1,2,3$.

(iii) Solve the four equations in part (ii) to find both the autocovariance function and the autocorrelation function for lags 0, 1, 2 and 3.

13.5 Calculate the autocorrelation function of the process $X_n = 1 + e_n - 5e_{n-1} + 6e_{n-2}$.

13.6 (i) The first differences of a time series X can be modelled by the process:

$$\nabla X_n = 0.5\nabla X_{n-1} + e_n$$

Determine the model for X_n.

(ii) Show that the process X_n is non-stationary.

13.7 (i) Show that the relationship $Y_t = 0.7Y_{t-1} + 0.3Y_{t-2} + Z_t + 0.7Z_{t-1}$ (where the Z's denote white noise) defines an $ARIMA(1,1,1)$ process.

(ii) Show carefully that the relationship $S_t = 1.5S_{t-1} + 0.5S_{t-3} + Z_t + 0.5Z_{t-1}$ *cannot* be expressed as an $ARIMA(1,2,1)$ process.

13.8 Consider the process with defining equation:

$$X_n = 5X_{n-1} - 4X_{n-2} + X_{n-3} + e_n$$

Write this as a vector process that possesses the Markov property.

13.9 Let $X_n = e_n + e_{n-2}$ be an MA(2) process where $e_n \sim N(0,1)$.

(i) Calculate $P\left(X_n \geq 0 \middle| X_{n-1} \leq 0\right)$.

(ii) Compare your answer to (i) with $P\left(X_n \geq 0 \middle| X_{n-1} \leq 0, X_{n-2} \leq 0\right)$ and hence comment on whether the process is Markov.

13.10 Calculate the values of ρ_1 and ρ_2, the autocorrelation function at lags 1 and 2, for the stationary $AR(2)$ process defined by the equation:

Exam style

$$X_n = -0.8X_{n-1} + 0.1X_{n-2} + \varepsilon_n$$ [4]

13.11 Consider the time series model defined by:

Exam style

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \varepsilon_t$$

where $\{\varepsilon_t\}$ is white noise.

(i) Show that the autocorrelation coefficient with lag 1 for this process is:

$$\rho_1 = \frac{\alpha_1 + \alpha_2\alpha_3}{1 - \alpha_2 - \alpha_1\alpha_3 - \alpha_3^2}$$ [3]

(ii) Consider the case where $\alpha_1 = \alpha_2 = \alpha_3 = 0.2$.

(a) Comment on the stationarity of this model.

Hint: $5 - x - x^2 - x^3 \approx (1.278 - x)(3.912 + 2.278x + x^2)$

(b) Calculate ρ_1 and ρ_2.

(c) Calculate the partial autocorrelation coefficients ϕ_1 and ϕ_2.

(d) Sketch correlograms of the autocorrelation function and the partial autocorrelation function. (You are not required to calculate the coefficients for higher lags.) [9]

[Total 12]

13.12 Exam style

$\{X_t\}$ is a stationary $ARMA(1,2)$ time series defined at integer times by the relationship:

$$X_t = \alpha X_{t-1} + Z_t + \beta Z_{t-2}$$

where α, β are constants and $\{Z_t\}$ is a purely random process with mean 0 and constant variance σ^2.

(i) Define the term 'weakly stationary process'. [2]

(ii) Assuming that the above process has a very long history, state the conditions on α and β needed to ensure that it is:

(a) stationary

(b) invertible. [2]

(iii) Show that for any integer s:

$$\text{cov}(X_s, Z_s) = \sigma^2 \quad \text{cov}(X_s, Z_{s-1}) = \alpha\sigma^2 \quad \text{cov}(X_s, Z_{s-2}) = (\alpha^2 + \beta)\sigma^2$$ [3]

(iv) (a) By considering $\text{cov}(X_t, X_t)$, $\text{cov}(X_t, X_{t-1})$ and $\text{cov}(X_t, X_{t-2})$, write down three equations involving γ_0, γ_1 and γ_2.

(b) Hence find expressions for γ_0, γ_1 and γ_2 in terms of the parameters α, β and σ^2. [7]

(v) (a) Calculate the values of ρ_0, ρ_1, ρ_2 and ρ_3 in the case where $\alpha = -0.4$ and $\beta = -0.9$.

(b) Hence sketch a graph of the autocorrelation function ρ_k for lags $k = 0,1,2,\ldots,10$ in this case. [5]

[Total 19]

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 13 Solutions

13.1 Using the difference operator, we can write:

$$\begin{aligned}2X_t - 5X_{t-1} + 4X_{t-2} - X_{t-3} &= 2(X_t - X_{t-1}) - 3(X_{t-1} - X_{t-2}) + (X_{t-2} - X_{t-3})\\ &= 2\nabla X_t - 3\nabla X_{t-1} + \nabla X_{t-2}\\ &= 2(\nabla X_t - \nabla X_{t-1}) - (\nabla X_{t-1} - \nabla X_{t-2})\\ &= 2\nabla^2 X_t - \nabla^2 X_{t-1}\end{aligned}$$

13.2 The series is defined by the equation:

$$X_t = X_{t-1} + Z_t$$

If we take variances of both sides, we get:

$$\text{var}(X_t) = \text{var}(X_{t-1}) + \sigma^2$$

Here we have used the fact that Z_t and X_{t-1} are uncorrelated. This is because X_{t-1} depends only on the past values of X and Z (ie values for times up to $t-1$), whereas Z_t is a new random value at time t that is not influenced in any way by these earlier values.

This implies that $\text{var}(X_t) > \text{var}(X_{t-1})$, which contradicts the stationarity requirement, which requires (amongst other things) that the variance is independent of time. So the series cannot be stationary.

13.3 The characteristic equation of the AR part is:

$$1 - \lambda + 0.5\lambda^2 = 0$$

The roots of this equation are:

$$\lambda = \frac{-(-1) \pm \sqrt{(-1)^2 - 4 \times 0.5 \times 1}}{2 \times 0.5} = \frac{1 \pm \sqrt{-1}}{1} = 1 \pm i$$

Since $|1 \pm i| = \sqrt{2} > 1$, the process is stationary.

13.4 (i) W_t is only dependent on past Z's, which are not correlated with the future Z's. Therefore $\text{cov}\left(W_t, Z_{t+1}\right) = 0$. Also, since W_t is stationary, only the lag matters so the second equation holds.

(ii) If we take the covariance of both sides of the defining equation with W_t, W_{t-1}, W_{t-2} and W_{t-3} in turn, then we obtain the four equations:

$$\gamma_0 = 0.6\gamma_1 + 0.4\gamma_2 - 0.1\gamma_3 + \sigma^2 \qquad (1)$$

$$\gamma_1 = 0.6\gamma_0 + 0.4\gamma_1 - 0.1\gamma_2 \qquad (2)$$

$$\gamma_2 = 0.6\gamma_1 + 0.4\gamma_0 - 0.1\gamma_1 \qquad (3)$$

$$\gamma_3 = 0.6\gamma_2 + 0.4\gamma_1 - 0.1\gamma_0 \qquad (4)$$

The σ^2 term in the first equation comes from:

$$\text{cov}(W_t, Z_t) = \text{cov}(0.6W_{t-1} + 0.4W_{t-2} - 0.1W_{t-3} + Z_t, Z_t) = \text{cov}(Z_t, Z_t) = \sigma^2$$

(iii) Equation (2) gives $\gamma_2 = 6\gamma_0 - 6\gamma_1$.

Likewise, from Equation (3) we get $\gamma_2 = 0.4\gamma_0 + 0.5\gamma_1$.

Equating these two expressions for γ_2 gives $\gamma_1 = \frac{56}{65}\gamma_0$ and hence $\gamma_2 = \frac{54}{65}\gamma_0$.

Equation (4) can then be applied to get $\gamma_3 = \frac{483}{650}\gamma_0$.

Finally Equation (1) can be applied to show that $\sigma^2 = 0.22508\gamma_0$.

Rearranging these expressions in terms of σ^2 we obtain the first four autocovariance values:

$$\gamma_0 = 4.4429\sigma^2, \quad \gamma_1 = 3.8278\sigma^2, \quad \gamma_2 = 3.6910\sigma^2, \quad \gamma_3 = 3.3014\sigma^2$$

The autocorrelations are then obtained by dividing by γ_0:

$$\rho_0 = 1, \quad \rho_1 = 0.862, \quad \rho_2 = 0.831, \quad \rho_3 = 0.743$$

13.5 The autocovariance function is:

$$\text{cov}(X_n, X_{n+k}) = \text{cov}(e_n - 5e_{n-1} + 6e_{n-2}, e_{n+k} - 5e_{n+k-1} + 6e_{n+k-2})$$

$$= \begin{cases} 62 & k = 0 \\ -35 & |k| = 1 \\ 6 & |k| = 2 \\ 0 & |k| > 2 \end{cases}$$

For example:

$$\begin{aligned}\gamma_0 &= \text{cov}(X_n, X_n)\\ &= \text{cov}(e_n - 5e_{n-1} + 6e_{n-2}, e_n - 5e_{n-1} + 6e_{n-2})\\ &= \text{cov}(e_n, e_n) + \text{cov}(-5e_{n-1}, -5e_{n-1}) + \text{cov}(6e_{n-2}, 6e_{n-2})\\ &= 1^2 + (-5)^2 + 6^2 = 62\end{aligned}$$

and similarly for the other values of k.

Since the autocovariance and autocorrelation functions are even, we can give the results for non-negative values of k. The autocorrelation function is:

$$\rho_0 = 1 \qquad \rho_{\pm 1} = -\frac{35}{62} \qquad \rho_{\pm 2} = \frac{6}{62} \qquad \rho_k = 0 \text{ for } |k| > 2$$

13.6 (i) ***Integrated process***

The series $\nabla X_n = 0.5\nabla X_{n-1} + e_n$, can be written:

$$X_n - X_{n-1} = 0.5\left(X_{n-1} - X_{n-2}\right) + e_n$$

Rearranging we end up with:

$$X_n = 1.5X_{n-1} - 0.5X_{n-2} + e_n$$

(ii) ***Show that X is non-stationary***

We have:

$$X_n - 1.5X_{n-1} + 0.5X_{n-2} = e_n$$

which is an $AR(2)$ process. Its characteristic equation is:

$$1 - 1.5\lambda + 0.5\lambda^2 = 0$$

This has roots $\dfrac{1.5 \pm \sqrt{2.25 - 2}}{1}$, *ie* 1 and 2. For stationarity, we require both roots to be strictly greater than 1 in magnitude. This is not the case here, so the process is not stationary.

13.7 (i) ***ARIMA*(1,1,1)**

The characteristic equation of the AR part is:

$$1 - 0.7\lambda - 0.3\lambda^2 = 0$$

The roots of this equation are 1 and $-\dfrac{10}{3}$. Both roots must be strictly greater than 1 in magnitude for the series to be stationary. This is not the case here.

Differencing once removes the root of 1 as we can see by subtracting Y_{t-1} from both sides of the defining equation. This gives:

$$\begin{aligned} Y_t - Y_{t-1} &= -0.3Y_{t-1} + 0.3Y_{t-2} + Z_t + 0.7Z_{t-1} \\ &= -0.3(Y_{t-1} - Y_{t-2}) + Z_t + 0.7Z_{t-1} \end{aligned}$$

This can be expressed using the difference operator as:

$$\nabla Y_t = -0.3\nabla Y_{t-1} + Z_t + 0.7Z_{t-1}$$

The characteristic equation of the AR part of the differenced series is:

$$1 + 0.3\lambda = 0$$

and the root of this equation is $-\dfrac{10}{3}$. So ∇Y is stationary $ARMA(1,1)$ and hence Y is $ARIMA(1,1,1)$.

(ii) ***Not an ARIMA*(1,2,1)**

The general form of an $ARIMA(1,2,1)$ model is:

$$\nabla^2 S_t = \alpha\nabla^2 S_{t-1} + Z_t + \beta Z_{t-1}$$

Expanding the ∇ 's gives:

$$S_t - 2S_{t-1} + S_{t-2} = \alpha(S_{t-1} - 2S_{t-2} + S_{t-3}) + Z_t + \beta Z_{t-1}$$

ie $$S_t = (2+\alpha)S_{t-1} - (1+2\alpha)S_{t-2} + \alpha S_{t-3} + Z_t + \beta Z_{t-1}$$

The relationship given is:

$$S_t = 1.5S_{t-1} + 0.5S_{t-3} + Z_t + 0.5Z_{t-1}$$

Comparing the coefficients of S_{t-1} implies that $\alpha = -0.5$. This correctly makes the S_{t-2} term vanish, but gives the wrong sign for S_{t-3}. So this definition is not consistent with an $ARIMA(1,2,1)$ model.

13.8 We can write $X_n = 5X_{n-1} - 4X_{n-2} + X_{n-3} + e_n$ in vector form as follows:

$$\begin{pmatrix} X_n \\ X_{n-1} \\ X_{n-2} \end{pmatrix} = \begin{pmatrix} 5 & -4 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{n-1} \\ X_{n-2} \\ X_{n-3} \end{pmatrix} + \begin{pmatrix} e_n \\ 0 \\ 0 \end{pmatrix}$$

This vector-valued process has the Markov property.

13.9 (i) If e_n and e_{n-2} are independent standard normal random variables, then $e_n + e_{n-2} \sim N(0,2)$. Hence:

$$\begin{aligned} P\left(X_n \geq 0 \middle| X_{n-1} \leq 0\right) &= P\left(e_n + e_{n-2} \geq 0 \middle| e_{n-1} + e_{n-3} \leq 0\right) \\ &= P\left(e_n + e_{n-2} \geq 0\right) \\ &= 0.5 \end{aligned}$$

(ii) We have:

$$\begin{aligned} P\left(X_n \geq 0 \middle| X_{n-1} \leq 0, X_{n-2} \leq 0\right) &= P\left(e_n + e_{n-2} \geq 0 \middle| e_{n-1} + e_{n-3} \leq 0, e_{n-2} + e_{n-4} \leq 0\right) \\ &= P\left(e_n + e_{n-2} \geq 0 \middle| e_{n-2} + e_{n-4} \leq 0\right) \\ &= \frac{P\left(e_n + e_{n-2} \geq 0 \cap e_{n-2} + e_{n-4} \leq 0\right)}{P\left(e_{n-2} + e_{n-4} \leq 0\right)} \\ &= \frac{P\left(e_n + e_{n-2} \geq 0 \cap e_{n-2} + e_{n-4} \leq 0\right)}{0.5} \end{aligned}$$

The process is Markov if $P\left(X_n \geq 0 \middle| X_{n-1} \leq 0\right) = P\left(X_n \geq 0 \middle| X_{n-1} \leq 0, X_{n-2} \leq 0\right)$. Hence, we require $P\left(e_n + e_{n-2} \geq 0 \cap e_{n-2} + e_{n-4} \leq 0\right) = 0.25$.

Equivalently, we require $P\left(e_n \geq -e_{n-2} \cap e_{n-4} \leq -e_{n-2}\right) = 0.25$.

For any given observation ε of e_{n-2}, this is:

$$\left(1 - \Phi(-\varepsilon)\right)\Phi\left(-\varepsilon\right) = 0.25$$

and the solution to this quadratic is at $\Phi\left(-\varepsilon\right) = 0.5$ or $e_{n-2} = 0$.

In other words, $P\left(X_n \geq 0 \middle| X_{n-1} \leq 0\right) = P\left(X_n \geq 0 \middle| X_{n-1} \leq 0, X_{n-2} \leq 0\right)$ only if $e_{n-2} = 0$. So the process cannot be Markov.

13.10 The defining equation is:

$$X_n = -0.8X_{n-1} + 0.1X_{n-2} + \varepsilon_n$$

The autocovariance at lag 1 is:

$$\begin{aligned} \gamma_1 &= \text{cov}(X_n, X_{n-1}) \\ &= \text{cov}(-0.8X_{n-1} + 0.1X_{n-2} + \varepsilon_n, X_{n-1}) \\ &= -0.8\,\text{cov}(X_{n-1}, X_{n-1}) + 0.1\,\text{cov}(X_{n-2}, X_{n-1}) + \text{cov}(\varepsilon_n, X_{n-1}) \\ &= -0.8\gamma_0 + 0.1\gamma_1 \end{aligned}$$ [1]

Dividing through by γ_0 and using the fact that $\rho_0 = 1$ gives:

$$\rho_1 = -0.8 + 0.1\rho_1 \quad \Rightarrow \rho_1 = -\tfrac{8}{9} \qquad [1]$$

Similarly, the autocovariance at lag 2 is:

$$\begin{aligned}\gamma_2 &= \text{cov}(X_n, X_{n-2})\\ &= \text{cov}(-0.8X_{n-1} + 0.1X_{n-2} + \varepsilon_n, X_{n-2})\\ &= -0.8\gamma_1 + 0.1\gamma_0 \qquad [1]\end{aligned}$$

Dividing through by γ_0 gives:

$$\rho_2 = -0.8\rho_1 + 0.1 = -0.8 \times (-\tfrac{8}{9}) + 0.1 = \tfrac{73}{90} \qquad [1]$$

The value of γ_0 is not required here.

13.11 (i) ***Autocorrelation coefficient***

Applying the Yule-Walker method by taking covariances with X_{t-1}, we get:

$$\text{cov}(X_t, X_{t-1}) = \text{cov}(\alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \varepsilon_t, X_{t-1})$$

ie: $\gamma_1 = \alpha_1\gamma_0 + \alpha_2\gamma_1 + \alpha_3\gamma_2 + 0$ [½]

Dividing by γ_0:

$$\rho_1 = \alpha_1 + \alpha_2\rho_1 + \alpha_3\rho_2$$

ie: $(1-\alpha_2)\rho_1 = \alpha_1 + \alpha_3\rho_2$ [½]

Then taking covariances with X_{t-2}:

$$\text{cov}(X_t, X_{t-2}) = \text{cov}(\alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \varepsilon_t, X_{t-2})$$

ie: $\gamma_2 = \alpha_1\gamma_1 + \alpha_2\gamma_0 + \alpha_3\gamma_1 + 0$ [½]

Dividing by γ_0:

$$\rho_2 = \alpha_1\rho_1 + \alpha_2 + \alpha_3\rho_1$$

ie: $\rho_2 = (\alpha_1 + \alpha_3)\rho_1 + \alpha_2$ [½]

If we eliminate ρ_2 from these two simultaneous equations, we get:

$$(1-\alpha_2)\rho_1 = \alpha_1 + \alpha_3[(\alpha_1 + \alpha_3)\rho_1 + \alpha_2] \qquad [½]$$

Rearranging:

$$\big[(1-\alpha_2) - \alpha_3(\alpha_1 + \alpha_3)\big]\rho_1 = \alpha_1 + \alpha_3\alpha_2$$

So:

$$\rho_1=\frac{\alpha_1+\alpha_3\alpha_2}{(1-\alpha_2)-\alpha_3(\alpha_1+\alpha_3)}=\frac{\alpha_1+\alpha_2\alpha_3}{1-\alpha_2-\alpha_1\alpha_3-\alpha_3^2}$$ [½]

(ii)(a) ***Stationarity***

In this case the model can be written as:

$$X_t-0.2X_{t-1}-0.2X_{t-2}-0.2X_{t-3}=\varepsilon_t$$

Stationarity is determined by the nature of the roots of the characteristic equation:

$$1-0.2\lambda-0.2\lambda^2-0.2\lambda^3=0$$ [1]

Multiplying by 5, this is:

$$5-\lambda-\lambda^2-\lambda^3=0$$

Using the hint given, we can write this as:

$$(1.278-\lambda)(3.912+2.278\lambda+\lambda^2)=0$$

$\Rightarrow$ $\lambda=1.278$ or $3.912+2.278\lambda+\lambda^2=0$ [½]

The quadratic equation has roots:

$$\lambda=\frac{-2.278\pm\sqrt{(2.278)^2-4(1)(3.912)}}{2}=-1.139\pm1.617i$$ [1]

The real root (1.278) and the two complex roots are all strictly greater than 1 in magnitude. (Recall that the magnitude of a complex root $a+bi$ is given by $\sqrt{a^2+b^2}$.)

So this model is stationary. [½]

(ii)(b) ***ACF at lags 1 and 2***

Using the formula derived in part (i):

$$\rho_1=\frac{\alpha_1+\alpha_2\alpha_3}{1-\alpha_2-\alpha_1\alpha_3-\alpha_3^2}=\frac{0.2+(0.2)^2}{1-0.2-(0.2)^2-(0.2)^2}=\frac{0.24}{0.72}=\frac{1}{3}$$ [½]

From the second Yule-Walker equation, derived in part (i), we also have:

$$\rho_2=(\alpha_1+\alpha_3)\rho_1+\alpha_2=(0.2+0.2)\times\frac{1}{3}+0.2=\frac{1}{3}$$ [½]

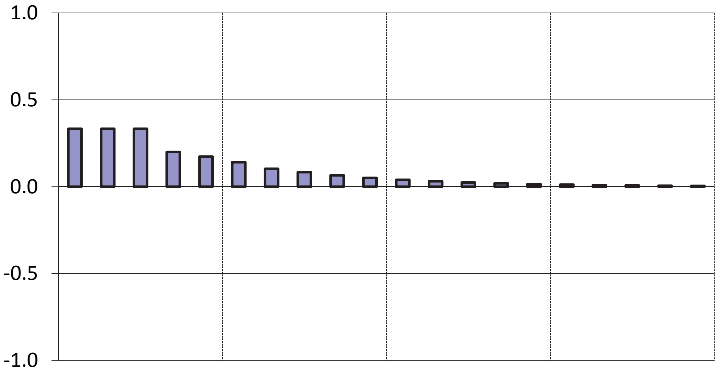
(ii)(c) ***PACFs at lags 1 and 2***

Using the formulae on page 40 of the *Tables*:

$$\phi_1 = \rho_1 = \frac{1}{3} \qquad \text{and} \qquad \phi_2 = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2} = \frac{1}{4}$$ [1]

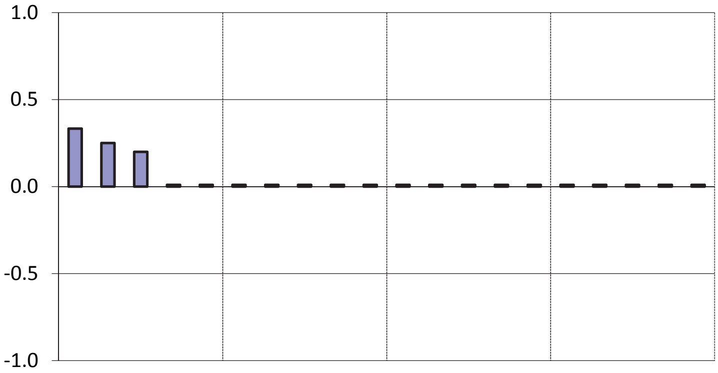
(ii)(d) ***Correlograms***

We've already calculated $\rho_1 = \frac{1}{3}$ and $\rho_2 = \frac{1}{3}$. Higher-lag autocorrelations will satisfy the Yule-Walker equation $\rho_k = 0.2(\rho_{k-1} + \rho_{k-2} + \rho_{k-3})$. So the values will tail off quite quickly to zero, always taking positive values. The correlogram (*ie* graph of ρ_k versus k) will look like this:



[2]

We've already calculated $\phi_1 = \frac{1}{3}$ and $\phi_2 = \frac{1}{4}$. Since the process is $AR(3)$, the partial autocorrelations will all equal zero from lag 4 onwards. So the partial correlogram (*ie* graph of ϕ_k versus k) will look like this:



[2]

13.12 (i) ***Weakly stationary process***

A weakly stationary process has constant mean, and the covariance is constant for each fixed lag. The variance (a special case of the covariance) is also constant. [2]

(ii)(a) ***Condition for stationarity***

The root of the characteristic equation of the AR part, *ie*:

$$1-\alpha\lambda=0$$

must be strictly greater than 1 in magnitude. So, for stationarity, we must have $|\alpha|<1$.

(ii)(b) ***Condition for invertibility***

The root of the characteristic equation of the MA part, *ie*:

$$1-\beta\lambda^2=0$$

must be strictly greater than 1 in magnitude. So, for invertibility, we must have $|\beta|<1$. [2]

(iii) ***Covariances***

Using the definition of the series with $t=s$, we have:

$$\begin{aligned}\text{cov}(X_s,Z_s)&=\text{cov}(\alpha X_{s-1}+Z_s+\beta Z_{s-2},Z_s)\\&=\alpha\,\text{cov}(X_{s-1},Z_s)+\text{cov}(Z_s,Z_s)+\beta\,\text{cov}(Z_{s-2},Z_s)\end{aligned}$$ [½]

The first term on the RHS is zero since Z_s is uncorrelated with earlier values of the series. The third term is also zero since $\{Z_t\}$ is a purely random process. So:

$$\text{cov}(X_s,Z_s)=\text{cov}(Z_s,Z_s)=\text{var}(Z_s)=\sigma^2$$ [½]

Similarly using the result just proved, with s replaced by $s-1$, we have:

$$\begin{aligned}\text{cov}(X_s,Z_{s-1})&=\text{cov}(\alpha X_{s-1}+Z_s+\beta Z_{s-2},Z_{s-1})\\&=\alpha\,\text{cov}(X_{s-1},Z_{s-1})=\alpha\sigma^2\end{aligned}$$ [1]

Finally:

$$\begin{aligned}\text{cov}(X_s,Z_{s-2})&=\text{cov}(\alpha X_{s-1}+Z_s+\beta Z_{s-2},Z_{s-2})\\&=\alpha\,\text{cov}(X_{s-1},Z_{s-2})+\text{cov}(Z_s,Z_{s-2})+\beta\,\text{cov}(Z_{s-2},Z_{s-2})\\&=\alpha\,\text{cov}(X_s,Z_{s-1})+0+\beta\sigma^2\\&=(\alpha^2+\beta)\sigma^2\end{aligned}$$ [1]

(iv)(a) ***Yule-Walker equations***

If we replace one of the X_t's in $\text{cov}(X_t,X_t)$ with the definition given for X_t, we get:

$$\text{cov}(X_t,X_t)=\text{cov}(\alpha X_{t-1}+Z_t+\beta Z_{t-2},X_t)$$

Using the first and third results from part (iii) gives:

$$\gamma_0 = \alpha\gamma_1 + \sigma^2 + \beta(\alpha^2 + \beta)\sigma^2 = \alpha\gamma_1 + (1 + \alpha^2\beta + \beta^2)\sigma^2 \qquad \text{... (1)} \qquad [1]$$

If we replace the X_t in $\text{cov}(X_t, X_{t-1})$ with the definition given for X_t, we get:

$$\text{cov}(X_t, X_{t-1}) = \text{cov}(\alpha X_{t-1} + Z_t + \beta Z_{t-2}, X_{t-1}) \qquad [1]$$

Using the second result from part (iii), this simplifies to:

$$\gamma_1 = \alpha\gamma_0 + 0 + \alpha\beta\sigma^2 = \alpha\gamma_0 + \alpha\beta\sigma^2 \qquad \text{... (2)} \qquad [1]$$

Similarly:

$$\text{cov}(X_t, X_{t-2}) = \text{cov}(\alpha X_{t-1} + Z_t + \beta Z_{t-2}, X_{t-2})$$

Using the first result from part (iii), this simplifies to:

$$\gamma_2 = \alpha\gamma_1 + 0 + \beta\sigma^2 = \alpha\gamma_1 + \beta\sigma^2 \qquad \text{... (3)} \qquad [1]$$

(iv)(b) ***Obtain autocovariances***

Substituting the value of γ_1 from Equation (2) into Equation (1), rearranging and simplifying gives:

$$\gamma_0 = \frac{1 + \beta^2 + 2\alpha^2\beta}{1 - \alpha^2}\sigma^2 \qquad [1]$$

Substituting this into Equation (2) gives:

$$\gamma_1 = \frac{\alpha(1 + \beta + \beta^2 + \alpha^2\beta)}{1 - \alpha^2}\sigma^2 \qquad [1]$$

Substituting this into Equation (3) gives:

$$\gamma_2 = \frac{\beta + \alpha^2 + \alpha^2\beta^2 + \alpha^4\beta}{1 - \alpha^2}\sigma^2 \qquad [1]$$

(v)(a) ***ACF***

Using the values given for α and β, the numerators in the expressions for γ_0, γ_1 and γ_2 are 1.522, –0.3064 and –0.63344.

So:

$$\rho_0 = 1 \quad \rho_1 = -\frac{0.3064}{1.522} = -0.201 \quad \rho_2 = -\frac{0.63344}{1.522} = -0.416 \qquad [2]$$

If we use the same method as in (iv)(a), we see that subsequent values of γ (and hence ρ) will just include an extra factor of α each time.

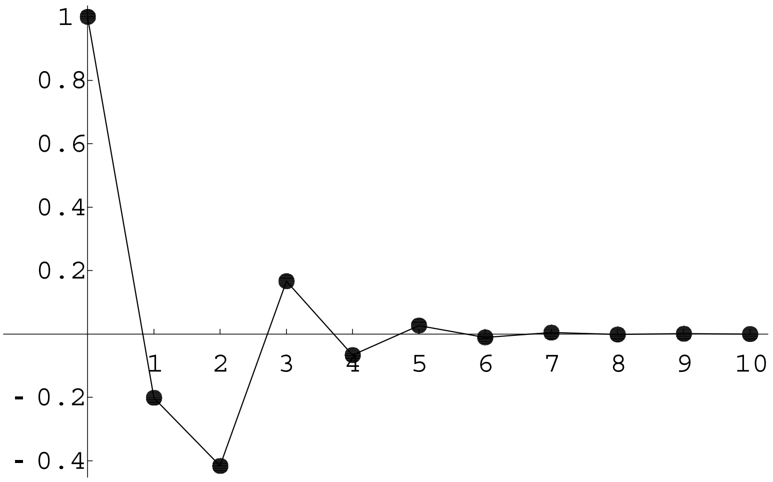
So:

$$\rho_k = -0.416 \times (-0.4)^{k-2} \text{ for } k \geq 2$$

$$\Rightarrow \quad \rho_3 = -0.416 \times (-0.4) = 0.166$$ [1]

(v)(b) ***Graph of ACF***

So the graph of ρ_k looks like this (with the values for $k \geq 2$ alternating in sign and reducing in magnitude):



[2]

14

Time series 2

Syllabus objectives

2.1 Concepts underlying time series models

2.1.3 Explain the concept of a filter applied to a stationary random series.

2.1.7 Explain the basic concept of a multivariate autoregressive model.

2.1.8 Explain the concept of cointegrated time series.

2.2 Applications of time series models

2.2.1 Outline the process of identification, estimation and diagnosis of a time series, the criteria for choosing between models and the diagnostic tests that might be applied to the residuals of a time series after estimation.

2.2.2 Describe briefly other non-stationary, non-linear time series models.

2.2.3 Describe simple applications of a time series model, including random walk, autoregressive and cointegrated models as applied to security prices and other economic variables.

2.2.4 Develop deterministic forecasts from time series data, using simple extrapolation and moving average models, applying smoothing techniques and seasonal adjustment when appropriate.

0 Introduction

We have seen in Chapter 13 that stationary time series are easier to analyse than non-stationary ones. In this chapter we look more closely at the causes of non-stationarity, and we consider a number of related questions:

1. How can we tell whether a time series is stationary or not?

2. If it is not, how do we go about turning it into a stationary series?

We shall use the characteristics of the different types of time series that we studied in Chapter 13 to help to determine an appropriate approach. The ACF and PACF will be key tools here.

We also look at the Box-Jenkins approach to fitting and forecasting.

Once we have chosen an appropriate time series model, two further questions arise.

1. How do we use the model to determine estimates of the future values of our time series?

2. Is the model we have chosen a good one? How well does it fit the data that we already have?

We shall look at possible answers to both these questions in this chapter.

1 Compensating for trend and seasonality

All the methods which we shall investigate in Sections 3 and 4 apply only to a time series which gives the appearance of stationarity. In this section, therefore, we deal with possible sources of non-stationarity and how to compensate for them.

A simple time series plot in R can be generated as:

```
ts.plot(x)
```

where *x* is some (vector) time series data.

Lack of stationarity may be caused by the presence of deterministic effects in the quantity being observed. Monthly sales figures for a company which is expanding rapidly would be expected to show a steady underlying increase, possibly linear or perhaps even exponential. A company which sells greetings cards will find that the sales in some months of the year will be much higher than in others. In both cases there is an underlying deterministic pattern and some (possibly stationary) random variation on top of that. In order to predict sales figures in future months it is necessary to extrapolate the deterministic trends as well as to analyse the stationary random variation.

A further cause of non-stationarity may be that the process observed is the integrated version of a more fundamental process; in these cases, differencing the observed time series may produce a series which is more likely to be a realisation of some stationary process.

In summary, we have identified three possible causes of non-stationarity:

1. a deterministic trend (*eg* exponential or linear growth)
2. a deterministic cycle (*eg* seasonal effect)
3. the time series is integrated.

It is worth pointing out that this list is not exhaustive. For example, the first two causes are just specific cases of general deterministic behaviour. In theory, this could take many different forms, but trends and cycles are the most likely to be met in practice.

The items on the list are not mutually exclusive either. Consider a simple random walk with probability 0.6 of stepping up, and 0.4 of stepping down. This can be represented by the equation:

$$X_n = X_{n-1} + Z_n$$

where:

$$Z_n = \begin{cases} +1 & \text{with probability } 0.6 \\ -1 & \text{with probability } 0.4 \end{cases}$$

together with the condition that $X_0 = 0$.

This process is $I(1)$ since the first difference is stationary, but the process itself is not.

On the other hand, the process also has a increasing deterministic trend:

$$E[X_n] = E[Z_1 + Z_2 + \cdots + Z_n] = E[Z_1] + E[Z_2] + \cdots + E[Z_n] = 0.2n$$

We consider in the following sections how to detect and remove such causes of non-stationarity.

1.1 Detecting non-stationary series

The most useful tools in identifying non-stationarity are the simplest: a plot of the series against *t*, and the sample ACF.

The sample ACF is an estimate of the ACF based on sample data. It is defined in Section 2.1.

Plotting the series will highlight any obvious trends in the mean and will show up any cyclic variation which could also form evidence of non-stationarity. This should always be the first step in any practical time series analysis.

The R code below uses the `ts.plot` function.

(It also assumes we have a list of FTSE100 data values.)

```
ts.plot(log(FTSE100$Close))
points(log(FTSE100$Close),cex=.4)
```

generates Figure 14.1, which shows the time series of the logs of 300 successive closing values of FTSE100 index.

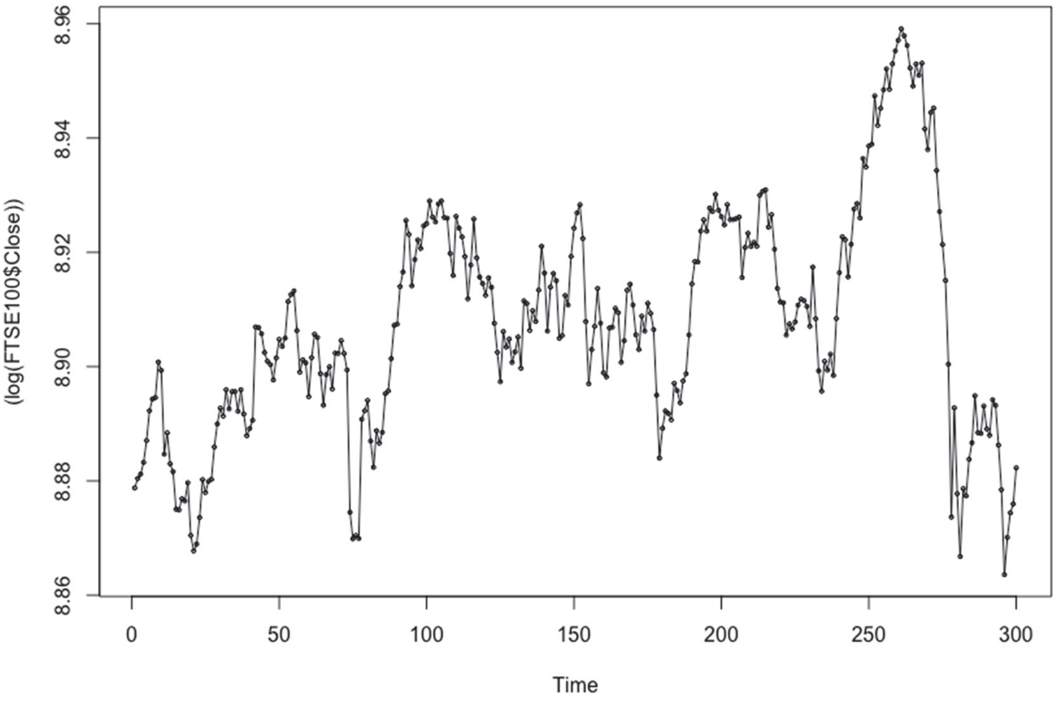


Figure 14.1: 300 successive closing values of the FTSE100 index, Jan 2017 – Mar 2018; log-transformed

The corresponding sample ACF and sample PACF are produced using:

```
par(mfrow=c(1,2))
acf(log(FTSE100$Close))
pacf(log(FTSE100$Close))
```

These are shown in Figure 14.2 below.

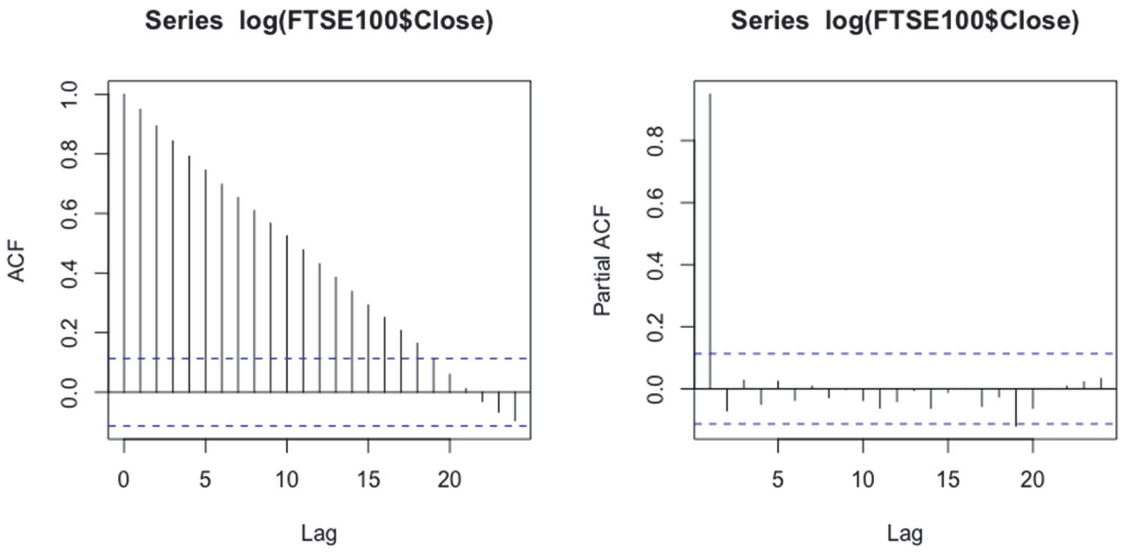


Figure 14.2: Sample ACF and sample PACF of the log(FTSE100) data; dotted lines indicate cut-offs for significance if data came from some white noise process.

The sample ACF should, in the case of a stationary time series, ultimately converge towards zero exponentially fast, as for $AR(1)$ where $\rho_s = \alpha^s$.

The function $\rho_s = \alpha^s$, $s = 0,1,2,...$ has *exponential decay* if $|\alpha| < 1$ as this is a power function whose values are:

$$1, \alpha, \alpha^2, \alpha^3, ...$$

and this sequence tends to 0 as the power s tends to ∞.

If the sample ACF decreases slowly but steadily from a value near 1, we would conclude that the data need to be differenced before fitting the model. If the sample ACF exhibits a periodic oscillation, however, it would be reasonable to conclude that there is some underlying cause of the variation.

Figure 14.2 shows the sample ACF of a time series which is clearly non-stationary as the values decrease in some linear fashion; differencing is therefore required before fitting a stationary model.

See, for example, the change of ACF and PACF for the differenced data:

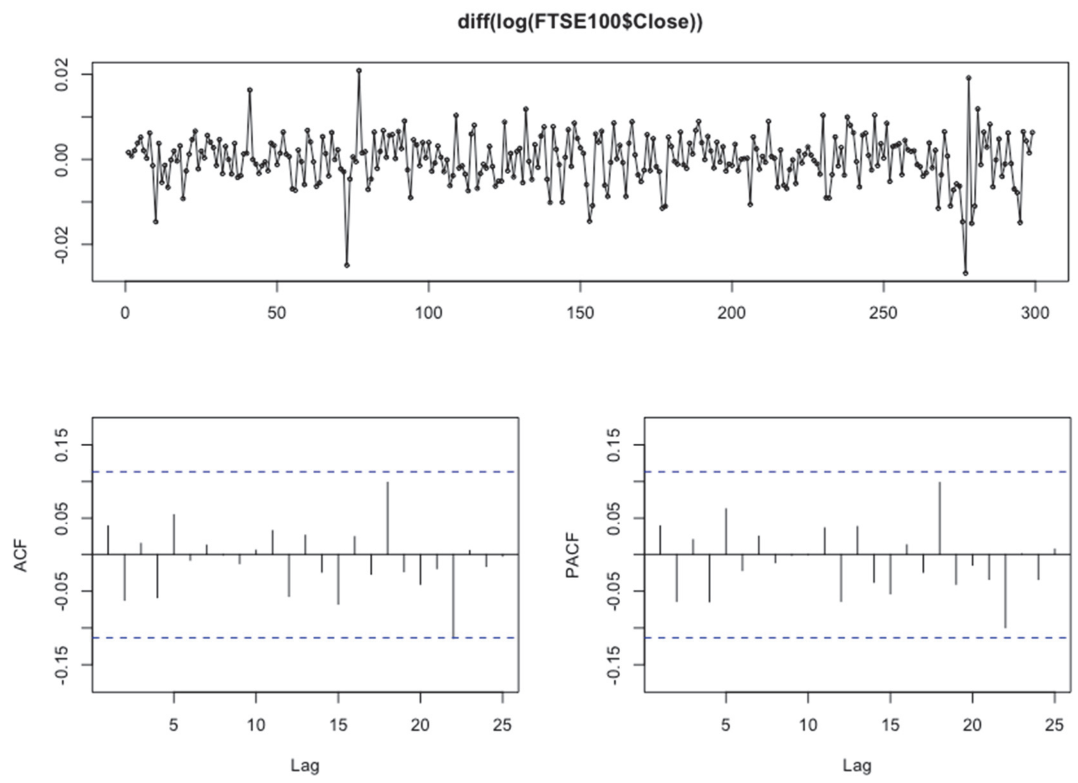


Figure 14.3: Data plot, sample ACF and sample PACF of ∇ln(FTSE100).

The ACF for a stationary $ARMA(p,q)$ process decays fairly rapidly. For example, we have seen that the ACF of a moving average process cuts off sharply, and the ACF of an $AR(1)$ process has exponential decay: $\rho(k)=\alpha^{|k|}$. In theory, this could still actually lead to a slow decay if the values of p and/or q were high, (*eg* the autocorrelation of an $MA(100)$ process wouldn't cut off until lag 100), but in practice the parameter values will be fairly small. If many parameters are used then the resulting model may give a good fit to the sample data, but it is unlikely to be useful for forecasting. A fairly slow decay of the sample autocorrelation function is therefore more likely to be interpreted as an indication that the time series needs to be differenced before being modelled.

If the sample autocorrelation function oscillates without decaying rapidly, as we will see in the hotel example in Section 1.4, then we might conclude that there is an underlying deterministic cycle. This would have to be removed before fitting a model to the residuals.

We now look at two methods for removing a linear trend.

1.2 Least squares trend removal

The simplest way to remove a linear trend is by ordinary least squares. This is equivalent to fitting the model:

$$\boldsymbol{x_t = a + bt + y_t}$$

where a and b are constants and y is a zero-mean stationary process. The parameters a and b can be estimated by linear regression prior to fitting a stationary model to the residuals y_t.

The formulae for estimating a and b are given on page 24 of the *Tables*.

1.3 Differencing

Differencing may well be beneficial if the sample ACF decreases slowly from a value near 1, but has useful effects in other instances as well. If, for instance, $x_t = a + bt + y_t$, then:

$$\nabla \boldsymbol{x_t = b + \nabla y_t}$$

so that the differencing has removed the trend in the mean.

Differencing a series d times will make an $I(d)$ series stationary. In addition however, differencing once will also remove any linear trend, as above.

On the other hand, we could remove the linear trend by using linear regression, as in Section 1.2. However, if the series is actually $I(1)$ with a trend, then least squares regression will only remove the trend. We will still be left with an $I(1)$ process that is non-stationary.

For example, consider the simple random walk discussed earlier, which has probability 0.6 of stepping up, and 0.4 of stepping down:

$$X_n = X_{n-1} + Z_n$$

where:

$$Z_n = \begin{cases} +1 & 0.6 \\ -1 & 0.4 \end{cases}$$

We have seen that this process has an increasing trend and that $E[X_n] = 0.2n$. If we let $Y_n = X_n - 0.2n$, then $E[Y_n] = 0$. So we have removed the trend. However, since:

$$Y_n = Y_{n-1} + Z_n - 0.2$$

we are still left with an $I(1)$ process that needs to be differenced in order to be stationary.

We now look at three methods for removing cycles (seasonal variation).

1.4 Seasonal differencing

Where seasonal variation is present in the data, one way of removing it is to take a seasonal difference.

Example 1

Suppose that the time series x records the monthly average temperature in London. A model of the form:

$$x_t = \mu + \theta_t + y_t \qquad \textbf{(14.1)}$$

might be applied, where θ is a periodic function with period 12 and y is a stationary series. Then the seasonal difference of x is defined as $(\nabla_{12}x)_t = x_t - x_{t-12}$ and we see that:

$$(\nabla_{12}x)_t = x_t - x_{t-12} = (\mu + \theta_t + y_t) - (\mu + \theta_{t-12} + y_{t-12}) = y_t - y_{t-12}$$

is a stationary process.

We can then model the seasonal difference of x as a stationary process and reconstruct the original process x itself afterwards.

We can use R to plot the time series data and to remove seasonal variation.

Figure 14.4 below is generated from the following lines in R, where functions `ts.plot`, `acf` and `pacf` are used:

```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
ts.plot(manston1$tmax,ylab="", main="Max temperatures observed at
each month (2010-2017), Manston, UK")
points(manston1$tmax,cex=0.4)
acf(manston1$tmax,main="")
pacf(manston1$tmax,main="")
```

This code assumes that there is a list of data stored as ‘manston1’.

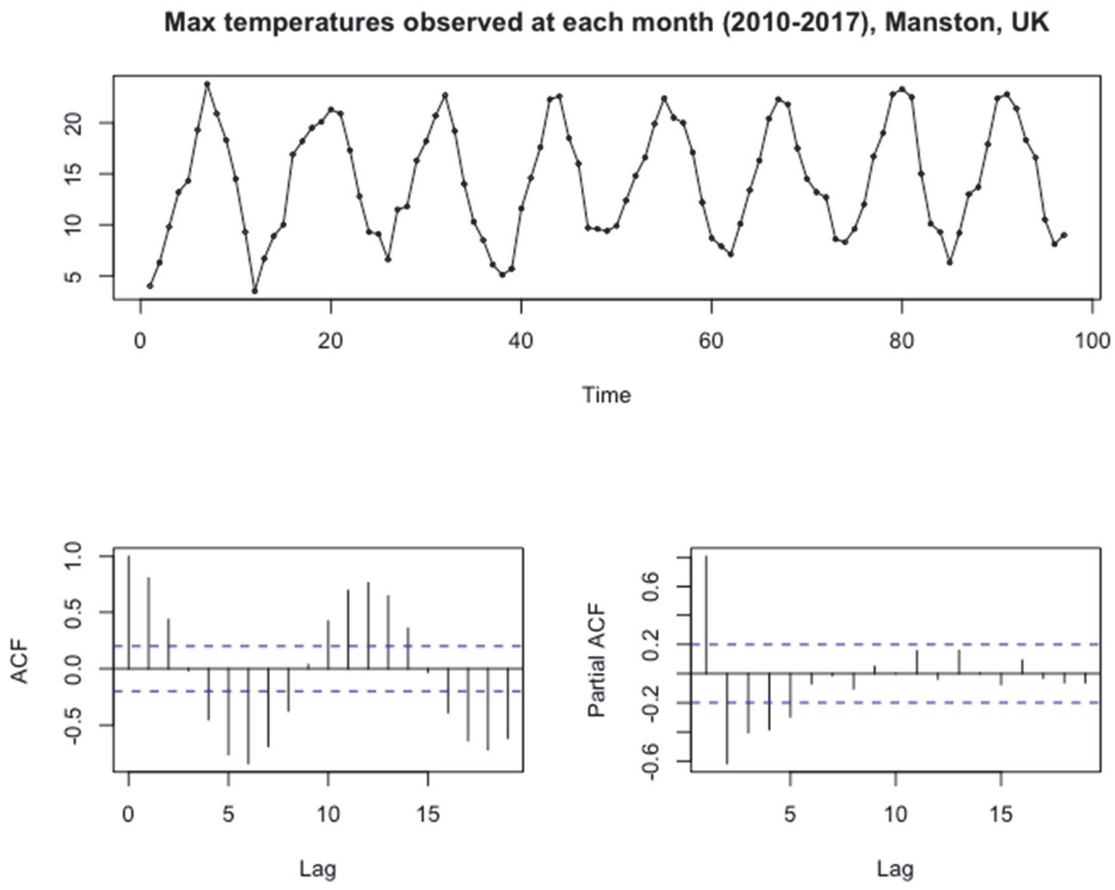


Figure 14.4: Data plot, sample ACF and PACF of temperature data.

Seasonal differencing ∇_{12} seems to have removed the seasonal behaviour of the data. See Figure 14.5 generated from:

```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
ts.plot(diff(manston1$tmax,lag=12),ylab="", main="Seasonal
differenced temperature data")
points(diff(manston1$tmax,lag=12),cex=0.4)
acf(diff(manston1$tmax,lag=12),main="")
pacf(diff(manston1$tmax,lag=12),,main="")
```

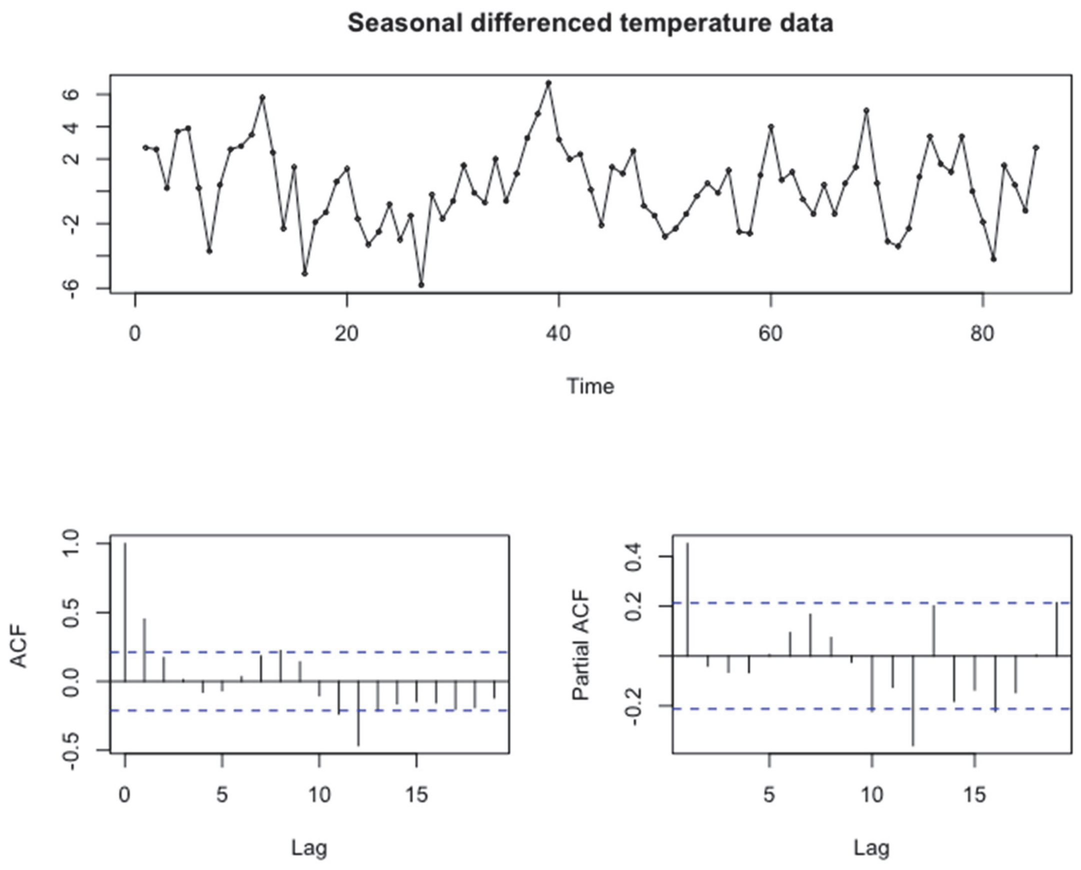


Figure 14.5: Temperature data after appropriate differencing

Example 2

The monthly inflation figures are obtained by seasonal differencing of the Retail Prices Index. If x_t is the value of the RPI in month t, the annual inflation figure reported is:

$$\frac{x_t - x_{t-12}}{x_{t-12}} \times 100\%$$

1.5 Method of moving averages

The method of moving averages makes use of a simple linear filter to eliminate the effects of periodic variation.

A linear filter is a transformation of a time series x (the input series) to create an output series y that satisfies:

$$y_t = \sum_{k=-\infty}^{\infty} a_k x_{t-k}$$

The collection of weights $\{a_k : k \in Z\}$ forms a complete description of the filter. The objective of the filtering is to modify the input series to meet particular objectives, or to display specific features of the data. For example, an important problem in analysis of economic time series is detection, isolation, and removal of deterministic trends.

In practice a filter $\{a_k : k \in Z\}$ normally contains only a relatively small number of non-zero components.

A very simple example of a linear filter is the difference operator $\nabla = 1 - B$. Using this filter produces:

$$y_t = (1-B)x_t = x_t - x_{t-1}$$

So, in this case, we have $a_0 = 1$, $a_1 = -1$ and $a_k = 0$ for all other values of k.

As a second example, suppose that the input series is a white noise process e, and the filter takes the form:

$$a_0 = 1, a_1 = \beta_1, \ldots, a_q = \beta_q \text{ and } a_k = 0 \text{ for all other values of } k$$

Then the output series is $MA(q)$, since we have:

$$y_t = \sum_{k=0}^{q} \beta_k e_{t-k}$$

Conversely, applying a filter of the form:

$$a_0 = 1, a_1 = -\alpha_1, \ldots, a_p = -\alpha_p \text{ and } a_k = 0 \text{ for all other values of } k$$

to an input series x that is $AR(p)$ recovers the original white noise series:

$$y_t = x_t - \sum_{k=1}^{p} \alpha_k x_{t-k} = e_t$$

If x is a time series with seasonal effects with even period $d = 2h$, then we define a smoothed process y by:

$$y_t = \frac{1}{2h}\left(\tfrac{1}{2}x_{t-h} + x_{t-h+1} + \dots + x_{t-1} + x_t + \dots + x_{t+h-1} + \tfrac{1}{2}x_{t+h}\right)$$

This ensures that each period makes an equal contribution to y_t.

For example, with quarterly data a yearly period will have $d = 4 = 2h$, so $h = 2$ and we have:

$$y_t = \frac{1}{4}\left(\frac{1}{2}x_{t-2} + x_{t-1} + x_t + x_{t+1} + \frac{1}{2}x_{t+2}\right)$$

In this case the filter has weights:

$$a_k = \begin{cases} \frac{1}{8} & \text{for } k = -2,2 \\ \frac{1}{4} & \text{for } k = -1,0,1 \\ 0 & \text{otherwise} \end{cases}$$

This is a *centred* moving average since the average is taken symmetrically around the time t. Such a centred moving average introduces the practical problem that the average can only be calculated in retrospect, *ie* there will be a natural delay.

The same can be done with odd periods $d = 2h + 1$, but the end terms x_{t-h} and x_{t+h} do not need to be halved.

For example, with data every 4 months, a yearly period will have $d = 3 = 2h + 1$, so $h = 1$ and we have:

$$y_t = \frac{1}{3}\left(x_{t-1} + x_t + x_{t+1}\right)$$

In this case the filter has weights:

$$a_k = \begin{cases} \frac{1}{3} & \text{for } k = -1,0,1 \\ 0 & \text{otherwise} \end{cases}$$

As with most filtering techniques, care must be taken lest the smoothing of the data obscure the very effects which the procedure is intended to uncover.

The method of taking moving averages is one example of a series of approaches known as filtering techniques. We lose some of our knowledge about the variation in the data in exchange for (hopefully) a clearer picture of the underlying process.

1.6 Method of seasonal means

The simplest method for removing seasonal variation is to subtract from each observation the estimated mean for that period, obtained by simply averaging the corresponding observations in the sample.

Recall that the model in Equation 14.1 has the form:

$$x_t = \mu + \theta_t + y_t$$

where θ is a periodic function with period 12 and y is a stationary series. The term θ_t contains the deviation of the model at time t due to the seasonal effect. So:

$$y_t = x_t - \mu - \theta_t$$

When fitting the model in Equation 14.1 to a monthly time series x extending over 10 years from January 1990 the estimate for μ is $\bar{x}$ and the estimate for θ_{January} is:

$$\hat{\theta}_{\text{January}} = \frac{1}{10}(x_1 + x_{13} + x_{25} + \cdots + x_{109}) - \hat{\mu}$$

So, in this case, we can remove the seasonal variation by deducting the January average, $\bar{x}_{\text{January}} = \frac{1}{10}(x_1 + x_{13} + x_{25} + \cdots + x_{109})$, from all the January values, deducting the February average, $\bar{x}_{\text{February}} = \frac{1}{10}(x_2 + x_{14} + x_{26} + \cdots + x_{110})$, from all the February values, *etc*.

Alternatively, we could deduct $\hat{\theta}_{January}$ from all the January values, $\hat{\theta}_{February}$ from all the February values, *etc*.

In R the function `decompose` **can be used to obtain both the moving average and seasonal means described in Sections 1.5 and 1.6.**

```
ts.plot(manston1$tmax,ylab="", main="Max temperatures")
points(manston1$tmax,cex=0.4)
```

The time series data is plotted as in Figure 14.6 below.

```
decomp=decompose(ts(manston1$tmax,frequency = 12),type="additive")
```

The decomposition is saved as `decomp`**.**

The moving average can be added (in red) using the code:

```
lines(as.vector(decomp$trend),col="red")
```

The sum of seasonal and moving average trends can be added (in blue) as follows:

```
lines(as.vector(decomp$seasonal+decomp$trend),col="blue")
```

The resulting graph is shown below. A colour version is also available online in the tuition materials for the R part of CS2.

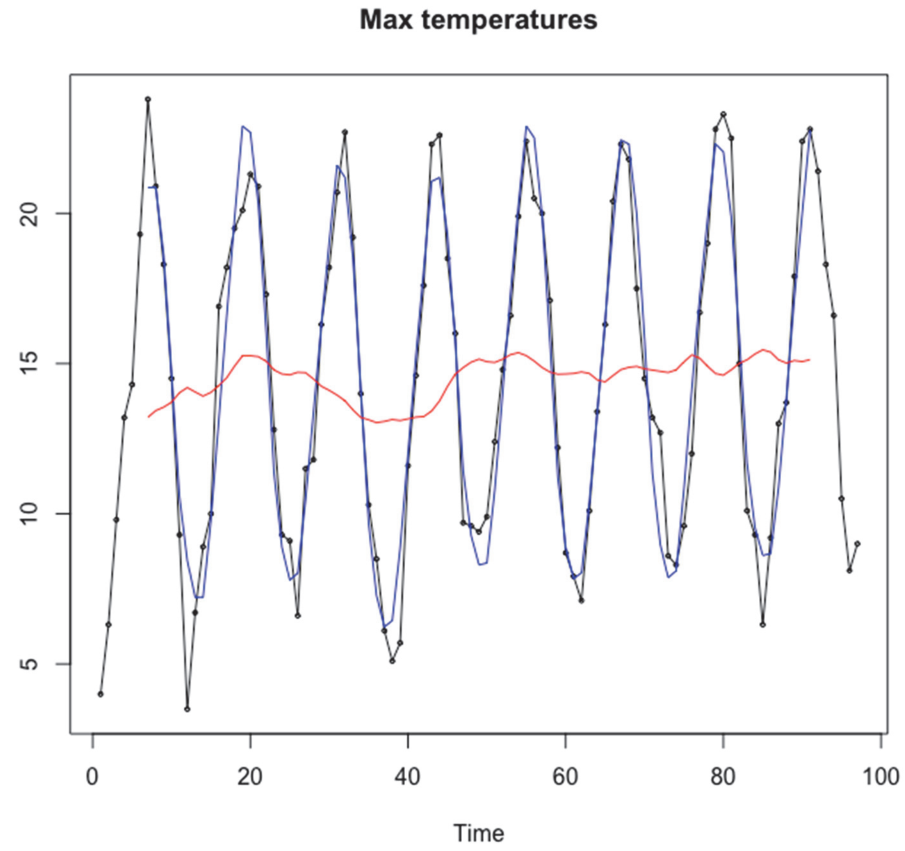


Figure 14.6: Temperature data and its decomposition into moving average (in red) and seasonal trend (in blue) added.

1.7 Transformation of the data

Diagnostic procedures such as an inspection of a plot of the residuals may suggest that even the best-fitting standard linear time series model is failing to provide an adequate fit to the data. Before attempting to use more advanced non-linear models it is often worth attempting to transform the data in some straightforward way in an attempt to find a data set on which the linear theory will work properly.

An example of a simple transformation would be $Y_t = \log X_t$, which would be used to remove an exponential growth effect.

Variance-stabilising transformations

Transformations are most commonly used when a dependence is suspected between the variance of the residuals and the size of the fitted values. If, for example, the standard deviation of $X_{t+1} - X_t$ appears to be proportional to X_t, then it would be appropriate to use the logarithmic transformation, to work on the time series $Y = \ln X$.

Transformations to increase normality

In certain applications it may be found that most residuals are small and negative, with a few large positive values to offset them. This may be taken to indicate that the distribution of the error terms is non-normal, leading to doubts as to whether the standard time series procedures, designed for normal errors, are applicable. It may be possible to find a transformation which will improve the normality of the error terms of the transformed process, but care should be taken that this does not lead to instability in the variance. A further caution when using transformed data involves the final step of turning forecasts for the transformed process into forecasts for the original process, as some transformations introduce a systematic bias.

2 Identification of *MA*(*q*) and *AR*(*p*) models

The treatment of this section assumes that the sequence of observations $\{x_1, x_2, \dots, x_n\}$ may be presumed to come from a stationary time series process. The problems of how to tell if the assumption of stationarity is reasonable and what to do if it is not have been treated in the previous section.

2.1 Estimation of the ACF and PACF

The autocovariance and autocorrelation functions, as seen above, play a central role in the analysis of time series. Other descriptive tools, such as the partial autocorrelation function, are derived from the ACF. Faced, then, with a sequence of observations $\{x_1, x_2, \dots, x_n\}$ and the task of finding a time series model to fit the sequence, a primary concern must be to estimate the ACF of the time series process of which the data form a realisation.

The common mean of a stationary model can be estimated using the *sample mean*:

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

The autocovariance function γ_k can be estimated using the sample autocovariance function, denoted c_k or $\hat{\gamma}_k$, given by:

$$\hat{\gamma}_k = \frac{1}{n}\sum_{t=k+1}^{n} (x_t - \hat{\mu})(x_{t-k} - \hat{\mu})$$

from which are derived estimates r_k for the autocorrelation function ρ_k:

$$r_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$$

The notation $\hat{\rho}_k$ is sometimes used instead of r_k. The formulae for $\hat{\mu}$, $\hat{\gamma}_k$ and $\hat{\rho}_k$ are given on page 40 of the *Tables*.

The collection $\{r_k : k \in Z\}$ is called the *sample autocorrelation function* (SACF). Every time series analysis involves at least one plot of r_k against k. Such a plot is called a *correlogram*.

It might seem that a more natural choice for the denominator in the definition of the sample autocovariance function would be $n-k$, since there are $n-k$ terms being summed. However, the definition given is the most common one used. One theoretical reason for this is that it has a smaller mean square error at large lags, and at smaller lags the difference between the two is negligible anyway. In any case, this estimator is consistent, *ie* the bias disappears as the sample size, *n*, gets large.

The partial autocorrelation function ϕ_k can be estimated using the formula involving the ratio of determinants to which reference was made in Chapter 13, but with the ρ_k replaced by their estimates $\hat{\rho}_k$. The resulting function $\hat{\phi}_k$, called the *sample partial autocorrelation function* (SPACF), and the plot of $\hat{\phi}_k$ against k, called the *partial correlogram*, are as important as the SACF and the correlogram in the analysis of time series.

As we have seen before, R functions `acf` and `pacf` can be used for generating these values.

For example, the following lines simulate observations from an *ARMA*(1,1) model.

Set the seed to guarantee reproducibility. The code is:

```
set.seed(123)
```

Call the simulated data x:

```
x=arima.sim(n=300,model=list(ar=0.7,ma=0.5))
```

Then:

```
par(mfrow=c(1,2))
acf(x,main="Sample ACF")
pacf(x,main="Sample PACF")
```

produces the graphs below:

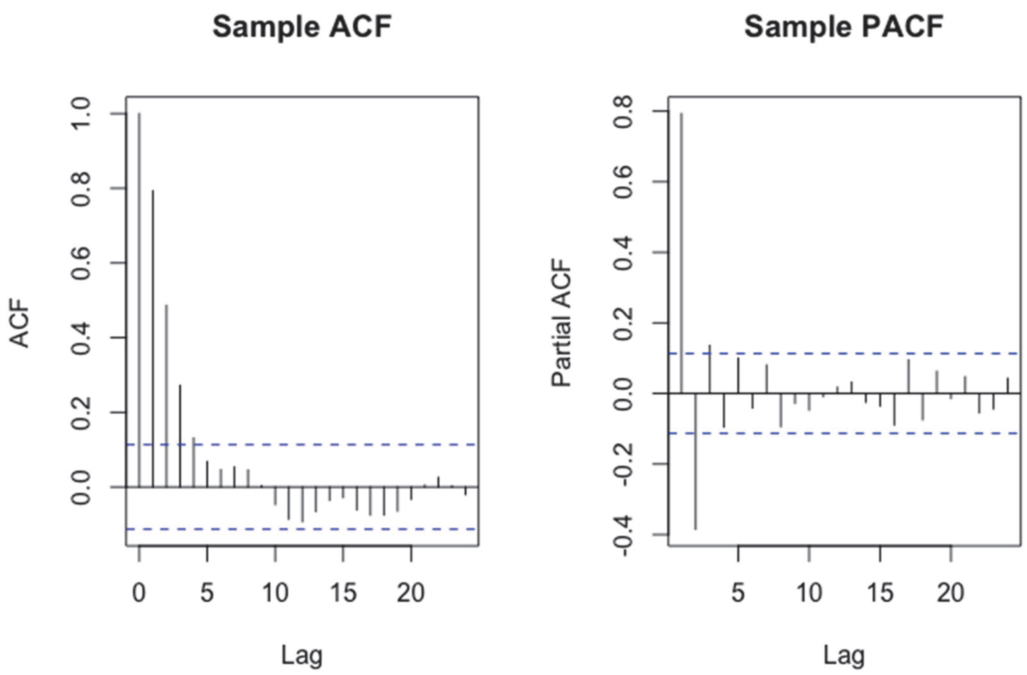


Figure 14.7: ACF and PACF of some simulated data from *ARMA*(1,1).

2.2 Identification of white noise

A test for whether a particular sequence of observations forms a standard white noise process may seem of doubtful usefulness, but one of the techniques of residual analysis suggests that the verification of goodness of fit of any model should include a test as to whether the residuals form a white noise process. A suitable test, or portfolio of tests, is therefore a valuable asset.

We are already familiar with this idea from (normal) linear regression – we always have a look at a plot of the residuals and carry out other tests to check that the residuals do form a set of independent normal random variables.

There are many tests that could be carried out to see if a sequence of observations is a likely realisation of a white noise process. Some of these tests will be discussed later in the context of the diagnostic checking stage of the Box-Jenkins method – see Section 3.5. For the moment we will concentrate on tests associated with the SACF and SPACF.

Clearly the SACF and SPACF are random, being simple functions of the observations. In particular, even if the original process was a perfectly standard white noise the SACF and SPACF would not be identically zero. The question is what scale of deviation from zero is to be expected?

An asymptotic result states that, if the original model is white noise:

$$X_t = \mu + e_t$$

then the estimators $\tilde{\rho}_k$ and $\tilde{\phi}_k$ are approximately normally distributed with mean 0, variance $1/n$ for each k.

For large samples, *ie* large values of n, we have the approximate distributions:

$$\tilde{\rho}_k \div N\left(0,\frac{1}{n}\right) \qquad \text{and} \qquad \tilde{\phi}_k \div N\left(0,\frac{1}{n}\right)$$

Values of the SACF or SPACF falling outside the range from $-2/\sqrt{n}$ to $2/\sqrt{n}$ can be taken as suggesting that the white noise model is inappropriate. This range is indicated by dashed lines in the standard output in R for ACF and PACF.

This range is an approximate 95% confidence interval based on the critical value 1.96.

But some care should be exercised: the cut-off points of $\pm 2/\sqrt{n}$ give approximate 95% limits, implying that about one value in 20 will fall outside the range even when the white noise model is correct. This means that one single value of r_k or $\hat{\phi}_k$ outside the specified range would not be regarded as significant on its own, but three such values might well be significant.

Rather than testing to see if each individual value of the SACF or SPACF lies outside a confidence interval, we can alternatively consider an overall goodness-of-fit test, much like the standard χ^2 test. In other words, we can carry out a test to see if some measure of the overall deviation over several lags lies outside some confidence interval.

A 'portmanteau' test is due to Ljung and Box, who state that, if the white noise model is correct, then:

$$n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k} \sim \chi_m^2$$

for each m.

In the result above, the notation r_k is being used to represent the estimator $\tilde{\rho}_k$ (rather than the estimate $\hat{\rho}_k$). This is a one-sided test. A large test statistic indicates that the data do *not* conform to a white noise process.

The standard commands for running these tests in R on some observations (simulated white noise here) are:

```
x <- rnorm (100)
Box.test (x, lag = 1, type = "Ljung")
```

Question

An analysis of the first 369 draws of the National Lottery gave the following SACF values:

$r_1 = 0.100$ $r_2 = 0.056$ $r_3 = 0.059$ $r_4 = 0.054$ $r_5 = -0.005$ $r_6 = 0.003$

Use the portmanteau test to ascertain whether these data can be considered to be white noise.

Solution

We are testing:

H_0 : the residuals form a white noise process

against:

H_1 : the residuals do not form a white noise process

We have $m = 6$ and $n = 369$. So the observed value of the test statistic is:

$$n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k} = 369\times 371\left(\frac{0.100^2}{369-1}+\frac{0.056^2}{369-2}+\frac{0.059^2}{369-3}+\frac{0.054^2}{369-4}+\frac{(-0.005)^2}{369-5}+\frac{0.003^2}{369-6}\right)$$

$$= 7.298$$

We compare this with the χ_6^2 distribution. Since 7.298 is less than 12.59, the upper 5% point of χ_6^2, there is insufficient evidence to reject H_0 at the 5% level. So we conclude that the residuals are consistent with white noise.

2.3 Identification of $MA(q)$

The distinguishing characteristic of $MA(q)$ is that $\rho_k = 0$ for all $k > q$.

A test for the appropriateness of an $MA(q)$ model, therefore, is that r_k is close to 0 for all $k > q$. If the data really do come from a $MA(q)$ model, the estimators $\tilde{\rho}_k$ for $k > q$ will be roughly normally distributed with mean 0, variance $n^{-1}\left(1+2\sum_{i=1}^{q}\rho_i^2\right)$.

In other words:

$$\tilde{\rho}_k \div N\left(0,\frac{1}{n}\left(1+2\sum_{i=1}^{q}\rho_i^2\right)\right)$$

This asymptotic result enables a test to be formulated.

If we assume that the data do come from a $MA(q)$ process, then an approximate 95% confidence interval for ρ_k is:

$$\left[-1.96\sqrt{\frac{1}{n}\left(1+2\sum_{i=1}^{q}\hat{\rho}_i^2\right)},+1.96\sqrt{\frac{1}{n}\left(1+2\sum_{i=1}^{q}\hat{\rho}_i^2\right)}\right]$$

So no more than about 1 in 20 values of the sample autocorrelation function should lie outside the interval.

2.4 Identification of $AR(p)$

The corresponding diagnostic procedure for an autoregressive model is based on the sample partial ACF, since the PACF of an $AR(p)$ is distinctive, being equal to zero for $k > p$.

The asymptotic variance of $\tilde{\phi}_k$ is $1/n$ for each $k > p$. Again a normal approximation can be used, so that values of the SPACF outside the range $\pm 2/\sqrt{n}$ may suggest that the $AR(p)$ model is inappropriate.

So:

$$\tilde{\phi}_k \div N\left(0,\frac{1}{n}\right)$$

and hence an approximate 95% confidence interval for ϕ_k is:

$$\left[-1.96\sqrt{\frac{1}{n}},+1.96\sqrt{\frac{1}{n}}\right]$$

As suggested by the Core Reading, the value 2 may be used as an approximation to 1.96.

3 Fitting a time series model using the Box-Jenkins methodology

In this section we consider the general class of autoregressive integrated moving average models – the $ARIMA(p,d,q)$ models. As usual we assume that historical data, comprising a time series $\{x_t : t = 1,2,\dots,n\}$, are given.

We will also assume that deterministic trends and seasonal effects have been removed from the data, as in Section 1, although no differencing of the process is assumed – that is part of the Box-Jenkins method.

3.1 The Box-Jenkins methodology

The Box-Jenkins approach allows one to find an ARIMA model which is reasonably simple and provides a sufficiently accurate description of the behaviour of the historical data.

Main steps in the Box-Jenkins approach to modelling

The main steps of the approach are:

- **tentative identification of a model from the ARIMA class**
- **estimation of parameters in the identified model**
- **diagnostic checks.**

If the tentatively identified model passes the diagnostic tests, the model is ready to be used for forecasting. If it does not, the diagnostic tests should indicate how the model ought to be modified, and a new cycle of identification, estimation and diagnosis is performed.

The identification process is carried out in Sections 3.2 and 3.3. In Section 3.4 we look at the estimation stage, and finally, in Section 3.5 we describe some diagnostic tests that can be performed.

3.2 Differencing

An $ARIMA(p,d,q)$ model is completely identified by the choice of non-negative integer values for the parameters p, d and q. The parameter d is the number of times we have to difference the time series x to convert it to some stationary level.

In other words, once we have differenced the sample data d times, the resulting values look like a realisation of a stationary process.

We have already discussed how to detect non-stationary series in Section 1.1. Recall that there are three basic causes of non-stationarity with which we are concerned. To help identify these, we should look at plots of the data values and the SACF. The plot of the data values should highlight any obvious trends or cycles and the latter should also show up as cycles in the SACF. We are now assuming that these kind of effects have already been removed. The other cause of non-stationarity is that the time series could be the realisation of an integrated process. To remove this source of non-stationarity the sample series needs to be differenced d times.

The following principles can be used to choose the appropriate value of d :

1. **A time series x can be modelled by a stationary ARMA model if the sample autocorrelation function r_k decays rapidly to zero with k. If, on the other hand, a slowly decaying positive sample autocorrelation function r_k is observed, this should be taken to indicate that the time series needs to be differenced to convert it into a likely realisation of a stationary random process.**

   (This is also mentioned in Section 1.1.)

2. **Let $\hat{\sigma}_d^2$ denote the sample variance of the process $z^{(d)} = \nabla^d x$, *ie* the sample variance of the data values after they have been differenced d times. It is normally the case that $\hat{\sigma}_d^2$ first decreases with d until stationarity is achieved and then starts to increase. Therefore d can be set to the value which minimises $\hat{\sigma}_d^2$. This could be $d = 0$ if the original time series x is already stationary.**

Question

The time series Q_t for the monthly RPI given on page 152 of the *Tables* has SACF:

$$r_1 = 0.977 \qquad r_2 = 0.954 \qquad r_3 = 0.930 \qquad r_4 = 0.908 \qquad \ldots$$

Find an appropriate value for d in this case.

Solution

This SACF decays slowly, suggesting that it needs to be differenced. If we look at the sample variances given in the *Tables*, we find that:

$$\text{var}(Q_t) = 11.9^2$$

$$\text{var}(\nabla Q_t) = 0.6^2$$

$$\text{var}(\nabla^2 Q_t) = 0.8^2$$

This suggests that $d = 1$ would be an appropriate value.

3.3 Fitting an *ARMA*(*p*,*q*) model

Suppose now that the appropriate value for the parameter d has been found, and the time series $\{z_{d+1}, z_{d+2}, \ldots, z_n\}$ is adequately stationary. (Notice that a differenced series has d fewer observation than the original series.) We shall assume throughout this section that the sample mean of the z sequence is zero; if this is not the case, obtain a new sequence by subtracting $\hat{\mu} = \bar{z}$ from each value in the sequence. We shall also assume, for the sake of simplicity in setting down the lower and upper limits of sums, that $d = 0$.

In the framework of the Box-Jenkins approach we try to find an $ARMA(p,q)$ model which fits the data z.

If either the correlogram or the partial correlogram appears to be close to zero for sufficiently large k, an $MA(q)$ or $AR(p)$ model is indicated.

We have already seen several times that the ACF of an $MA(q)$ process cuts off (*ie* is equal to 0) after lag q, and similarly, the PACF of an $AR(p)$ process cuts off after lag p. We can therefore test the SACF and SPACF to see if it likely that they share one of these cut-off properties. The exact tests for this are given in Sections 2.3 and 2.4.

Otherwise we should look for an $ARMA(p,q)$ model with non-zero values of p and q.

A good indicator for possible values of p and q in an $ARMA(p,q)$ is the number of spikes in the ACF and PACF until some geometrical decay to zero is observed. Since models can be readily fitted in R, it is not hard to start with a simple model like $ARMA(1,1)$ and to work up to more complicated models if the simpler ones are deemed inadequate.

Every additional parameter improves the fit of the model by reducing the residual sum of squares. Taking this to extremes, a model with n parameters could be found to fit the data exactly.

Recall that n is the number of observed values of the time series. If we have the same number of parameters as observations, then the optimised model is a perfect fit to the data.

But this will result in some spurious model with insignificant t-values of parameter estimates and the forecasts made with such a model will be found to be practically useless. This is known as the problem of *overfitting*. The question of when to stop adding new parameters is addressed by *Akaike's information criterion* (AIC), which states that we should only consider adding an extra parameter if this results in a reduction of the residual sum of squares by a factor of at least $e^{-2/n}$, or alternatively, one can evaluate for each possible model the value of:

$$AIC(\text{model}) = \log(\hat{\sigma}^2) + 2 \times \frac{\text{number of parameters}}{n}$$

and choose as the most appropriate the one corresponding to the lowest such value.

3.4 Parameter estimation

Once the values of p and q have been identified, the problem becomes to estimate the values of parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$ and $\beta_1, \beta_2, \ldots, \beta_q$ for the $ARMA(p,q)$ model:

$$Z_t = \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \cdots + \alpha_p Z_{t-p} + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \cdots + \beta_q e_{t-q}$$

Least squares estimation suggests itself; this is equivalent to maximum likelihood estimation if the e_t may be assumed normally distributed.

Question

Consider a linear regression model of the form $Y_i = \alpha + \beta x_i + e_i$ where $e_i \sim N(0, \sigma^2)$. Show that maximum likelihood estimation and least squares estimation must result in the same parameter values.

Solution

The log-likelihood function is given by:

$$\ln L(\alpha,\beta) = const + \sum_{i=1}^{n} \left\{ -\frac{1}{2}\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right)^2 \right\}$$

$$= const - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

It follows that maximising the log-likelihood must be equivalent to minimising the squared error:

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

In the case of an $AR(p)$ we have:

$$\boldsymbol{e_t = Z_t - \alpha_1 Z_{t-1} - \cdots - \alpha_p Z_{t-p}}$$

and the estimators $\hat{\alpha}_1,\ldots,\hat{\alpha}_p$ are chosen to minimise:

$$\boldsymbol{\sum_{t=p+1}^{n} \left(z_t - \alpha_1 z_{t-1} - \cdots - \alpha_p z_{t-p}\right)^2}$$

We don't need to make any distributional assumptions in order to calculate the least squares estimates of $\alpha_1, \alpha_2, \ldots, \alpha_p$. However, we do need to make a distributional assumption to calculate their maximum likelihood estimates. The two methods coincide when the errors are normally distributed.

In the case of a more general ARMA process we encounter the difficulty that the e_t cannot be deduced from the z_t. For example, in the case of $ARMA(1,1)$ we have:

$$\boldsymbol{e_t = z_t - \alpha_1 z_{t-1} - \beta_1 e_{t-1}}$$

an equation which can be solved iteratively for e_t as long as some starting value e_0 is assumed. For an $ARMA(p,q)$ the list of starting values is $(e_0,\ldots,e_{q-1})$.

The starting values need to be estimated, which is usually carried out by a recursive technique. First assume they are all equal to zero and estimate the α_i and β_j on that basis, then use standard forecasting techniques on the time-reversed process $\{z_n,\ldots,z_1\}$ to obtain predicted values for $(e_0,\ldots,e_{q-1})$, a method known as *backforecasting*. These new values can be used as the starting point for another application of the estimation procedure; this continues until the estimates have converged.

This would be a time consuming process to carry out by hand but such iterative procedures are easy to implement on a computer.

In Figure 14.7, the ACF and PACF plots show some significant spikes in the early lags suggesting some presence of autoregressive and moving average.

The code:

```
      fit=arima(x,order=c(1,0,1));fit
```

fits the $ARIMA(1,0,1)$ to this data set, with standard output:

```
arima(x = x, order = c(1, 0, 1))
Coefficients:
          ar1     ma1   intercept
       0.6118  0.5849      0.0911
s.e.   0.0530  0.0600      0.2224
sigma^2 estimated as 0.9016:  log likelihood = -410.9,  aic = 829.8
```

where the estimated parameters ar1 and ma1 correspond to α and β in the $ARMA(1,1)$ model. The fitted model has AIC = 829.8, which is the smallest value among other possible models like $AR(1)$, $MA(1)$, $ARMA(1,2)$ and $ARMA(2,2)$.

An alternative method of estimation is based on method of moments estimation. There are $p+q$ parameters to be estimated. We can calculate the theoretical ACF $\{\rho_k\}$ of an ARMA (p,q) process, which will be a function of the α's and β's. Then the method of moments estimators are those values of α and β such that the theoretical ACF $\rho_1,\ldots,\rho_{p+q}$ coincides with the observed sample ACF $r_1,\ldots,r_{p+q}$. This method is easily available for $AR(p)$ models since the corresponding Yule-Walker equations are linear, therefore moment estimation requires solving them with respect to the unknown parameters α_i.

Question

It is believed that a set of data values is the realisation of a $MA(1)$ process $X_n = e_n + \beta e_{n-1}$ where the errors are standard normal. Given that $\hat{\gamma}_0 = 1$ and $\hat{\gamma}_1 = -0.25$, use the method of moments to estimate the parameter β, ensuring that the fitted process is invertible.

Solution

We have $r_1 = -0.25$. The theoretical values of the autocovariance at lags 0 and 1 are:

$$\gamma_0 = \text{cov}(X_t, X_t) = \text{cov}(e_t + \beta e_{t-1}, e_t + \beta e_{t-1}) = 1 + \beta^2$$

$$\gamma_1 = \text{cov}(X_t, X_{t+1}) = \text{cov}(e_t + \beta e_{t-1}, e_{t+1} + \beta e_t) = \beta$$

So $\rho_1 = \dfrac{\beta}{1+\beta^2}$. Equating this with the corresponding sample correlation at lag 1 gives:

$$\rho_1 = \frac{\beta}{1+\beta^2} = \frac{-0.25}{1} \quad \Rightarrow \quad \beta^2 + 4\beta + 1 = 0$$

Solving this quadratic we get $\beta = -0.268$ or -3.732.

Setting $\beta = -0.268$ gives an invertible process; setting $\beta = -3.732$ does not. So we take -0.268 to be our estimated value of β.

The final parameter of the model is σ^2, the variance of the e_t, which may be estimated using:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=p+1}^{n} \hat{e}_t^2 = \frac{1}{n}\sum_{t=p+1}^{n} (z_t - \hat{\alpha}_1 z_{t-1} - \cdots - \hat{\alpha}_p z_{t-p} - \hat{\beta}_1 \hat{e}_{t-1} - \cdots - \hat{\beta}_q \hat{e}_{t-q})^2$$

where $\hat{e}_t$ denotes the residual at time t.

The residuals are the realised values of the white noise terms.

The parameter σ^2 needs to be estimated last as the formula above involves the estimated values of the α's and β's. σ^2 has not been used to estimate the other parameters, whichever method is used, so the overall estimation procedure is well-defined. However, if we use the method of moments to estimate the α's and β's, we would have to estimate the errors separately (as we would not have estimated these as part of the procedure for the estimation of the α's and β's).

If the number of observations n of the time series is sufficiently large there will be little difference between the least squares estimates and the method of moments estimates of the parameters.

3.5 Diagnostic checking

After the tentative identification of an $ARIMA(p,d,q)$ model and calculation of the estimates $\hat{\mu}, \hat{\sigma}, \hat{\alpha}_1, \ldots, \hat{\alpha}_p, \hat{\beta}_1, \ldots, \hat{\beta}_q$ we have to perform diagnostic checking. The principle of this is that, if the $ARMA(p,q)$ model is a good approximation to the underlying time series process, then the residuals $\hat{e}_t$ will form a good approximation to a white noise process.

One test that we could perform on the residuals is the Ljung and Box portmanteau test, which is covered in Section 2.2. There are other tests, however, some of which are outlined below.

The following checks are frequently used.

Inspection of the graph of the residuals

The visual inspection of the graph of the residuals against t or the graph of $\hat{e}_t$ against z_t can help to highlight a poorly fitting model. If any pattern is evident, whether in the average level of the residuals or in the magnitude of the fluctuations about 0, this should be taken to mean that the model is inadequate.

The question is really whether the errors are a likely realisation of a set of independent normal random variables. So a 'pattern' is anything that may suggest non-independence.

We could be slightly more quantitative with this. Assuming the errors are independent normal random variables there should be a certain proportion of the sample errors within any given range. For example, if the errors were thought to be standard normal, then we would expect 34% of the sample values to lie between 0 and 1.

Inspection of the sample autocorrelation functions of the residuals

The behaviour of the sample ACF and sample PACF of a white noise sequence have already been described.

This was done in Section 2.2.

If the SACF or SPACF of the sequence of residuals has too many values outside the range $\pm 2/\sqrt{N}$, we conclude that the fitted model does not have enough parameters and a new model with additional parameters should be fitted. The Ljung-Box chi-squared statistic may also be used for this purpose, but the degrees of freedom of the test statistics needs to be reduced by the number of parameters $p+q$ of the *ARMA* model.

Here the Core Reading is using N rather than n to denote the number of recorded values of the time series.

Counting turning points

If $y_1, y_2, \ldots, y_N$ is a sequence of numbers, then we say that the sequence has a *turning point* at time k if either $y_{k-1} < y_k$ and $y_k > y_{k+1}$, or $y_{k-1} > y_k$ and $y_k < y_{k+1}$.

If $Y_1, Y_2, \ldots, Y_N$ is a sequence of independent random variables with continuous distribution, then the probability of a turning point at time k is $^2\!/_3$, the expected number of turning points is $\frac{2}{3}(N-2)$, and the variance is $\dfrac{16N-29}{90}$.

This is a result for a sequence of *independent* random variables. It is therefore usually applied to the *residuals* of the time series, not to the original time series itself, which will not be independent. The procedure is therefore to calculate the residuals, and then to apply the turning point test to them to see if the residuals have a reasonable number of turning points.

A proof of this result is beyond the scope of Subject CS2. The formulae for the mean and variance of the number of turning points are both on page 42 of the *Tables*.

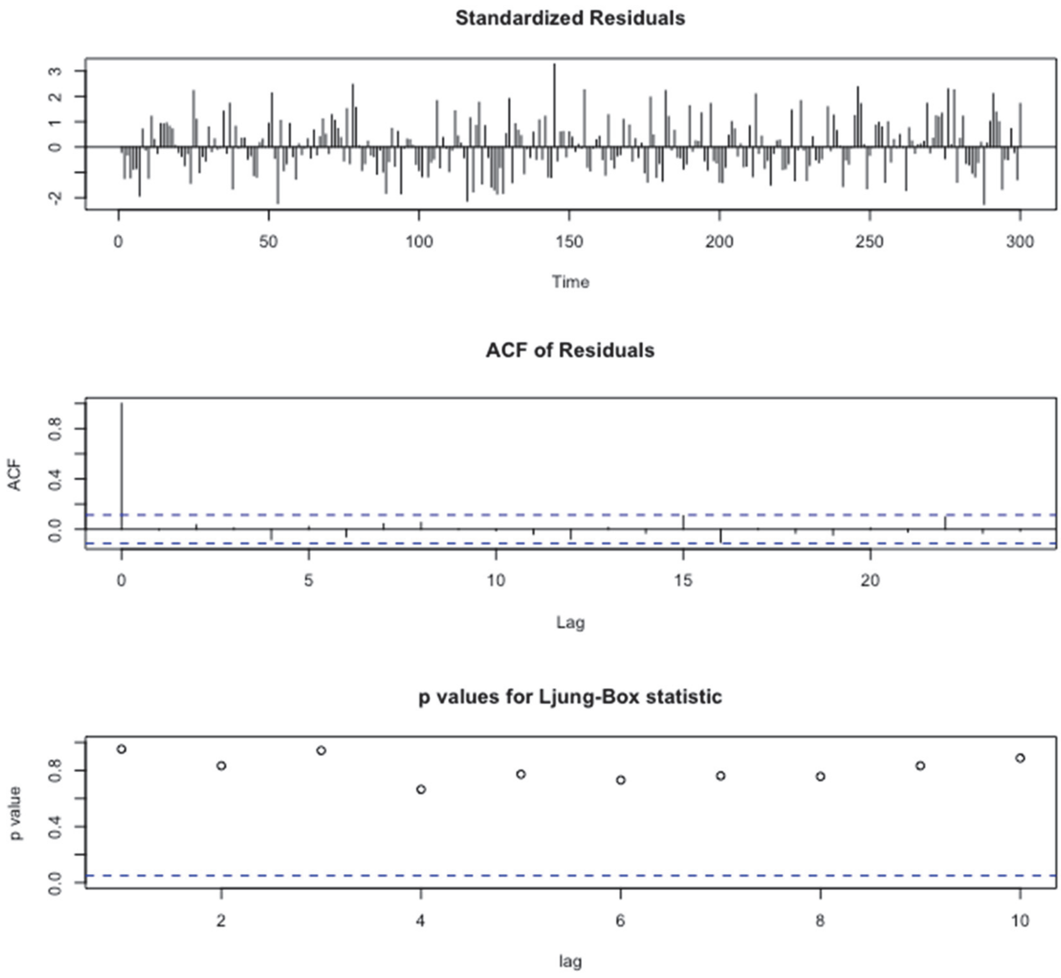
Therefore the number of turning points in a realisation of $Y_1, Y_2, \ldots, Y_N$ should be within the 95% confidence interval:

$$\left[\frac{2}{3}(N-2) - 1.96\sqrt{\frac{16N-29}{90}}\,,\ \frac{2}{3}(N-2) + 1.96\sqrt{\frac{16N-29}{90}}\ \right]$$

The command:

```
tsdiag(fit)
```

generates a graphical summary of the diagnostic checks of the residuals.



where the last plot shows a sequence of *p*-values of the Ljung-Box test, high values observed suggesting good fit, *ie* residuals close to white noise.

Figure 14.8: Diagnostic checks of residuals

4 Forecasting

4.1 Box-Jenkins approach to forecasting stationary time series

Using the Box-Jenkins approach, forecasting is relatively straightforward. Having fitted an ARMA model to the data $\{x_1,\ldots,x_n\}$ we have the equation:

$$X_{n+k} = \mu + \alpha_1(X_{n+k-1} - \mu) + \cdots + \alpha_p(X_{n+k-p} - \mu) + e_{n+k} + \beta_1 e_{n+k-1} + \cdots + \beta_q e_{n+k-q}$$

Recall that at the start of Section 3.3, it was assumed that the data set to which we were fitting the ARMA model had zero mean. In the expression above the mean μ has now been put back in.

Forecasting future values of an ARMA process

The forecast value of X_{n+k} given all observations up until time n, known as the k*-step ahead forecast* and denoted $\hat{x}_n(k)$, is obtained from this equation by:

- **replacing all (unknown) parameters by their estimated values;**
- **replacing the random variables $X_1,\ldots,X_n$ by their observed values $x_1,\ldots,x_n$;**
- **replacing the random variables $X_{n+1},\ldots,X_{n+k-1}$ by their forecast values $\hat{x}_n(1),\ldots,\hat{x}_n(k-1)$;**
- **replacing the innovations $e_1,\ldots,e_n$ by the residuals $\hat{e}_1,\ldots,\hat{e}_n$;**
- **replacing the random variables $e_{n+1},\ldots,e_{n+k-1}$ by their expectations, 0.**

For example, the one-step ahead and two-step ahead forecasts for an $AR(2)$ are given by:

$$\hat{x}_n(1) = \hat{\mu} + \hat{\alpha}_1(x_n - \hat{\mu}) + \hat{\alpha}_2(x_{n-1} - \hat{\mu})$$

$$\hat{x}_n(2) = \hat{\mu} + \hat{\alpha}_1(\hat{x}_n(1) - \hat{\mu}) + \hat{\alpha}_2(x_n - \hat{\mu})$$

Question

Write down an expression for the two-step ahead forecast of a general $ARMA(2,2)$ process.

Solution

The two-step ahead forecast is:

$$\hat{x}_n(2) = \hat{\mu} + \hat{\alpha}_1\left(\hat{x}_n(1) - \hat{\mu}\right) + \hat{\alpha}_2\left(x_n - \hat{\mu}\right) + \hat{\beta}_2\,\hat{e}_n$$

Thus the k-step ahead forecast is essentially the conditional expectation of the future value of the process given all the information currently available at time n.

A point estimate of X_{n+k} is less useful than a confidence interval, for which an estimate of the variance is required. A comparison of X_{n+1} with $\hat{x}_n(1)$ shows that the difference between them arises from numerous sources, including e_{n+1}, differences between true values of parameters and their estimates, and differences between true values of the e_t and the residuals $\hat{e}_t$ which are used to estimate them. Calculation of the prediction variance in any given case is complicated and is best left to a computer. In general, though, it is possible to state that the variance of the k-step ahead estimator is relatively small for small values of k and converges, for large k, to γ_0, the variance of the stationary process X.

Question

Explain why the *k*-step ahead variance converges toward the variance of the process.

Solution

The process is stationary. So starting from time n and projecting into the future, the process eventually settles down into the equilibrium distribution.

4.2 Forecasting ARIMA processes

If X is an $ARIMA(p,d,q)$ process, then $Z = \nabla^d X$ is $ARMA(p,q)$, so the techniques of Section 4.1 can be used to produce forecasts and confidence intervals for future values of Z. By reversing the differencing procedure these can be translated into forecasts of future values of X.

For example, suppose that X is $ARIMA(0,1,1)$.

Then $Z_n = \nabla X_n = X_n - X_{n-1}$ is $ARMA(0,1)$, and $X_n = X_{n-1} + Z_n$.

Hence $X_{n+1} = X_n + Z_{n+1}$, and $\hat{x}_n(1) = x_n + \hat{z}_n(1)$.

Question

Give an expression for the two-step ahead forecast of an $ARIMA(1,2,1)$ process.

Solution

If we define:

$$Z_n = \nabla^2 X_n = (1-B)^2 X_n = (1-2B+B^2)X_n = X_n - 2X_{n-1} + X_{n-2}$$

This implies that:

$$X_n = Z_n + 2X_{n-1} - X_{n-2}$$

Hence:

$$X_{n+2} = 2X_{n+1} - X_n + Z_{n+2}$$

So the two-step ahead forecast is:

$$\hat{x}_n(2) = 2\hat{x}_n(1) - x_n + \hat{z}_n(2)$$

Since Z is $ARMA(1,1)$, it has a defining equation of the form:

$$Z_n = \mu + \alpha\left(Z_{n-1} - \mu\right) + e_n + \beta e_{n-1}$$

So:

$$Z_{n+2} = \mu + \alpha\left(Z_{n+1} - \mu\right) + e_{n+2} + \beta e_{n+1}$$

and:

$$\hat{z}_n(2) = \hat{\mu} + \hat{\alpha}\left(\hat{z}_n(1) - \hat{\mu}\right)$$

Hence the two-step ahead forecast can be expressed as:

$$\hat{x}_n(2) = 2\hat{x}_n(1) - x_n + \hat{\mu} + \hat{\alpha}\left(\hat{z}_n(1) - \hat{\mu}\right)$$

An $ARIMA(p,d,q)$ process with $d > 0$ is not stationary and therefore has no stationary variance. It should come as no surprise, then, that the prediction variance for the k-step ahead forecast increases to infinity as k increases. This is easily seen in the case of the random walk process.

For predicting three steps ahead:

```
predict(fit,n.ahead=3)
```

4.3 Exponential smoothing

The Box-Jenkins methodology is demanding, requiring a skilled operator to produce reliable results. There are many instances in which a company needs no more than a simple forecast of some future value without having to employ a trained statistician to provide it. A much simpler forecasting technique, introduced by Holt in 1958, uses a weighted combination of past values to predict future observations.

One-step ahead forecast using exponential smoothing

$$\hat{x}_n(1) = \alpha(x_n + (1-\alpha)x_{n-1} + (1-\alpha)^2 x_{n-2} + \cdots)$$

The weights used here are $\alpha, \alpha(1-\alpha), \alpha(1-\alpha)^2, \dots$.

Here α is a single parameter, either chosen by the user or estimated by least squares from past data. Typically a value in the range 0.2 to 0.3 is used.

A value of α between 0 and 1 will give a weighted average of historic values with less emphasis on values that are further back in time.

The geometrically decreasing weights give rise to the name *exponential smoothing*.

Since the weights sum to 1, the exponential smoothing filter is a weighted average of historic values, with the weights decreasing geometrically as we go further back in time.

Question

Show that the weights sum to 1 when $0<\alpha<1$.

Solution

The weights form a geometric progression with first term α and common ratio $1-\alpha$. Using the formula for the sum to infinity of a geometric progression, we see that:

$$\alpha+\alpha(1-\alpha)+\alpha(1-\alpha)^2+\cdots=\frac{\alpha}{1-(1-\alpha)}=1$$

The method lends itself easily to regular updating. It is easy to see that:

$$\hat{x}_n(1)=(1-\alpha)\hat{x}_{n-1}(1)+\alpha x_n=\hat{x}_{n-1}(1)+\alpha\left[x_n-\hat{x}_{n-1}(1)\right]$$

so that the current forecast is obtained by taking the previous forecast and compensating for the error observed when the actual figure became available.

This technique works for stationary series, but clearly cannot be applied to series exhibiting a trend or seasonal variation. There are more sophisticated versions of exponential smoothing which are able to cope with trends or seasonal variation, and are even well equipped to handle slowly varying trends or multiplicative, rather than additive, seasonal variation.

5 Multivariate time series models

5.1 Vector autoregressions

An m-dimensional multivariate time series $\{\underline{x}_1,\ldots,\underline{x}_n\}$ is a sequence of m-dimensional vectors. Each vector $\underline{x}_t$ is a set of observations of the values of m variables of interest at time t. A multivariate time series is modelled by a sequence of random vectors $\{\underline{X}_1,\underline{X}_2,\ldots\}$. The components of $\underline{X}_t$ will be denoted $X_t^{(1)},\ldots,X_t^{(m)}$.

The second-order properties of a sequence of random vectors are summarised by:

- **the *vectors of expected values* $\underline{\mu}_t = E\left[\underline{X}_t\right]$, and**
- **the *covariance matrices* for all pairs of random vectors, $\text{cov}\left(\underline{X}_t,\underline{X}_{t+k}\right)$.**

The definition of stationarity is the same in the multidimensional case as it is for univariate time series: the vector process is (weakly) stationary if both $E\left[\underline{X}_t\right]$ and $\text{cov}\left(\underline{X}_t,\underline{X}_{t+k}\right)$ are independent of t. In the stationary case, the notation $\underline{\mu}$ will be used to represent the common mean vector, Σ_k the covariance matrix $\text{cov}\left(\underline{X}_t,\underline{X}_{t+k}\right)$.

The diagonal elements of the covariance matrix Σ_k are clearly the autocovariances at lag k of the individual components of the random vector $\underline{X}_t$. The off-diagonal elements $\Sigma_k(i,j)$ are called the lag k *cross-covariances* of $X^{(i)}$ with $X^{(j)}$, $\text{cov}(X_t^{(i)},X_{t+k}^{(j)})$.

Example 1

A multivariate white noise process is the simplest example of a multivariate random process. Suppose $\underline{e}_1,\underline{e}_2,\ldots$ is a sequence of independent zero-mean random vectors, each having the same covariance matrix Σ.

Σ need not be a diagonal matrix, though it must be symmetrical. In other words, the components of the innovations vector need not be independent of one another. This is a multivariate analogue of the zero-mean white noise.

A *vector autoregressive process* of order p, denoted $VAR(p)$, is a sequence of m-component random vectors $\{\underline{X}_1,\underline{X}_2,\ldots\}$ satisfying:

$$\underline{X}_t = \underline{\mu} + \sum_{j=1}^{p} A_j\left(\underline{X}_{t-j} - \underline{\mu}\right) + \underline{e}_t \qquad \textbf{(14.2)}$$

where $\underline{e}$ is an *m*-dimensional white noise process and the A_j are $m \times m$ matrices.

So a $VAR(1)$ process has the following structure:

$$\underline{X}_t - \underline{\mu} = A(\underline{X}_{t-1} - \underline{\mu}) + \underline{e}_t$$

Example 2

We might believe that interest rates, i_t, and tendency to invest, I_t, are related to one another by the equations:

$$\begin{cases} i_t - \mu_i = \alpha_{11}(i_{t-1} - \mu_i) + e_t^{(i)} \\ I_t - \mu_I = \alpha_{21}(i_{t-1} - \mu_i) + \alpha_{22}(I_{t-1} - \mu_I) + e_t^{(I)} \end{cases} \qquad (14.3)$$

where $e^{(i)}$ and $e^{(I)}$ are zero-mean (univariate) white noises. They may have different variances and are not necessarily uncorrelated; that is, we do not require $\text{cov}\left(e^{(i)}, e^{(I)}\right) = 0$, although we do require $\text{cov}\left(e_t^{(i)}, e_s^{(I)}\right) = 0$ for $s \neq t$.

This model can be expressed as a 2-dimensional $VAR(1)$ process:

$$\begin{pmatrix} i_t - \mu_i \\ I_t - \mu_I \end{pmatrix} = \begin{pmatrix} \alpha_{11} & 0 \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} i_{t-1} - \mu_i \\ I_{t-1} - \mu_I \end{pmatrix} + \begin{pmatrix} e_t^{(i)} \\ e_t^{(I)} \end{pmatrix}$$

The theory and analysis of a $VAR(1)$ closely parallels that of a univariate $AR(1)$. Iterating from Equation (14.2) in the case $p = 1$, it is clear that:

$$\underline{X}_t = \underline{\mu} + \sum_{j=0}^{t-1} A^j \underline{e}_{t-j} + A^t \left(\underline{X}_0 - \underline{\mu}\right)$$

In order that $\underline{X}$ should represent a stationary time series, the powers of A should converge to zero in some sense. The appropriate requirement is that all eigenvalues of the matrix A should be less than 1 in absolute value.

Condition for stationarity of a $VAR(1)$ process

A process of the form:

$$\underline{X}_t - \underline{\mu} = A(\underline{X}_{t-1} - \underline{\mu}) + \underline{e}_t$$

is stationary if all the eigenvalues of the matrix A are strictly less than 1 in magnitude.

The eigenvalues of matrix A are the values λ such that $\det(A - \lambda I) = 0$ where I is the identity matrix.

***Eg* for a 2-dimensional time series this equation reduces to:**

$$(\alpha_{11} - \lambda)(\alpha_{22} - \lambda) - \alpha_{12}\alpha_{21} = 0$$

where $A[i,j] = \alpha_{ij}$ $(i = 1,2, j = 1,2)$.

This is because we require:

$$\det(A-\lambda I)=\det\left(\begin{pmatrix}\alpha_{11} & \alpha_{12}\\ \alpha_{21} & \alpha_{22}\end{pmatrix}-\lambda\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)=\det\begin{pmatrix}\alpha_{11}-\lambda & \alpha_{12}\\ \alpha_{21} & \alpha_{22}-\lambda\end{pmatrix}=0$$

The appendix to this chapter briefly revises the definition of an eigenvalue.

Question

Determine whether the following multivariate time series is stationary:

$$\begin{pmatrix}X_t\\ Y_t\end{pmatrix}=\begin{pmatrix}0.3 & 0.5\\ 0.2 & 0.2\end{pmatrix}\begin{pmatrix}X_{t-1}\\ Y_{t-1}\end{pmatrix}+\begin{pmatrix}e_t^X\\ e_t^Y\end{pmatrix}$$

Solution

To check stationarity, we need to calculate the eigenvalues of the matrix $\begin{pmatrix}0.3 & 0.5\\ 0.2 & 0.2\end{pmatrix}$, *ie* we need to determine the values of λ for which:

$$\det\begin{pmatrix}0.3-\lambda & 0.5\\ 0.2 & 0.2-\lambda\end{pmatrix}=(0.3-\lambda)(0.2-\lambda)-0.1=\lambda^2-0.5\lambda-0.04=0$$

The solutions of this equation are $\lambda=0.57$ and $\lambda=-0.07$. Since both eigenvalues are strictly less than 1 in magnitude, the process is stationary.

The matrix equation given in the previous question can be written as the pair of equations:

$$X_t=0.3X_{t-1}+0.5Y_{t-1}+e_t^X$$

$$Y_t=0.2X_{t-1}+0.2Y_{t-1}+e_t^Y$$

Rearranging the first equation we get:

$$Y_{t-1}=2\left(X_t-0.3X_{t-1}-e_t^X\right)$$

We can substitute this in the second equation for Y_{t-1} and the similar expression for Y_t that it implies. After tidying up we have:

$$X_{t+1}=0.5X_t+0.04X_{t-1}+e_{t+1}^X-0.2e_t^X+0.5e_t^Y$$

We can check for stationarity by looking at the roots of the characteristic equation of the autoregressive part:

$$1-0.5\lambda-0.04\lambda^2=0$$

This has roots –14.25 and 1.75. Since both roots are strictly greater than 1 in magnitude, the process X is stationary in its own right. Note that these roots are the reciprocals of the eigenvalues calculated in the previous question.

Similar, though more complicated, requirements can be set out under which a more general $VAR(p)$ process is stationary.

Fitting a vector autoregression is very similar to the process of fitting a univariate autoregression. Parameter estimation can be carried out by least squares or by method of moments. Some elements of the univariate theory, such as the use of Akaike's Information Criterion, do not translate unchanged into a multivariate setting, but other topics carry across relatively easily.

Example

The following simple dynamic Keynesian model provides an example of a multivariate autoregressive process.

Keynesian models are studied in economics. Here we are not really interested in the economic theory that leads to this model. The important point is to understand the vector time series equations.

Denote by Y_t the national income over a certain period of time, and denote by C_t and I_t the total consumption and investment over the same period. It is assumed that the consumption, C_t, depends on the income over the previous period:

$$C_t = \alpha Y_{t-1} + e_t^{(1)}$$

where $e^{(1)}$ is a zero-mean white noise. The investment, I_t, is determined by the 'accelerator' mechanism:

$$I_t = \beta(C_{t-1} - C_{t-2}) + e_t^{(2)}$$

where $e^{(2)}$ is another zero-mean white noise. Finally, any part of the national income is either consumed or invested; therefore:

$$Y_t = C_t + I_t$$

Eliminating the national income we arrive at the following two-dimensional $VAR(2)$ process:

$$C_t = \alpha C_{t-1} + \alpha I_{t-1} + e_t^{(1)}$$

$$I_t = \beta(C_{t-1} - C_{t-2}) + e_t^{(2)}$$

Using matrix notation we can rewrite the above equation as:

$$\begin{pmatrix} C_t \\ I_t \end{pmatrix} = \begin{pmatrix} \alpha & \alpha \\ \beta & 0 \end{pmatrix} \begin{pmatrix} C_{t-1} \\ I_{t-1} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ -\beta & 0 \end{pmatrix} \begin{pmatrix} C_{t-2} \\ I_{t-2} \end{pmatrix} + \begin{pmatrix} e_t^{(1)} \\ e_t^{(2)} \end{pmatrix}$$

5.2 Cointegrated time series

Recall that a time series process X is called integrated of order d, abbreviated as $I(d)$, if the process $Y = \nabla^d X$ is stationary.

Cointegrated series

Two time series processes X and Y are called *cointegrated* if:

(i) X and Y are $I(1)$ random processes,

(ii) there exists a non-zero vector (α, β) such that $\alpha X + \beta Y$ is stationary.

The vector (α, β) is called a *cointegrating vector*.

In other words, two processes are themselves non-stationary (technically $I(1)$), but their movements are correlated in such a way that a certain weighted average of the two processes is stationary.

There are a number of circumstances when it is reasonable to expect that two processes may be cointegrated:

- **if one of the processes is driving the other**
- **if both are being driven by the same underlying process.**

We now consider an example of cointegrated processes. Understanding the underlying economic theory is not so important here. However, it is important to understand why the processes are cointegrated.

Example

The following simple model of evolution of the USDollar/GBPound exchange rate X_t provides an example of a cointegrated model. It is assumed that the exchange rate fluctuates around the purchasing power P_t/Q_t, where P_t and Q_t are the consumer price indices for US and UK, respectively.

This is described by the following model:

$$\ln X_t = \ln \frac{P_t}{Q_t} + Y_t$$

$$Y_t = \mu + \alpha(Y_{t-1} - \mu) + e_t + \beta e_{t-1}$$

where e is a zero-mean white noise process.

The evolution of $\ln P$ and $\ln Q$ is described by $ARIMA(1,1,0)$ models:

$$\mathbf{(1-B)\ln P_t = \mu_1 + \alpha_1\left[(1-B)\ln P_{t-1} - \mu_1\right] + e_t^{(1)}}$$

$$\mathbf{(1-B)\ln Q_t = \mu_2 + \alpha_2\left[(1-B)\ln Q_{t-1} - \mu_2\right] + e_t^{(2)}}$$

where $e^{(1)}$ and $e^{(2)}$ are zero-mean white noises, possibly correlated.

ln P and ln Q are both $ARIMA(1,1,0)$ processes. The logarithm of the exchange rate is also non-stationary. However:

$$\mathbf{\ln X - \ln P + \ln Q}$$

is the $ARMA(1,1)$ random process Y and, therefore, is a stationary random process.

Here we are assuming that $|\alpha|<1$.

It follows that the sequence of random vectors:

$$\mathbf{\{(\ln X_t, \ln P_t, \ln Q_t) : t = 1,2,\ldots\}}$$

is described by a cointegrated model with the cointegrating vector $(1,-1,1)$.

Question

Two time series X and Y are defined by the equations:

$$X_t = 0.65X_{t-1} + 0.35Y_{t-1} + e_t^X$$

$$Y_t = 0.35X_{t-1} + 0.65Y_{t-1} + e_t^Y$$

where e^X and e^Y are independent white noise processes.

Show that X and Y are cointegrated, with cointegrating vector $(1,-1)$.

Solution

We begin by showing that the processes X and Y are $I(1)$.

From the first equation we have:

$$Y_{t-1} = \frac{1}{0.35}\left(X_t - 0.65X_{t-1} - e_t^X\right)$$

Using this in the second equation gives:

$$\frac{1}{0.35}\left(X_{t+1} - 0.65X_t - e_{t+1}^X\right) = 0.35X_{t-1} + 0.65\frac{1}{0.35}\left(X_t - 0.65X_{t-1} - e_t^X\right) + e_t^Y$$

Tidying up we have:

$$X_{t+1} = 1.3X_t - 0.3X_{t-1} + e_{t+1}^X - 0.65e_t^X + 0.35e_t^Y$$

or equivalently:

$$X_t = 1.3X_{t-1} - 0.3X_{t-2} + e_t^X - 0.65e_{t-1}^X + 0.35e_{t-1}^Y$$

The characteristic polynomial of the AR part of this equation is:

$$1 - 1.3\lambda + 0.3\lambda^2$$

The roots of this equation are $\frac{10}{3}$ and 1. So X is not stationary. Differencing once will eliminate the root of 1. Since the only other root is strictly greater than 1 in magnitude, ∇X is a stationary process. Hence X is $I(1)$.

The process Y has a similar structure to that of X, so Y is also $I(1)$.

We now need to show that $X - Y$ is stationary.

We have equations:

$$X_t = 0.65X_{t-1} + 0.35Y_{t-1} + e_t^X$$

$$Y_t = 0.35X_{t-1} + 0.65Y_{t-1} + e_t^Y$$

Subtracting the second equation from the first we see that:

$$X_t - Y_t = 0.3X_{t-1} - 0.3Y_{t-1} + e_t^X - e_t^Y = 0.3\left(X_{t-1} - Y_{t-1}\right) + e_t^X - e_t^Y$$

Setting $W_t = X_t - Y_t$, this is:

$$W_t = 0.3W_{t-1} + (e_t^X - e_t^Y)$$

The process W is a stationary $AR(1)$ process since the root of its characteristic equation is $\frac{10}{3}$, and this is greater than 1 in magnitude. (As always, the white noise terms don't affect the stationarity.)

6 Some special non-stationary and non-linear time series models

Although the ARIMA class of processes is the most important for us, there are many other types of model. This section briefly discusses a few of them.

6.1 Bilinear models

The general class of bilinear models can be exemplified by its simplest representative, the random process X defined by the relation:

$$X_n - \alpha(X_{n-1} - \mu) = \mu + e_n + \beta e_{n-1} + b(X_{n-1} - \mu)e_{n-1}$$

Considered only as a function of X, this relation is linear; it is also linear when considered as a function of e only. This is why it is called 'bilinear'.

The main qualitative difference between the bilinear model and models from the ARMA class is that many bilinear models exhibit 'bursty' behaviour: when the process is far from its mean it tends to exhibit larger fluctuations. The difference between this model and an $ARMA(1,1)$ process may be seen to lie in the last term on the right-hand side: when X_{n-1} is far from μ and e_{n-1} is far from 0 – events which are far from being independent – the final term assumes a much greater significance.

6.2 Threshold autoregressive models

A simple representative of the class of threshold autoregressive models is the random process X defined by the relation:

$$X_n = \mu + \begin{cases} \alpha_1(X_{n-1} - \mu) + e_n, & \text{if } X_{n-1} \le d \\ \alpha_2(X_{n-1} - \mu) + e_n, & \text{if } X_{n-1} > d \end{cases}$$

The distinctive feature of some models from the threshold autoregressive class is the limit cycle behaviour. This makes the threshold autoregressive models suitable for the description of 'cyclic' phenomena.

In an extreme case we might set $\alpha_2 = 0$ for example. Then X_n follows an autoregressive process until it passes the threshold value d. At this point X_n returns to μ and the process effectively starts again. Thus we get cyclic behaviour as the process keeps resetting.

6.3 Random coefficient autoregressive models

Another modification of the AR class of models is that of autoregressive models for which the coefficient is random. In other words:

$$X_t = \mu + \alpha_t(X_{t-1} - \mu) + e_t$$

where $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ is a sequence of independent random variables.

Such a model could be used to represent the behaviour of an investment fund, with $\mu = 0$ and $\alpha_t = 1 + i_t$ with i_t being the random rate of return.

The behaviour of these processes can vary widely, depending on the distribution chosen for the α_t, but is in general more irregular than that of the corresponding $AR(1)$.

6.4 Autoregressive models with conditional heteroscedasticity

Financial assets often display the following behaviour. After a large change in the asset price there follows a period of high volatility, which can be in either direction. Following small changes there tend to be further small changes. In other words, the variance of the process is dependent upon the size of the previous value. This is the property of *conditional heteroscedasticity*.

The words 'homoscedastic' and 'heteroscedastic' just mean having equal (*ie* constant) or different variances respectively. The 'c' is pronounced like a 'k' in these words.

The class of autoregressive models with conditional heteroscedasticity of order p – the $ARCH(p)$ – is defined by the relation:

$$X_t = \mu + e_t \sqrt{\alpha_0 + \sum_{k=1}^{p} \alpha_k \left(X_{t-k} - \mu\right)^2}$$

where e is a sequence of independent standard normal random variables. The simplest representative of the $ARCH(p)$ class is the $ARCH(1)$ model defined by the relation:

$$X_t = \mu + e_t \sqrt{\alpha_0 + \alpha_1 \left(X_{t-1} - \mu\right)^2}$$

If μ is zero, it can be shown that $\text{cov}(X_t, X_s) = 0$ for $s \neq t$ confirming that X_t is white noise with uncorrelated but not independent components.

The ARCH models have been used for modelling financial time series. If Z_t is the price of an asset at the end of the t-th trading day, it is found that the ARCH model can be used to model $X_t = \ln\left(Z_t / Z_{t-1}\right)$, interpreted as the daily return on day t.

The ARCH family of models captures the feature frequently observed in asset price data that a significant change in the price of an asset is often followed by a period of high volatility. As may be seen from the $ARCH(1)$ model, a significant deviation of X_{t-1} from the mean μ gives rise to an increase in the conditional variance of X_t given X_{t-1}.

7 Appendix – Eigenvalues

λ is an *eigenvalue* of an $n \times n$ matrix A if there is a non-zero vector $\underline{x}$, such that:

$$A\underline{x} = \lambda \underline{x}$$

The vector $\underline{x}$ is known as the *eigenvector.* This equation is equivalent to $(A - \lambda I)\underline{x} = 0$ where I is the identity matrix, *ie* the matrix whose diagonal entries are 1 and whose off-diagonal entries are 0.

Hence we have a set of n linear equations in n unknowns, $\{x_1, \ldots, x_n\}$. These equations have a non-zero solution, if and only if the matrix $A - \lambda I$ has zero determinant:

$$\det(A - \lambda I) = 0$$

The equation $\det(A - \lambda I) = 0$ can be solved for λ to find the eigenvalues.

Question

Calculate the eigenvalues of the matrix $A = \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$.

Solution

Here we have:

$$A - \lambda I = \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} = \begin{pmatrix} 2-\lambda & 1 \\ 4 & 2-\lambda \end{pmatrix}$$

We have to solve:

$$\det \begin{pmatrix} 2-\lambda & 1 \\ 4 & 2-\lambda \end{pmatrix} = 0$$

The determinant of the 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $ad - bc$. So the eigenvalues of the matrix A are the solutions of the equation:

$$(2-\lambda)^2 - 4 = \lambda^2 - 4\lambda = \lambda(\lambda - 4) = 0$$

These are 0 and 4.

Chapter 14 Summary

Box-Jenkins methodology

The Box-Jenkins methodology gives us a way of fitting an $ARIMA(p,d,q)$ time series model to an actual data set. The method consists of the following steps:

- removing trends and cycles from the data set
- identifying a model from the $ARIMA(p,d,q)$ class
- estimating parameters
- diagnostic checks
- forecasting.

Removing trends

Time series data can be modelled efficiently only if stationary. In particular, any deterministic trends or cycles must be removed before applying the modelling procedure. There are various ways of doing this. In addition, a time series may still be non-stationary because it is integrated. In this case the time series must be differenced.

Let the set of observed values of the time series process be $\{x_t\}$.

Linear trends in the data can be removed by:

- least squares trend removal, *ie* we calculate $y_t = x_t - \hat{a} - \hat{b}t$ where $\hat{a}$ and $\hat{b}$ have been determined using linear regression
- differencing, *ie* we calculate $y_t = \nabla x_t = x_t - x_{t-1}$.

Seasonal trends in the data can be removed by:

- seasonal differencing, *eg* if seasonal variation is observed in monthly data, we calculate $y_t = x_t - x_{t-12}$
- method of moving averages (applying a filter), *eg* if seasonal variation is observed in monthly data, we calculate:

$$y_t = \frac{1}{12}\left(0.5x_{t-6} + x_{t-5} + \cdots + x_{t-1} + x_t + x_{t+1} + \cdots + x_{t+5} + 0.5x_{t+6}\right)$$

- method of seasonal means, *eg* if seasonal variation is observed in monthly data, we subtract from each observation the estimated mean for that month.

It may be possible to remove other trends in the data via a transformation. For example, if the $\{x_t\}$ values appear to have an exponential trend, we could apply the transformation $y_t = \log x_t$.

Fitting an ARIMA process – choosing a value for *d*

The following principles can be used to choose an appropriate value for d.

1. If the sample autocorrelation function r_k decays slowly to 0, this indicates that there are still trends in the data and that the data should be differenced again.

2. Let $\hat{\sigma}_d^2$ denote the sample variance of the process $\nabla^d x_t$, then d can be set to the value which minimises $\hat{\sigma}_d^2$.

Fitting an ARIMA process – choosing values for *p* and *q*

If the underlying time series process is $MA(q)$, then we would expect the sample autocorrelation function $\hat{\rho}_k$ (or r_k) to cut off for $k > q$. It can also be shown that the estimator $\tilde{\rho}_k$ has the following approximate distribution for $k > q$:

$$\tilde{\rho}_k \div N\left(0, \frac{1}{n}\left(1 + 2\sum_{i=1}^{q} \rho_i^2\right)\right)$$

We might conclude that the ACF cuts off for $k > q$ if 95% of the $\hat{\rho}_k$ values fall within the confidence interval:

$$\left[-1.96\sqrt{\frac{1}{n}\left(1+2\sum_{i=1}^{q}\hat{\rho}_i^2\right)}, +1.96\sqrt{\frac{1}{n}\left(1+2\sum_{i=1}^{q}\hat{\rho}_i^2\right)}\right]$$

If the underlying time series process is $AR(p)$, then we would expect the sample partial autocorrelation function $\hat{\phi}_k$ to cut off for $k > p$. It can also be shown that the estimator $\tilde{\phi}_k$ has the following approximate asymptotic distribution:

$$\tilde{\phi}_k \div N\left(0, \frac{1}{n}\right)$$

We might conclude that the PACF cuts off for $k > p$ if 95% of the $\hat{\phi}_k$ values fall within the confidence interval:

$$\left[-1.96\sqrt{\frac{1}{n}}, +1.96\sqrt{\frac{1}{n}}\right]$$

Otherwise, we look to fit an $ARMA(p,q)$ model. In practice, we might start with an $ARMA(1,1)$ model and then apply diagnostic tests on the residuals to see whether this is a reasonable fit. If not, we would try adding more parameters.

Akaike's Information Criterion (AIC) states that we should only consider adding an extra parameter if this results in a reduction of the residual sum of squares by a factor of at least $e^{-2/n}$.

The formula for the $\hat{\rho}_k$ is given on page 40 of the *Tables*. The sample partial autocorrelation can then be calculated using the formulae (also on page 40 of the *Tables*) but with ρ_k replaced by $\hat{\rho}_k$.

Parameter estimation

Once we have identified p, d and q, we move forward with a time series of the form:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \cdots + \beta_q e_{t-q}$$

The parameters, the alphas and betas can be estimated as follows:

- least squares estimation (which is equivalent to maximum likelihood estimation if the error terms can be assumed to be normally distributed)
- method of moments, where we equate population autocorrelations ρ_k with sample autocorrelations $\hat{\rho}_k$.

The final parameter of the model is σ^2, the variance of the e_t, which may be estimated using:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=p+1}^{n} \hat{e}_t^{\,2} \quad \text{where } \hat{e}_t \text{ denotes the estimate of the residual at time } t$$

Diagnostic tests

If the model chosen is a good fit to the data, we would expect the estimates of the residuals $\{\hat{e}_t\}$ to show the characteristics of white noise (*ie* a set of uncorrelated random variables with zero mean). Examples of diagnostic tests include:

- checking that the graph of the $\{\hat{e}_t\}$ terms is patternless
- the $\{\hat{e}_t\}$ terms appear to be close to zero
- the turning point test: the number of points of inflexion in the graph of the $\{\hat{e}_t\}$ terms should fall within the 95% confidence interval:

$$\left[\frac{2}{3}(N-2) - 1.96\sqrt{\frac{16N-29}{90}}\,,\ \frac{2}{3}(N-2) + 1.96\sqrt{\frac{16N-29}{90}}\right]$$

- the sample autocorrelation $\hat{\rho}_k$ of the $\{\hat{e}_t\}$ terms is close to zero and has an approximate $N\left(0, \frac{1}{n}\right)$ distribution so that 95% of its values should fall within the confidence interval $\left[-1.96\sqrt{\frac{1}{n}}, +1.96\sqrt{\frac{1}{n}}\right]$
- the Ljung and Box 'portmanteau' test: if the $\{\hat{e}_t\}$ terms are white noise then they should be uncorrelated. Under the null hypothesis that the residuals are white noise, the sample autocorrelation r_k of the $\{\hat{e}_t\}$ terms satisfies:

$$n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k} \sim \chi_m^2 \text{ for each } m$$

This is a one-sided test.

Forecasting

Future values of the time series can be forecast using k-step ahead forecasting. We use the notation $\hat{x}_n(k)$ to be the estimate of the expected value of X_{n+k} (given the observations up to X_n). To determine $\hat{x}_n(k)$, we take our time series equation and:

- replace all unknown parameters by their estimated values
- replace the random variables $X_1,\ldots,X_n$ by their observed values $x_1,\ldots,x_n$
- replace the random variables $X_{n+1},\ldots,X_{n+k-1}$ by their forecast values $\hat{x}_n(1),\ldots,\hat{x}_n(k-1)$
- replace the innovations $e_1,\ldots,e_n$ by the residuals $\hat{e}_1,\ldots,\hat{e}_n$
- replace the random variables $e_{n+1},\ldots,e_{n+k-1}$ by 0 (their expectations).

An alternative to 1-step ahead forecasting is exponential smoothing. We use the notation $\hat{x}_n(1)$ to be the estimate of the expected value of X_{n+1} (given the observations up to X_n).

$$\hat{x}_n(1) = \alpha\left[x_n + (1-\alpha)x_{n-1} + (1-\alpha)^2 x_{n-2} + \cdots\right]$$

This is a weighted average of the past values but there is less emphasis on older values. The parameter α is called the smoothing parameter. Rearrangements include:

$$\hat{x}_n(1) = \alpha x_n + (1-\alpha)\hat{x}_{n-1}(1)$$

and:

$$\hat{x}_n(1) = \hat{x}_{n-1}(1) + \alpha\left[x_n - \hat{x}_{n-1}(1)\right]$$

Multivariate time series

We can write a univariate time series in multivariate (or vector) form.

For example, the time series $X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + e_t + \beta e_{t-1}$ can be written as

$$\begin{pmatrix} X_t \\ X_{t-1} \\ X_{t-2} \end{pmatrix} = \begin{pmatrix} 0 & \alpha_1 & \alpha_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_t \\ X_{t-1} \\ X_{t-2} \end{pmatrix} + \begin{pmatrix} 1 & \beta & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} e_t \\ e_{t-1} \\ e_{t-2} \end{pmatrix}$$

ie $\underline{X}_t = A\underline{X}_{t-1} + B\underline{e}_t$

The advantage of the vector form is that it displays the Markov property.

The vector process is stationary if the eigenvalues λ of the matrix A are all strictly less than 1 in magnitude. The eigenvalues are found by solving $\det(A - \lambda I) = 0$ where I is the identity matrix.

Cointegrated series

Two time series processes *X* and *Y* are called cointegrated if:

(i) *X* and *Y* are *I*(1) random processes

(ii) there exists a non-zero vector (called the cointegrating vector) (α, β) such that $\alpha X + \beta Y$ is stationary.

We might expect that two processes are cointegrated if one of the processes is driving the other or if both are being driven by the same underlying process.

Other non-linear, non-stationary time series

Other examples of time series include:

- bilinear models, which exhibit 'bursty' behaviour:

$$X_n - \alpha(X_{n-1} - \mu) = \mu + e_n + \beta e_{n-1} + b(X_{n-1} - \mu)e_{n-1}$$

- threshold autoregressive models, which are used to model 'cyclical' behaviour:

$$X_n = \mu + \begin{cases} \alpha_1(X_{n-1} - \mu) + e_n, & \text{if } X_{n-1} \le d \\ \alpha_2(X_{n-1} - \mu) + e_n, & \text{if } X_{n-1} > d \end{cases}$$

- random coefficient, autogressive models:

$$X_t = \mu + \alpha_t(X_{t-1} - \mu) + e_t$$

where $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ is a sequence of independent random variables.

- autoregressive conditional heteroscedasticity (ARCH) models, which are used to model asset prices, where we require the volatility to depend on the size of the previous value:

$$X_t = \mu + e_t \sqrt{\alpha_0 + \sum_{k=1}^{p} \alpha_k \left(X_{t-k} - \mu\right)^2}$$

Chapter 14 Practice Questions

14.1 The following table shows some data for the d th order differences of an observed time series x_t, $t=1,2,\ldots,100$:

Properties of $\nabla^d x_t$		$d=0$	$d=1$	$d=2$	$d=3$	$d=4$
Sample autocorrelation coefficients	r_1	0.97	0.41	0.03	–0.24	–0.41
	r_2	0.92	–0.19	–0.49	–0.42	–0.25
	r_3	0.88	–0.17	–0.16	–0.01	0.06
	r_4	0.85	0.04	0.19	0.26	0.22
Sample variance		162.3	7.4	8.5	16.4	40.7

State, with reasons, the most appropriate value of d if this series is to be modelled using an $ARIMA(p,d,q)$ model.

14.2 A time series, X, is believed to conform to the $AR(1)$ model $X_n = \alpha X_{n-1} + \varepsilon_n$, where ε is a white noise process. The value of the parameter α is unknown.

(i) The table below shows an extract of the calculation of the residuals for this model when α is assumed to equal 0.6 and 0.7.

n	…	15	16	17	18	19	20	…
x_n	…	371	507	B	449	272	76	…
Residuals ($\alpha = 0.6$)	…	357	A	–48	295	3	–87	…
Residuals ($\alpha = 0.7$)	…	354	247	–99	C	–42	–114	…

Complete the table by calculating the values of A, B and C.

(ii) List the tests that could be applied to the residuals to test the model for goodness of fit.

14.3 An $ARIMA(p,d,q)$ model is to be fitted to the RPI data on page 152 of the *Tables*.

(i) Explain why the value $d=1$ would be considered the most appropriate choice for the parameter d.

(ii) Comment on whether there is any evidence of seasonal variation in this dataset.

14.4 An $ARIMA(0,1,1)$ model of the form $\nabla E_t = \mu + \varepsilon_t + \beta\varepsilon_{t-1}$, where $\{\varepsilon_t\}$ is zero-mean white noise and $|\beta| < 1$, is to be fitted to the NAEI data on page 153 of the *Tables*.

(i) State why the condition $|\beta| < 1$ has been imposed.

(ii) Estimate the value of μ.

(iii) Estimate the value of β by equating the sample and theoretical autocorrelations for lag 1 for the series $\{\nabla E_t\}$.

(iv) Use this model to estimate E_{120}, the value of the series for Jan-02, given that $\hat{\varepsilon}_{119} = 3.1$.

(v) Explain why this estimate might not be reliable.

14.5 Exam style

From a sample of 50 consecutive observations from a stationary process, the table below gives values for the sample autocorrelation function (ACF) and the sample partial autocorrelation function (PACF):

Lag	ACF	PACF
1	0.854	0.854
2	0.820	0.371
3	0.762	0.085

The sample variance of the observations is 1.253.

(i) Suggest, giving reasons, an appropriate model based on this information. [2]

(ii) Consider the $AR(1)$ model:

$$Y_t = a_1 Y_{t-1} + e_t$$

where e_t is a white noise error term with mean zero and variance σ^2.

Calculate method of moments (Yule-Walker) estimates for the parameters of a_1 and σ^2 on the basis of the observed sample. [4]

(iii) Consider the $AR(2)$ model:

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + e_t$$

where e_t is a white noise error term with mean zero and variance σ^2.

Calculate method of moments (Yule-Walker) estimates for the parameters of a_1, a_2 and σ^2 on the basis of the observed sample. [7]

(iv) List two statistical tests that could be applied to the residuals after fitting a model to time series data. [2]

[Total 15]

14.6 Exam style

The following data is observed from $n = 500$ realisations from a time series:

$$\sum_{i=1}^{n} x_i = 13{,}153.32\,,\quad \sum_{i=1}^{n}(x_i - \bar{x})^2 = 3{,}153.67 \quad\text{and}\quad \sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x}) = 2{,}176.03$$

(i) Estimate, using the data above, the parameters μ, α_1 and σ from the model:

$$X_t - \mu = \alpha_1(X_{t-1} - \mu) + \varepsilon_t$$

where ε_t is a white noise process with variance σ^2. [7]

(ii) After fitting the model with the parameters found in (i), it was calculated that the number of turning points of the residuals series $\hat{\varepsilon}_t$ is 280.

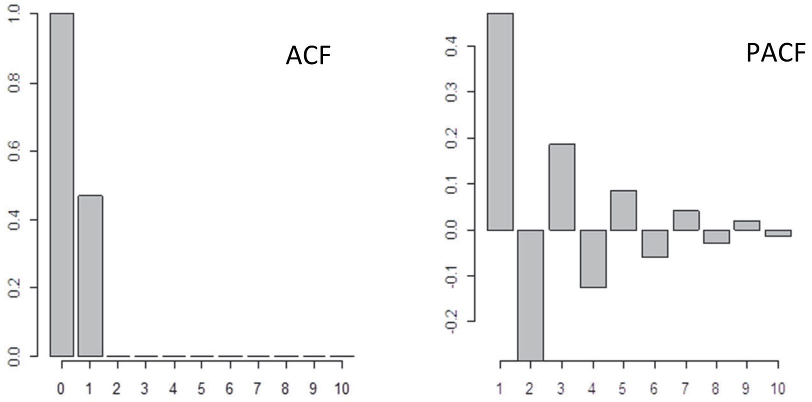
Perform a statistical test to check whether there is evidence that $\hat{\varepsilon}_t$ is not generated from a white noise process. [3]

[Total 10]

14.7 Exam style

(i) State the three main stages in the Box-Jenkins approach to fitting an ARIMA time series model. [3]

(ii) Explain, with reasons, which ARIMA time series would fit the observed data in the charts below. [2]



Now consider the time series model given by:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \beta_1 e_{t-1} + e_t$$

where e_t is a white noise process with variance σ^2.

(iii) Derive the Yule-Walker equations for this model. [6]

(iv) Explain whether the partial autocorrelation function for this model can ever give a zero value. [2]

[Total 13]

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 14 Solutions

14.1 The strongest clue in this question is the sample variance, which should have a small value when an appropriate value of d is found. We can rule out $d = 0$ because the sample variance is much bigger here. $d = 1$ or $d = 2$ seem likely candidates, since the sample variance is low for these, but then starts to increase when $d = 3$.

We also expect the sample autocorrelation coefficients to decay rapidly to zero. Again, this rules out $d = 0$. $d = 1$ looks good and has a more regular pattern than the higher values of d.

Putting these clues together suggests that we should select $d = 1$, as this value satisfies the required criteria, and going to $d = 2$ doesn't result in any worthwhile improvement.

In fact the series used here was a simulation of an $ARIMA(2,1,0)$ *process. So* d *really is equal to 1.*

14.2 (i) ***Completed table***

The completed table looks like this:

n	...	15	16	17	18	19	20	...
x_n	...	371	507	256	449	272	76	...
Residuals ($\alpha = 0.6$)	...	357	284	–48	295	3	–87	...
Residuals ($\alpha = 0.7$)	...	354	247	–99	270	–42	–114	...

The missing numbers are calculated as follows:

$$A = 507 - 0.6 \times 371 = 284$$

$$B - 0.6 \times 507 = -48 \quad \Rightarrow \quad B = 256$$

$$C = 449 - 0.7 \times 256 = 270$$

(ii) ***Tests on residuals***

The tests described in the Core Reading are:

- inspection of the graph of the residuals
- inspection of the SACF and SPACF
- the portmanteau test
- counting turning points.

14.3 (i) ***Explain why $d = 1$***

The parameter d should be the smallest non-negative integer for which the series $\{\nabla^d Q_t\}$ can be considered to be stationary.

The graph in the *Tables* strongly suggests that $\{Q_t\}$ itself has an upward trend and is therefore not stationary.

The sample autocorrelations of $\{Q_t\}$ decay slowly from 1 which indicates differencing is required. However, the sample autocorrelations of $\{\nabla Q_t\}$ do not decay slowly from 1 which indicates that no further differencing is required.

The sample standard deviation of the values is minimised for $\{\nabla Q_t\}$. This also indicates that $d = 1$.

If there is any conflict between the two criteria then we should use the principle of parsimony in choosing the value for d.

(ii) ***Seasonal variation***

Seasonal variation means that an annual cycle is present in the data.

The sample autocorrelation for lag 12 (*ie* 0.637) for the series $\{\nabla Q_t\}$, which we would otherwise consider to be stationary, is positive and significantly different from zero. This suggests the presence of a 12-monthly cycle.

Note that, because $\{Q_t\}$ doesn't appear to be stationary, the value of r_{12} for this series would be quite high, whether or not seasonality was present.

14.4 (i) ***Reason for condition***

This condition ensures that the model is invertible.

(ii) ***Estimate μ***

Since the white noise has mean zero, we know that $E[\nabla E_t] = \mu$.

If we equate $E[\nabla E_t]$ to the sample mean of the first differences shown in the *Tables*, this gives $\hat{\mu} = 0.4$.

(iii) ***Estimate β***

We can calculate the autocovariances of $\{\nabla E_t\}$ for lags 0 and 1 for this model to be:

$$\gamma_0 = (1+\beta^2)\sigma^2 \quad \text{and} \quad \gamma_1 = \beta\sigma^2$$

where σ^2 is the variance of the white noise series.

So:

$$\rho_1 = \frac{\beta}{1+\beta^2}$$

Equating this to the sample value gives:

$$\frac{\beta}{1+\beta^2} = -0.245$$

This leads to the quadratic equation:

$$0.245\beta^2 + \beta + 0.245 = 0 \quad \Rightarrow \hat{\beta} = -0.262 \text{ or } -3.82$$

Since $|\beta| < 1$, we conclude that $\hat{\beta} = -0.262$.

(iv) ***Estimate E_{120}***

We first need to estimate ∇E_{120}. The actual value will be:

$$\nabla E_{120} = \mu + \varepsilon_{120} + \beta\varepsilon_{119}$$

We estimate this using the equation:

$$\nabla\hat{E}_{120} = \hat{\mu} + 0 + \hat{\beta}\hat{\varepsilon}_{119}, \text{ so } \nabla\hat{E}_{120} = 0.4 + 0 - 0.262(3.1) = -0.4$$

We can then calculate the required estimate as:

$$\hat{E}_{120} = E_{119} + \nabla\hat{E}_{120} = 134.1 + (-0.4) = 133.7$$

(v) ***Why estimate may not be reliable***

We can see from the graph (and from the value of r_{12} for $\{\nabla E_t\}$) that the series contains a strong seasonal component, which we have not allowed for in this model.

The actual value for Jan-02 (not shown in the Tables) was 132.4, which is significantly different from our estimate.

14.5 *This is Subject CT6, September 2008, Question 10.*

(i) ***Appropriate model***

From the figures given it looks like the ACF is decaying slowly and the PACF is cutting off after lag 2. This is a characteristic of an $AR(2)$ model. [2]

(ii) ***Parameter estimates***

As a starter step:

$$\text{cov}(Y_t, e_t) = \text{cov}(a_1 Y_{t-1} + e_t, e_t) = \sigma^2 \qquad [½]$$

Consider the autocovariance with a lag of 1:

$$\gamma_1 = \text{cov}(Y_t, Y_{t-1}) = \text{cov}(a_1 Y_{t-1} + e_t, Y_{t-1}) = a_1 \gamma_0 \quad \Rightarrow \quad \rho_1 = a_1$$ [1]

Because we are told in the question that the sample ACF with lag 1 is 0.854, this is our estimate of ρ_1, so we have:

$$0.854 = \hat{a}_1$$ [½]

Consider the autocovariance with a lag of 0:

$$\gamma_0 = \text{cov}(Y_t, Y_t) = \text{cov}(a_1 Y_{t-1} + e_t, Y_t) = a_1 \gamma_1 + \sigma^2$$ [½]

Because we are told in the question that the sample ACF with lag 1 is 0.854 (our estimate of ρ_1) and the sample variance is 1.253 (our estimate of γ_0), we have:

$$\hat{\gamma}_0 = 1.253$$

$$\frac{\hat{\gamma}_1}{\hat{\gamma}_0} = 0.854 \quad \Rightarrow \quad \hat{\gamma}_1 = 0.854 \times 1.253$$ [1]

From $\gamma_0 = a_1 \gamma_1 + \sigma^2$:

$$1.253 = 0.854^2 \times 1.253 + \hat{\sigma}^2 \quad \Rightarrow \quad \hat{\sigma}^2 = 0.339$$ [½]

(iii) ***Parameter estimates***

Consider the autocovariance with a lag of 1:

$$\gamma_1 = \text{cov}(Y_t, Y_{t-1}) = \text{cov}(a_1 Y_{t-1} + a_2 Y_{t-2} + e_t, Y_{t-1}) = a_1 \gamma_0 + a_2 \gamma_1$$

$$\Rightarrow \quad \rho_1 = \frac{a_1}{1 - a_2}$$ [1]

Consider the autocovariance with a lag of 2:

$$\gamma_2 = \text{cov}(Y_t, Y_{t-2}) = \text{cov}(a_1 Y_{t-1} + a_2 Y_{t-2} + e_t, Y_{t-2}) = a_1 \gamma_1 + a_2 \gamma_0$$

$$= \frac{a_1^2}{1 - a_2} \gamma_0 + a_2 \gamma_0 = \frac{a_1^2 + (1 - a_2) a_2}{1 - a_2} \gamma_0$$ [1]

From this:

$$\rho_2 = \frac{a_1^2 + (1 - a_2) a_2}{1 - a_2}$$ [½]

Because we are told in the question that the sample ACF with lag 1 is 0.854 (our estimate for ρ_1) and that the sample ACF with lag 2 is 0.820 (our estimate of ρ_2), we have:

$$0.854 = \frac{\hat{a}_1}{1-\hat{a}_2}$$

$$0.820 = \frac{\hat{a}_1^2 + (1-\hat{a}_2)\hat{a}_2}{1-\hat{a}_2}$$ [1]

Replacing $\hat{a}_1$ by $0.854(1-\hat{a}_2)$ in the second equation, we get:

$$0.820 = \frac{0.854^2(1-\hat{a}_2)^2 + (1-\hat{a}_2)\hat{a}_2}{1-\hat{a}_2} = 0.854^2(1-\hat{a}_2) + \hat{a}_2$$

$$\Rightarrow\ 0.820 - 0.854^2 = \hat{a}_2(1-0.854^2)$$

$$\Rightarrow\ \hat{a}_2 = 0.335$$ [1]

By substituting this back into the first equation above, we get $\hat{a}_1 = 0.568$. [½]

Consider the autocovariance with a lag of 0:

$$\gamma_0 = \text{cov}(Y_t, Y_t) = \text{cov}(a_1 Y_{t-1} + a_2 Y_{t-2} + e_t, Y_t) = a_1\gamma_1 + a_2\gamma_2 + \sigma^2$$ [½]

Because we are told in the question that the sample ACF with lag 1 is 0.854, the sample ACF with lag 2 is 0.820 and the sample variance is 1.253, we have:

$$\hat{\gamma}_0 = 1.253$$

$$\frac{\hat{\gamma}_1}{\hat{\gamma}_0} = 0.854 \quad\Rightarrow\quad \hat{\gamma}_1 = 0.854 \times 1.253$$

$$\frac{\hat{\gamma}_2}{\hat{\gamma}_0} = 0.820 \quad\Rightarrow\quad \hat{\gamma}_2 = 0.820 \times 1.253$$ [1]

From $\gamma_0 = a_1\gamma_1 + a_2\gamma_2 + \sigma^2$:

$$1.253 = 0.568 \times 0.854 \times 1.253 + 0.335 \times 0.820 \times 1.253 + \hat{\sigma}^2$$

$$\Rightarrow\ \hat{\sigma}^2 = 0.301$$ [½]

(iv) ***Tests***

We could mention any two of the following:

- Portmanteau (Ljung and Box) test
- Turning points test
- Inspection of the graph of the residuals

- Inspection of the values of the sample autocorrelation function based on their 95% confidence intervals
- Inspection of the values of the sample partial autocorrelation function based on their 95% confidence intervals [2]

14.6 *This is Subject CT6, September 2009, Question 6.*

(i) ***Estimation of parameters***

The value of μ is estimated by the sample mean:

$$\hat{\mu} = \overline{x} = \frac{13,153.32}{500} = 26.31 \qquad [1]$$

To estimate α_1 and σ, we consider the Yule-Walker equations:

$$\begin{aligned}\gamma_0 &= \text{var}(X_t) = \text{cov}(X_t, X_t)\\ &= \text{cov}(\alpha_1 X_{t-1} + \varepsilon_t, X_t)\\ &= \alpha_1 \text{cov}(X_{t-1}, X_t) + \text{cov}(\varepsilon_t, X_t)\\ &= \alpha_1 \gamma_1 + \sigma^2 \qquad [1]\end{aligned}$$

since:

$$\text{cov}(\varepsilon_t, X_t) = \text{cov}(\varepsilon_t, \alpha_1 X_{t-1} + \varepsilon_t) = \alpha_1 \text{cov}(\varepsilon_t, X_{t-1}) + \text{cov}(\varepsilon_t, \varepsilon_t) = 0 + \sigma^2$$

In addition:

$$\begin{aligned}\gamma_1 &= \text{cov}(X_t, X_{t-1})\\ &= \text{cov}(\alpha_1 X_{t-1} + \varepsilon_t, X_{t-1})\\ &= \alpha_1 \text{cov}(X_{t-1}, X_{t-1}) + \text{cov}(\varepsilon_t, X_{t-1})\\ &= \alpha_1 \gamma_0 + 0 \qquad [1]\end{aligned}$$

$$\Rightarrow \rho_1 = \frac{\gamma_1}{\gamma_0} = \alpha_1 \qquad [½]$$

Now using the formulae for $\hat{\gamma}_k$ and $\hat{\rho}_k$ given on page 40 of the *Tables*:

$$\hat{\gamma}_0 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{3,153.67}{500} \qquad [½]$$

$$\hat{\gamma}_1 = \frac{1}{n}\sum_{i=1}^{n-1}(x_i - \overline{x})(x_{i+1} - \overline{x}) = \frac{2,176.03}{500} \qquad [½]$$

So:

$$\hat{\rho}_1 = \frac{\hat{\gamma}_1}{\hat{\gamma}_0} = \frac{2,176.03}{3,153.67} = 0.6900$$ [½]

$$\hat{\alpha}_1 = \hat{\rho}_1 = 0.6900$$ [1]

and:

$$\hat{\gamma}_0 = \hat{\alpha}_1 \hat{\gamma}_1 + \hat{\sigma}^2$$

$$\Rightarrow \frac{3,153.67}{500} = 0.6900 \times \frac{2,176.03}{500} + \hat{\sigma}^2$$

$$\Rightarrow \hat{\sigma}^2 = \frac{3,153.67}{500} - 0.6900 \times \frac{2,176.03}{500} = 3.3044$$

$$\Rightarrow \hat{\sigma} = 1.8178$$ [1]

(ii) ***Turning point test***

The null and alternative hypotheses are:

H_0 : the residuals are from a white noise process

H_1 : the residuals are not from a white noise process [1]

Using the formulae from page 42 of the *Tables*, we have:

$$E(T) = \frac{2}{3} \times 498 = 332$$

$$\text{var}(T) = \frac{16 \times 500 - 29}{90} = 88.567$$

The value of the test statistic is:

$$\frac{280 + 0.5 - 332}{\sqrt{88.567}} = -5.47$$ [1]

which should be from a $N(0,1)$ distribution if H_0 holds. Since $-5.47 < -1.96$ we have very strong evidence to reject H_0. This suggests that the residuals are not from a white noise process. [1]

14.7 *This is Subject CT6, September 2013, Question 9.*

(i) ***Box-Jenkins approach***

The three main stages in the Box-Jenkins methodology are:

- Tentative identification of a model from the ARIMA class. [1]
- Estimation of parameters in the identified model. [1]
- Diagnostic checks. [1]

(ii) ***ARIMA time series to fit the observed data in the charts***

The ACF cuts off (becomes 0) at all lags greater than 1, whereas the PACF decays towards 0. Hence we have an $MA(1)$. [2]

(iii) ***Yule-Walker equations***

We start with some useful preliminary equations:

$$\begin{aligned}\text{cov}(X_t,e_t)&=\text{cov}(\alpha_1X_{t-1}+\alpha_2X_{t-2}+\beta_1e_{t-1}+e_t,e_t)\\&=\alpha_1\text{cov}(X_{t-1},e_t)+\alpha_2\text{cov}(X_{t-2},e_t)+\beta_1\text{cov}(e_{t-1},e_t)+\text{cov}(e_t,e_t)\\&=0+0+0+\sigma^2=\sigma^2\end{aligned}$$ [1]

and:

$$\begin{aligned}\text{cov}(X_t,e_{t-1})&=\text{cov}(\alpha_1X_{t-1}+\alpha_2X_{t-2}+\beta_1e_{t-1}+e_t,e_{t-1})\\&=\alpha_1\text{cov}(X_{t-1},e_{t-1})+\alpha_2\text{cov}(X_{t-2},e_{t-1})+\beta_1\text{cov}(e_{t-1},e_{t-1})+\text{cov}(e_t,e_{t-1})\\&=\alpha_1\text{cov}(X_t,e_t)+0+\beta_1\text{var}(e_t)+0\\&=\alpha_1\sigma^2+\beta_1\sigma^2=(\alpha_1+\beta_1)\sigma^2\end{aligned}$$ [1]

The Yule-Walker equations are as follows:

$$\begin{aligned}\gamma_0&=\text{cov}(X_t,X_t)\\&=\text{cov}(\alpha_1X_{t-1}+\alpha_2X_{t-2}+\beta_1e_{t-1}+e_t,X_t)\\&=\alpha_1\text{cov}(X_{t-1},X_t)+\alpha_2\text{cov}(X_{t-2},X_t)+\beta_1\text{cov}(e_{t-1},X_t)+\text{cov}(e_t,X_t)\\&=\alpha_1\gamma_1+\alpha_2\gamma_2+\beta_1\left(\alpha_1+\beta_1\right)\sigma^2+\sigma^2\end{aligned}$$ [1]

$$\begin{aligned}\gamma_1&=\text{cov}(X_t,X_{t-1})\\&=\text{cov}(\alpha_1X_{t-1}+\alpha_2X_{t-2}+\beta_1e_{t-1}+e_t,X_{t-1})\\&=\alpha_1\text{cov}(X_{t-1},X_{t-1})+\alpha_2\text{cov}(X_{t-2},X_{t-1})+\beta_1\text{cov}(e_{t-1},X_{t-1})+\text{cov}(e_t,X_{t-1})\\&=\alpha_1\gamma_0+\alpha_2\gamma_1+\beta_1\sigma^2\end{aligned}$$ [1]

$$\begin{aligned}\gamma_2&=\text{cov}(X_t,X_{t-2})\\&=\text{cov}(\alpha_1X_{t-1}+\alpha_2X_{t-2}+\beta_1e_{t-1}+e_t,X_{t-2})\\&=\alpha_1\text{cov}(X_{t-1},X_{t-2})+\alpha_2\text{cov}(X_{t-2},X_{t-2})+\beta_1\text{cov}(e_{t-1},X_{t-2})+\text{cov}(e_t,X_{t-2})\\&=\alpha_1\gamma_1+\alpha_2\gamma_0\end{aligned}$$ [1]

In general, for lags $k\geq 2$:

$$\gamma_k=\alpha_1\gamma_{k-1}+\alpha_2\gamma_{k-2}$$ [1]

(iv) ***Can the partial auto-correlation function ever give a zero value?***

For an $MA(q)$ process, where $q \geq 1$, the PACF tends towards 0 but does not completely cut off. [1]

Here, we have an $ARMA(2,1)$ process, *ie* $q \geq 1$. Hence the PACF tends towards 0 but does not completely cut off. There will always be a small partial autocorrelation. [1]

15

Loss distributions

Syllabus objectives

1.1 Loss distributions, with and without risk sharing

1.1.1 Describe the properties of the statistical distributions which are suitable for modelling individual and aggregate losses.

1.1.5 Estimate the parameters of a failure time or loss distribution when the data is complete, or when it is incomplete, using maximum likelihood and the method of moments.

1.1.6 Use R to fit a statistical distribution to a dataset and calculate appropriate goodness-of-fit measures.

0 Introduction

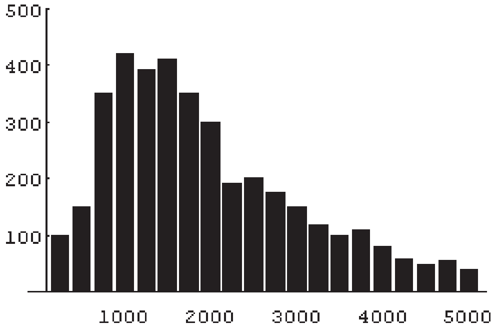
General insurance companies need to investigate claims experience and apply mathematical techniques for many purposes. These include:

- premium rating (*ie* deciding what premium rates to charge policyholders)
- reserving (*ie* assessing how much money should be set aside to cover the cost of claims)
- reviewing reinsurance arrangements
- testing for solvency (*ie* assessing the company's financial position).

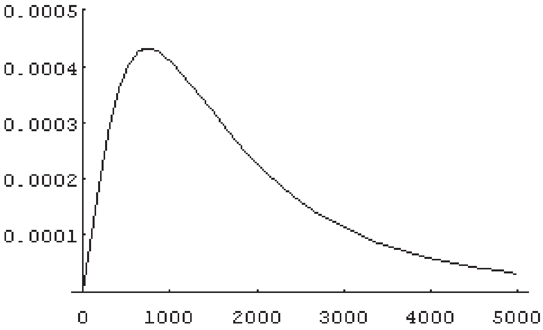
In this chapter we will look at loss distributions. These are statistical distributions that are used to model individual claim amounts. We will introduce some new distributions and we see how these can be fitted to observed claims data. We can then test for goodness of fit, and use the fitted loss distributions to estimate probabilities.

The total amount of claims in a particular time period is a quantity of fundamental importance to the proper management of an insurance company. The key assumption in all the models studied here is that the occurrence of a claim and the amount of a claim can be studied separately. Thus, a claim occurs according to some simple model for events occurring in time, then the amount of the claim is chosen from a distribution describing the claim amount.

Note carefully the distinction here. The frequency of claim amounts when plotted against size might look like this:



The statistical distributions in this chapter are used to approximate this distribution, which is called a *loss distribution*. For example, we might to decide to use a loss distribution like this as an approximation to the claims arising in the graph above:



A range of statistical techniques can be used to describe the distribution of random variables. The object is to describe the variation in claim amounts by finding a loss distribution that adequately describes the claims that actually occur. As usual this can be done at two levels.

At a first level, it can be assumed that the claims arise as realisations from a known distribution. For example, it may be possible to assume that the logarithm of the claim amount follows, to a reasonable approximation, a normal distribution with known mean and known standard deviation. Knowledge of the claim amount process would be complete, and interest would then centre on the consequences for insurance. For example, claims above a certain level might trigger some reinsurance arrangements or claims below a certain level might never be lodged if a policy excess was in force.

A policy excess means that the policyholder has to pay the first part of any claim. For example, with car insurance in the UK the policyholder often has to pay the first £200 of any claim. The insurer pays the rest.

In practice the exact claims distribution will hardly ever be known. At this second level a standard method of proceeding is to assume that the claims distribution is a member of a certain family. The parameters of the family must now be estimated using the claim amount records by an appropriate method such as maximum likelihood. Complications will arise if large claims have been limited (reinsurance) or some small claims have not been lodged (policy excess).

We will consider the effects of reinsurance and policy excesses in Chapter 18.

Many studies have been made of the kind of distribution that can be used to describe the variation in claim amounts.

The typical pattern is as shown in the histogram above, with a few small claims, rising to a peak, then tailing off gradually with a few very large claims.

The general conclusion is that claims distributions tend to be positively skewed and long tailed.

1 Simple loss distributions

In this section we will review some of the properties of the statistical distributions that are used to model claim amounts. In most cases we use a positively skewed, continuous distribution.

Recall that, for a continuous random variable, X :

- the cumulative distribution function (CDF) is:

  $$F_X(x) = P(X \le x)$$

- the probability density function (PDF) is:

  $$f_X(x) = F'_X(x)$$

- probabilities can be expressed in terms of the PDF or CDF:

  $$P(a \le X \le b) = \int_a^b f_X(x)\,dx = F_X(b) - F_X(a)$$

- the moment generating function (MGF) is:

  $$M_X(t) = E(e^{tX}) = \int_x e^{tx} f_X(x)\,dx$$

The CDF, PDF and MGF may also be denoted without the subscript as $F(x)$, $f(x)$ and $M(t)$, respectively, provided the meaning is clear.

The formulae for the densities, the moments and the moment generating functions (where they exist) for the distributions discussed in this chapter are given in the *Formulae and Tables for Actuarial Examinations*.

In the *Tables*, the abbreviation DF is used for cumulative distribution function.

1.1 The exponential distribution

A random variable X has the exponential distribution with parameter $\lambda > 0$ if it has CDF:

$$\boldsymbol{F(x) = 1 - e^{-\lambda x}\ ,\ x > 0}$$

In that case we write $X \sim Exp(\lambda)$.

The PDF is:

$$f(x) = \lambda e^{-\lambda x},\ x > 0$$

The mean and variance are $\dfrac{1}{\lambda}$ and $\dfrac{1}{\lambda^2}$ respectively.

The PDF can also be written as:

$$f(x) = \frac{1}{\mu} e^{-x/\mu}$$

where μ is the mean.

The MGF is:

$$M(t) = \left(1 - \frac{t}{\lambda}\right)^{-1}, \quad t < \lambda$$

All of these formulae are given on page 11 of the *Tables*.

Question

A portfolio of insurance policies contains two types of risk. Type I risks make up 70% of claims and give rise to loss amounts that are exponentially distributed with mean 500. Type II risks give rise to loss amounts that are exponentially distributed with mean 1,000.

Let X denote the amount of a randomly chosen loss. Determine $E(X)$, $\text{var}(X)$ and $M_X(t)$.

Solution

Since the amount of a loss depends on the type of risk from which it arises, we calculate $E(X)$ using the conditional expectation formula (from Subject CS1). This formula is given on page 16 of the *Tables*. In this case:

$$E(X) = E(E(X \mid \text{Type})) = E(X \mid \text{Type I})P(\text{Type I}) + E(X \mid \text{Type II})P(\text{Type II})$$

$$= 500 \times 0.7 + 1{,}000 \times 0.3 = 650$$

Similarly:

$$E(X^2) = E(E(X^2 \mid \text{Type})) = E(X^2 \mid \text{Type I})P(\text{Type I}) + E(X^2 \mid \text{Type II})P(\text{Type II})$$

Now:

$$E(X^2 \mid \text{Type I}) = \text{var}(X \mid \text{Type I}) + \left[E(X \mid \text{Type I})\right]^2 = 500^2 + 500^2 = 500{,}000$$

Here we are using the fact that the variance of an exponential random variable is the square of its mean. So the variance of losses from Type I risks is 500^2. We can use the same approach for Type II risks:

$$E(X^2 \mid \text{Type II}) = \text{var}(X \mid \text{Type II}) + \left[E(X \mid \text{Type II})\right]^2 = 1{,}000^2 + 1{,}000^2 = 2{,}000{,}000$$

So:

$$E(X^2) = 500{,}000 \times 0.7 + 2{,}000{,}000 \times 0.3 = 950{,}000$$

Hence:

$$\text{var}(X) = 950{,}000 - 650^2 = 527{,}500$$

Alternatively, we could calculate $\text{var}(X)$ using the conditional variance formula, which is also given on page 16 of the *Tables*:

$$\text{var}(X) = E[\text{var}(X \mid \text{Type})] + \text{var}[E(X \mid \text{Type})]$$

For notational convenience, let $V = \text{var}(X \mid \text{Type})$ and let $W = E(X \mid \text{Type})$. Then V has the following distribution:

v	500^2	$1{,}000^2$
$P(V = v)$	0.7	0.3

So:

$$E(V) = 500^2 \times 0.7 + 1{,}000^2 \times 0.3 = 475{,}000$$

In addition, we can calculate $\text{var}(W)$ from the distribution of W:

w	500	1,000
$P(W = w)$	0.7	0.3

We have:

$$E(W) = 500 \times 0.7 + 1{,}000 \times 0.3 = 650$$

$$E(W^2) = 500^2 \times 0.7 + 1{,}000^2 \times 0.3 = 475{,}000$$

and hence:

$$\text{var}(W) = 475{,}000 - 650^2 = 52{,}500$$

So:

$$\text{var}(X) = E[V] + \text{var}[W] = 475{,}000 + 52{,}500 = 527{,}500$$

Finally, we will consider the moment generating function of X. Again, we will use the conditional expectation formula:

$$M_X(t) = E(e^{tX}) = E\left[E(e^{tX} \mid \text{Type})\right]$$

$$= E(e^{tX} \mid \text{Type I})P(\text{Type I}) + E(e^{tX} \mid \text{Type II})P(\text{Type II})$$

$E(e^{tX} \mid \text{Type I})$ is the MGF of the exponential distribution with mean 500, *ie*:

$$E(e^{tX} \mid \text{Type I}) = (1-500t)^{-1}$$

Similarly, $E(e^{tX} \mid \text{Type II})$ is the MGF of the exponential distribution with mean 1,000, *ie*:

$$E(e^{tX} \mid \text{Type II}) = (1-1{,}000t)^{-1}$$

So:

$$M_X(t) = 0.7(1-500t)^{-1} + 0.3(1-1{,}000t)^{-1}$$

We can use R to simulate values from statistical distributions, plot their PDFs, and calculate probabilities and percentiles. An example involving the exponential distribution is given below.

Suppose we have an exponential distribution with parameter $\lambda = 0.5$. The R code for simulating 100 values is given by:

```
rexp(100,rate=0.5)
```

The PDF is obtained by `dexp(x, rate=0.5)` **and is useful for graphing. For example:**

```
plot(seq(0:5000),dexp(seq(0:5000), rate=0.5),type="l")
```

To calculate probabilities for a continuous distribution we use the CDF which is obtained by `pexp`**. For example, to calculate $P(X \le 2) = 0.6321206$ we use the R code:**

```
pexp(2,rate=0.5)
```

Similarly, the quantiles can be calculated with `qexp`**.**

This code can be adapted to deal with other statistical distributions.

1.2 The gamma distribution

The random variable X has a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$ if it has PDF:

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x), \quad x > 0$$

In that case we write $X \sim Ga(\alpha, \lambda)$.

This may also be written as $Gamma(\alpha, \lambda)$.

The gamma function, $\Gamma(\alpha)$, appears in the denominator of this PDF. The definition and properties of this function are given on page 5 of the *Tables*.

The mean and variance of X are:

$$\boldsymbol{E(X)} = \frac{\alpha}{\lambda}$$

$$\textbf{var}\boldsymbol{(X)} = \frac{\alpha}{\lambda^2}$$

Question

If $X \sim Gamma(\alpha, \lambda)$, show that the MGF of X is:

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-\alpha}$$

Solution

Using the definition of the MGF, we have:

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}\, dx = \int_0^\infty \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-(\lambda-t)x}\, dx$$

We can make the integrand look like the PDF of the $Gamma(\alpha, \lambda - t)$ distribution by writing:

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha \int_0^\infty \frac{1}{\Gamma(\alpha)} (\lambda - t)^\alpha x^{\alpha-1} e^{-(\lambda-t)x} dx$$

This integral is equal to 1 provided $\lambda - t > 0$, so:

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha = \left(\frac{\lambda - t}{\lambda}\right)^{-\alpha} = \left(1 - \frac{t}{\lambda}\right)^{-\alpha}, \quad \text{for } t < \lambda$$

By differentiating the MGF, we can obtain the non-central moments, $E(X^k)$, $k = 1, 2, 3, \ldots$:

$$E(X) = M_X'(0), \quad E(X^2) = M_X''(0), \quad etc$$

The variance and skewness can be obtained more quickly using the cumulant generating function (CGF). Recall that:

$$C_X(t) = \ln M_X(t)$$

Question

Suppose that $X \sim Gamma(\alpha, \lambda)$.

Derive formulae for the skewness and coefficient of skewness of X.

Solution

The skewness of X is its third central moment, $E\left[(X-E(X))^3\right]$. It can be obtained by differentiating the CGF three times and evaluating the third derivative when $t=0$.

Since $X \sim Gamma(\alpha,\lambda)$:

$$C_X(t)=\ln\left(1-\frac{t}{\lambda}\right)^{-\alpha}=-\alpha\ln\left(1-\frac{t}{\lambda}\right)$$

Differentiating using the chain rule:

$$C_X'(t)=-\alpha\left(-\frac{1}{\lambda}\right)\left(1-\frac{t}{\lambda}\right)^{-1}=\frac{\alpha}{\lambda}\left(1-\frac{t}{\lambda}\right)^{-1}$$

$$C_X''(t)=\frac{\alpha}{\lambda}\left(-\frac{1}{\lambda}\right)(-1)\left(1-\frac{t}{\lambda}\right)^{-2}=\frac{\alpha}{\lambda^2}\left(1-\frac{t}{\lambda}\right)^{-2}$$

$$C_X'''(t)=\frac{\alpha}{\lambda^2}\left(-\frac{1}{\lambda}\right)(-2)\left(1-\frac{t}{\lambda}\right)^{-3}=\frac{2\alpha}{\lambda^3}\left(1-\frac{t}{\lambda}\right)^{-3}$$

So:

$$\text{skew}(X)=C_X'''(0)=\frac{2\alpha}{\lambda^3}\left(1-\frac{0}{\lambda}\right)^{-3}=\frac{2\alpha}{\lambda^3}$$

The coefficient of skewness of X is:

$$\text{coeff of skew}(X)=\frac{\text{skew}(X)}{\left[\text{var}(X)\right]^{3/2}}$$

The variance can also be obtained from the CGF:

$$\text{var}(X)=C_X''(0)=\frac{\alpha}{\lambda^2}\left(1-\frac{0}{\lambda}\right)^{-2}=\frac{\alpha}{\lambda^2}$$

So:

$$\text{coeff of skew}(X)=\frac{\text{skew}(X)}{\left[\text{var}(X)\right]^{3/2}}=\frac{2\alpha/\lambda^3}{(\alpha/\lambda^2)^{3/2}}=\frac{2}{\alpha^{1/2}}=\frac{2}{\sqrt{\alpha}}$$

Formulae for the PDF, MGF, mean, variance, non-central moments and coefficient of skewness of the gamma distribution are all given on page 12 of the *Tables*.

There is no closed form (*ie* no simple formula) for the CDF of a gamma random variable, which means that it is not easy to find gamma probabilities directly without using a computer package such as R. However, these probabilities can be obtained using the relationship between the gamma and chi-squared distributions.

Relationship between gamma and chi-squared distributions

If $X \sim Gamma(\alpha, \lambda)$ and 2α is an integer, then:

$$2\lambda X \sim \chi^2_{2\alpha}$$

This result is also given on page 12 of the *Tables*.

As an illustration of how this relationship can be used, suppose that $X \sim Gamma(10,4)$ and we want to calculate $P(X > 4.375)$. Using the result above, we know that $8X \sim \chi^2_{20}$, so:

$$P(X > 4.375) = P(8X > 8 \times 4.375) = P(\chi^2_{20} > 35)$$

From page 166 of the *Tables*, we see that:

$$P(\chi^2_{20} \leq 35) = 0.9799$$

So:

$$P(X > 4.375) = 1 - 0.9799 = 0.0201$$

The R code for simulating a random sample of 100 values from the gamma distribution with $\alpha = 2$ and $\lambda = 0.25$ is:

```
    rgamma(100, 2, 0.25)
```

Similarly, the PDF, CDF and quantiles can be obtained using the R functions `dgamma`**,** `pgamma` **and** `qgamma`**.**

1.3 The normal distribution

The normal distribution arises in a variety of contexts. It is of limited use for modelling loss distributions because of its symmetry (as loss distributions tend to be positively skewed).

Question

Derive the formula for the MGF of a standard normal random variable.

Solution

Suppose that $X \sim N(0,1)$. Then the PDF of X is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-½x^2}$$

and the MGF is:

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-½x^2} \; dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-½(x^2-2tx)} dx$$

Completing the square gives:

$$M_X(t) = e^{½t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-½(x-t)^2} dx$$

The integrand in the expression immediately above is the PDF of a $N(t,1)$ random variable. Integrating this over all possible values of x gives the total probability, which is 1. So:

$$M_X(t) = e^{½t^2}$$

2 Other loss distributions

The distributions given in Section 1 (exponential, gamma, normal) all have easily derivable MGFs. However, there is a wide variety of other distributions that may also be used to model losses. We consider some of these here. None of the distributions in this section have an MGF that is easy to derive or use.

2.1 The lognormal distribution

The definition of the lognormal distribution is very simple: X has a lognormal distribution if $\log X$ has a normal distribution. When $\log X \sim N(\mu, \sigma^2)$, $X \sim \log N(\mu, \sigma^2)$.

So the range of values taken by the lognormal distribution is 0 to ∞.

As usual, log here refers to the natural logarithm, *ie* log base e.

The mean and variance of a lognormal random variable can be obtained from the MGF of the corresponding normal distribution. If $X \sim \log N(\mu, \sigma^2)$, then:

$$E(X) = E(e^{\ln X}) = E(e^Y)$$

where $Y = \ln X \sim N(\mu, \sigma^2)$. However:

$$E(e^Y) = M_Y(1) = e^{\mu + \frac{1}{2}\sigma^2}$$

So $E(X) = e^{\mu + \frac{1}{2}\sigma^2}$.

Similarly:

$$E(X^2) = E(e^{2Y}) = M_Y(2) = e^{2\mu + 2\sigma^2}$$

and hence:

$$\text{var}(X) = e^{2\mu + 2\sigma^2} - \left(e^{\mu + \frac{1}{2}\sigma^2}\right)^2 = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2} = e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$$

Alternatively, the mean and variance can be derived using integration.

Formulae for the PDF, mean, variance, non-central moments and coefficient of skewness of the lognormal distribution are given on page 14 of the *Tables*.

Lognormal probabilities can be evaluated by expressing them as standard normal probabilities and looking up the values given on pages 160 and 161 of the *Tables*.

The R code for simulating values and obtaining the PDF, CDF and quantiles from the lognormal distribution is similar to the R code used for other continuous distributions using the R functions `rlnorm`, `dlnorm`, `plnorm` and `qlnorm`.

2.2 The two-parameter Pareto distribution

A random variable X has the Pareto distribution with parameters $\alpha > 0$ and $\lambda > 0$ if it has CDF:

$$F(x) = 1 - \left(\frac{\lambda}{\lambda + x}\right)^{\alpha} \ , \quad x > 0$$

In that case we write $X \sim Pa(\alpha, \lambda)$.

It is easily checked by differentiating $F(x)$ with respect to x that the Pareto distribution has PDF:

$$f(x) = \frac{\alpha\lambda^{\alpha}}{(\lambda + x)^{\alpha+1}} \ , \quad x > 0$$

Question

Suppose that $X \sim Pa(\alpha, \lambda)$. Derive a formula for $E(X)$.

Solution

The expected value is:

$$E(X) = \int_x x\, f(x)\, dx = \int_0^\infty x \frac{\alpha\lambda^{\alpha}}{(\lambda + x)^{\alpha+1}} dx$$

One way to simplify this is to use the substitution $t = \lambda + x$. Using this substitution:

$$E(X) = \int_\lambda^\infty (t - \lambda) \frac{\alpha\lambda^{\alpha}}{t^{\alpha+1}} dt = \alpha\lambda^{\alpha} \int_\lambda^\infty t^{-\alpha} dt - \alpha\lambda^{\alpha+1} \int_\lambda^\infty t^{-\alpha-1} dt$$

Integrating gives:

$$E(X) = \alpha\lambda^{\alpha} \left[\frac{t^{-\alpha+1}}{-\alpha+1}\right]_\lambda^\infty - \alpha\lambda^{\alpha+1} \left[\frac{t^{-\alpha}}{-\alpha}\right]_\lambda^\infty = \frac{\alpha\lambda}{\alpha - 1} - \lambda = \frac{\lambda}{\alpha - 1}$$

This expression is valid only if the powers in the bracketed terms are both negative, *ie* if $\alpha > 1$.

Alternatively, this formula could be derived using integration by parts.

Formulae for the CDF, PDF, mean, variance, non-central moments and coefficient of skewness of the Pareto distribution are given on page 14 of the *Tables*.

Let's now consider the median of the Pareto distribution. By definition, the median m is the point where $F(m) = P(X \leq m) = ½$. So, in this case, we have:

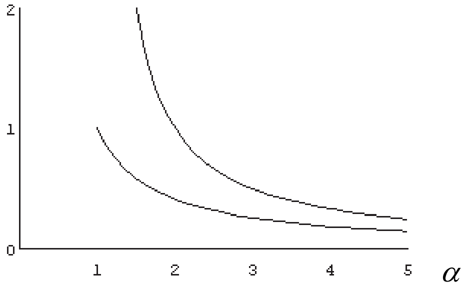
$$1 - \left(\frac{\lambda}{\lambda + m}\right)^{\alpha} = \frac{1}{2}$$

and this can be rearranged to give:

$$m = \lambda(2^{1/\alpha} - 1)$$

We can compare the median and the mean by drawing a graph. We have just seen that the mean, μ, is equal to $\dfrac{\lambda}{\alpha - 1}$.

A sketch of the graphs of $\dfrac{m}{\lambda}$ (bottom curve) and $\dfrac{\mu}{\lambda}$ (top curve) for values of $\alpha > 1$ is shown below:



From this we see that the mean is always greater than the median, *ie* the Pareto distribution is always positively skewed.

There is no built in R code for the Pareto distribution so we would have to define the functions `rpareto`, `dpareto`, `ppareto` and `qpareto` from first principles as follows:

```
rpareto <- function(n,a,l){
rp <- l*((1-runif(n))^(-1/a)-1)
rp}

dpareto <- function(x,a,l){
a*l^(a)/((l+x)^(a+1))}

ppareto <- function(q,a,l){
1-(l/(l+q))^a}

qpareto <- function(p,a,l){
q <- l*((1-p)^(-1/a)-1)
q}
```

2.3 The Burr distribution

The CDF of the Pareto distribution $Pa(\alpha, \lambda)$ is:

$$F(x) = 1 - \frac{\lambda^\alpha}{(\lambda + x)^\alpha}, x > 0$$

A further parameter $\gamma > 0$ can be introduced by setting:

$$F(x) = 1 - \frac{\lambda^\alpha}{(\lambda + x^\gamma)^\alpha}, \quad x > 0$$

This is the CDF of the transformed Pareto or Burr distribution. The additional parameter gives extra flexibility when a fit to data is required.

Formulae for the CDF, PDF and non-central moments of the Burr distribution are given on page 15 of the *Tables*.

There is no built in R code for the Burr distribution so we would have to define the functions `rburr`, `dburr`, `pburr` **and** `qburr` **from first principles as follows:**

```
rburr <- function(n,a,l,g){
rp <- (l*((1-runif(n))^(-1/a)-1))^(1/g)
rp}

dburr <- function(x,a,l,g){
((a*g*l^a)*x^(g-1))/((l+x^g)^(a+1))}

pburr <- function(q,a,l,g){
1-(l/(l+q^g))^a}

qburr <- function(p,a,l,g){
q <- (l*((1-p)^(-1/a)-1))^(1/g)
q}
```

2.4 The three-parameter Pareto distribution

The PDF of the Pareto distribution $Pa(\alpha,\lambda)$ is:

$$f(x) = \frac{\alpha\lambda^\alpha}{(\lambda + x)^{\alpha+1}}, \quad x > 0$$

Another generalisation of the Pareto distribution is to add a further parameter k so that the PDF becomes:

$$f(x) = \frac{\Gamma(\alpha + k)\,\delta^\alpha}{\Gamma(\alpha)\,\Gamma(k)} \frac{x^{k-1}}{(\delta + x)^{\alpha+k}}, \quad x > 0$$

Formulae for the PDF, mean, variance and non-central moments of the three-parameter Pareto distribution are given on page 15 of the *Tables*.

The three-parameter Pareto distribution is equivalent to the two-parameter Pareto distribution when $k=1$.

The moments of the generalised Pareto can be obtained either directly by evaluating $E(X^n)=\int_x x^n\, f(x)\, dx$ or by using a conditional expectation argument.

Here the Core Reading is using the phrase 'generalised Pareto distribution' to refer to the three-parameter Pareto distribution. However, this is not the same as the generalised Pareto distribution that we will meet in Chapter 16.

The easiest way to evaluate the integral expression:

$$\int_x x^n\, f(x)\, dx$$

is to make it look like the PDF of another three-parameter Pareto distribution.

Question

Suppose that X has a three-parameter Pareto distribution with parameters α, λ and k. Derive formulae for $E(X)$ and $\operatorname{var}(X)$.

Solution

The mean is:

$$E(X)=\int_0^\infty x\,\frac{\Gamma(\alpha+k)\lambda^\alpha x^{k-1}}{\Gamma(\alpha)\Gamma(k)(\lambda+x)^{\alpha+k}}\,dx=\int_0^\infty \frac{\Gamma(\alpha+k)\lambda^\alpha x^{k}}{\Gamma(\alpha)\Gamma(k)(\lambda+x)^{\alpha+k}}\,dx$$

This expression can be simplified by making the integrand look like the PDF of the Pareto distribution with parameters $\alpha-1$, λ and $k+1$:

$$E(X)=\lambda\frac{\Gamma(\alpha-1)}{\Gamma(\alpha)}\frac{\Gamma(k+1)}{\Gamma(k)}\int_0^\infty \frac{\Gamma(\alpha+k)\lambda^{\alpha-1}x^{k}}{\Gamma(\alpha-1)\Gamma(k+1)(\lambda+x)^{\alpha+k}}\,dx$$

Since this integrand is a PDF, integrating it over all possible values of x gives us 1. So:

$$E(X)=\lambda\frac{\Gamma(\alpha-1)}{\Gamma(\alpha)}\frac{\Gamma(k+1)}{\Gamma(k)}=\frac{\lambda k}{\alpha-1}$$

Here we are using the result $\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$, which is given on page 5 of the *Tables*.

We can use a similar method to calculate the second moment:

$$E(X^2)=\int_0^\infty x^2\,\frac{\Gamma(\alpha+k)\lambda^\alpha x^{k-1}}{\Gamma(\alpha)\Gamma(k)(\lambda+x)^{\alpha+k}}\,dx=\int_0^\infty \frac{\Gamma(\alpha+k)\lambda^\alpha x^{k+1}}{\Gamma(\alpha)\Gamma(k)(\lambda+x)^{\alpha+k}}\,dx$$

Making the integrand look like the PDF of the Pareto distribution with parameters $\alpha-2$, λ and $k+2$, we have:

$$E(X^2)=\lambda^2\frac{\Gamma(\alpha-2)}{\Gamma(\alpha)}\frac{\Gamma(k+2)}{\Gamma(k)}\int_0^\infty\frac{\Gamma(\alpha+k)\lambda^{\alpha-2}x^{k+1}}{\Gamma(\alpha-2)\Gamma(k+2)(\lambda+x)^{\alpha+k}}dx=\frac{k(k+1)\lambda^2}{(\alpha-1)(\alpha-2)}$$

So the variance is:

$$\begin{aligned}\frac{\lambda^2k(k+1)}{(\alpha-1)(\alpha-2)}-\left(\frac{\lambda k}{\alpha-1}\right)^2&=\frac{\lambda^2k(k+1)(\alpha-1)-\lambda^2k^2(\alpha-2)}{(\alpha-1)^2(\alpha-2)}\\&=\frac{\lambda^2k\left[(k\alpha+\alpha-k-1)-(k\alpha-2k)\right]}{(\alpha-1)^2(\alpha-2)}\\&=\frac{\lambda^2k(k+\alpha-1)}{(\alpha-1)^2(\alpha-2)}\end{aligned}$$

There is no built in R code for the three-parameter Pareto distribution, so we would have to define the function `dgpareto` **from first principles as we did for the Pareto. However, since the CDF does not exist in closed form it is not easy to create functions to obtain probabilities, percentage points or simulated values.**

2.5 The Weibull distribution

The Pareto distribution is a distribution with an upper tail that tends to 0 as a power of x. This gives a distribution with a much heavier tail than the exponential. The expressions for the upper tails of the exponential and the Pareto distributions are:

exponential $P(X>x)=\exp(-\lambda x)$

Pareto $P(X>x)=(\lambda/(\lambda+x))^\alpha$

So, if we want to choose a model with a thick tail so as not to underestimate the probability of a large claim, we might well choose the Pareto distribution to model our claims (assuming that it is a suitable distribution in other respects).

However, these are not the only types of tail.

There is a further possibility. Set:

$$P(X>x)=\exp(-\lambda x^\gamma),\quad \gamma>0$$

There are now two cases. If $\gamma<1$, a distribution with a tail intermediate in weight between the exponential and the Pareto will be obtained, while if $\gamma>1$, the upper tail will be lighter than the exponential ($\gamma=1$ is the exponential distribution).

This distribution is called the Weibull distribution, a very flexible distribution, which can be used as a model for losses in insurance, usually with $\gamma < 1$. A random variable X has a Weibull distribution with parameters $c > 0$ and $\gamma > 0$ if it has CDF:

$$F(x) = 1 - \exp(-cx^{\gamma}),\ x > 0$$

In that case we write $X \sim W(c, \gamma)$. (Note the change from λ to c; this is the notation used in the *Tables for Actuarial Examinations*).

The PDF of the $W(c, \gamma)$ distribution is:

$$f(x) = c\gamma x^{\gamma-1}\exp(-cx^{\gamma}),\ x > 0$$

Formulae for the CDF, PDF and non-central moments of the Weibull distribution are given on page 15 of the *Tables*.

Question

Suppose that X has a Weibull distribution with parameters c and γ. Derive a formula for $E(X)$.

Solution

The mean is:

$$E(X) = \int_0^{\infty} x\, c\gamma x^{\gamma-1} e^{-cx^{\gamma}}\, dx$$

Making the substitution $u = cx^{\gamma}$, so that $\dfrac{du}{dx} = c\gamma\, x^{\gamma-1}$ and $x = \left(\dfrac{u}{c}\right)^{1/\gamma}$ gives:

$$E(X) = \int_0^{\infty} \left(\frac{u}{c}\right)^{1/\gamma} e^{-u} du$$

Now, manipulating the integrand so that it looks like the PDF of a $Gamma\left(1 + \dfrac{1}{\gamma}, 1\right)$ random variable, we have:

$$E(X) = \Gamma\left(1 + \frac{1}{\gamma}\right)\frac{1}{c^{1/\gamma}}\int_0^{\infty}\frac{1}{\Gamma(1 + 1/\gamma)}u^{1/\gamma}e^{-u}du$$

The integral is now equal to 1 (as we're integrating a PDF over all possible values of the random variable). So:

$$E(X) = \Gamma\left(1 + \frac{1}{\gamma}\right)\frac{1}{c^{1/\gamma}}$$

The R code for simulating a random sample of 100 values from the Weibull distribution with $c = 2$ and $\gamma = 0.25$ is:

```
rweibull(100, 0.25, 2^(-1/0.25))
```

R uses a different parameterisation for the scale parameter, *c*.

Similarly, the PDF, CDF and quantiles can be obtained using the R functions `dweibull`**,** `pweibull` **and** `qweibull`**.**

Alternatively, we could redefine them from first principles as follows:

```
rweibull <- function(n,c,g){
rp <- (log(1-runif(n))/c)^(1/g)
rp}

dweibull <- function(x,c,g){
c*g*x^(g-1)*exp(-c*x^g)}

pweibull <- function(q,c,g){
1-exp(-c*x^g)}

qweibull <- function(p,c,g){
q <- (log(1-p)/c)^(1/g)
q}
```

2.6 Illustration of tail weights

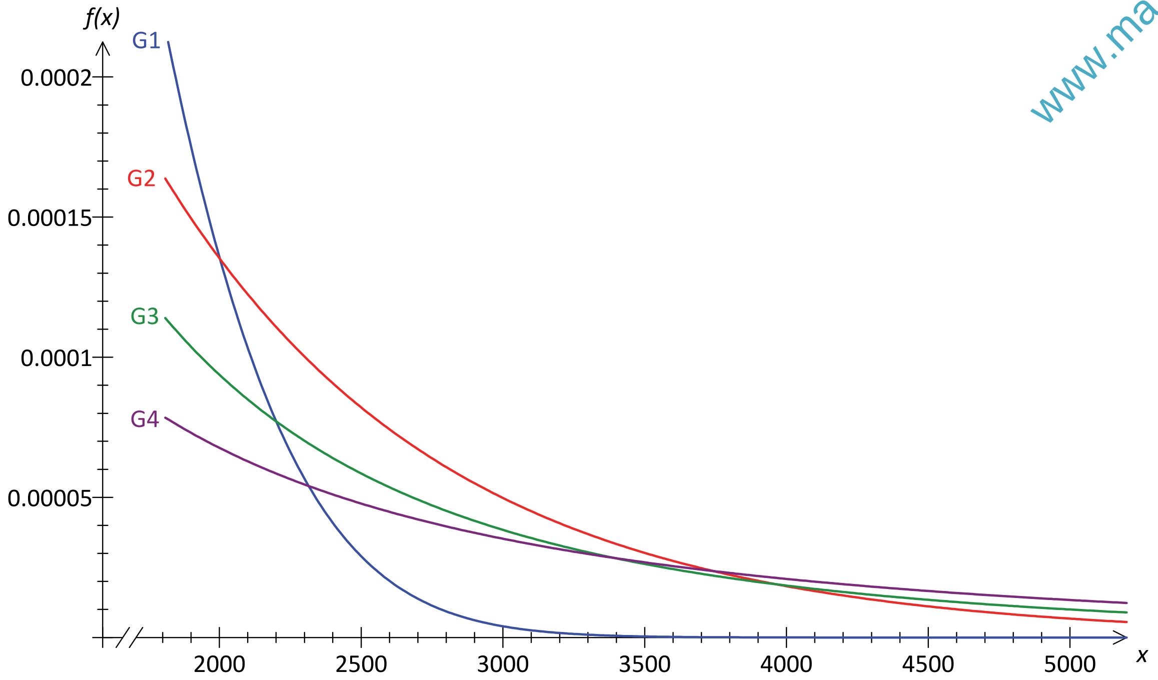
The PDFs shown in the diagram below illustrate the difference in the tails of the exponential, Pareto and Weibull distributions. All four of the distributions have a mean of 1,000.

G1 is a Weibull distribution with parameters $\gamma = 2$ and $c = \dfrac{\pi}{2{,}000^2} = 7.854 \times 10^{-7}$ (standard deviation = 522.7).

G2 is an exponential distribution with $\lambda = \dfrac{1}{1{,}000} = 0.001$ (standard deviation = 1,000).

G3 is a (two-parameter) Pareto distribution with $\alpha = 3$ and $\lambda = 2{,}000$ (standard deviation = 1,732).

G4 is a Weibull distribution with $\gamma = ½$ and $c = \dfrac{1}{\sqrt{500}} = 0.04472$ (standard deviation = 2,236).


f(x)
G1
G2
G3
G4
0.0002
0.00015
0.0001
0.00005
2000
2500
3000
3500
4000
4500
5000
x

3 Estimation

The methods of maximum likelihood, moments and percentiles can be used to fit distributions to sets of data.

We can check the fit in R by plotting a histogram of the data and superimposing the density function of the fitted distribution. Better yet, we can plot an empirical density function from the data using the function `density` **and add the true density function of the fitted distribution.**

A better way is to use the `qqplot` **function to compare the sample data to simulated values from the fitted model distribution. A straight diagonal line indicates perfect fit:**

```
qqplot(<simulated theoretical values>, <sample values>)
abline(0,1)
```

The fit of the distribution can also be tested formally by using a χ^2 test. The method of percentiles is outlined in Section 3.3; the other methods and the χ^2 test have been covered in Subject CS1, Actuarial Statistics 1.

We will now give a summary of the method of moments and maximum likelihood estimation. We will also introduce the method of percentiles and give a brief reminder of how the chi-squared test can be used to check the fit of a statistical distribution to a data set.

3.1 The method of moments

For a distribution with r parameters, the moments are as follows:

$$m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j \qquad j = 1, 2 \dots r$$

where:

$m_j = E(X^j \mid \theta)$, a function of the unknown parameter, θ, being estimated

$n =$ the sample size

$x_i =$ the i *th* value in the sample

The estimate for the parameter, θ, can be determined by solving the equation above. Where there is more than one parameter, they can be determined by solving the simultaneous equations for each m_j.

So, for example, if we are trying to estimate the value of a single parameter, and we have a sample of n claims whose sizes are $x_1, x_2, \dots, x_n$, we would solve the equation:

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

ie we would equate the first non-central moments for the population and the sample.

If we are trying to find estimates for two parameters (for example if we are fitting a gamma distribution and need to obtain estimates for both parameters), we would solve the simultaneous equations:

$$E(X)=\frac{1}{n}\sum_{i=1}^{n}x_i \quad \text{and} \qquad E(X^2)=\frac{1}{n}\sum_{i=1}^{n}x_i^2$$

In fact, in the two-parameter case, estimates are often obtained by equating sample and population means and variances. If we use the n-denominator sample variance:

$$s_n^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n}\left[\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right]$$

this will give the same estimates as would be obtained by equating the first two non-central moments.

More generally, we use as many equations of the form $E(X^k)=\frac{1}{n}\sum_{i=1}^{n}x_i^k$, $k=1,2,\dots$ as are needed to determine estimates of the relevant parameters.

3.2 Maximum likelihood estimation

The likelihood function of a random variable, X, is the probability (or PDF) of observing what was observed given a hypothetical value of the parameter, θ. The maximum likelihood estimate (MLE) is the one that yields the highest probability (or PDF), *ie* that maximises the likelihood function.

For the sample in Section 3.1 above, the likelihood function $L(\theta)$ can be expressed as:

$$L(\theta)=\prod_{i=1}^{n}P\left(X=x_i \mid \theta\right) \text{ for a discrete random variable, } X$$

or:

$$L(\theta)=\prod_{i=1}^{n}f\left(x_i \mid \theta\right) \text{ for a continuous random variable, } X$$

To determine the MLE the likelihood function needs to be maximised. Often it is practical to consider the log-likelihood function:

$$l(\theta)=\log L(\theta)=\sum_{i=1}^{n}\log P\left(X=x_i \mid \theta\right) \text{ for a discrete random variable, } X$$

or:

$$l(\theta)=\log L(\theta)=\sum_{i=1}^{n}\log f\left(x_i \mid \theta\right) \text{ for a continuous random variable, } X$$

If $l(\theta)$ can be differentiated with respect to θ, the MLE, expressed as $\hat{\theta}$, satisfies the expression:

$$\frac{d}{d\theta} l(\hat{\theta}) = 0$$

Where there is more than one parameter, the MLEs for each parameter can be determined by taking partial derivatives of the log-likelihood function and setting each to zero.

The determination of MLEs when the data are incomplete is covered in Chapter 18.

We will now look at the distributions described earlier in this chapter and consider how the parameters can be estimated in each case.

The exponential distribution

It is possible to use the method of maximum likelihood (ML) or the method of moments to estimate the parameter of the exponential distribution.

For example, suppose that an insurance company uses an exponential distribution to model the cost of repairing insured vehicles that are involved in accidents, and the average cost of repairing a random sample of 1,000 vehicles is £2,200.

We can calculate the maximum likelihood estimate of the exponential parameter as follows.

Let $x_1, x_2, \ldots, x_{1,000}$ denote the individual repair costs.

The likelihood of obtaining these values for the costs, if they come from an exponential distribution with parameter λ, is:

$$L = \prod_{i=1}^{1,000} \lambda e^{-\lambda x_i} = \lambda^{1,000} e^{-\lambda \Sigma x_i} = \lambda^{1,000} e^{-1,000\lambda \bar{x}}$$

(where $\bar{x} = \dfrac{1}{1,000} \sum_{i=1}^{1,000} x_i$ denotes the average claim amount).

We want to determine the value of λ that maximises the likelihood, or equivalently the value that maximises the log-likelihood:

$$\ln L = 1,000 \ln \lambda - 1,000 \lambda \bar{x}$$

Differentiating with respect to λ:

$$\frac{\partial}{\partial \lambda} \ln L = \frac{1,000}{\lambda} - 1,000 \bar{x}$$

This is equal to 0 when:

$$\lambda = \frac{1}{\bar{x}}$$

The second derivative is:

$$\frac{\partial^2}{\partial \lambda^2}\log L = -\frac{1{,}000}{\lambda^2}$$

Since the second derivative is negative when $\lambda = \frac{1}{\bar{x}}$, the stationary point is a maximum. So, $\hat{\lambda}$, the maximum likelihood estimate of λ is $\frac{1}{\bar{x}}$, or $\frac{1}{2{,}200}$.

Alternatively, we could argue that the likelihood function is continuous and is always positive (by necessity) and that $\lambda^n e^{-\lambda n \bar{x}} \to 0$ as $\lambda \to 0$ or $\lambda \to \infty$. So any stationary point that we find must be a maximum.

To obtain ML estimates in R, we could use the `fitdistr` **in the MASS package as follows:**

```
      fitdistr(<data vector>,"exponential")
```

Or we could define the log-likelihood function and use the function `nlm` **on the negative value of the log-likelihood function.**

```
      nlm(-<log likelihood function>, <vector of parameters>)
```

So to fit an exponential distribution to a vector `<data>` **with initial estimate of $\lambda = 0.5$ we would use:**

```
      params <- 0.5
      n <- length(<data>)
      sx <- sum(<data>)
      fMLE <- function(params) {n*log(params[1])-params[1]*sx}
      nlm(-fMLE,params)
```

The gamma distribution

The moments have a simple form and so the method of moments is very easy to apply. The MLEs for the gamma distribution cannot be obtained in closed form (*ie* in terms of elementary functions) but the moment estimators can be used as initial estimators in the search for the MLEs.

It is more convenient to obtain MLEs for the gamma distribution using a different parameterisation. Set $\mu = \alpha / \lambda$ and estimate the parameters α and μ. Then recover the MLE of λ by setting $\hat{\lambda} = \hat{\alpha} / \hat{\mu}$. This uses the invariance property of maximum likelihood estimators.

The invariance property says that if $\hat{\theta}$ is the maximum likelihood estimator of θ and $f(\theta)$ is a function of θ, then $f(\hat{\theta})$ is the maximum likelihood estimator of $f(\theta)$.

For the gamma distribution, λ is a function of both α and μ.

To obtain ML estimates in R, we could use the `fitdistr` in the MASS package as follows:

```
fitdistr(<data vector>,"gamma")
```

However, it is better to include the initial estimates obtained from the method of moments (and put a lower limit of say, 0.001 > 0, to prevent invalid answers). For example:

```
fitdistr(<data vector>,dgamma, list(shape = <alpha>,
rate = <lambda>), lower = 0.001)
```

Alternatively, we could define the log-likelihood function and use the function `nlm` on the negative value of the log-likelihood function as before.

The normal distribution

The method of moments and maximum likelihood estimation are both straightforward to apply in this case. Both give the following estimates:

$$\hat{\mu} = \bar{x} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The estimate for the population variance is $\dfrac{n-1}{n} \times$ the usual sample variance. Of course, provided the sample size is large, there will be little difference between estimates calculated using the two different sample variance formulae.

The lognormal distribution

Estimation for the lognormal distribution is straightforward since μ and σ^2 may be estimated using the log-transformed data. Let $x_1, x_2, ..., x_n$ be the observed values and let $y_i = \log x_i$. The MLEs of μ and σ^2 are $\bar{y}$ and s_y^2, where the subscript y signifies a sample variance (n-denominator) computed on the y values.

In other words, the maximum likelihood estimate of σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Alternatively the method of moments can be used to estimate the parameters.

As an example, suppose that based on an analysis of past claims, an insurance company believes that individual claims in a particular category for the coming year will be lognormally distributed with a mean size of £5,000 and a standard deviation of £7,500. The company wants to estimate the proportion of claims that will exceed £25,000.

To do this, it needs to estimate the parameters, μ and σ^2, of the lognormal distribution. Equating the formulae for the mean and standard deviation of the lognormal distribution to the values given gives:

$$e^{\mu+\frac{1}{2}\sigma^2}=5{,}000 \quad \text{and} \quad e^{\mu+\frac{1}{2}\sigma^2}\sqrt{e^{\sigma^2}-1}=7{,}500$$

Dividing the second equation by the first gives:

$$\sqrt{e^{\sigma^2}-1}=\frac{7{,}500}{5{,}000}=1.5$$

$$\Rightarrow \quad \sigma^2=1.179$$

We can now solve for μ:

$$\mu=\log 5{,}000-½(1.179)=7.928$$

So the proportion of claims expected to exceed £25,000 is:

$$\begin{aligned} P(X>25{,}000)&=P(\ln X>\ln 25{,}000)\\ &=P(N(7.928,1.179)>\ln 25{,}000)\\ &=P\left(N(0,1)>\frac{\ln 25{,}000-7.928}{\sqrt{1.179}}\right)\\ &=1-\Phi(2.025)=0.021 \end{aligned}$$

ie 2.1% of claims are expected to exceed £25,000.

To obtain ML estimates in R, we could use the `fitdistr` in the MASS package as follows:

```
fitdistr(<data vector>,"log-normal")
```

Alternatively, we could define the log-likelihood function and use the function `nlm` on the negative value of the log-likelihood function as before.

The two-parameter Pareto distribution

The method of moments is very easy to apply in the case of the two-parameter Pareto distribution, but the estimators obtained in this way will tend to have rather large standard errors, mainly because S^2, the sample variance, has a very large variance. However, the method does provide initial estimates for more efficient methods of estimation that may not be so simple to apply, like maximum likelihood, where numerical methods may need to be used.

Question

Claims arising from a particular group of policies are believed to follow a Pareto distribution with parameters α and λ. A random sample of 20 claims gives values such that $\sum x = 1{,}508$ and $\sum x^2 = 257{,}212$. Estimate α and λ using the method of moments.

Solution

Suppose that X is the claim amount random variable. Then:

$$E(X) = \frac{\lambda}{\alpha - 1}$$

Rearranging the variance formula to find $E(X^2)$, we have:

$$E(X^2) = \text{var}(X) + [E(X)^2] = \frac{2\lambda^2}{(\alpha-1)(\alpha-2)}$$

So we set:

$$\frac{\lambda}{\alpha-1} = \frac{1{,}508}{20} = 75.4 \quad \text{and} \quad \frac{2\lambda^2}{(\alpha-1)(\alpha-2)} = \frac{257{,}212}{20} = 12{,}860.6$$

Squaring the first of these equations and substituting into the second, we see that:

$$\frac{2 \times 75.4^2(\alpha-1)}{\alpha-2} = 12{,}860.6$$

Solving this equation, we find that the method of moments estimates of α and λ are 9.630 and 650.7, respectively.

The three-parameter Pareto distribution

Things are not quite so easy for the three-parameter Pareto distribution.

As for estimation, the CDF does not exist in closed form, so the method of percentiles is not available.

The method of percentiles is described in Section 3.3.

ML can be used, but again suitable computer software is required; the method of moments can provide initial estimates for any iterative scheme.

We will need to define the log-likelihood function and use the function `nlm` on the negative value of the log-likelihood function as before.

The Weibull and Burr distributions

Neither the method of moments nor maximum likelihood is elementary to apply if both c and γ are unknown (although if a computer is available, as would be the case in practice, the equations are simple enough).

To obtain ML estimates in R, we could use the `fitdistr` in the MASS package as follows:

```
fitdistr(<data vector>,"weibull")
```

However, it is better to include the initial estimates obtained from the method of percentiles (and put a lower limit of say, 0.001 > 0, to prevent invalid answers). For example:

```
fitdistr(<data vector>,dweibull,
list(shape = <gamma>, scale = <c^(-1/gamma)>, lower = 0.001)
```

Alternatively, we could define the log-likelihood function and use the function `nlm` on the negative value of the log-likelihood function as before.

In the case where γ has the known value γ^*, maximum likelihood is easy enough.

To do this, we let $y_i = x_i^{\gamma}$. If the original distribution is Weibull, the y values now have an exponential distribution. If the original distribution is Burr, the y values now come from a Pareto distribution. This can be seen by comparing the CDFs.

In the case of the Weibull, the MLE of c can now be determined in the usual way. In the case of the Burr distribution, the estimates have to be calculated numerically (since the MLEs of the parameters of the Pareto distribution cannot be calculated algebraically).

3.3 The method of percentiles

The distribution function of the $W(c,\gamma)$ distribution is an elementary function, and a simple method of estimation of both c and γ is based on this. The method involves equating selected sample percentiles to the distribution function; for example, equate the sample quartiles, the 25th and 75th sample percentiles, to the population quartiles. This corresponds to the way in which sample moments are equated to population moments in the method of moments. This method will be referred to as the method of percentiles.

In the method of moments, the first two moments are used if there are two unknown parameters, and this seems intuitively reasonable (although the theoretical basis for this is not so clear). In a similar fashion, when using the method of percentiles, the median would be used if there were one parameter to estimate. With two parameters, the best procedure is less clear, but the lower and upper quartiles seem a sensible choice.

Example

Estimate c and γ in the Weibull distribution using the method of percentiles, where the first sample quartile is 401 and the third sample quartile is 2,836.75.

Solution

The two equations for c and γ are:

$$\boldsymbol{F(401) = 1 - \exp(-c \times 401^{\gamma}) = 0.25}$$

$$\boldsymbol{F(2{,}836.75) = 1 - \exp(-c \times 2{,}836.75^{\gamma}) = 0.75}$$

which can be rewritten as:

$$\boldsymbol{-c \times 401^{\gamma} = \ln 0.75}$$

and: $\boldsymbol{-c \times 2{,}836.75^{\gamma} = \ln 0.25}$

Dividing, it is found that $\tilde{\gamma}$ = 0.8038, and hence $\tilde{c}$ = 0.002326, where ~ denotes the percentile estimate. Note that $\tilde{\gamma}$ is less than 1, indicating a fatter tail than the exponential distribution gives.

We can apply the method of percentiles to any distribution for which it is possible to calculate a closed form for the cumulative distribution function, although the resulting algebra can be messy.

Question

Claims arising from a particular group of policies are believed to follow a Pareto distribution with parameters α and λ. A random sample of 20 claims has a lower quartile of 11 and an upper quartile of 85. Estimate the values of α and λ using the method of percentiles.

Solution

The cumulative distribution function of the Pareto distribution is $F(x) = 1 - \left(\frac{\lambda}{\lambda + x}\right)^{\alpha}$.

Q_1, the lower quartile of the distribution, satisfies the equation:

$$F(Q_1) = 1 - \left(\frac{\lambda}{\lambda + Q_1}\right)^{\alpha} = 0.25$$

and Q_3, the upper quartile of the distribution, satisfies the equation:

$$F(Q_3) = 1 - \left(\frac{\lambda}{\lambda + Q_3}\right)^{\alpha} = 0.75$$

So:

$$Q_1 = \lambda\left[(3/4)^{-1/\alpha} - 1\right] \qquad \text{and} \qquad Q_3 = \lambda\left[(1/4)^{-1/\alpha} - 1\right]$$

The method of percentiles estimates $\tilde{\alpha}$ and $\tilde{\lambda}$ are obtained by setting $Q_1 = 11$ and $Q_3 = 85$:

$$11 = \tilde{\lambda}\left[(3/4)^{-1/\tilde{\alpha}} - 1\right]$$

$$85 = \tilde{\lambda}\left[(1/4)^{-1/\tilde{\alpha}} - 1\right]$$

We can eliminate $\tilde{\lambda}$ by dividing these equations. This gives:

$$\frac{11}{85} = \frac{(3/4)^{-1/\tilde{\alpha}} - 1}{(1/4)^{-1/\tilde{\alpha}} - 1}$$

We cannot solve this algebraically but it can easily be done on a computer, *eg* using the goalseek function in Excel. Doing this, we find that $\tilde{\alpha} = 1.284$ and hence $\tilde{\lambda} = 43.790$.

These estimates are very different from those obtained using the method of moments. (In an earlier question, we calculated the method of moments estimates of α and λ to be 9.630 and 650.7, respectively.)

The method of percentiles is very unreliable for estimating the parameters of a Pareto distribution unless we use extremely large samples. In this particular case, the method of percentiles is unlikely to give us reasonable estimates unless we use samples of, say, 1,000 or more.

We now turn to the Burr distribution.

Since the CDF exists in closed form, it may be possible to fit the Burr distribution to data by using the method of percentiles; ML will certainly require the use of computer software that allows non-linear optimisation.

We will need to define the log-likelihood function and use the function `nlm` on the negative value of the log-likelihood function as before.

4 Goodness-of-fit tests

As mentioned earlier, one way of testing whether a given loss distribution provides a good model for the observed claim amounts is to apply a chi-squared goodness-of-fit test.

Recall that the formula for the test statistic is $\sum \frac{(O-E)^2}{E}$, where:

- O is the observed number in a particular category
- E is the corresponding expected number predicted by the assumed probabilities
- the sum is over all possible categories.

A high value for the total indicates that the overall discrepancy is quite large and would lead us to reject the model.

As an example, suppose that an insurance company uses an exponential distribution to model the cost of repairing insured vehicles that are involved in accidents, and the average cost of repairing a random sample of 1,000 vehicles is £2,200. A breakdown of the repair costs revealed the following numbers in different bands:

Repair cost, £	**Observed number**
0 – 1,000	200
1,000 – 2,000	300
2,000 – 3,000	250
3,000 – 4,000	150
4,000 – 5,000	100
5,000+	0

We can use this information to test whether the exponential distribution provides a good model for the individual repair costs.

Here we are testing:

H_0 : repair costs are exponentially distributed

against:

H_1 : repair costs are not exponentially distributed

In order to apply the chi-squared test, we need to calculate the expected number of repair costs in each interval based on the assumption that the null hypothesis is true.

Using the maximum likelihood estimate of the value of λ (*ie* $1/2,200$), the probability that an individual repair cost will fall in the interval £2,000 - £3,000 is:

$$\int_{2,000}^{3,000} \lambda e^{-\lambda x}\,dx = \left[-e^{-\lambda x}\right]_{2,000}^{3,000} = e^{-2,000\lambda} - e^{-3,000\lambda} = 0.1472$$

and the expected number for this band is: $1,000 \times 0.1472 = 147.2$

The expected numbers for all the intervals can be calculated in a similar way, giving the following results:

365.3, 231.8, 147.2, 93.4, 59.3, 103.0

The value of the test statistic is:

$$\sum \frac{(O-E)^2}{E} = \frac{(200-365.3)^2}{365.3} + \frac{(300-231.8)^2}{231.8} + \cdots + \frac{(0-103.0)^2}{103.0} = 331.89$$

We have 6 intervals, but we have equated the totals and estimated one parameter. So there are $6-1-1=4$ degrees of freedom.

The observed value of the chi-squared statistic far exceeds 14.86, the upper 99.95% point of the chi-squared distribution with 4 degrees of freedom (given on p169 of the *Tables*). So we can reject H_0 with almost total confidence and conclude that the repair costs do not conform to an exponential distribution.

In fact we need only work out the value of the first term in the chi-squared statistic to see that we will reject the null hypothesis.

This conclusion is supported by the observation that, if the values did come from an exponential distribution, we would expect the numbers in each band to decline steadily. However, we recorded 100 fewer values in the first band than in the second.

Chapter 15 Summary

Loss distributions

Individual claim amounts can be modelled using a loss distribution. Loss distributions are often positively skewed and long-tailed.

The (cumulative) distribution function of X is denoted by $F_X(x)$. It is defined by the equation: $F_X(x) = P(X \leq x)$.

The (probability) density function of X is denoted by $f_X(x)$. It is defined by the equation: $f_X(x) = F'_X(x)$, wherever this derivative exists.

Distributions such as the exponential, normal, lognormal, gamma, Pareto, Burr and Weibull distributions are commonly used to model individual claim amounts.

Once the form of the loss distribution has been decided upon, the values of the parameters must be estimated. This might be done using the method of maximum likelihood, the method of moments, or the method of percentiles. Goodness of fit can then be checked using a chi-squared test.

Method of moments

The method of moments involves equating population and sample moments to solve for the unknown parameter values. If there is one parameter to estimate, we equate the population mean with the sample mean. If there are two parameters to estimate, we could equate the first two non-central population moments with the equivalent non-central sample moments. Equivalently, we could equate the first two central population moments with the equivalent central sample moments, noting that (for equivalence) we would need to use the *n*-denominator sample variance.

Method of maximum likelihood

The steps involved in finding a maximum likelihood estimate (MLE) are as follows:

- write down the likelihood function L – this the probability/PDF of obtaining the values we have observed
- take logs and simplify the resulting expression
- differentiate the log-likelihood with respect to each parameter to be estimated – this will involve partial differentiation if there is more than one parameter to be estimated
- set the derivatives equal to 0 and solve the equations simultaneously
- check that the resulting values are maxima.

Method of percentiles

The method of percentiles involves equating population and sample percentiles to solve for the unknown parameter values. If there is just one parameter to estimate, we equate the population median with the sample median. If there are two parameters to estimate, we equate the population lower and upper quartiles with the sample lower and upper quartiles.

Testing goodness of fit

We can test whether a given loss distribution provides a good model for the observed claim amounts by applying a chi-squared goodness-of-fit test.

The formula for the test statistic is $\sum \frac{(O-E)^2}{E}$, where:

- O is the observed number in a particular category
- E is the corresponding expected number predicted by the assumed probabilities
- the sum is over all possible categories.

Under the null hypothesis (that the model is correct), the test statistic has a chi-squared distribution.

Chapter 15 Practice Questions

15.1 Losses arising from a portfolio follow a Pareto distribution with parameters $\alpha = 3$ and $\lambda = 2{,}000$.

Calculate the probability that a randomly chosen loss amount exceeds the mean loss amount.

15.2 Suppose that X has a Weibull distribution with parameters c and γ.

(i) Using the formula for $E(X^r)$ given in the *Tables,* write down an expression for $\text{var}(X)$.

(ii) Show that, when $\gamma = 1$, this reduces to the formula for the variance of an exponential random variable.

15.3 Show that if $\gamma = ½$, the standard deviation of the Weibull distribution is greater than the mean, whereas if $\gamma = 2$ the opposite is true.

15.4 The random variable X follows a gamma distribution with parameters $\alpha = 20$ and $\lambda = 0.1$. Determine the value of a such that:

$$P(X > a) = 0.05$$

15.5 The random variable X has a Burr distribution with parameters $\gamma = 2$ and $\lambda = 500$.

(i) Show that the maximum likelihood estimate of the parameter α, based on a random sample $x_1, x_2, \dots, x_n$ is:

$$\hat{\alpha} = \frac{n}{\Sigma \log(500 + x_i^2) - n \log 500}$$

You may assume that this is a maximum.

(ii) Evaluate this based on a sample consisting of the five values 52, 109, 114, 163 and 181.

15.6 Claim amounts from a particular group of policies have the following distribution:

Amount	£200	£500
Probability	p	$1-p$

In a random sample of 40 claims, 25 were for £200 and the other 15 were for £500.

Calculate the maximum likelihood estimate of p.

15.7 A loss amount random variable has MGF:

$$M(t) = 0.4(1-20t)^{-2} + 0.6(1-30t)^{-3}$$

Calculate the expected loss amount.

15.8 Exam style

Individual claim amounts on a portfolio of motor insurance policies follow a gamma distribution with parameters α and λ. It is known that $\lambda = 0.8$ for all drivers, but the value of α varies across the population.

Given that $\alpha \sim Gamma(200, 0.5)$, calculate the mean and variance of a randomly chosen claim amount. [5]

15.9 Exam style

(i) The distribution of claims on a portfolio of general insurance policies is a Weibull distribution, with density function $f_1(x)$ where:

$$f_1(x) = 2cxe^{-cx^2} \quad (x > 0)$$

It is expected that one claim out of every 100 will exceed £1,000. Use this information to estimate *c*. [2]

(ii) An alternative suggestion is that the density function is $f_2(x)$, where:

$$f_2(x) = \lambda e^{-\lambda x} \qquad (x > 0)$$

Use the same information as in part (i) to estimate λ. [2]

(iii) (a) For each of $f_1(x)$ and $f_2(x)$ calculate the value of M such that:

$$P(X > M) = 0.001$$

(b) Comment on these results. [3]

[Total 7]

15.10 Exam style

A random sample of 100 claim amounts $x_1, x_2, \dots, x_{100}$ is observed from a Weibull distribution with parameter $\gamma = 2$, where c is unknown. For these data:

$$\sum x_i = 487{,}926 \qquad \sum x_i^2 = 976{,}444{,}000 \qquad \text{sample median} = 4{,}500$$

(i) Show that the maximum likelihood estimate for c based on a sample of size n is given by:

$$\hat{c} = n \Big/ \sum x_i^2$$

and hence estimate the value of c. [4]

(ii) Estimate the value of c using the method of moments. [2]

(iii) Calculate the method of percentiles estimate of c. [2]

[Total 8]

15.11 Exam style

Claims arising from a certain type of insurance policy are believed to follow an exponential distribution. The lower quartile claim is 200.

Calculate the mean claim size. [3]

Chapter 15 Solutions

15.1 Let X denote the loss amount random variable. Then $X \sim Pa(3,2000)$ and $E(X)=\dfrac{2,000}{3-1}=1,000$.

So the required probability is:

$$P(X>1,000)=1-F_X(1,000)=1-\left[1-\left(\frac{2,000}{2,000+1,000}\right)^3\right]=\left(\frac{2,000}{3,000}\right)^3=0.29630$$

15.2 (i) ***Variance***

We have:

$$E(X)=\Gamma\left(1+\frac{1}{\gamma}\right)\frac{1}{c^{1/\gamma}} \quad \text{and} \quad E(X^2)=\Gamma\left(1+\frac{2}{\gamma}\right)\frac{1}{c^{2/\gamma}}$$

So:

$$\text{var}(X)=E(X^2)-[E(X)]^2=\Gamma\left(1+\frac{2}{\gamma}\right)\frac{1}{c^{2/\gamma}}-\left[\Gamma\left(1+\frac{1}{\gamma}\right)\frac{1}{c^{1/\gamma}}\right]^2$$

(ii) ***Simplification when $\gamma=1$***

When $\gamma=1$, this becomes:

$$\Gamma(3)\times\frac{1}{c^2}-\left[\Gamma(2)\times\frac{1}{c}\right]^2=\frac{2}{c^2}-\left[\frac{1}{c}\right]^2=\frac{1}{c^2}$$

which is the formula for the variance of an $Exp(c)$ random variable.

15.3 When $\gamma=½$:

$$E(X)=\frac{\Gamma(1+2)}{c^2}=\frac{2!}{c^2}=\frac{2}{c^2}$$

$$E(X^2)=\frac{\Gamma(1+4)}{c^4}=\frac{4!}{c^4}=\frac{24}{c^4}$$

and:

$$\text{var}(X)=E(X^2)-[E(X)]^2=\frac{24}{c^4}-\left(\frac{2}{c^2}\right)^2=\frac{20}{c^4}$$

So the standard deviation is $\dfrac{\sqrt{20}}{c^2}$, which is greater than $E(X)$.

When $\gamma = 2$:

$$E(X) = \frac{\Gamma(1+½)}{c^{½}} = \frac{\Gamma(1.5)}{c^{½}}$$

Using the properties of the gamma function given on page 5 of the *Tables*:

$$\Gamma(1.5) = 0.5\Gamma(0.5) = 0.5\sqrt{\pi}$$

So:

$$E(X) = \frac{0.5\sqrt{\pi}}{c^{½}} = \frac{0.886227}{c^{½}}$$

Also:

$$E(X^2) = \frac{\Gamma(1+2\times½)}{c^{2\times½}} = \frac{\Gamma(2)}{c} = \frac{1}{c}$$

and:

$$\text{var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{c} - \left(\frac{0.5\sqrt{\pi}}{c^{½}}\right)^2 = \frac{1-0.25\pi}{c}$$

So the standard deviation is $\sqrt{\frac{1-0.25\pi}{c}} = \frac{0.463251}{c^{½}}$, which is less than $E(X)$.

In fact the mean and standard deviation are equal when $\gamma = 1$.

15.4 We can calculate the value of a using the relationship between the gamma distribution and the chi-squared distribution:

$$X \sim Gamma(20, 0.1) \Leftrightarrow 2\times 0.1X \sim \chi^2_{2\times 20}$$

So:

$$P(X > a) = P(0.2X > 0.2a) = P(\chi^2_{40} > 0.2a) = 0.05$$

ie $0.2a$ is the upper 5% point of χ^2_{40}. From page 169 of the *Tables*, we see that the upper 5% point of this chi-squared distribution is 55.76. So:

$$a = \frac{55.76}{0.2} = 278.8$$

The value of a *can also be determined in R using the command* $qgamma(0.95, 20, 0.1)$. *The R command* q *gives us the percentiles of a distribution. We follow the letter* q *with the name of the distribution. Here we want the upper 5% point, ie the 95th percentile, so the first argument is 0.95. The second and third arguments are the parameters of the gamma distribution.*

15.5 (i) ***Maximum likelihood estimate***

The likelihood function is:

$$L(\alpha,\lambda,\gamma)=\prod_{i=1}^{n} f(x_i)$$

$$=\prod_{i=1}^{n}\frac{\alpha\gamma\lambda^{\alpha}x_i^{\gamma-1}}{(\lambda+x_i^{\gamma})^{\alpha+1}}$$

$$=\alpha^n\gamma^n\lambda^{n\alpha}\prod_{i=1}^{n}\frac{x_i^{\gamma-1}}{(\lambda+x_i^{\gamma})^{\alpha+1}}$$

Taking logs:

$$\ln L=n\ln\alpha+n\ln\gamma+n\alpha\ln\lambda+(\gamma-1)\sum_{i=1}^{n}\ln x_i-(\alpha+1)\sum_{i=1}^{n}\ln(\lambda+x_i^{\gamma})$$

Differentiating with respect to α :

$$\frac{\partial}{\partial\alpha}\ln L=\frac{n}{\alpha}+n\ln\lambda-\sum_{i=1}^{n}\ln(\lambda+x_i^{\gamma})$$

This is equal to 0 when:

$$\alpha=\frac{n}{\sum_{i=1}^{n}\ln(\lambda+x_i^{\gamma})-n\ln\lambda}$$

Since we can assume that this is a maximum, we can say that the maximum likelihood estimate of α is:

$$\hat{\alpha}=\frac{n}{\sum_{i=1}^{n}\ln(\lambda+x_i^{\gamma})-n\ln\lambda}$$

(ii) ***Numerical value***

For the sample given, we have $\sum_{i=1}^{5}\ln(500+x_i^2)=47.6245$. So:

$$\hat{\alpha}=\frac{5}{47.6245-5\ln 500}=0.3021$$

α is the easy parameter to estimate. λ and γ are much more difficult to estimate using MLE because of the form of the last term in the log-likelihood function.

15.6 Let X denote the claim amount random variable. Then the likelihood function is:

$$L = C\left[P(X=200)\right]^{25}\left[P(X=500)\right]^{15} = C\,p^{25}(1-p)^{15}$$

where C is a constant.

The log-likelihood function is:

$$\ln L = \ln C + 25\ln p + 15\ln(1-p)$$

Differentiating this with respect to p gives:

$$\frac{d\ln L}{dp} = \frac{25}{p} - \frac{15}{1-p}$$

Now:

$$\begin{aligned}\frac{d\ln L}{dp} = 0 \Leftrightarrow \frac{25}{p} &= \frac{15}{1-p}\\ \Leftrightarrow 25-25p &= 15p\\ \Leftrightarrow 25 &= 40p\\ \Leftrightarrow p &= \frac{25}{40} = 0.625\end{aligned}$$

So we have a stationary point when $p = 0.625$. To determine the nature of the stationary point, we check the sign of the second derivative:

$$\frac{d^2\ln L}{dp^2} = -\frac{25}{p^2} - \frac{15}{(1-p)^2}$$

This is negative when $p = 0.625$. (In fact, this second derivative is always negative.) So the maximum likelihood estimate of p is 0.625.

15.7 Differentiating the MGF:

$$\begin{aligned}M'(t) &= 0.4(-2)(-20)(1-20t)^{-3} + 0.6(-3)(-30)(1-30t)^{-4}\\ &= 16(1-20t)^{-3} + 54(1-30t)^{-4}\end{aligned}$$

The expected loss amount is:

$$M'(0) = 16 + 54 = 70$$

15.8 Let X denote the amount of a randomly chosen claim. We know that $X \mid \alpha \sim Gamma(\alpha, 0.8)$. So, using the conditional expectation formula:

$$E(X) = E(E(X \mid \alpha)) = E\left(\frac{\alpha}{0.8}\right) = \frac{1}{0.8}E(\alpha)$$ [1]

Then using the fact that $\alpha \sim Gamma(200, 0.5)$:

$$E(X) = \frac{1}{0.8} \times \frac{200}{0.5} = 500$$ [1]

Using the conditional variance formula:

$$\text{var}(X) = E(\text{var}(X \mid \alpha)) + \text{var}(E(X \mid \alpha)) = E\left(\frac{\alpha}{0.8^2}\right) + \text{var}\left(\frac{\alpha}{0.8}\right) = \frac{1}{0.8^2}E(\alpha) + \frac{1}{0.8^2}\text{var}(\alpha)$$ [2]

Then using the fact that $\alpha \sim Gamma(200, 0.5)$:

$$\text{var}(X) = \frac{1}{0.8^2} \times \frac{200}{0.5} + \frac{1}{0.8^2} \times \frac{200}{0.5^2} = 1{,}875$$ [1]

15.9 *This is Subject 106, September 2003, Question 5.*

(i) ***Estimate c***

The random variable X has a Weibull distribution. Comparing the given PDF with the Weibull PDF from page 15 of the *Tables*:

$$f_1(x) = 2cxe^{-cx^2} = c\gamma x^{\gamma-1}e^{-cx^\gamma} \quad \Rightarrow \quad \gamma = 2$$

We are told that $P(X > 1{,}000)$ is expected to be 0.01 and we know that:

$$P(X > 1{,}000) = 1 - F(1{,}000) = e^{-c \times 1{,}000^2}$$ [1]

Setting this equal to 0.01 gives the estimated value of c to be:

$$\hat{c} = -\frac{\ln 0.01}{1{,}000^2} = 4.605 \times 10^{-6}$$ [1]

(ii) ***Estimate λ***

Here, X has an exponential distribution and:

$$P(X > 1{,}000) = 1 - F(1{,}000) = e^{-1{,}000\lambda}$$ [1]

Setting this equal to 0.01 gives the estimated value of λ to be:

$$\hat{\lambda} = -\frac{\ln 0.01}{1{,}000} = 4.605 \times 10^{-3}$$ [1]

(iii)(a) ***Calculate M***

We require M such that:

$$P(X>M)=1-F(M)=0.001$$

For the Weibull distribution with $c=4.605\times 10^{-6}$:

$$e^{-4.605\times 10^{-6}\times M^2}=0.001$$

$$\Rightarrow M^2=-\frac{\ln 0.001}{4.605\times 10^{-6}}=1{,}500{,}000$$

$$\Rightarrow M=1{,}225$$ [1]

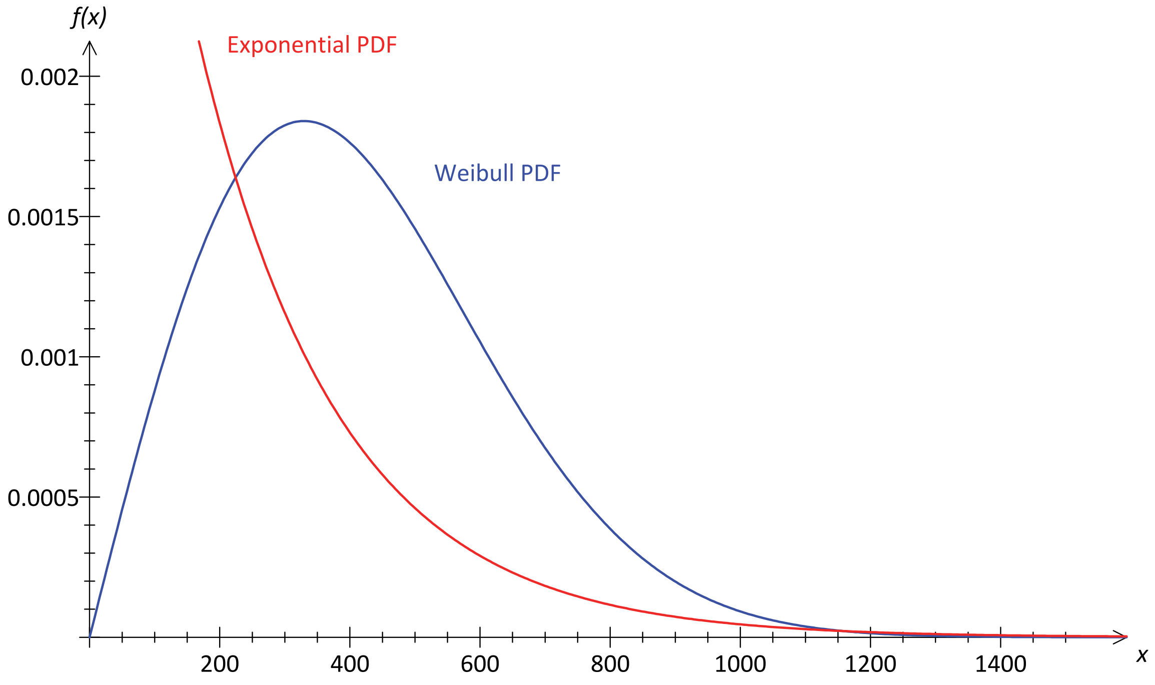
For the exponential distribution with $\lambda=4.605\times 10^{-3}$:

$$e^{-4.605\times 10^{-3}\times M}=0.001 \quad\Rightarrow\quad M=-\frac{\ln 0.001}{4.605\times 10^{-3}}=1{,}500$$ [1]

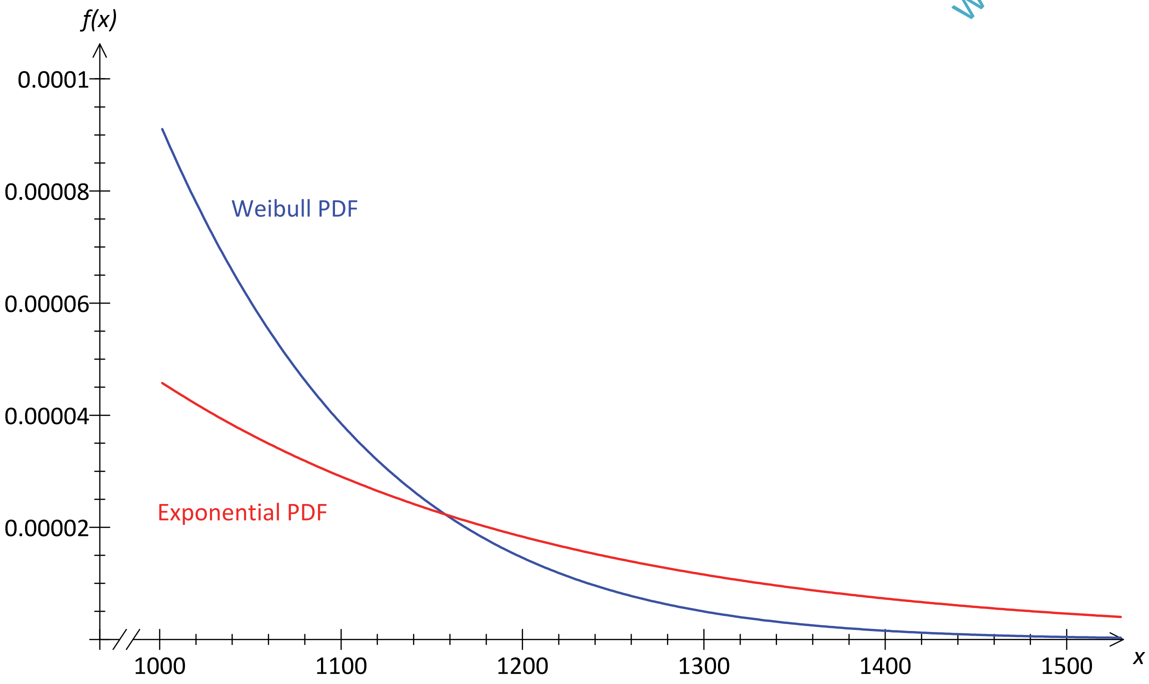
(iii)(b) ***Comment***

The probability that the Weibull random variable exceeds 1,225 is 0.001 but the probability that the exponential random variable exceeds 1,225 is more than 0.001. This is because the exponential distribution has a heavier tail than the Weibull. [1]

A graph of the distributions is shown below:



Looking more closely at the tails, it is clear that the exponential distribution has a heavier tail than the Weibull distribution:



15.10 (i)(a) ***Maximum likelihood estimate***

The PDF of the Weibull distribution is:

$$f(x) = c\gamma x^{\gamma-1} e^{-cx^{\gamma}} \ , \ x > 0$$

So the likelihood function in this case is:

$$L(c) = 2cx_1 e^{-cx_1^2} \times \cdots \times 2cx_n e^{-cx_n^2} = \text{constant} \times c^n e^{-c\sum x_i^2}$$ [½]

Taking logs, we obtain:

$$\ln L = \text{constant} + n \ln c - c\sum_{i=1}^{n} x_i^2$$ [½]

Differentiating this with respect to c, we obtain:

$$\frac{d}{dc} \ln L = \frac{n}{c} - \sum x_i^2$$ [½]

This is equal to 0 when:

$$c = \frac{n}{\sum x_i^2}$$ [½]

We can check that this does give us a maximum by examining the second derivative of the log-likelihood:

$$\frac{d^2}{dc^2}\ln L = -\frac{n}{c^2}$$ [½]

This is negative when $c=\dfrac{n}{\sum x_i^2}$. So, $\hat{c}$, the maximum likelihood estimate of c is $\dfrac{n}{\sum x_i^2}$. [½]

Substituting the sample data results into the formula from part (i)(a) gives:

$$\hat{c}=\frac{100}{976{,}444{,}000}=1.0241\times 10^{-7}$$ [1]

(ii) ***Method of moments estimate***

We obtain the corresponding method of moments estimate for c by equating the sample and population means.

Using the formula for the mean of the Weibull distribution with $\gamma = 2$ (given on page 15 of the *Tables*) and the properties of the gamma function (given on page 5 of the *Tables*), we have:

$$E(X)=\frac{\Gamma(1+½)}{c^{½}}=\frac{0.5\Gamma(0.5)}{c^{½}}=\frac{0.5\sqrt{\pi}}{c^{½}}$$ [1]

From the data we have:

$$\bar{x}=\frac{487{,}926}{100}=4{,}879.26$$ [½]

Equating $E(X)$ and $\bar{x}$ gives:

$$\hat{c}=\left(\frac{0.5\sqrt{\pi}}{\bar{x}}\right)^2=\left(\frac{0.5\sqrt{\pi}}{4{,}879.26}\right)^2=3.299\times 10^{-8}$$ [½]

(iii) ***Method of percentiles estimate***

The median of the distribution is the value of M such that $F(M)=½$.

Equating this to the sample median of 4,500 gives the method of percentiles estimate, $\tilde{c}$:

$$F(4{,}500)=1-e^{-\tilde{c}\times 4{,}500^2}=½ \quad\Rightarrow\quad \tilde{c}=-\frac{\ln ½}{4{,}500^2}=3.423\times 10^{-8}$$ [2]

15.11 Since the lower quartile is 200, we have:

$$F(200) = 0.25$$

Also, using the fact that claim amounts follow an exponential distribution:

$$F(200) = 1 - e^{-200\lambda}$$ [½]

So:

$$e^{-200\lambda} = 1 - 0.25 = 0.75$$ [½]

Taking logs:

$$\lambda = -\frac{1}{200}\ln 0.75 = 0.0014384$$ [1]

So the mean claim amount is:

$$\frac{1}{\lambda} = 695.21$$ [1]

16

Extreme value theory

Syllabus objectives

1.4 Introduction to extreme value theory.

1.4.1 Recognise extreme value distributions, suitable for modelling the distribution of severity of loss and their relationships.

1.4.2 Calculate various measures of tail weight and interpret the results to compare the tail weights.

0 Introduction

In this chapter we look at how we can model extreme events. In the context of insurance, these are events that are very unlikely but can have a large financial impact. Examples include natural catastrophes such as earthquakes, man-made catastrophes such as aeroplane crashes and financial events such as stock market crashes.

Our first thought in modelling extreme events might be to fit a distribution to past data and then to use the tails of the distribution to estimate the probability of future extreme events. For example, consider a data set of past claim amounts on an insurer's motor insurance portfolio. An actuary could fit a loss distribution to these data values. However, the estimation of the parameters of the distribution would be heavily influenced by the bulk of the past claims data, which is likely to be non-extreme. Relatively little weight would be placed on the extreme data in the fitting process. Therefore, if the insurer uses the fitted distribution to estimate the probability of future extreme events, such events may be underestimated.

Better modelling of extreme events can be done by considering distributions that are fitted specifically to the tail of a dataset rather than to the entire dataset. We consider two such distributions:

1. the generalised extreme value distribution is studied in Section 2
2. the generalised Pareto distribution in studied Section 3.

In Section 4, we look at measures of tail weight, *ie* how likely extreme values are to occur.

1 Extreme events and extreme value theory

1.1 Extreme events

Question

Define an 'extreme' event in terms of its frequency and severity.

Solution

An extreme event is one that occurs with very low frequency and very high severity.

1.2 Difficulties modelling extreme events

Low frequency events involving large losses can have a devastating impact on companies and investment funds. The 'credit crunch' that started in 2007 was an example of this. It generated more extreme movements in share prices than had been seen for over 20 years previously.

The credit crunch highlighted many deficiencies in the modelling of extreme events by financial providers. For example:

- Financial risk events such as asset price movements were often modelled using normal distributions. However, empirical evidence suggests that this is not the case in practice. We discuss this further below.
- Distributions were typically fitted to whole (rather than extreme) datasets of asset price movements. This resulted in fitted distributions that understated the probability of extreme events.
- Many financial providers focused on the losses predicted to occur at the 99th percentile or 99.5th percentile point of a distribution. Providers held capital to meet the losses at these points. Little attention was paid to the expected loss (or range of losses) beyond these points.
- The models failed to recognise that, in times of financial crisis, correlations between risk events increase. Hence the models underestimated the *joint* probability of multiple extreme risk events happening at once.

Since the credit crunch, there has been increased focus on the techniques used to model extreme events.

So it is important to ensure that we model the form of the distribution in the tails correctly. However, the low frequency of these events also means that there is relatively little data to model their effects accurately.

Hence the irony of the situation – the tail data with which to fit the distribution is sparse, yet the distribution is being used to model tail risk events.

Many types of financial data tend to be much more narrowly peaked in the centre of the distribution and to have fatter tails than the normal distribution. This shape of distribution is known as leptokurtic. For example, when share prices are modelled, large price movements occur more frequently than predicted by the normal distribution. So the normal distribution may be unsuitable for modelling the large movements in the tails.

The word 'leptokurtic' is a measure of the kurtosis of a distribution, which is the fourth *standardised* central moment of a distribution:

$$\kappa = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

All (univariate) normal distributions have kurtosis equal to three and are described as mesokurtic. A distribution with kurtosis greater than three is leptokurtic (more peaked with fatter tails). A distribution with kurtosis less than three is platykurtic (a broader peak with more slender tails).

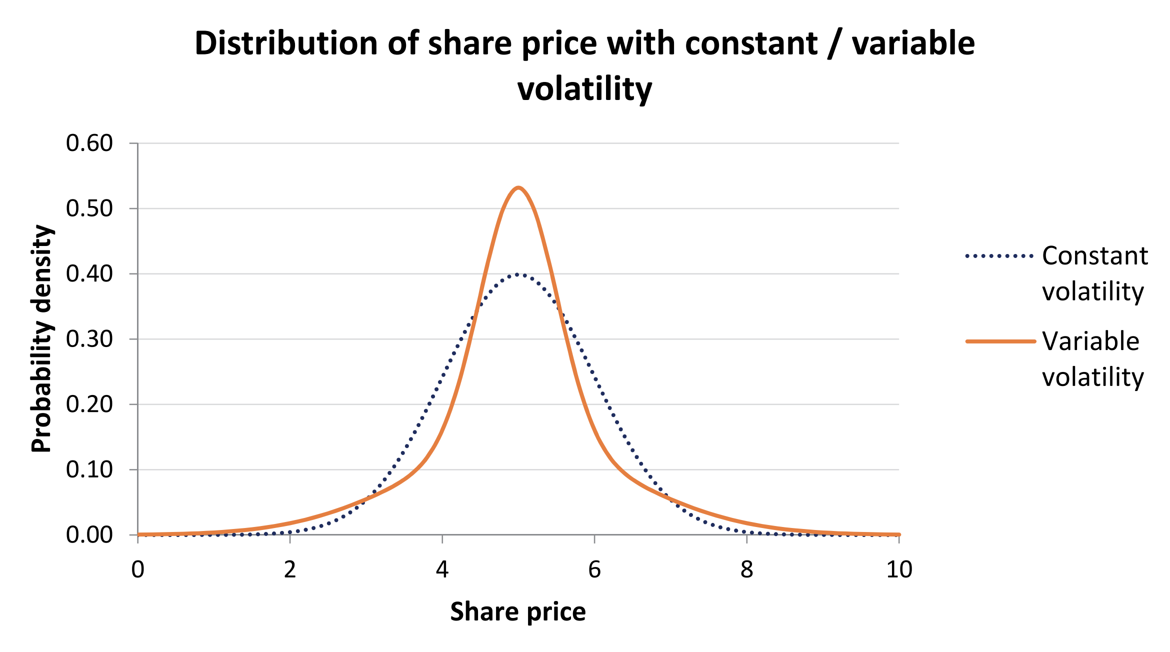
One reason for these fat tails is that the volatility of financial variables does not remain constant, but varies stochastically over time. This property is known as *heteroscedasticity*.

The volatility of a random variable is equivalent to its standard deviation.

Question

The graph below compares two distributions for the price of a share in one year's time:

- a $N(5,\sigma^2)$ distribution with constant volatility, $\sigma = 1$
- a $N(5,\sigma^2)$ distribution where the volatility is heteroscedastic, *ie* $\sigma = 0.5$ and $\sigma = 1.5$ with equal probability.



Comment on the relative shape of the graphs.

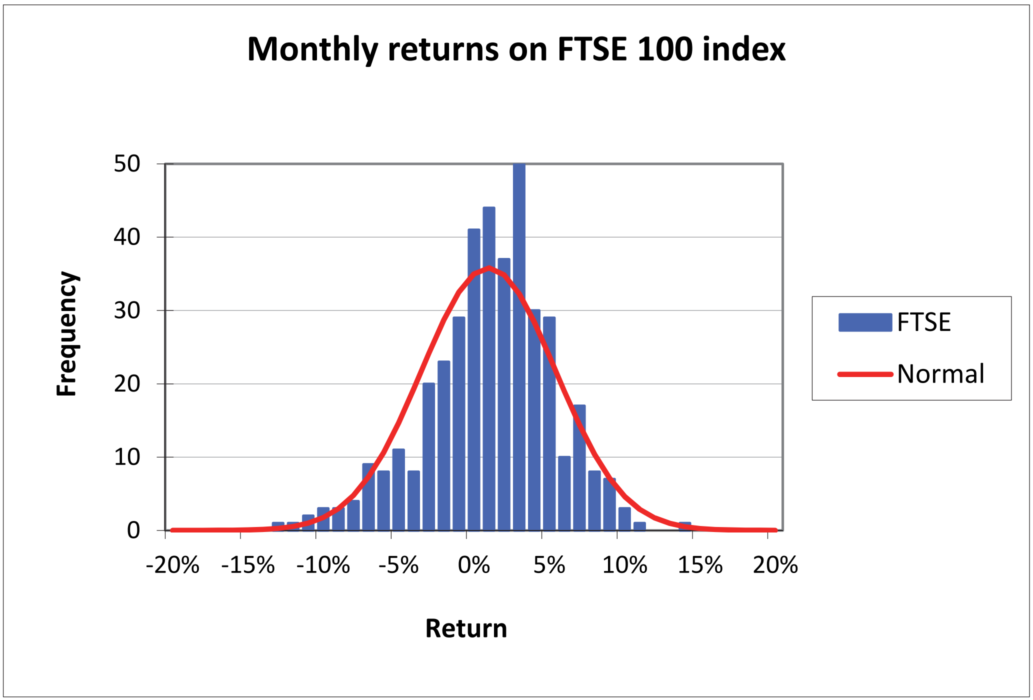
Solution

In the case where the volatility is variable, the resultant probability density function is more peaked with fatter tails, *ie* it is leptokurtic.

When empirical asset return data values are analysed, they exhibit volatility clustering, *ie* periods of sustained high volatility and periods of sustained low volatility. This suggests that volatility of asset return data is not constant but heteroscedastic.

Even if we select an appropriate form of fat-tailed distribution, if we attempt to fit the distribution using the whole of our dataset, this is unlikely to result in a good model for the tails, since the parameter estimates will be heavily influenced by the main bulk of the data in the central part of the distribution.

This is illustrated by the graph below, which shows the frequency of (log) returns on the FTSE-100 between April 1984 and September 2017. A normal distribution has been fitted to the whole dataset. It can be seen, from the graph, that the normal distribution underestimates the probability of the extreme events in the lower tail and overestimates extreme events in the upper tail.



1.4 Extreme value theory

Fortunately, better modelling of the tails of the data can be done through the application of extreme value theory. The key idea of extreme value theory is that the asymptotic behaviour of the tails of most distributions can be accurately described by certain families of distributions.

Question

Explain what is meant by the phrase 'asymptotic behaviour of the tails of a distribution' in the paragraph above.

Solution

The phrase is referring to how the distribution behaves in the limit, as a certain parameter (such as the number of observations in a sample) tends to infinity.

More specifically, the maximum values of a distribution (when appropriately standardised) and the values exceeding a specified threshold (called threshold exceedances) converge to two particular families of distributions as the sample size increases.

These two families of distributions are:

- generalised extreme value distributions, and
- generalised Pareto distributions.

We look at these in Sections 2 and 3 respectively.

2 Generalised extreme value (GEV) distribution

If we are dealing with losses that have typical sizes, *ie* ones whose values come from the central part of the distribution, we can make use of the Central Limit Theorem.

This tells us that, if we calculate the mean, $\bar{X}$, of a set of n values taken from a loss distribution that has mean μ and variance σ^2, the standardised value, $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, can be approximated using the standard normal distribution.

However, the most financially significant part of a loss distribution is usually the right-hand tail where the large losses occur. These are the *extreme values* of the distribution. So, is there a similar way to approximate the behaviour of the extreme values in the tail of the distribution?

The Core Reading here is referring to a distribution of losses (rather than returns or profits) and so we are concerned with the extreme right-hand tail of the loss distribution.

2.1 Maximum values

One approach is to look at $X_M = \max\{X_1, X_2, \ldots, X_n\}$, the maximum value in a set of n values. This is referred to as a *block maximum*.

Question

The dataset below shows the claim amounts in £000s in respect of a commercial property portfolio over a period of a year.

Claim number	*Claim amount*	*Claim number*	*Claim amount*	*Claim number*	*Claim amount*	*Claim number*	*Claim amount*
1	9	17	12	33	19	49	118
2	28	18	35	34	17	50	55
3	20	19	12	35	66	51	14
4	8	20	75	36	55	52	94
5	102	21	80	37	81	53	54
6	152	22	42	38	140	54	81
7	23	23	9	39	64	55	62
8	108	24	122	40	9	56	83
9	42	25	145	41	9	57	23
10	12	26	13	42	36	58	19
11	110	27	16	43	185	59	55
12	9	28	113	44	135	60	104
13	22	29	9	45	25		
14	37	30	8	46	16		
15	147	31	12	47	55		
16	128	32	84	48	31		

(i) Determine the values of X_M where the block size is:

(a) $n=5$

(b) $n=10$

(ii) Comment on the trade-off between the block size and the values of X_M that will be used to fit the extreme value distribution.

Solution

(i)(a) The values of X_M are $\{102, 152, 147, 128, 145, 113, 84, 140, 185, 118, 94, 104\}$.

(i)(b) The values of X_M are $\{152, 147, 145, 140, 185, 104\}$.

(ii) The larger the block size, the fewer the number of blocks (*eg* when $n=10$ there are six blocks whereas when $n=5$ there are twelve blocks). The fewer the number of blocks, the fewer (but more 'extreme') the values of X_M that will be used to fit the extreme value distribution.

If we look at a number of such blocks, we find that these maximum values can be standardised in a similar way, *ie* we can calculate expressions of the form $\frac{X_M - \alpha_n}{\beta_n}$ that can be approximated by a particular type of distribution – called an *extreme value distribution*.

As with the Central Limit Theorem, we are interested in determining the distribution of the *standardised* quantity:

$$\frac{X_M - \alpha_n}{\beta_n}$$

where $X_M = \max\{X_1, X_2, ..., X_n\}$ with each X_i representing an observed loss. The α_n and β_n are appropriately chosen constants. We will give an example of these in the next section.

2.2 Distribution of the (standardised) maximum values

In order to determine the distribution of this standardised quantity, we consider its CDF.

If the values are independent and identically distributed (IID), each with cumulative distribution function, $F(x)$, the cumulative distribution function of the block maximum is:

$$\begin{aligned} P(X_M \le x) &= P(X_1 \le x,\ X_2 \le x,\ \dots,\ X_n \le x) \\ &= P(X_1 \le x)P(X_2 \le x)\dots P(X_n \le x) \\ &= \left[P(X \le x)\right]^n \\ &= \left[F(x)\right]^n \end{aligned}$$

We can attempt to standardise the values of X_M by finding a sequence of constants $\alpha_1, \alpha_2, \ldots$ and $\beta_1, \beta_2, \ldots > 0$ so that the limiting distribution:

$$\lim_{n\to\infty} P\left(\frac{X_M - \alpha_n}{\beta_n} \le x\right) = \lim_{n\to\infty}\left[F\left(\beta_n x + \alpha_n\right)\right]^n$$

depends only on *x*.

The Extreme Value Theorem tells us that it is possible to find such values of α_n and β_n for most common distributions.

For example, if the individual losses are distributed exponentially with $F(x) = 1 - e^{-\lambda x}$, we can set $\alpha_n = \frac{1}{\lambda}\ln n$ and $\beta_n = \frac{1}{\lambda}$.

We don't need to concern ourselves with how these values for α_n and β_n are chosen for the exponential distribution, only that they can be chosen.

Question

Let $X \sim Exp(\lambda)$, $\alpha_n = \frac{1}{\lambda}\ln n$ and $\beta_n = \frac{1}{\lambda}$ for all n.

By substituting in for α_n and β_n and by using the CDF of the exponential distribution, determine the CDF of the limiting distribution of the *standardised* values of X_M:

$$\lim_{n\to\infty} P\left(\frac{X_M - \alpha_n}{\beta_n} \le x\right) = \lim_{n\to\infty}\left[F\left(\beta_n x + \alpha_n\right)\right]^n$$

(*Hint:* $\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x$.)

Solution

(i) The solution below forms part of the Core Reading.

$$\begin{aligned}
\lim_{n\to\infty}\left[F\left(\beta_n x + \alpha_n\right)\right]^n &= \lim_{n\to\infty}\left[F\left(\frac{1}{\lambda}x + \frac{1}{\lambda}\ln n\right)\right]^n \\
&= \lim_{n\to\infty}\left\{1 - \exp\left[-\lambda\left(\frac{1}{\lambda}x + \frac{1}{\lambda}\ln n\right)\right]\right\}^n \\
&= \lim_{n\to\infty}\left\{1 - \exp\left(-x - \ln n\right)\right\}^n \\
&= \lim_{n\to\infty}\left\{1 - \frac{e^{-x}}{n}\right\}^n = e^{-e^{-x}}
\end{aligned}$$

The last line in the Core Reading above follows from the hint:

$$\lim_{n\to\infty}\left\{1-\frac{e^{-x}}{n}\right\}^n=\lim_{n\to\infty}\left\{1+\left(\frac{-e^{-x}}{n}\right)\right\}^n=e^{-e^{-x}}$$

This distribution is known as the standard Gumbel distribution.

The standard Gumbel distribution is a particular type of extreme value distribution, which we consider in more detail later on.

2.3 GEV distribution

More generally, whatever the underlying distribution of the data, the distribution of the standardised maximum values will converge to a distribution called the *generalised extreme value* (GEV) distribution as *n* increases, *ie* $\lim\limits_{n\to\infty}\left[F(\beta_n x+\alpha_n)\right]^n=H(x)$.

In Section 2.2, we considered the specific case where the underlying distribution is $X \sim Exp(\lambda)$. However, the generalised extreme value distribution is more generic than this and caters for a number of different underlying distributions for X.

The cumulative distribution function of the GEV distribution is:

$$H(x)=\begin{cases}\exp\left(-\left(1+\dfrac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right) & \gamma\neq 0\\[2ex] \exp\left(-\exp\left(-\dfrac{(x-\alpha)}{\beta}\right)\right) & \gamma=0\end{cases}$$

This distribution has three parameters:

- **a location parameter α**
- **a scale parameter $\beta>0$**
- **a shape parameter γ.**

The parameters α and β just rescale (shift and stretch) the distribution. They are analogous to (but do not usually correspond to) the mean and standard deviation.

The parameter γ determines the overall shape of the distribution (analogous to the skewness) and its sign (positive, negative or zero) results in three different shaped distributions.

Question

Derive the PDF for the GEV distribution.

Solution

In the case where $\gamma \neq 0$:

$$H(x)=\exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right)$$

We must differentiate this expression to obtain the PDF. Let $u=\dfrac{\gamma(x-\alpha)}{\beta}$ and $v=(1+u)^{-\frac{1}{\gamma}}$ so that $H(x)=\exp(-v)$. Then:

$$\frac{du}{dx}=\frac{\gamma}{\beta}$$

$$\frac{dv}{du}=\left(-\frac{1}{\gamma}\right)(1+u)^{-\frac{1}{\gamma}-1}=-\frac{1}{\gamma}\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\left(1+\frac{1}{\gamma}\right)}$$

and $$\frac{dH(x)}{dv}=-\exp(-v)=-\exp\left(-(1+u)^{-\frac{1}{\gamma}}\right)=-\exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right).$$

So the PDF is:

$$\begin{aligned}
h(x)&=\frac{dH(x)}{dx}\\
&=\frac{dH(x)}{dv}\times\frac{dv}{du}\times\frac{du}{dx}\\
&=-\exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right)\times-\frac{1}{\gamma}\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\left(1+\frac{1}{\gamma}\right)}\times\frac{\gamma}{\beta}\\
&=\frac{1}{\beta}\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\left(1+\frac{1}{\gamma}\right)}\exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right)
\end{aligned}$$

In the case where $\gamma=0$:

$$H(x)=\exp\left(-\exp\left(-\frac{(x-\alpha)}{\beta}\right)\right)$$

Again, we must differentiate the above expression to obtain the PDF. Let $u=\dfrac{x-\alpha}{\beta}$ and $v=\exp(-u)$ so that $H(x)=\exp(-v)$. Then:

$$\frac{du}{dx}=\frac{1}{\beta}$$

$$\frac{dv}{du}=-\exp\left(-u\right)=-\exp\left(-\frac{\left(x-\alpha\right)}{\beta}\right)$$

and $$\frac{dH(x)}{dv}=-\exp\left(-v\right)=-\exp\left(-\exp\left(-u\right)\right)=-\exp\left(-\exp\left(-\frac{\left(x-\alpha\right)}{\beta}\right)\right).$$

So the PDF is:

$$\begin{aligned}h(x)&=\frac{dH(x)}{dx}\\&=\frac{dH(x)}{dv}\times\frac{dv}{du}\times\frac{du}{dx}\\&=-\exp\left(-\exp\left(-\frac{\left(x-\alpha\right)}{\beta}\right)\right)\times-\exp\left(-\frac{\left(x-\alpha\right)}{\beta}\right)\times\frac{1}{\beta}\\&=\frac{1}{\beta}\exp\left(-\left[\frac{\left(x-\alpha\right)}{\beta}+\exp\left(-\frac{\left(x-\alpha\right)}{\beta}\right)\right]\right)\end{aligned}$$

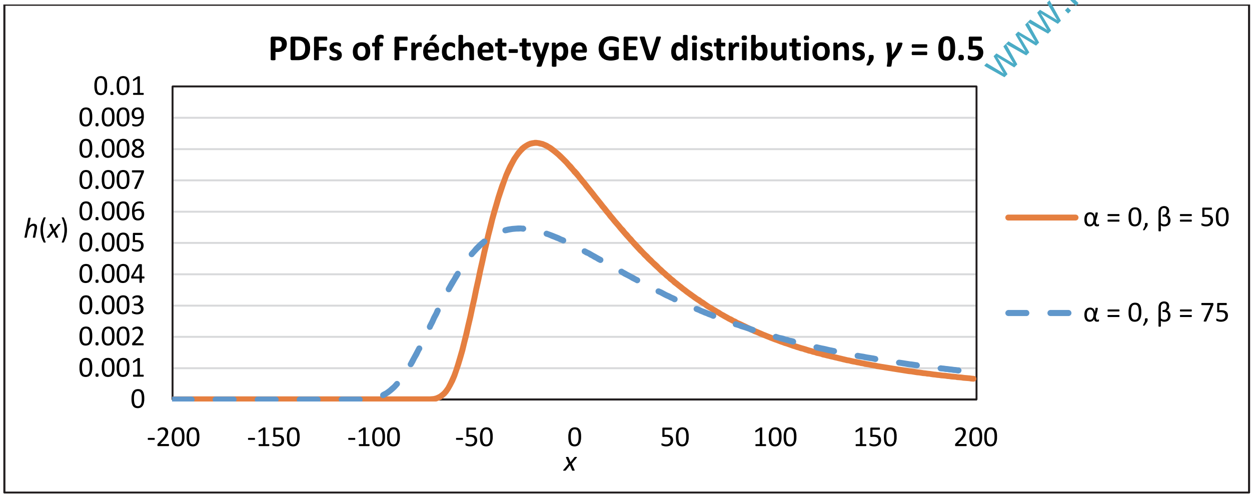
2.4 Fréchet, Weibull and Gumbel GEV distributions

The GEV family of distributions subdivides into three distinct classes depending on the value of the shape parameter, γ. These are called:

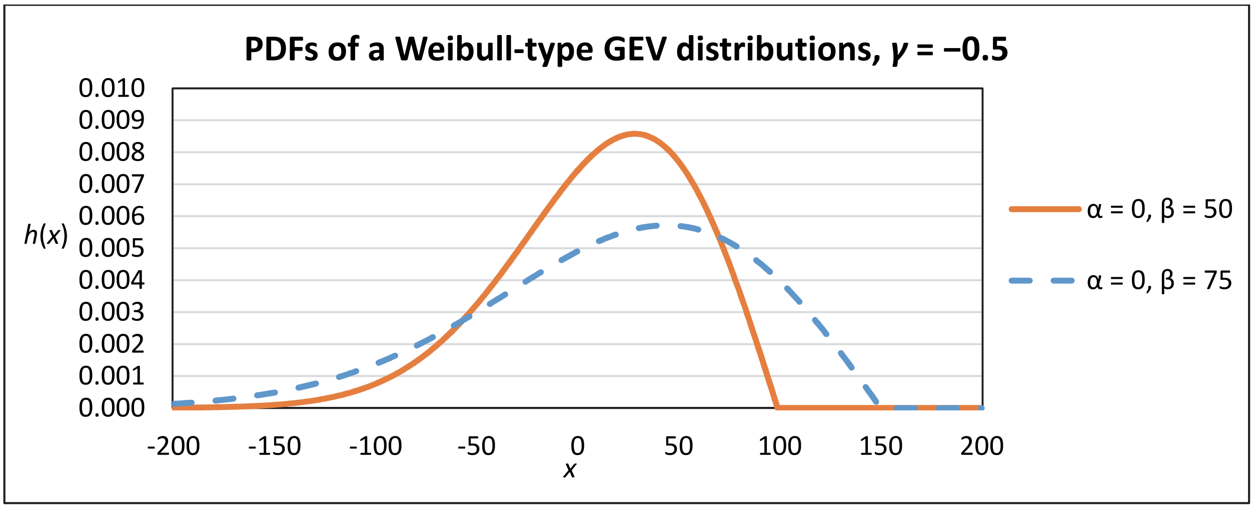
- Fréchet-type GEV distributions – when $\gamma>0$
- Weibull-type GEV distributions – when $\gamma<0$
- Gumbel-type GEV distributions – when $\gamma=0$.

Each type gives rise to a different shape of PDF as can be seen in the three graphs below.

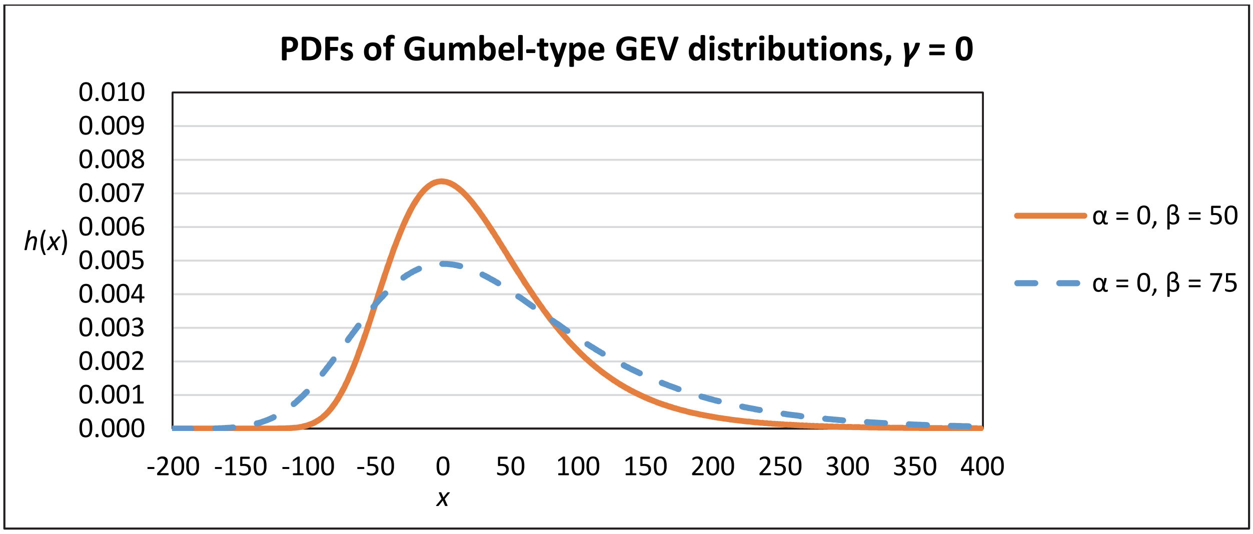
Fréchet-type GEV distributions



Weibull-type GEV distribution



Gumbel-type GEV distributions



Question

Comment briefly on the range of values that x takes for each of the Fréchet-type, Weibull-type and Gumbel-type GEV distributions, with reference to the three graphs above.

Solution

- When $\gamma > 0$, x is bounded below and when $\gamma < 0$, x is bounded above.
- When $\gamma = 0$, x is unbounded.

We will now explain algebraically how these bounds arise.

Fréchet-type GEV distribution

For $\gamma > 0$, the distribution is a Fréchet-type GEV distribution. Earlier, we derived the PDF as:

$$h(x) = \frac{1}{\beta}\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\left(1+\frac{1}{\gamma}\right)} \exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right)$$

In order for the PDF to be non-negative, we require the expression $\left(1+\dfrac{\gamma(x-\alpha)}{\beta}\right)$ to be positive.

(The other factors in the PDF are positive since $\beta > 0$ and the exponential function takes positive values only.)

Since $\gamma > 0$, this results in a lower bound on the values that x can take:

$$1+\frac{\gamma(x-\alpha)}{\beta} > 0 \;\Rightarrow\; x-\alpha > \frac{-\beta}{\gamma} \;\Rightarrow\; x > \alpha - \frac{\beta}{\gamma}$$

The Fréchet-type GEV distributions tend to be those most suitable for modelling extreme financial (loss) events. This is because there is no upper bound to the loss events but also because of the tail of the distribution. Fréchet-type GEV distributions have a heavier tail (*ie* a tail that decays more slowly to 0) than other types of GEV distribution. This is due to the behaviour of the PDF:

- The factor $\exp\left(-\left(1+\dfrac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right)$ tends to 1 as $x \to \infty$.
- The factor $\dfrac{1}{\beta}\left(1+\dfrac{\gamma(x-\alpha)}{\beta}\right)^{-\left(1+\frac{1}{\gamma}\right)}$ tends to 0 as $x \to \infty$. It decays in accordance with what is known as the negative power law, *ie* in proportion to x^{-k} where $k > 0$.

Weibull-type GEV distribution

For $\gamma < 0$, the distribution is a Weibull-type GEV distribution. The PDF is of the same form as the Fréchet-type GEV distribution, *ie* it is given by:

$$h(x) = \frac{1}{\beta}\left(1 + \frac{\gamma(x-\alpha)}{\beta}\right)^{-\left(1+\frac{1}{\gamma}\right)} \exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right)$$

Question

Determine a bound for values of x for the Weibull-type GEV distribution.

Solution

In order for PDF to be positive, we require the factor $\left(1+\dfrac{\gamma(x-\alpha)}{\beta}\right)$ to be positive. Since $\gamma < 0$, this results in an upper bound on the values that x can take:

$$1+\frac{\gamma(x-\alpha)}{\beta} > 0 \quad \Rightarrow \frac{\gamma(x-\alpha)}{\beta} > -1 \;\Rightarrow\; x-\alpha < \frac{-\beta}{\gamma} \;\Rightarrow\; x < \alpha - \frac{\beta}{\gamma}$$

We might expect to fit such a distribution to, for example, the ages of a human population (indicating an upper bound to possible age) or where a loss is certain not to exceed a certain value (for example, if such losses are reinsured).

Gumbel-type GEV distribution

When $\gamma = 0$, the GEV distribution reduces to the Gumbel distribution.

In this case, the PDF is given by:

$$h(x) = \frac{1}{\beta}\exp\left(-\left[\frac{(x-\alpha)}{\beta} + \exp\left(-\frac{(x-\alpha)}{\beta}\right)\right]\right)$$

This has a tail that decays exponentially. The decay is more rapid (*ie* the tail is lighter) than that for the Fréchet-type GEV distribution given the same values of α and β.

If $\alpha = 0$ and $\beta = 1$ this becomes the PDF of the standard Gumbel distribution (whose CDF we saw in the example in Section 2.2).

The standard Gumbel distribution is the extreme value distribution arising from an exponential distribution.

2.5 Choosing the form of the GEV distribution

If we know the form of the underlying distribution, it is possible to work out the limiting distribution of the maximum value. We can then use the appropriate member of the GEV family to model the tail of the distribution.

Note the distinction between the distribution that applies to the full dataset (the underlying distribution) and the distribution that we are using to model the extreme values (the GEV distribution). Different underlying distributions are closely related to different GEV distributions. The relationship depends on factors such as whether the underlying distribution has a finite end point (*ie* whether there is a lower or upper bound to the values that can be taken by the underlying random variable), and whether the tail of the PDF of the underlying distribution exhibits exponential or power decay.

The underlying distribution will determine which of the three different types of GEV distribution will arise, as indicated in the table below. The three types are distinguished by the sign of the shape parameter γ and are named after their original discoverers.

	GEV distributions (for the maximum value) corresponding to common loss distributions		
Type **Shape parameter**	**WEIBULL** $\gamma < 0$	**GUMBEL** $\gamma = 0$	**FRÉCHET** $\gamma > 0$
Underlying distribution	**Beta** **Uniform** **Triangular**	**Chi-square** **Exponential** **Gamma** **Lognormal** **Normal** **Weibull**	**Burr** ***F*** **Log-gamma*** **Pareto** ***t***
Range of values permitted	$x < \alpha - \dfrac{\beta}{\gamma}$	$-\infty < x < \infty$	$x > \alpha - \dfrac{\beta}{\gamma}$

* Note that $X \sim loggamma$ if $\ln X \sim Gamma$.

Unhelpfully, the extreme value distribution corresponding to the Weibull distribution from the *Tables* is actually of the Gumbel type (rather than the Weibull type).

Mathematicians have determined criteria that can be used to predict which family a particular distribution belongs to. As a rough guide:

- **underlying distributions that have finite upper limits (*eg* the uniform distribution) are of the Weibull type (which also has a finite upper limit).**
- **'light tail' distributions that have finite moments of all orders (*eg* exponential, normal, lognormal) are typically of the Gumbel type**
- **'heavy tail' distributions whose higher moments can be infinite are of the Fréchet type.**

For example, suppose we have monthly claim data stored in a matrix `data` with the first column `month` and the second column `claim`.

To calculate the block maxima for these claims using block sizes of 12 months, we would use the following R code:

```
block<-(data-1)%/%12+1
blockmax<-aggregate(claim~block,data,max)
```

We can plot a histogram of the block maxima using the `hist` function and an empirical density function using `density` in the `plot` function (if there is enough data). We can then superimpose a GEV distribution to see if it is a good approximation.

```
GEV <- function(z) {1/beta*(1+gamma*(z-alpha)/beta)^-
(1+1/gamma)*exp(-((1+gamma*(z-alpha)/beta)^(-1/gamma))) }
lines(<sequence of x values>,GEV(<sequence of x values))
```

The `qqplot` function is used to compare the sample data to simulated values from a fitted GEV model.

We can estimate the maximum likelihood values as we did in Chapter 15 by defining the log-likelihood function and using the function `nlm` on the negative value of the log-likelihood function as before.

Question

In the question in Section 2.1, the block maximum, X_M took the values:

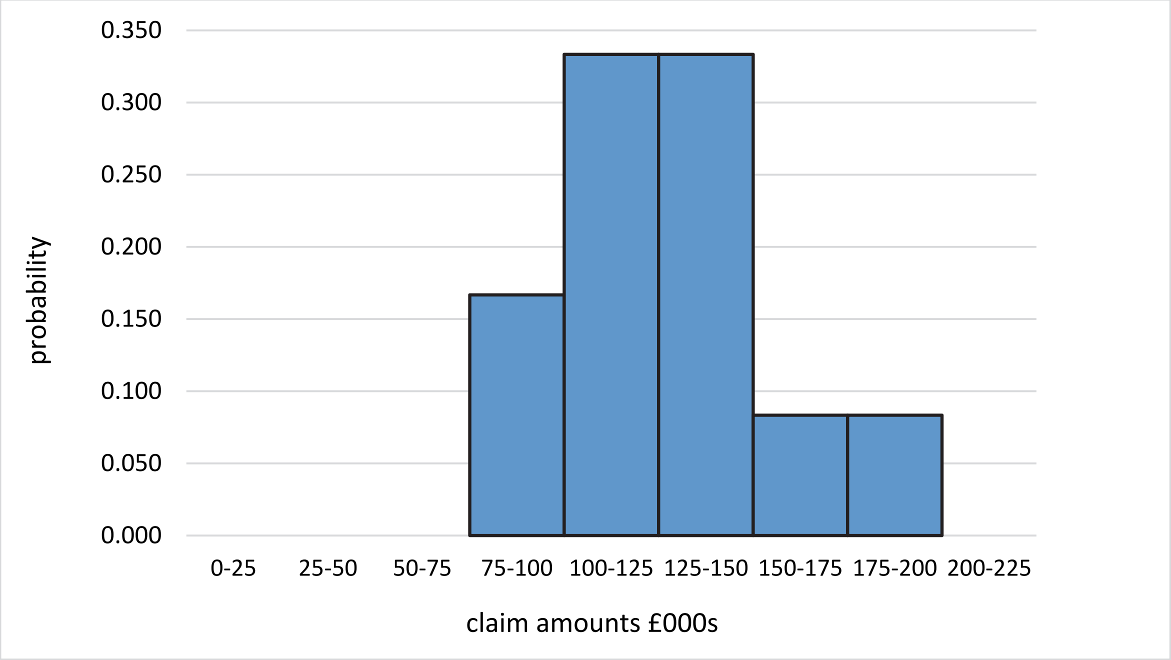
$$\{102,152,147,128,145,113,84,140,185,118,94,104\}$$

when the block size was 5.

Plot these points on a frequency diagram and suggest a type of GEV distribution that might be appropriate.

Solution

A frequency diagram representing the above block maximum data is given below:



The data set is too small to be able to suggest a type of GEV distribution with any confidence. However, the upper tail decay appears to be rapid, which might lead us to consider a Gumbel-type GEV distribution.

3 Generalised Pareto distribution (GPD)

3.1 Threshold exceedances

As an alternative to focusing on the maximum value, we can consider the distribution of all the values of the variable that exceed some (large) specified threshold, *eg* all claims exceeding £1 million. For large samples, the distribution of these extreme values converges to the *generalised Pareto distribution*. This enables us to model the tail of a distribution by selecting a suitably high threshold and then fitting a generalised Pareto distribution to the observed values in excess of that threshold.

For example, under excess of loss reinsurance, this distribution could be used to model the claim amounts above a retention limit, *ie* the amounts that will pass to the reinsurer. Excess of loss reinsurance is studied in detail in Chapter 18.

If we let X be a random variable with cumulative distribution function, F, then the excess over the threshold, u, is $X-u\,|\,X>u$.

Question

The dataset below shows the claim amounts in £000s in respect of a commercial property portfolio over a period of a year. (This is the dataset from the question in Section 2.1.)

Claim number	*Claim amount*
1	9
2	28
3	20
4	8
5	102
6	152
7	23
8	108
9	42
10	12
11	110
12	9
13	22
14	37
15	147
16	128
17	12
18	35
19	12
20	75
21	80
22	42
23	9
24	122
25	145
26	13
27	16
28	113
29	9
30	8
31	12
32	84
33	19
34	17
35	66
36	55
37	81
38	140
39	64
40	9
41	9
42	36
43	185
44	135
45	25
46	16
47	55
48	31
49	118
50	55
51	14
52	94
53	54
54	81
55	62
56	83
57	23
58	19
59	55
60	104

(i) Calculate the values of $X-u\,|\,X>u$ when:

(a) $u=100$

(b) $u=125$

(ii) Comment on the trade-off in the choice of the threshold, u.

Solution

(i)(a) The values of $X-100|X>100$ are $\{2,52,8,10,47,28,22,45,13,40,85,35,18,4\}$.

(i)(b) The values of $X-125|X>125$ are $\{27,22,3,20,15,60,10\}$.

(ii) The higher the value of the threshold, the more extreme the values of X. However, using a higher threshold means that we have fewer values with which to fit the extreme value distribution.

For example, suppose we have monthly claim data stored in a matrix `data` with the first column `month` and the second column `claim`.

To calculate the threshold exceedances, `xe`, for these claims, at the threshold `u` we would use the following R code:

```
x<-data[,-1]
xe<-x[x>u]-u
```

Sometimes, the value of the threshold, u, may be specified, *eg* if it is a reinsurance retention limit. Other times, we may need to make a judgement as to where the threshold should be. Typically we may choose the threshold to be say the 90th or 95th percentile of the underlying distribution. The choice of u also depends on there being a sufficient volume of data available above the selected threshold.

If the maximum possible value of X is $x_F \le \infty$, the cumulative distribution function of the excess is (for $0 \le x < x_F - u$):

$$
\begin{aligned}
F_u(x) = P(X-u \le x \mid X>u) &= \frac{P(X-u \le x, X>u)}{P(X>u)} \\
&= \frac{P(X \le x+u, X>u)}{P(X>u)} \\
&= \frac{P(X \le x+u) - P(X \le u)}{P(X>u)} \\
&= \frac{F(x+u)-F(u)}{1-F(u)}
\end{aligned}
$$

For example, if the individual losses are distributed exponentially with $F(x)=1-e^{-\lambda x}$, we have:

$$F_u(x) = \frac{\left\{1-e^{-\lambda(x+u)}\right\} - \left\{1-e^{-\lambda u}\right\}}{1-\left\{1-e^{-\lambda u}\right\}} = \frac{e^{-\lambda u} - e^{-\lambda(x+u)}}{e^{-\lambda u}} = 1-e^{-\lambda x}$$

So, in this case, the threshold exceedances follow the same exponential distribution as X , irrespective of the threshold applied.

This is the memoryless property of the exponential distribution.

Question

With reference to the question above, use the method of maximum likelihood to fit a distribution to the threshold exceedances when the underlying claims distribution is exponential and the threshold is chosen to be 100.

Solution

The values of $X-100\,|\,X>100$ are $\{2,52,8,10,47,28,22,45,13,40,85,35,18,4\}$.

The result in the Core Reading above tells us that, if $X \sim Exp(\lambda)$ then $W = X-u\,|\,X>u \sim Exp(\lambda)$.

The likelihood function is given by:

$$L(\lambda)=\prod_{i=1}^{n} f_W(w_i)=\prod_{i=1}^{n}\lambda\exp(-\lambda w_i)=\lambda^n\exp\left(-\lambda\sum_{i=1}^{n}w_i\right)$$

Using the data values above, we have $n=14$ and $\sum_{i=1}^{14}w_i=409$, so that:

$$L(\lambda)=\lambda^{14}\exp(-409\lambda)$$

Taking natural logs:

$$\ln L(\lambda)=14\ln\lambda-409\lambda$$

Differentiating with respect to λ:

$$\frac{d\ln L(\lambda)}{d\lambda}=\frac{14}{\lambda}-409$$

This is equal to 0 when:

$$\lambda=\frac{14}{409}$$

Differentiating a second time

$$\frac{d^2\ln L(\lambda)}{d\lambda^2}=-\frac{14}{\lambda^2}<0 \Rightarrow \text{max}$$

So the maximum likelihood estimate of λ is $\dfrac{14}{409}$ or 0.0342, and hence the fitted distribution for the threshold exceedances is $Exp(0.0342)$.

3.2 Generalised Pareto distribution

More generally we find that, whatever the underlying distribution of the data, the distribution of the threshold exceedances will converge to a *generalised Pareto distribution* as the threshold u increases, *ie* $\lim_{u\to\infty} F_u(x) = G(x)$.

The generalised Pareto distribution is a two-parameter distribution with CDF:

$$G(x) = \begin{cases} 1-\left(1+\dfrac{x}{\gamma\beta}\right)^{-\gamma} & \gamma \neq 0 \\ 1-\exp\left(-\dfrac{x}{\beta}\right) & \gamma = 0 \end{cases}$$

This distribution has two parameters:

- **a scale parameter $\beta > 0$**
- **a shape parameter γ.**

When $\gamma = 0$, this distribution reduces to the exponential distribution.

When $\gamma = 0$, the CDF is:

$$G(x) = 1-\exp\left(-\frac{x}{\beta}\right)$$

which is the CDF of the $Exp\left(\frac{1}{\beta}\right)$ distribution.

When $\gamma > 0$, the CDF is:

$$G(x) = 1-\left(1+\frac{x}{\gamma\beta}\right)^{-\gamma} = 1-\left(\frac{\gamma\beta+x}{\gamma\beta}\right)^{-\gamma} = 1-\left(\frac{\gamma\beta}{\gamma\beta+x}\right)^{\gamma}$$

which is the CDF of the $Pareto(\gamma, \gamma\beta)$ distribution.

Question

Derive the PDF for the generalised Pareto distribution.

Solution

In the case when $\gamma \neq 0$:

$$G(x) = 1 - \left(1 + \frac{x}{\gamma\beta}\right)^{-\gamma}$$

Let $v = 1 + \dfrac{x}{\gamma\beta}$ so that $G(x) = 1 - v^{-\gamma}$. Then:

$$\frac{dv}{dx} = \frac{1}{\gamma\beta} \quad \text{and} \quad \frac{dG(x)}{dv} = \gamma v^{-\gamma-1} = \gamma v^{-(\gamma+1)} = \gamma\left(1 + \frac{x}{\gamma\beta}\right)^{-(\gamma+1)}$$

So the PDF is:

$$\begin{aligned} g(x) &= \frac{dG(x)}{dx} \\ &= \frac{dG(x)}{dv} \times \frac{dv}{dx} \\ &= \gamma\left(1 + \frac{x}{\gamma\beta}\right)^{-\gamma-1} \times \frac{1}{\gamma\beta} \\ &= \frac{1}{\beta}\left(1 + \frac{x}{\gamma\beta}\right)^{-(\gamma+1)} \end{aligned}$$

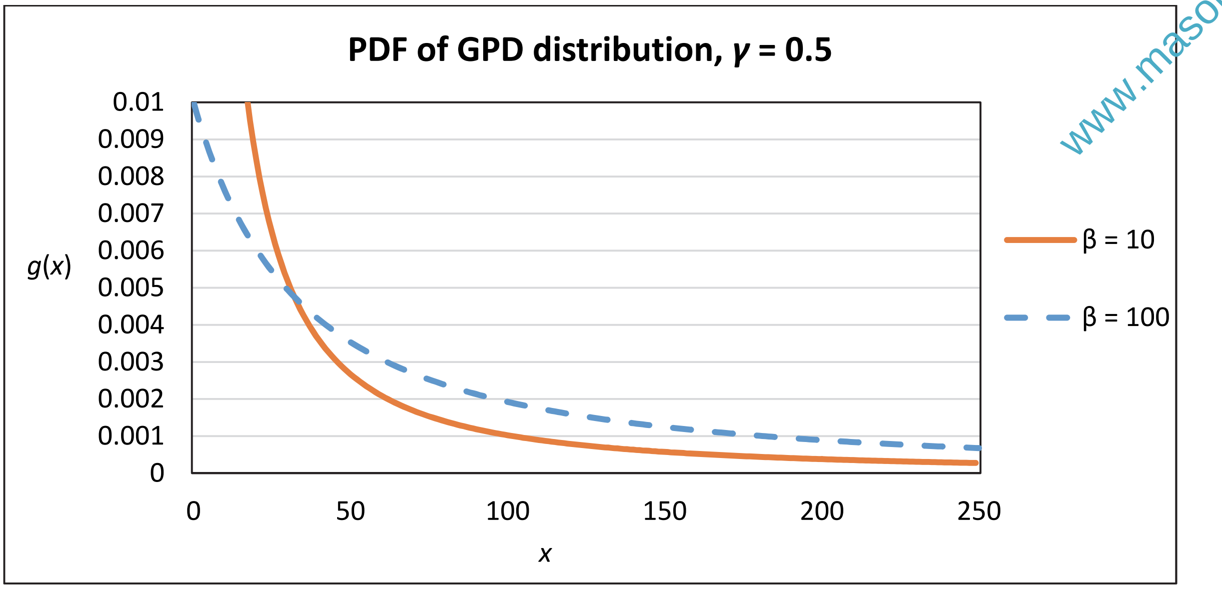
In the case when $\gamma = 0$:

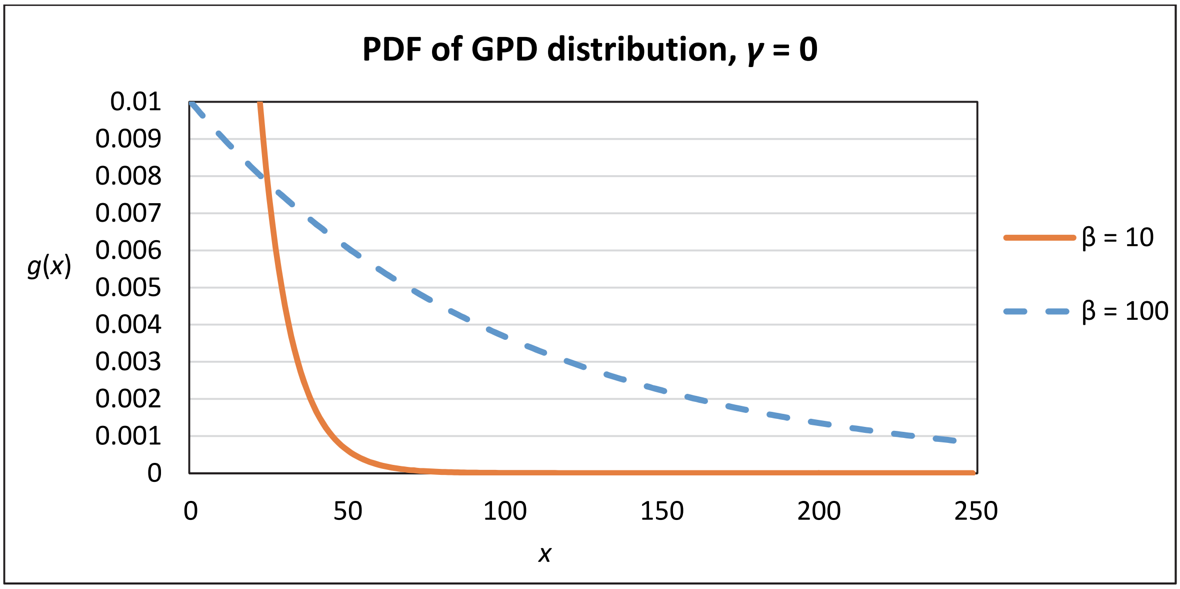
$$G(x) = 1 - \exp\left(-\frac{x}{\beta}\right)$$

Differentiating with respect to x, we see that the PDF is:

$$g(x) = \frac{1}{\beta}\exp\left(-\frac{x}{\beta}\right)$$

The two graphs below illustrate the PDF of the generalised Pareto distribution, using different values for the scale parameter, β , and the shape parameter, γ .


PDF of GPD distribution, γ = 0.5
g(x)
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
0
50
100
150
200
250
x
β = 10
β = 100


PDF of GPD distribution, γ = 0
g(x)
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
0
50
100
150
200
250
x
β = 10
β = 100

4 Measures of tail weight

There are a number of measures we can use to quantify the *tail weight* of a particular distribution, *ie* how likely very large values are to occur.

Depending on the context, an exponential, normal or lognormal distribution may be considered to be a suitable baseline to use for comparison.

Tail weight is a measure of how quickly the (upper) tail of a PDF tends to 0. If the PDF of random variable, X_A, tends to 0 as $x \to \infty$ more slowly than the PDF of random variable, X_B, then X_A is said to have a *heavier* tail than X_B. We will consider four ways of measuring tail weight:

1. the existence of moments
2. limiting density ratios
3. the hazard rate
4. the mean residual life.

4.1 Existence of moments

Recall that the *k*th moment of a continuous positive-valued distribution with density function $f(x)$ is:

$$\int_0^\infty x^k f(x)\,dx$$

The quantity that is being determined here is the *k*th non-central moment, $E(X^k)$.

In order for the *k*th moment to exist, then the integral expression above must converge (*ie* take a finite value).

For example, for the gamma distribution with density function:

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

the *k*th moment exists for all values of *k*, indicating that it has a relatively light tail.

Question

Derive a formula for $E(X^k)$ for $k = 1,2,3,\ldots$ when $X \sim Gamma(\alpha,\lambda)$.

Solution

If $X \sim Gamma(\alpha,\lambda)$, then:

$$E(X^k) = \int_0^\infty x^k \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\,dx = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{k+\alpha-1} e^{-\lambda x}\,dx$$

The integrand can be transformed into the PDF of a $Gamma(k+\alpha,\lambda)$ distribution by adjusting terms inside and outside of the integral as follows:

$$E(X^k)=\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)\lambda^k}\int_0^\infty \frac{\lambda^{k+\alpha}}{\Gamma(k+\alpha)}x^{k+\alpha-1}e^{-\lambda x}\,dx$$

Since the integral of a PDF over all values of x is equal to 1, the expression reduces to:

$$E(X^k)=\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)\lambda^k}\quad\text{for } k=1,2,3,\ldots$$

(This expression is given on page 12 of the *Tables*.)

The moments can also be obtained from the moment generating function by differentiation.

However, for some distributions, the value of the *k*th moment does not exist beyond a certain value of *k* (*ie* its value becomes infinite).

For the Pareto distribution with density function:

$$f(x)=\frac{\alpha\lambda^\alpha}{(\lambda+x)^{\alpha+1}}$$

the *k*th moment only exists when $k<\alpha$.

So a Pareto distribution (with a low value of the parameter α) will have a much thicker tail.

For $X\sim Pa(\alpha,\lambda)$, the mean and variance are given in the *Tables* as follows:

$$E(X)=\frac{\lambda}{\alpha-1}\quad\text{and}\quad \text{var}(X)=\frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)}$$

From the denominators of these two expressions, we see that the mean is defined only for $\alpha>1$ and the variance is defined only for $\alpha>2$.

4.2 Limiting density ratios

We can compare the thickness of the tail of two distributions by calculating the relative values of their density functions at the far end of the upper tail. For example, if we compare the Pareto distributions with parameters $\alpha=2$ and $\alpha=3$ (both with the same value of λ), we find that:

$$\lim_{x\to\infty}\frac{f_{\alpha=2}(x)}{f_{\alpha=3}(x)}=\lim_{x\to\infty}\left\{\frac{2\lambda^2}{(\lambda+x)^3}\bigg/\frac{3\lambda^3}{(\lambda+x)^4}\right\}=\frac{2}{3\lambda}\lim_{x\to\infty}(\lambda+x)=\infty$$

This confirms that the distribution with $\alpha=2$ has a much thicker tail.

If we compare the gamma distribution with the Pareto distribution, we find that the presence of the exponential factor in the gamma density results in a limiting density ratio of zero, which confirms that the gamma distribution has a lighter tail.

Question

Consider the *Gamma*$(0.5, 0.005)$ and *Pa*$(4, 300)$ distributions, both of which have a mean of 100 and a variance of 20,000.

Use the 'limiting density ratio' method to compare these two distributions.

Solution

Comparing the two density functions and taking the limit as $x \to \infty$, we have:

$$\lim_{x\to\infty} \frac{f_{gamma}(x)}{f_{Pareto}(x)} = \lim_{x\to\infty} \left\{ \left(\frac{0.005^{0.5}}{\Gamma(0.5)} x^{-0.5} e^{-0.005x} \right) \Big/ \left(\frac{4 \times 300^4}{(300+x)^5} \right) \right\}$$

$$= \lim_{x\to\infty} \left\{ C\, x^{-0.5} (300+x)^5 e^{-0.005x} \right\}$$

In the above expression, C is a (very small) constant. (We know that $\Gamma(0.5) = \sqrt{\pi}$ from the properties of the gamma function given on page 5 of the *Tables*.)

As $x \to \infty$, the $e^{-0.005x}$ factor tends to 0, the $x^{-0.5}$ factor tends to 0 and the $(300+x)^5$ factor tends to infinity, and so the overall expression tends to 0. (To see, this, try some large values of x.) Therefore:

$$\lim_{x\to\infty} \frac{f_{gamma}(x)}{f_{Pareto}(x)} = 0$$

This suggests that the gamma distribution has a lighter tail than the Pareto distribution.

We can obtain the values of the PDF of two distributions `X1` **and** `X2` **for, say,** x **values 1 to 1,000 and then calculate the ratio,** R**, using:**

```
R = X1/X2
```

We can then plot the graph of R **against** x **to determine which of** `X1` **and** `X2` **has the thicker tail.**

4.3 Hazard rate

The *hazard rate* of a distribution with density function $f(x)$ and distribution function $F(x)$ is defined as:

$$h(x) = \frac{f(x)}{1 - F(x)}$$

We have already seen this formula in Chapter 8. Recall that the hazard rate is the rate of failure given survival up until that point.

Question

(i) Determine the hazard rate for:

(a) the $Exp(\lambda)$ distribution

(b) the $Pa(\alpha,\lambda)$ distribution.

(ii) Comment on the differences between these hazard rates.

Solution

(i)(a) The hazard rate for the exponential distribution is:

$$h(x)=\frac{f(x)}{1-F(x)}=\frac{\lambda e^{-\lambda x}}{e^{-\lambda x}}=\lambda$$

(i)(b) The hazard rate for the Pareto distribution is:

$$h(x)=\frac{f(x)}{1-F(x)}=\frac{\alpha\lambda^{\alpha}}{(\lambda+x)^{\alpha+1}}\Bigg/\left(\frac{\lambda}{\lambda+x}\right)^{\alpha}=\frac{\alpha}{\lambda+x}$$

(ii) The exponential hazard rate is constant (*ie* it is independent of x, which demonstrates the memoryless property of the exponential distribution), whereas the Pareto hazard rate is a decreasing function of x.

We can interpret the hazard rate by analogy with μ_x, the force of mortality at age x.

The force of mortality has been discussed in detail in earlier chapters.

(The force of mortality at age x ($0 \le x < \omega$) is defined as:

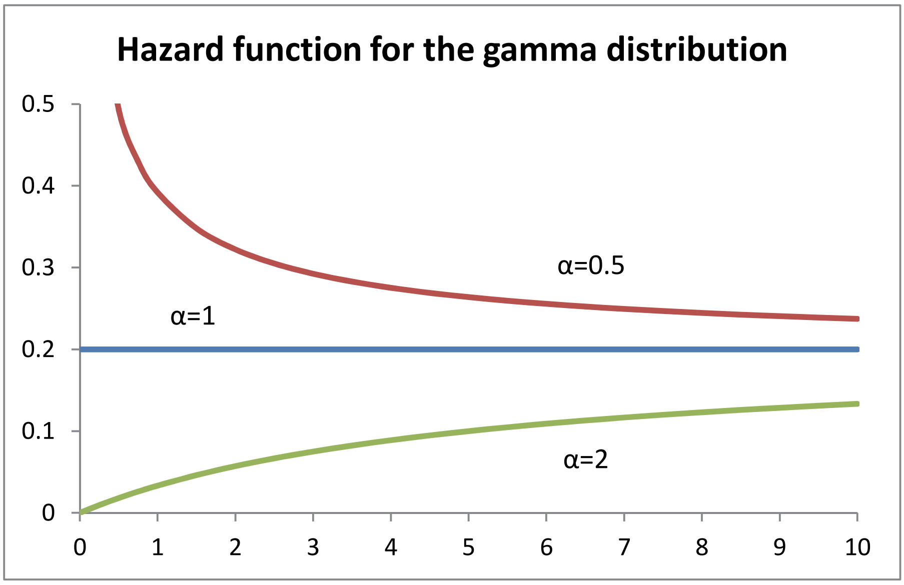
$$\mu_x = \lim_{h\to 0^+}\frac{1}{h}\times P[T \le x+h \mid T > x]$$

where T is a random variable measuring a person's length of life. See Chapter 6, Section 1.3.)

If the force of mortality increases as a person's age increases, relatively few people will live to old age (corresponding to a light tail). If, on the other hand, the force of mortality decreases as the person's age increases, there is the potential to live to a very old age (corresponding to a heavier tail).

For the gamma distribution we find that, if $\alpha=1$ (*ie* it is an exponential distribution), the hazard rate is constant, but if $\alpha<1$, it is a decreasing function of x (indicating a heavier tail than the exponential distribution) and if $\alpha>1$, it is an increasing function (indicating a lighter tail than the exponential distribution).

The graph below illustrates the hazard function for the $Gamma(\alpha,\lambda)$ distribution for different values of α and $\lambda = 0.2$. The axes of the graph are the hazard function $h(x) = \dfrac{f(x)}{1-F(x)}$ against x.



The formula for hazard function in each of the three cases is:

- $$h(x) = \frac{f(x)}{1-F(x)} = \frac{\frac{\lambda^2}{\Gamma(2)}x^{2-1}e^{-\lambda x}}{1-\int_0^x \frac{\lambda^2}{\Gamma(2)}t^{2-1}e^{-\lambda t}dt} = \frac{\lambda^2 x e^{-\lambda x}}{1-\int_0^x \lambda^2 t e^{-\lambda t}dt} \text{ when } \alpha = 2$$

- $$h(x) = \frac{f(x)}{1-F(x)} = \frac{\frac{\lambda}{\Gamma(1)}x^{1-1}e^{-\lambda x}}{1-\int_0^x \frac{\lambda}{\Gamma(1)}t^{1-1}e^{-\lambda t}dt} = \frac{\lambda e^{-\lambda x}}{1-\int_0^x \lambda e^{-\lambda t}dt} \text{ when } \alpha = 1$$

- $$h(x) = \frac{f(x)}{1-F(x)} = \frac{\frac{\lambda^{0.5}}{\Gamma(0.5)}x^{0.5-1}e^{-\lambda x}}{1-\int_0^x \frac{\lambda^{0.5}}{\Gamma(0.5)}t^{0.5-1}e^{-\lambda t}dt} = \frac{\frac{\lambda^{0.5}}{\sqrt{\pi}}x^{-0.5}e^{-\lambda x}}{1-\int_0^x \frac{\lambda^{0.5}}{\sqrt{\pi}}t^{-0.5}e^{-\lambda t}dt} \text{ when } \alpha = 0.5\,.$$

For the Pareto distribution, we find that the hazard rate is always a decreasing function of x, confirming that it has a heavy tail.

We derived these results for the exponential and Pareto distributions in the previous question.

The R code to calculate the hazard rate, H, for a Weibull distribution with parameters g and $b = c^{-1/g}$ is given by:

```
H<-dweibull(x,g,b)/(1-pweibull(x,g,b))
```

We can then plot the graph of H against x to determine the thickness of its tail.

4.4 Mean residual life

The *mean residual life* of a distribution with density function $f(x)$ and distribution function $F(x)$ is defined as:

$$e(x) = \frac{\int_x^\infty (y-x)f(y)dy}{\int_x^\infty f(y)dy} = \frac{\int_x^\infty \{1-F(y)\}dy}{1-F(x)}$$

This function gives the expected remaining survival time given survival up until this point.

Question

(i) Determine the mean residual life for:

(a) the $Exp(\lambda)$ distribution

(b) the $Pa(\alpha,\lambda)$ distribution where $\alpha > 1$.

(ii) Comment on the behaviour of these functions.

Solution

(i)(a) The mean residual life for the exponential distribution is:

$$e(x) = \frac{\int_x^\infty \{1-F(y)\}dy}{1-F(x)} = \frac{\int_x^\infty e^{-\lambda y}\,dy}{e^{-\lambda x}}$$

$$= \frac{\left[-\frac{1}{\lambda}e^{-\lambda y}\right]_x^\infty}{e^{-\lambda x}} = \frac{1}{\lambda}$$

(i)(a) The mean residual life for the Pareto distribution with $\alpha > 1$ is:

$$e(x) = \frac{\int_x^\infty \{1-F(y)\}dy}{1-F(x)} = \frac{\int_x^\infty \left(\frac{\lambda}{\lambda+y}\right)^\alpha dy}{\left(\frac{\lambda}{\lambda+x}\right)^\alpha} = (\lambda+x)^\alpha \int_x^\infty (\lambda+y)^{-\alpha}\,dy$$

$$= (\lambda+x)^\alpha \left[\frac{1}{-\alpha+1}(\lambda+y)^{-\alpha+1}\right]_x^\infty = (\lambda+x)^\alpha \left[0-\left(\frac{1}{-\alpha+1}\right)(\lambda+x)^{-\alpha+1}\right]$$

$$= \frac{\lambda+x}{\alpha-1}$$

(ii) The exponential mean residual life is constant whereas the Pareto mean residual life is an increasing function of x.

When $x = 0$ the mean residual life is equal to the mean of the underlying distribution, as we would expect.

Again we can interpret this in terms of mortality as the expected future lifetime, $\overset{\circ}{e}_x$.

We discussed expected future lifetimes in Chapter 6. There we derived the formula:

$$\overset{\circ}{e}_x = \int_0^\infty {}_t p_x \, dt$$

However since:

$${}_t p_x = \frac{{}_{t+x}p_0}{{}_x p_0} = \frac{S(t+x)}{S(x)} = \frac{1-F(t+x)}{1-F(x)}$$

we have:

$$\overset{\circ}{e}_x = \frac{\int_0^\infty \{1-F(t+x)\}dt}{1-F(x)}$$

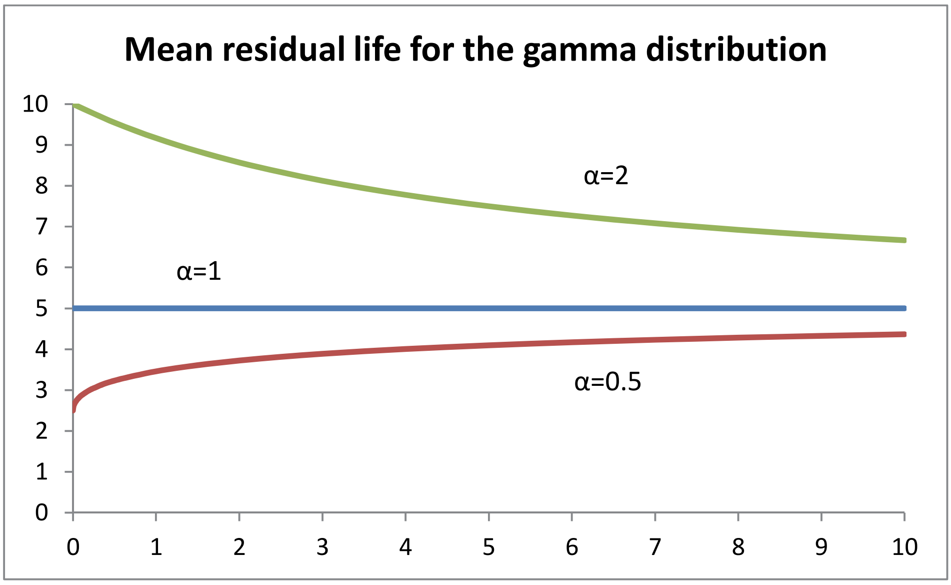
Then making the substitution $y = t + x$, we see that:

$$\overset{\circ}{e}_x = \frac{\int_x^\infty \{1-F(y)\}dy}{1-F(x)} = e(x)$$

If the expected future lifetime decreases with age, relatively few people will live to old age (corresponding to a light tail), but if it increases, there is the potential to live to a very old age (corresponding to a heavier tail).

For the gamma distribution we find that, if $\alpha = 1$ (*ie* it is an exponential distribution), the mean residual life is constant, but if $\alpha < 1$, it is an increasing function of x (indicating a heavier tail than the exponential distribution) and if $\alpha > 1$, it is a decreasing function (indicating a lighter tail than the exponential distribution).

The graph below illustrates the mean residual life for the $Gamma(\alpha, \lambda)$ distribution for different values of α and $\lambda = 0.2$. The axes of the graph are the mean residual life $e(x) = \dfrac{\int_x^\infty \{1-F(y)\} dy}{1-F(x)}$ against x.



For the Pareto distribution, we find that the mean residual life is always an increasing function of *x*, confirming that it has a heavy tail.

We derived these results for the exponential and Pareto distributions in the previous question.

The R code for the survival function of a Weibull distribution with parameters g and $b = c^{-1/g}$ is given by:

```
Sy<-function(y) {(1-pweibull(y,g,b))}
```

Hence, the mean residual life for `x` **is given by** `ex` **as follows:**

```
int<-integrate(Sy,x,Inf)
ex<-int$value/(1-pweibull(x,g,b))
```

We can then plot the graph of `ex` **against** `x`**.**

Chapter 16 Summary

Extreme events

An extreme event is one with a very low frequency and very high severity.

Modelling extreme financial events is difficult due to:

- lack of historic data in the tails of the distribution
- the 'true' distribution of many types of financial data being more leptokurtic (more peaked with fatter tails) than the normal distribution
- the volatility of financial variables being heteroscedastic (varying over time)
- parameter estimates being inappropriately influenced by the main bulk of the data in the middle of the distribution.

Better modelling of the tails of the data can be done using extreme value theory.

Extreme value theory

Extreme value theory attempts to model the asymptotic behaviour of the tails of distributions. There are two main approaches:

1. modelling the maximum values of a distribution – using the generalised extreme value family of distributions
2. modelling the values exceeding a threshold – using the generalised Pareto family of distributions.

Generalised extreme value distributions

Let:

- losses X_i be IID with cumulative distribution $F(x_i)$
- $X_M = \max\{X_1, X_2, \dots, X_n\}$ be the block maxima
- $\alpha_1, \dots, \alpha_n$ and $\beta_1, \dots, \beta_n > 0$ be suitable sequences of real constants.

Then, if n is sufficiently large, the distribution of the standardised block maxima, $\dfrac{X_M - \alpha_n}{\beta_n}$, is asymptotically described by the generalised extreme value (GEV) family of distributions with CDF:

$$H(x) = \lim_{n \to \infty} P\left(\frac{X_M - \alpha_n}{\beta_n} \le x\right) = \lim_{n \to \infty} \left[F(\beta_n x + \alpha_n)\right]^n$$

The cumulative distribution function of the GEV distribution is:

$$H(x)=\begin{cases}\exp\left(-\left(1+\dfrac{\gamma(x-\alpha)}{\beta}\right)^{-1/\gamma}\right) & \gamma\neq 0\\[2ex] \exp\left(-\exp\left(-\dfrac{(x-\alpha)}{\beta}\right)\right) & \gamma=0\end{cases}$$

The three parameters of the GEV family are:

- a location parameter, α
- a scale parameter, $\beta > 0$
- a shape parameter, γ .

There are three types of GEV distributions, which are distinguished by the sign of the shape parameter γ. Each type corresponds to different underlying loss distributions:

	GEV distributions (for the maximum value) corresponding to common loss distributions		
Type Shape parameter	WEIBULL $\gamma<0$	GUMBEL $\gamma=0$	FRÉCHET $\gamma>0$
Underlying distribution	Beta Uniform Triangular	Chi-square Exponential Gamma Lognormal Normal Weibull	Burr *F* Log-gamma Pareto *t*
Range of values permitted	$x<\alpha-\dfrac{\beta}{\gamma}$	$-\infty<x<\infty$	$x>\alpha-\dfrac{\beta}{\gamma}$

Generalised Pareto distributions

Let losses X_i be IID with cumulative distribution $F(x_i)$. Then the distribution of the conditional losses above a threshold, u, will converge (whatever the underlying distribution of the data) to a generalised Pareto distribution (GPD) with CDF:

$$G(x) = \lim_{u\to\infty} F_u(x) = \lim_{u\to\infty} P\left(X-u \le x \middle| X>u\right) = \lim_{u\to\infty} \frac{F(x+u)-F(u)}{1-F(u)}$$

The CDF of the GPD is of the form:

$$G(x) = \begin{cases} 1-\left(1+\dfrac{x}{\gamma\beta}\right)^{-\gamma} & \gamma \ne 0 \\ 1-\exp\left(-\dfrac{x}{\beta}\right) & \gamma = 0 \end{cases}$$

The two parameters of the GPD family are:

- a scale parameter, $\beta > 0$
- a shape parameter, γ.

Measures of tail weight

There are a number of measures we can use to quantify the tail weight of a particular distribution, *ie* how likely very large values are to occur:

- the existence of moments
  - The existence of all moments, $E(X^k)$, for all positive integers, k, indicates a light tail. If moments exist only up to a positive integer, k, this is an indication that the distribution has a heavy tail.
- limiting density ratios
  - The limiting value, as $x\to\infty$, of the ratio of two PDFs can be used to determine which distribution has the lighter or heavier tail.
- the hazard rate
  - An increasing (decreasing) hazard rate, $h(x)=\dfrac{f(x)}{1-F(x)}$, corresponds to a lighter (heavier) tail.
- the mean residual life
  - An increasing (decreasing) mean residual life, $e(x)=\dfrac{\int_x^\infty \{1-F(y)\}\,dy}{1-F(x)}$, corresponds to a heavier (lighter) tail.

The practice questions start on the next page so that you can keep the chapter summaries together for revision purposes.

Chapter 16 Practice Questions

16.1 (i) Explain what is meant by an extreme event and give two examples in an insurance context.

(ii) Explain why it is important to model extreme events separately from other events.

16.2 (i) Describe the generalised extreme value (GEV) distribution. [7]

Exam style

(ii) Outline an alternative approach that can be used in place of the GEV distribution to model extreme events. [3]

(iii) State the key advantage of the method outlined in (ii) over that described in (i). [1]

[Total 11]

16.3 The claim amounts in a general insurance portfolio are independent and follow an exponential distribution with mean £2,500.

Exam style

(i) Calculate the probability that an individual claim will exceed £10,000. [1]

(ii) Calculate the probability that, in a sample of 100 claims, the largest claim will exceed £10,000 using:

(a) an exact method

(b) an approximation based on a Gumbel-type GEV distribution. [5]

You are given that, for an exponential distribution with parameter λ, the approximate distribution of $\max\{X_1,\dots,X_n\}$ for large n is a Gumbel-type GEV distribution with CDF:

$$H(x)=\exp\left(-\exp\left(-\left(\frac{x-\alpha_n}{\beta_n}\right)\right)\right) \quad \textit{where } \alpha_n=\frac{1}{\lambda}\ln n \textit{ and } \beta_n=\frac{1}{\lambda}$$

(iii) State the two key assumptions made in (ii)(a). [1]

[Total 7]

16.4 If individual losses, X, follow a $Pa(\alpha,\lambda)$ distribution, determine the distribution of the threshold exceedances, $W=X-u\,|\,X<u$.

16.5 Compare the limiting value of the density functions for a $Gamma(\alpha,\lambda)$ and an $Exp(\lambda)$ distribution when $\alpha>1$ and hence determine which has the heavier tail.

16.6 (i) Determine the hazard rate for the Weibull distribution with parameters $c>0$ and $\gamma>0$.

(ii) Comment on the behaviour of the hazard rate.

16.7 Exam style

(i) Show that:

$$\int_x^{\infty} e^{-3y^{\frac{1}{2}}}\,dy = \frac{2}{9}P\left(\chi_4^2 > 6x^{\frac{1}{2}}\right)$$ [5]

Hint: use the substitution $u = 3y^{\frac{1}{2}}$ *and transform the integrand into the PDF of the* $Gamma(2,1)$ *distribution.*

(ii) Hence deduce an expression involving a chi-squared probability for the mean residual life for the $W\left(3,\frac{1}{2}\right)$ distribution. [2]

(iii) By calculating the values of mean residual life function when $x=1$, $x=4$ and $x=9$, determine whether the mean residual life of the $W\left(3,\frac{1}{2}\right)$ distribution is an increasing or decreasing function of x. [2]

[Total 9]

Chapter 16 Solutions

16.1 (i) ***Extreme events***

Extreme events are outcomes that have a very low probability of occurrence but involve very large sums of money.

In an insurance context, they may arise as a result of a single cause that has a high financial cost (*eg* a bodily injury claim or complete destruction of a building) …

… or as an accumulation of events with a related cause (*eg* flood damage to a large number of houses in one town).

(ii) ***Why extreme events are modelled separately***

The majority of risk events fall within the main body of the fitted distribution and can usually be modelled reasonably accurately by one of the standard statistical distributions.

However, there is usually a lack of past data on extreme events.

If a distribution is fitted to the whole dataset, the parameter estimates will reflect where the bulk of the data values lie rather than the extreme events. This might mean the fitted distribution understates the probability of future extreme events.

Therefore, a different approach to modelling extreme events is taken, *eg* by considering the distribution of block maxima or the distribution of threshold exceedances.

16.2 (i) ***Describe the GEV distribution***

The maximum value, X_M, in a sample of n IID random variables $X_1, X_2, \ldots, X_n$ tends to a particular distribution as the sample size increases. This is called the generalised extreme value (GEV) distribution. [1]

The GEV distribution has CDF:

$$H(x) = \begin{cases} \exp\left(-\left(1+\dfrac{\gamma(x-\alpha)}{\beta}\right)^{-1/\gamma}\right) & \gamma \neq 0 \\ \exp\left(-\exp\left(-\dfrac{(x-\alpha)}{\beta}\right)\right) & \gamma = 0 \end{cases}$$ [1]

The key parameter is the shape parameter, γ. [½]

When $\gamma > 0$, we have the Fréchet-type GEV distribution. [½]

This is the limiting form for ‘heavy tail’ underlying distributions with a finite lower bound, such as the Pareto distribution. [1]

When $\gamma < 0$, we have the Weibull-type GEV distribution. [½]

This is the limiting form for underlying distributions with a finite upper bound, such as the uniform distribution. [1]

When $\gamma = 0$, we have the Gumbel-type GEV distribution. [½]

This is the limiting form for most other underlying distributions that have finite moments, such as the normal and lognormal distributions. [½]

The parameters α and β are the location and scale parameters, respectively. These will differ depending on the underlying distribution. [½]

(ii) ***Alternative approach***

As an alternative to focusing upon a single maximum value, we can consider the distribution of *all* the claim values that exceed some threshold, u. The distribution of $X - u \mid X < u$ is called the threshold exceedances distribution. [1]

A similar theory to GEV predicts that the limiting distribution, as $u \to \infty$, is a generalised Pareto distribution (GPD). [½]

The GPD has CDF:

$$G(x) = \begin{cases} 1 - \left(1 + \dfrac{x}{\gamma\beta}\right)^{-\gamma} & \gamma \neq 0 \\ 1 - \exp\left(-\dfrac{x}{\beta}\right) & \gamma = 0 \end{cases} \qquad [1]$$

The parameters β and γ are the scale and shape parameters, respectively.

In order to fit the tail of a distribution we need to select a suitably high threshold and then fit the GPD to the values in excess of that threshold. [½]

(iii) ***Key advantage of the GPD method***

The GPD method has the advantage that it uses a larger part of the data and models *all* the large claims above the threshold, not just the single highest value. [1]

16.3 (i) ***Probability of a claim greater than £10,000 using an exponential distribution***

The claims distribution is $Exp\left(\dfrac{1}{2,500}\right)$.

Using the formula for CDF of an exponential random variable given on page 11 of the *Tables*, we have:

$$P(X > 10,000) = 1 - P(X \le 10,000) = 1 - F_X(10,000) = e^{-10,000\lambda} = e^{-4} = 0.0183 \qquad [1]$$

(ii)(a) ***Probability that at least one claim will exceed £10,000 using an exact method***

The required probability is:

$$\begin{aligned} P\left(\max\{X_1,\dots,X_{100}\}>10{,}000\right) &= 1-P\left(\max\{X_1,\dots,X_{100}\}\le 10{,}000\right) \\ &= 1-P\left(X_1\le 10{,}000,\dots,X_{100}\le 10{,}000\right) \\ &= 1-P\left(X_1\le 10{,}000\right)\times\cdots\times P\left(X_{100}\le 10{,}000\right) \\ &= 1-\left(1-e^{-4}\right)^{100} = 0.8425 \end{aligned}$$ [2]

(ii)(b) ***Probability that at least one claim will exceed £10,000 using an approximate method***

The approximate distribution of $\max\{X_1,\dots,X_{100}\}$ is a Gumbel-type GEV distribution with CDF:

$$H(x)=\exp\left(-\exp\left(-\left(\frac{x-\alpha_{100}}{\beta_{100}}\right)\right)\right)$$

where:

$$\alpha_{100}=2{,}500\ \ln 100=11{,}512.93 \quad \text{and} \quad \beta_{100}=2{,}500$$ [1]

So:

$$\begin{aligned} P\left(\max\{X_1,\dots,X_{100}\}>10{,}000\right) &= 1-P\left(\max\{X_1,\dots,X_{100}\}\le 10{,}000\right) \\ &\approx 1-\exp\left(-\exp\left(-\left(\frac{10{,}000-\alpha_{100}}{\beta_{100}}\right)\right)\right) \\ &= 1-\exp\left(-\exp\left(-\left(\frac{10{,}000-11{,}512.93}{2{,}500}\right)\right)\right) \\ &= 1-0.1602=0.8398 \end{aligned}$$ [2]

(iii) ***Assumptions***

The two key assumptions are that all claims come from an exponential distribution with mean £2,500 and that they are statistically independent. [1]

16.4 The CDF of the threshold exceedances is given by:

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)}$$

If the individual losses follow a $Pa(\alpha,\lambda)$ distribution, then:

$$F_u(x) = \frac{\left\{1 - \left(\frac{\lambda}{\lambda + x + u}\right)^{\alpha}\right\} - \left\{1 - \left(\frac{\lambda}{\lambda + u}\right)^{\alpha}\right\}}{1 - \left\{1 - \left(\frac{\lambda}{\lambda + u}\right)^{\alpha}\right\}}$$

$$= \frac{\left(\frac{\lambda}{\lambda + u}\right)^{\alpha} - \left(\frac{\lambda}{\lambda + x + u}\right)^{\alpha}}{\left(\frac{\lambda}{\lambda + u}\right)^{\alpha}}$$

$$= 1 - \left(\frac{\lambda + u}{\lambda + u + x}\right)^{\alpha}$$

This is the CDF of the $Pa(\alpha,\lambda+u)$ distribution. So the distribution of the threshold exceedances is $Pa(\alpha,\lambda+u)$.

16.5 Comparing the two density functions and taking the limit as $x \to \infty$, we have:

$$\lim_{x\to\infty} \frac{f_{gamma}(x)}{f_{exp}(x)} = \lim_{x\to\infty} \left\{ \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \Big/ \lambda e^{-\lambda x} \right\}$$

$$= \lim_{x\to\infty} \left\{ C\, x^{\alpha-1} \right\}$$

where C is a constant.

When $\alpha > 1$:

$$\frac{f_{gamma}(x)}{f_{exp}(x)} \to \infty \quad \text{as} \quad x \to \infty$$

Hence the $Gamma(\alpha,\lambda)$ distribution has the heavier tail.

16.6 (i) ***Hazard rate for the Weibull distribution***

The hazard rate for the $W(c,\gamma)$ distribution is:

$$h(x)=\frac{f(x)}{1-F(x)}$$

$$=\frac{c\gamma x^{\gamma-1}e^{-cx^{\gamma}}}{e^{-cx^{\gamma}}}$$

$$=c\gamma x^{\gamma-1}$$

(ii) ***Comment on behaviour***

If $\gamma>1$, then this hazard rate is an increasing function of x, which corresponds to a light tail.

If $0<\gamma<1$, then the hazard rate is a decreasing function of x, which corresponds to a heavy tail.

16.7 (i) ***Show the integral expression***

Using the substitution, $u=3y^{\frac{1}{2}}$, we have:

$$y=\left(\frac{u}{3}\right)^2$$

and $\dfrac{dy}{du}=\dfrac{2}{9}u$ [1]

Hence the integral becomes:

$$\int_x^\infty e^{-3y^{\frac{1}{2}}}dy=\int_{3x^{\frac{1}{2}}}^\infty \frac{2}{9}ue^{-u}du$$ [1]

The integrand now resembles, but is not quite, that of the $Gamma(2,1)$ *distribution.*

We can transform the integrand into that of the $Gamma(2,1)$ distribution by adjusting the constants inside and outside the integral:

$$\int_x^\infty e^{-3y^{\frac{1}{2}}}dy=\frac{2}{9}\left(\frac{\Gamma(2)}{1^2}\right)\int_{3x^{\frac{1}{2}}}^\infty \frac{1^2}{\Gamma(2)}ue^{-u}du$$ [1]

Now $\Gamma(2)=1!=1$ and the integral is the probability that a $Gamma(2,1)$ random variable takes a value greater than $3x^{\frac{1}{2}}$. So:

$$\int_x^\infty e^{-3y^{\frac{1}{2}}}dy=\frac{2}{9}P\left(U>3x^{\frac{1}{2}}\right) \quad \text{where } U\sim Gamma(2,1)$$ [1]

Using the relationship between the gamma and chi-squared distributions (given on page 12 of the *Tables*), we see that that $2U \sim \chi_4^2$. Hence:

$$\int_x^\infty e^{-3y^{\frac{1}{2}}}\,dy = \frac{2}{9}P\left(2U > 6x^{\frac{1}{2}}\right) = \frac{2}{9}P\left(\chi_4^2 > 6x^{\frac{1}{2}}\right)$$ [1]

(ii) ***Mean residual life***

The mean residual life for the $Weibull\left(3, \frac{1}{2}\right)$ distribution is:

$$e(x) = \frac{\int_x^\infty \{1 - F(y)\}\,dy}{1 - F(x)}$$

$$= \frac{\int_x^\infty e^{-3y^{\frac{1}{2}}}\,dy}{e^{-3x^{\frac{1}{2}}}}$$

$$= \frac{\frac{2}{9}P\left(\chi_4^2 > 6x^{\frac{1}{2}}\right)}{e^{-3x^{\frac{1}{2}}}}$$ [2]

(iii) ***Nature of the mean residual life function***

When $x = 1$, we have $e(x) = \dfrac{\frac{2}{9}P\left(\chi_4^2 > 6\right)}{e^{-3}} = \dfrac{\frac{2}{9}\times(1 - 0.8009)}{e^{-3}} = 0.8887$. [½]

When $x = 4$, we have $e(x) = \dfrac{\frac{2}{9}P\left(\chi_4^2 > 12\right)}{e^{-6}} = \dfrac{\frac{2}{9}\times(1 - 0.9826)}{e^{-6}} = 1.5599$. [½]

When $x = 9$, we have $e(x) = \dfrac{\frac{2}{9}P\left(\chi_4^2 > 18\right)}{e^{-9}} = \dfrac{\frac{2}{9}\times(1 - 0.9988)}{e^{-9}} = 2.1608$. [½]

The mean residual life is an increasing function of x, suggesting that this distribution has a heavy tail. [½]

End of Part 4

What next?

1. Briefly **review** the key areas of Part 4 and/or re-read the **summaries** at the end of Chapters 13 to 16.
2. Ensure you have attempted some of the **Practice Questions** at the end of each chapter in Part 4. If you don't have time to do them all, you could save the remainder for use as part of your revision.
3. Attempt **Assignment X4**.

17

Copulas

Syllabus objectives

1.3 Introduction to copulas

1.3.1 Describe how a copula can be characterised as a multivariate distribution function which is a function of the marginal distribution functions of its variates, and explain how this allows the marginal distributions to be investigated separately from the dependency between them.

1.3.2 Explain the meaning of the terms dependence or concordance, upper and lower tail dependence; and state in general terms how tail dependence can be used to help select a copula suitable for modelling particular types of risk.

1.3.3 Describe the form and characteristics of the Gaussian copula and the Archimedean family of copulas.

0 Introduction

Insurance and investment companies are often interested in being able to compute the joint probability of events occurring, for example the joint probability of losses on different classes of business or on investments, or the joint probability of default on investments. One way of calculating a joint probability is to use a joint PDF (or probability function in the case of discrete random variables) and then to integrate (or sum) this to find the probability.

There are a number of drawbacks to this approach. First of all, we would need to specify fully the joint distribution. This is not usually easy to determine unless, for example, all the underlying individual distributions are normal. Even if we can specify the joint distribution, it is not typically clear how the joint PDF (or probability function) relates to the individual PDFs (or probability functions) and what the nature of the association is between them.

An alternative way of calculating a joint probability is to use a copula. A copula is a function that takes as inputs marginal CDFs and outputs a joint CDF.

For example, suppose an insurer wants to work out the joint probability that annual losses on its household portfolio, X_H, will be less than or equal to £5*m* and that annual losses on its motor portfolio, X_M, will be less than or equal to £3*m*. For simplicity of calculation, we assume that the two portfolios give rise to losses independently.

Measuring in £*m* and using the assumption of independence, we calculate the probability as:

$$P(X_H \leq 5, X_M \leq 3) = P(X_H \leq 5)P(X_M \leq 3)$$

This calculation involves a copula function. We have taken the marginal CDFs as inputs:

$$F_{X_H}(5) = P(X_H \leq 5) \text{ and } F_{X_M}(3) = P(X_M \leq 3)$$

and outputted a joint CDF:

$$F_{X_H,X_M}(5,3) = P(X_H \leq 5, X_M \leq 3)$$

The copula function that we've used here is called the *independence (or product) copula* and can be expressed as follows:

$$C\left[F_{X_H}(x_H), F_{X_M}(x_M)\right] = F_{X_H}(x_H) \times F_{X_M}(x_M) \text{ or as } C[u,v] = uv$$

where $u = F_{X_H}(x_H)$ and $v = F_{X_M}(x_M)$

Of course, it is not always the case that the relationship between the random variables is one of independence. However, there are many distinct copula functions, each of which expresses different types and levels of association between the variables.

In Sections 1 to 3 of this chapter we develop ideas leading to the definition of copula. In Sections 4 to 6 we look at different types of copula functions. In Sections 7 and 8 we introduce some practical applications of copulas.

1 Marginal and joint distributions

1.1 Joint distribution and density functions

The meaning of jointly distributed random variables, marginal distributions and conditional distributions was described in Subject CS1.

Recall from Subject CS1 that for two variables, the joint (cumulative) distribution function (CDF) is:

$$F_{X,Y}(x,y) = P(X \le x, Y \le y)$$

This can be extended from the bivariate case to the multivariate case in d dimensions:

$$F_{X_1,X_2,\ldots,X_d}(x_1,x_2,\ldots,x_d) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_d \le x_d)$$

In the context of joint distribution functions, the individual distribution of each of the variables in isolation is known as its *marginal distribution*.

In the bivariate case, where random variables are continuous, we can derive the joint CDF by integrating the joint PDF, as follows:

$$F_{X,Y}(x,y) = \int_{s \le x} \int_{t \le y} f_{X,Y}(s,t)\, ds\, dt$$

1.2 Marginal distribution and density functions

To compute the marginal PDF from a joint PDF, we ‘integrate out’ the other variable:

$$f_X(x) = \int_y f_{X,Y}(x,y)\, dy$$

To compute a marginal CDF, we integrate the marginal PDF as follows:

$$F_X(x) = \int_{s \le x} f_X(s)\, ds$$

The formulae above can be generalised from 2-dimensions (bivariate or joint case) to higher dimensions (multivariate case).

1.3 Standard joint distribution and density functions

Example (‘off the shelf’) multivariate statistical distributions are:

- the *multivariate normal distribution*, where the marginal distributions are all normal
- the *multivariate Student’s t distribution*, where the marginal distributions are all t distributions.

The question below acts as a refresher on joint and marginal CDFs and PDFs and how they relate to each other.

Question

The joint PDF for two continuous random variables X and Y is:

$$f_{X,Y}(x,y)=\tfrac{1}{20}(x+4y)\,,\quad 0<x\le 2\,,\ 0<y\le 2$$

(i) Derive a formula for the joint CDF, $F_{X,Y}(x,y)$.

(ii) Derive formulae for the marginal PDFs, $f_X(x)$ and $f_Y(y)$, and comment on whether X and Y are independent.

(iii) Derive formulae for the marginal CDFs, $F_X(x)$ and $F_Y(y)$.

Solution

(i) ***Joint CDF***

The joint CDF is obtained by integrating the joint PDF with respect to both variables:

$$\begin{aligned}
F_{X,Y}(x,y)&=\int_{s=0}^{x}\int_{t=0}^{y} f(s,t)\,ds\,dt\\
&=\int_{s=0}^{x}\left(\int_{t=0}^{y}\tfrac{1}{20}(s+4t)\,dt\right)ds\\
&=\tfrac{1}{20}\int_{s=0}^{x}\left[\left(st+2t^2\right)\right]_0^y ds\\
&=\tfrac{1}{20}\int_{s=0}^{x}\left(sy+2y^2\right)ds\\
&=\tfrac{1}{20}\left[\left(\tfrac{1}{2}s^2y+2sy^2\right)\right]_0^x\\
&=\tfrac{1}{20}\left(\tfrac{1}{2}x^2y+2xy^2\right)\\
&=\tfrac{1}{40}xy(x+4y)
\end{aligned}$$

(ii) ***Marginal PDFs***

The marginal PDFs are obtained by 'integrating out' the other variable in the joint PDF:

$$f_X(x)=\int_{y=0}^{2} f_{X,Y}(x,y)\,dy=\int_{y=0}^{2}\tfrac{1}{20}(x+4y)\,dy=\tfrac{1}{20}\Big[\big(xy+2y^2\big)\Big]_0^2=\tfrac{1}{20}(2x+8)=\tfrac{1}{10}(x+4)$$

$$f_Y(y)=\int_{x=0}^{2} f_{X,Y}(x,y)\,dx=\int_{x=0}^{2}\tfrac{1}{20}(x+4y)\,dx=\tfrac{1}{20}\Big[\big(\tfrac{1}{2}x^2+4xy\big)\Big]_0^2=\tfrac{1}{20}(2+8y)=\tfrac{1}{10}(1+4y)$$

For X and Y to be independent, we must have:

$$f_{X,Y}(x,y)=f_X(x)\times f_Y(y)$$

In this case, the product of the marginal PDFs in (ii) is not equal to the joint PDF in (i). We therefore deduce that X and Y are *not* independent.

(iii) ***Marginal CDFs***

The marginal CDFs are obtained by integrating the marginal PDFs:

$$F_X(x)=\int_0^x f_X(s)\,ds=\int_0^x\tfrac{1}{10}(s+4)\,ds=\tfrac{1}{10}\Big[\tfrac{1}{2}s^2+4s\Big]_0^x=\tfrac{1}{10}\big(\tfrac{1}{2}x^2+4x\big)=\tfrac{1}{20}x(x+8)$$

$$F_Y(y)=\int_0^y f_Y(t)\,dt=\int_0^y\tfrac{1}{10}(1+4t)\,dt=\tfrac{1}{10}\Big[t+2t^2\Big]_0^y=\tfrac{1}{10}\big(y+2y^2\big)=\tfrac{1}{10}y(1+2y)$$

In the above question we have shown that the two random variables are not independent. However, the nature of the association is only *implicit* in the formulae for the joint PDF and joint CDF. We can't immediately see the nature and extent of their association just by looking at the formula for the joint CDF.

Later in this chapter we will see how copula functions *explicitly* describe the *full* nature and extent of the association between random variables. However, before we do that let's look at some simpler statistical measures of association.

2 Association, concordance, correlation and tail dependence

2.1 Introduction to terminology

Variables are said to be *associated* if there is some form of statistical relationship between them – whether causal or not. To facilitate comparisons, measures of association can be constructed.

Coefficients of association are generally designed so that their values vary between −1 and +1. Their absolute values increase with the strength of the relationship. They take a value of +1 (or −1) when there is perfect positive (or negative) association.

Any one particular type of coefficient of association measures a particular *form* of association. For example, Pearson's correlation coefficient measures the degree to which there is a *linear* relationship between the variables.

Concordance is another particular form of association. Broadly speaking, two random variables are concordant if small values of one are likely to be associated with small values of the other, and *vice versa*.

Spearman's rho and Kendall's tau (discussed in Subject CS1) are two examples of measures of concordance.

Note that a positive association between two variables does not necessarily imply that one is *dependent* on the other. For example, both might be dependent on a third (perhaps unobserved) variable, with neither being directly dependent on the other. A common pitfall for journalists is to forget that 'correlation does not imply causation'.

2.2 Pearson's linear correlation coefficient

We have previously met the linear correlation coefficient (also known as Pearson's ρ), which measures how strongly the values of two variables are related.

Pearson's rho is defined as:

$$\rho_{X,Y}=\frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

Question

Pearson's rho takes values between –1 and +1. Outline what it means when:

(a) $\rho_{X,Y}=+1$

(b) $\rho_{X,Y}=-1$

(c) $\rho_{X,Y}=0$.

Solution

(a) When Pearson's rho is equal to +1, the random variables X and Y are perfectly positively correlated. This means that there is a perfectly increasing linear relationship between the values that the two random variables take.

(b) When Pearson's rho is equal to –1, the random variables X and Y are perfectly negatively correlated. This means that there is a perfectly decreasing linear relationship between the values that the two random variables take.

(c) When Pearson's rho is equal to 0, the random variables X and Y are uncorrelated. This means that there is no *linear* relationship between the values that the two random variables take.

Since Pearson's rho is a measure of linear association between variables, it remains unchanged under a linear transformation of the variables, *ie* Pearson's rho for the random variables $aX+b$ and $cY+d$ has the same value as Pearson's rho for X and Y.

However, if we apply a non-linear transformation to the random variables, it is not necessarily the case that the value of Pearson's rho will stay the same. For example, if we calculate Pearson's rho for the random variables e^X and e^Y, this will not necessarily give us the same value as Pearson's rho for X and Y. In many respects this is counter-intuitive: we have applied a monotonically increasing function to the random variables and hence we might expect the degree of association to stay the same.

Desirable properties of a measure of concordance / association

A good measure of the *concordance* (or *association*) between two variables should have a number of properties. These include *invariance*, which requires that the measure of concordance does not change if we apply the same monotone function to the value of each variable. Pearson's ρ does not have this property.

2.3 Rank correlation coefficients

Two commonly used measures of concordance that are more robust than Pearson's ρ and are invariant are Spearman's ρ (often called the *rank correlation*) and Kendall's τ.

Measures of rank correlation look at the association between the position (or rank) of observations in a data series when they are arranged in order. For example, recall from Subject CS1, that Spearman's ${}_S\rho$ can be calculated as:

$${}_S\rho = 1 - \frac{6}{T(T^2-1)}\sum_{i=1}^{T} d_i^2$$

where:

- T is the number of (pairs of) observations
- d_i is the difference in rank for the i th observation (pair).

Properties of rank correlation coefficients

Spearman's rho and Kendall's tau measure the degree of concordance between the rank (or position) rather than between the actual observations. This means that, even if the value of an observation changes, as long as its relative ranking stays the same, then the measure of concordance will remain the same. For example, we could apply an exponential transformation to the observations and both Spearman's rho and Kendall's tau would be unchanged.

2.4 Tail dependence

The correlation measures described above each try to 'summarise' the nature and extent of the association between variables into a single statistic. Two key shortcomings of such statistics are:

- information is lost in the summarisation process
- they capture the interdependency through the *whole* distribution, and this may be of less interest than the interdependency in the *tails* of the distribution.

It is often the case in insurance and investment applications that large losses tend to occur together. For example a hurricane could result in large losses on several different property insurance policies sold by the same company or a stock market crash could lead to large losses on a number of investments in the same investment portfolio.

So the relationships between the variables at the extremes of the distributions, *ie* in the upper and lower tails, are of particular importance. These can be measured using the coefficients of upper and lower tail dependence.

Coefficient of upper tail dependence

We can define the coefficient of upper tail dependence as:

$$\lambda_U = \lim_{u \to 1^-} P\left(X > F_X^{-1}(u) \,\middle|\, Y > F_Y^{-1}(u)\right)$$

This coefficient is a probability, so it takes a value between 0 and 1.

The coefficient of upper tail dependence indicates whether high values of one random variable, X, tend to be linked with high values of another random variable Y.

It considers the probability of the random variable X taking a value in the upper tail of its distribution (*eg* a tail with a probability mass of $5\% \Rightarrow u = 0.95$), given that the random variable Y takes a value in the same sized upper tail of its distribution.

Specifically, the coefficient of upper tail dependence is the limiting value of this probability as $u \to 1^-$, *ie* as we move further into the upper tail (from below).

Coefficient of lower tail dependence

The coefficient of lower tail dependence is defined as:

$$\lambda_L = \lim_{u \to 0^+} P\left(X \le F_X^{-1}(u) \mid Y \le F_Y^{-1}(u)\right)$$

Again this coefficient is a probability, so takes a value between 0 and 1.

The coefficient of lower tail dependence indicates whether low values of one random variable, X, tend to be linked with low values of another random variable Y.

It considers the probability of the random variable X taking a value in the lower tail of its distribution (*eg* a tail with a probability mass of $5\% \Rightarrow u = 0.05$), given that the random variable Y takes a value in the same sized lower tail of its distribution.

Specifically, the coefficient of lower tail dependence is the limiting value of this probability as $u \to 0^+$, *ie* as we move further into the lower tail (from above).

3 Copulas

3.1 Expressing the association between variables explicitly

The joint distribution combines the information from the marginal distributions and the way in which the variables depend on each another. However, it expresses this dependence *implicitly*. We cannot immediately see the nature of the interdependence simply by looking at the formula for the joint distribution function.

Copulas provide an alternative approach that expresses the interdependence between the variables *explicitly*. They allow us to deconstruct the joint distribution of a set of variables into components (the marginal distributions plus a copula) that can be adjusted individually.

3.2 Definition of a copula

A *copula* is a function that expresses a multivariate cumulative distribution function in terms of the individual marginal cumulative distributions.

It is important to remember that a copula is a *function*. It takes marginal probabilities of random variables as inputs, and outputs a corresponding joint probability.

Definition of a copula

For a bivariate distribution, the copula is a function C_{XY} defined by:

$$C_{XY}\left[F_X(x),F_Y(y)\right]=P(X\le x,Y\le y)=F_{X,Y}(x,y)$$

This is often written in the more compact form:

$$C[u,v]=F_{X,Y}(x,y) \text{ where } u=F_X(x) \text{ and } v=F_Y(y)$$

This definition can be extended to the multivariate case where we have:

$$C[u_1,u_2,\dots,u_d]=F_{X_1,X_2,\dots,X_d}(x_1,x_2,\dots,x_d) \text{ where } u_i=F_{X_i}(x_i)$$

Note that the arguments $u_1,u_2,\dots,u_d$ and the output value of the copula function are restricted to the range $[0,1]$, as they correspond to probabilities.

3.3 Three properties of copulas

Copulas must also satisfy three technical properties to ensure that they correctly capture the properties we would expect of a joint distribution in all circumstances.

Although not included in the Core Reading, we outline these additional three properties below:

1. A copula is an increasing function of its inputs:

$$C\left[u_1,\dots,u_i^*,\dots,u_d\right]>C\left[u_1,\dots,u_i,\dots,u_d\right] \text{ for } u_i^*>u_i \text{ and } i=1,\dots,d$$

This makes sense from a probabilistic perspective because, if $u_i^* > u_i$, then $P\left(X_i \le x_i^*\right) > P\left(X_i \le x_i\right)$ for corresponding $x_i^* = F_{X_i}^{-1}\left(u_i^*\right)$ and $x_i = F_{X_i}^{-1}\left(u_i\right)$, and hence:

$$P(X_1 \le x_1, \ldots, X_i \le x_i^*, \ldots, X_d \le x_d) > P(X_1 \le x_1, \ldots, X_i \le x_i, \ldots, X_d \le x_d)$$

2. If all the marginal CDFs are equal to 1 except for one of them, then the copula function is equal to the value of that one marginal CDF:

$$C\left(1,\ldots,1,u_i,1,\ldots,1\right) = u_i \text{ for } i = 1,2,\ldots,d \text{ and } u_i \in \left[0,1\right]$$

This makes sense because $u_k = 1 \Rightarrow P\left(X_k \le x_k\right) = 1$ (*ie* a certainty), for $x_k = F_{X_k}^{-1}\left(u_k\right)$, and the only uncertainty in the above joint probability is the marginal probability with respect to the i th variable.

3. A copula function always returns a valid probability:

$$C\left[u_1, u_2, \ldots, u_d\right] \in \left[0,1\right]$$

3.4 Sklar's theorem

Sklar demonstrated in 1959 that the dependence structure of a set of random variables can be captured using copulas. The theorem is as follows:

Sklar's theorem

Let F be a joint (cumulative) distribution function with marginal cumulative distribution functions $F_1, \ldots, F_d$. Then there exists a copula, C, such that for all $x_1, \ldots, x_d \in \left[-\infty, \infty\right]$:

$$F\left(x_1, \ldots, x_d\right) = C\left[F_1\left(x_1\right), \ldots, F_d\left(x_d\right)\right]$$

In the case of variables that have a *continuous* distribution, the copula is unique.

Sklar's theorem tells us that if we have a joint CDF and marginal CDFs, then these can be linked by a copula function, *ie* a copula function exists.

The converse also holds:

Converse of Sklar's theorem

If C is a copula and $F_1, \ldots, F_d$ are univariate cumulative distribution functions, then the function F defined above is a joint cumulative distribution function with marginal cumulative distribution functions $F_1, \ldots, F_d$.

Question

The joint probability density function for two continuous random variables X and Y is:

$$f_{X,Y}(x,y)=\tfrac{1}{20}(x+4y),\quad 0<x\leq 2,\ 0<y\leq 2$$

(i) Derive formulae for the inverse cumulative distribution functions $F_X^{-1}(u)$ and $F_Y^{-1}(v)$.

(ii) Hence derive a formula for the copula function $C(u,v)=F_{XY}(x,y)$.

Solution

(i) ***Inverse CDFs***

In the solution to the question at the end of Section 1.3, we showed that:

$$F_X(x)=\int_0^x f_X(s)\,ds=\int_0^x \tfrac{1}{10}(s+4)\,ds=\tfrac{1}{10}\left[\tfrac{1}{2}s^2+4s\right]_0^x=\tfrac{1}{10}\left(\tfrac{1}{2}x^2+4x\right)=\tfrac{1}{20}x(x+8)$$

$$F_Y(y)=\int_0^y f_Y(t)\,dt=\int_0^y \tfrac{1}{10}(1+4t)\,dt=\tfrac{1}{10}\left[t+2t^2\right]_0^y=\tfrac{1}{10}\left(y+2y^2\right)=\tfrac{1}{10}y(1+2y)$$

To find the inverse CDFs, we need to invert these CDFs by making x and y the subject of the equations, which in this case are quadratic equations.

We have:

$$u=F_X(x)=\tfrac{1}{20}x(x+8)$$

$$\Rightarrow \quad x^2+8x-20u=0$$

$$\Rightarrow \quad x=\frac{-8\pm\sqrt{8^2-4(1)(-20u)}}{2(1)}=\frac{-8\pm\sqrt{64+80u}}{2}=-4\pm\sqrt{16+20u}$$

Only the positive root will result in values of x in the range $0<x\leq 2$. So:

$$x=F^{-1}(u)=-4+\sqrt{16+20u}$$

Similarly:

$$v=F_Y(y)=\tfrac{1}{10}y(1+2y)$$

$$\Rightarrow \quad y=F^{-1}(v)=\tfrac{1}{4}\left(-1+\sqrt{1+80v}\right)$$

(ii) ***Copula function***

The copula function, which is defined on the range $0 \le u, v \le 1$, is then obtained by writing the joint CDF in terms of u and v:

$$
\begin{aligned}
C[u,v] &= F_{XY}(x,y) \\
&= \tfrac{1}{40} xy\left(x+80y\right) \\
&= \tfrac{1}{160}\left(-4+\sqrt{16+20u}\right)\left(-1+\sqrt{1+80v}\right)\left(-24+\sqrt{16+20u}+20\sqrt{1+80v}\right)
\end{aligned}
$$

3.5 Expressions of tail dependence and the survival copula

Lower tail dependence

Recall that the coefficient of lower tail dependence is defined as:

$$\lambda_L = \lim_{u \to 0^+} P\left(X \le F_X{}^{-1}(u) \,\middle|\, Y \le F_Y{}^{-1}(u)\right)$$

The coefficient of lower tail dependence measures the limit of this probability as the lower tail becomes smaller and smaller, as u tends to 0 from above. We can write:

$$\lambda_L = \lim_{u \to 0^+} P\left(X \le x \,\middle|\, Y \le y\right)$$

where x and y are the lower tail percentage points of the distributions of the random variables X and Y respectively:

$$P\left(X \le x\right) = P\left(Y \le y\right) = u$$

Using the definition of conditional probabilities, the equation for the coefficient of lower tail dependence becomes:

$$\lambda_L = \lim_{u \to 0^+} \frac{P\left(X \le x, Y \le y\right)}{P\left(Y \le y\right)}$$

The expression in the numerator of this fraction is a joint (cumulative) distribution function, *ie* it is the copula function $C\left[u,u\right]$. The expression in the denominator is the probability that Y takes a value less than or equal to y, which we know is u. This leads to the following useful formula:

Coefficient of lower tail dependence in terms of the copula function

$$\boldsymbol{\lambda_L = \lim_{u \to 0^+} \frac{C\left[u,u\right]}{u}}$$

***ie* the coefficient of lower tail dependence can be calculated directly from the copula function.**

The coefficient of lower tail dependence can take values between 0 (no dependence) and 1 (full dependence).

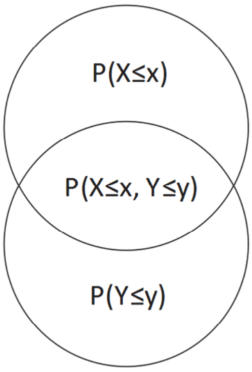
The survival copula

To define the upper tail dependence, we need to look at the opposite end of the marginal distributions. Associated with each copula function is a *survival copula* function (indicated with a bar), which is defined by:

$$\bar{F}(x,y) = P(X > x, Y > y) = \bar{C}\left[\bar{F}_X(x), \bar{F}_Y(y)\right]$$

where $\bar{F}_X(x) = 1 - F_X(x)$ and $\bar{F}_Y(y) = 1 - F_Y(y)$.

By the *principle of inclusion / exclusion*, we have:



$$P(X \le x \text{ and/or } Y \le y) = P(X \le x) + P(Y \le y) - P(X \le x, Y \le y)$$

ie $$1 - P(X > x, Y > y) = P(X \le x) + P(Y \le y) - P(X \le x, Y \le y)$$

$\Rightarrow$ $$P(X > x, Y > y) = 1 - P(X \le x) - P(Y \le y) + P(X \le x, Y \le y)$$

So the survival copula is related to the original copula function by:

$$\bar{C}[1-u, 1-v] = 1 - u - v + C[u,v]$$

Coefficient of upper tail dependence in terms of the survival copula function

We can then define the coefficient of upper tail dependence as:

$$\lambda_U = \lim_{u \to 1^-} P\left(X > F_X^{-1}(u) \mid Y > F_Y^{-1}(u)\right) = \lim_{u \to 0^+} \frac{\bar{C}[u,u]}{u}$$

Let's look more closely at how the above result is derived.

The coefficient of upper tail dependence measures the limit of this probability as the upper tail becomes smaller and smaller, as u tends to 100% from below.

We have:

$$\lambda_U = \lim_{u \to 1^-} P\left(X > x \mid Y > y\right)$$

where x and y correspond to upper tail percentage points of the distributions of random variables X and Y respectively:

$$P(X \le x) = P(Y \le y) = u$$

and:

$$P(X > x) = P(Y > y) = 1 - u$$

Using the definition of conditional probabilities, the equation for the coefficient of upper tail dependence becomes:

$$\lambda_U = \lim_{u \to 1^-} \frac{P(X > x, Y > y)}{P(Y > y)}$$

The expression in the numerator of this fraction can be worked out using the principle of inclusion / exclusion result in the Core Reading above:

$$\begin{aligned} P(X > x, Y > y) &= 1 - P(X \le x) - P(Y \le y) + P(X \le x, Y \le y) \\ &= 1 - u - u + C[u,u] \\ &= 1 - 2u + C[u,u] \end{aligned}$$

The expression in the denominator is the probability that Y takes a value greater than y, which we know is $1-u$. So we have the following formula:

Coefficient of upper tail dependence in terms of the copula function

$$\lambda_U = \lim_{u \to 1^-} \frac{1 - 2u + C[u,u]}{1-u}$$

***ie* the coefficient of upper tail dependence can be calculated directly from the copula function.**

Alternatively, we can express this using the survival copula function as:

$$\lambda_U = \lim_{u \to 1^-} \frac{\bar{C}[1-u, 1-u]}{1-u} = \lim_{u \to 0^+} \frac{\bar{C}[u,u]}{u}$$

3.6 Types of copula function

There are three main families of copula that we will go on to consider in the subsequent sections of this chapter:

(i) fundamental copulas

(ii) explicit copulas

(iii) implicit copulas.

4 Fundamental copulas

Fundamental copulas represent the three basic (or fundamental) dependencies that a set of variables can display, namely:

- independence
- perfect positive interdependence, and
- perfect negative interdependence.

These copulas are referred to as the:

- independence (or product) copula
- co-monotonic (or minimum) copula
- counter-monotonic (or maximum) copula.

Collectively these three copulas are referred to as *fundamental copulas*. They are specific cases of a more general family of copulas called *Fréchet-Höffding copulas*.

In the bivariate case, the co-monotonic and counter-monotonic copulas represent the extremes of the possible levels of association between variables. They are therefore the upper and lower boundaries for all copulas – known as the *Fréchet-Höffding bounds*. The co-monotonic copula is the upper bound copula and the counter-monotonic copula is the lower bound copula.

4.1 Independence (or product) copula

One example of a bivariate copula is the *product copula* $C[u,v]=uv$. Here we have:

$$F_{X,Y}(x,y)=C\left[F_X(x),F_Y(y)\right]=F_X(x)F_Y(y)$$

or: $$P(X\le x,Y\le y)=P(X\le x)P(Y\le y)$$

This captures the property of independence of the two variables X and Y, and so is also called the *independence (or product) copula*.

We looked at an example involving the independence (or product) copula in the introductory section of this chapter.

Question

Derive the coefficients of upper and lower tail dependence for the independence (or product) copula.

Solution

The independence (or product) copula is expressed as:

$$C\left[u,v\right]=uv$$

The coefficient of lower tail dependence is given by:

$$\begin{aligned}\lambda_L &= \lim_{u\to 0^+} \frac{C[u,u]}{u}\\ &= \lim_{u\to 0^+} \frac{u^2}{u}\\ &= \lim_{u\to 0^+} u\\ &=0\end{aligned}$$

The coefficient of upper tail dependence is given by:

$$\begin{aligned}\lambda_U &= \lim_{u\to 1^-} \frac{1-2u+C[u,u]}{1-u}\\ &= \lim_{u\to 1^-} \frac{1-2u+u^2}{1-u}\\ &= \lim_{u\to 1^-} \frac{(1-u)^2}{1-u}\\ &= \lim_{u\to 1^-} (1-u)\\ &=0\end{aligned}$$

4.2 Co-monotonic (or minimum) copula

This copula is used where random variable demonstrate perfect positive interdependence. The *co-monotonic copula* is defined in the bivariate case as:

$$\boldsymbol{C[u,v]=\min(u,v)}$$

Here we have:

$$C\left[F_X(x),F_Y(y)\right]=\min\left(F_X(x),F_Y(y)\right)$$

or: $$P(X\le x, Y\le y)=\min\left(P(X\le x),P(Y\le y)\right)$$

To help us understand why the co-monotonic copula is the minimum of the two marginal CDFs, let's consider an example where two random variables X and Y display perfect positive interdependence.

Question

A statistician believes that there is a perfect positive interdependency between the Consumer Price Index (CPI) inflation rate, X (per annum), and the Retail Price Index (RPI) inflation rate, Y (per annum), and that the relationship can be modelled by the equation:

$$Y = X + 0.01$$

Show that $P(X \le x, Y \le y) = \min(P(X \le x), P(Y \le y))$.

Solution

We have:

$$\begin{aligned} P(X \le x, Y \le y) &= P(X \le x, X + 0.01 \le y) \\ &= P(X \le x, X \le y - 0.01) \end{aligned}$$

Now, for X to be less than or equal to both x and $y - 0.01$, it must be less than or equal to the smallest of these. So:

$$\begin{aligned} P(X \le x, Y \le y) &= P(X \le \min(x, y-0.01)) \\ &= \min(P(X \le x), P(Y \le y - 0.01)) \\ &= \min(P(X \le x), P(Y - 0.01 \le y - 0.01)) \\ &= \min(P(X \le x), P(Y \le y)) \end{aligned}$$

The second line uses the fact that $P(X \le \min(a,b,c,...))$ is just the smallest of the probabilities $P(X \le a), P(X \le b), P(X \le c),...$ *ie* $\min(P(X \le a), P(X \le b), P(X \le c),...)$.

4.3 Counter-monotonic (or maximum) copula

The co-monotonic copula captures the relationship between two variables whose values are perfectly positively interdependent on each other, while the counter-monotonic copula captures the corresponding inverse relationship.

The *counter-monotonic copula* is defined in the bivariate case as:

$$\mathbf{C[u,v] = \max(u + v - 1, 0)}$$

Here we have:

$$C[F_X(x), F_Y(y)] = \max(F_X(x) + F_Y(y) - 1, 0)$$

or: $$P(X \le x, Y \le y) = \max(P(X \le x) + P(Y \le y) - 1, 0)$$

Question

Let X and Y be two random variables that are perfectly negatively related as follows:

$$Y=-X$$

Show that $P(X\le x, Y\le y)=P(X\le x)+P(Y\le y)-1$.

Solution

We have:

$$\begin{aligned}
P(X\le x, Y\le y) &= P(X\le x, -X\le y)\\
&= P(X\le x, X\ge -y)\\
&= P(-y\le X\le x)\\
&= P(X\le x)-P(X<-y)\\
&= P(X\le x)-P(-X>y)\\
&= P(X\le x)-P(Y>y)\\
&= P(X\le x)-\left[1-P(Y\le y)\right]\\
&= P(X\le x)+P(Y\le y)-1
\end{aligned}$$

4.4 The multivariate case

The independence and co-monotonic copulas can be extended in the obvious way to the multivariate case. However, the counter-monotonic copula cannot. This is because it is not possible to have three or more variables where each pair has a direct inverse relationship.

In the multivariate case, we can extend the independence and co-monotonic copulas to d dimensions as follows:

$$_{ind}C\left[F_{X_1}(x_1),\ldots,F_{X_d}(x_d)\right]=F_{X_1}(x_1)\times\ldots\times F_{X_d}(x_d)$$

$$_{\min}C\left[F_{X_1}(x_1),\ldots,F_{X_d}(x_d)\right]=\min\left(F_{X_1}(x_1),\ldots,F_{X_d}(x_d)\right)$$

However, it is impossible to have three or more variables, *eg* X_1, X_2 and X_3, each of which always move in the *opposite* direction to all of the others. This is why the counter-monotonicity copula is only defined in two dimensions.

4.5 Graphical representation of copulas

Bivariate copulas can be represented graphically in various ways:

- scatterplots
- 3D (perspective) representations, and corresponding contour plots.

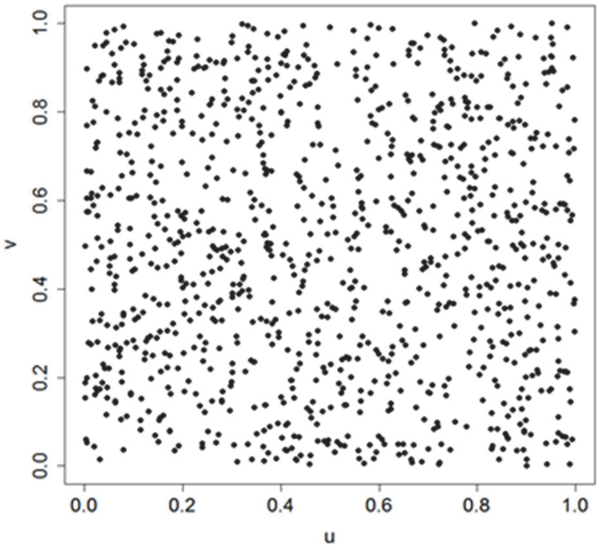
Scatterplots

The relationships implied by a copula can be illustrated using a scatterplot of simulated values of u and v.

Here we are recognising U and V as random variables, uniformly distributed on $[0,1]$, and illustrating the dependence structure between these two variables.

The diagrams below highlight the differences between the independence, co-monotonic and counter-monotonic copulas.

Scatterplot – Product (Independence) copula



This scatterplot of simulations illustrates the independence of the two variables.

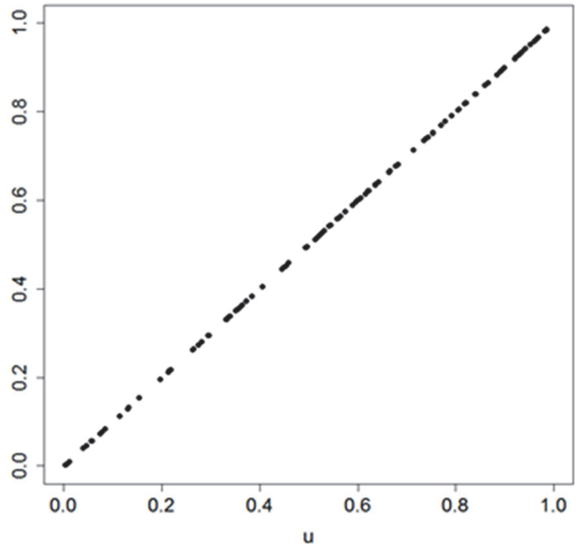
Each simulation (dot in the above diagram) has been created by:

- first simulating a particular value for U as a random number between 0 and 1
- then, independently, simulating the value for V as a random number between 0 and 1, because there is no association (dependence) between the two variables.

The variation in density of the dots in the above diagram is due to the random nature of the simulations, and the finite size of the sample. As the number of simulations increases the density will be become ever more uniform.

In this case, if we know the marginal distributions of X and Y, and observe $X=x$, then we know that $P(X \leq x)=F_X(x)=u$. However, knowing this value of u tells us nothing about the value of v as there is no relationship between u and v.

Scatterplot – Co-monotonic copula



This scatterplot of simulations illustrates a perfect *positive* interdependence between the two variables.

Each simulation (dot in the above diagram) has been created by:

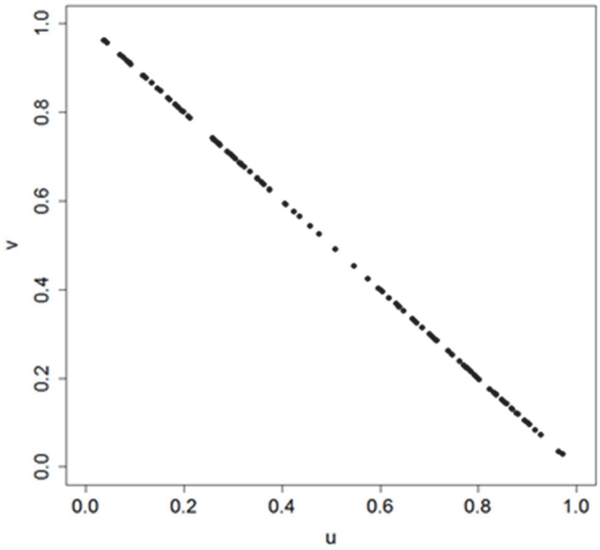
- first simulating a particular value for U as a random number between 0 and 1
- the value then taken by V is dictated by $u=v$.

As before, that the variation in density of the dots on the diagonal is due to the random nature of the simulations, and the finite size of the sample.

Since $x=F_X^{-1}(u)$ and $y=F_Y^{-1}(u)$, for all $0 \leq u \leq 1$, the diagram reinforces the fact that if X and Y are perfectly positively interdependent then:

- when X takes a value at a particular percentile of its distribution, Y also takes a value at the same percentile of its distribution
- when X takes a 'high' ('low') value in its possible range, Y also takes a 'high' ('low') value in its possible range.

Scatterplot – Counter-monotonic copula



This scatterplot of simulations illustrates a perfect *negative* interdependence between the two variables. Under such a relationship, given $U = u$ we can deduce that $V = v = 1 - u$.

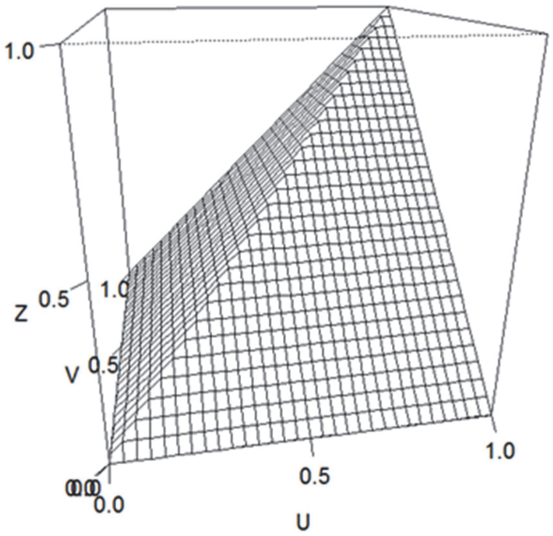
The diagram reinforces the fact that if X and Y are perfectly negatively interdependent, when X takes a 'high' ('low') value in its possible range, Y takes a 'low' ('high') value in its possible range.

These scatterplots are important, since, if a corresponding scatterplot based upon the underlying data shows features of independence, co-monotonicity or counter-monotonicity, it helps us to decide which copula function to fit to the data and, subsequently, use in our modelling.

3D representations and contour diagrams

The relationships described by copula functions, illustrated by the scatterplots (above), can also be represented in 3 dimensions: u, v and $C\left[u,v\right]$.

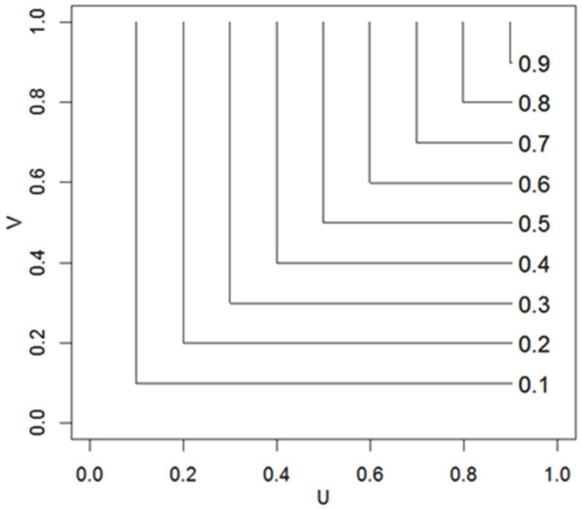
For example, the following two diagrams illustrate the co-monotonic copula. The first diagram shows the value of the copula plotted vertically in three dimensions. The second diagram shows contour lines of constant value of the copula function.



In general, we can interpret such a 3D-diagram as showing:

$$z = C[u,v] = P(X \leq x, Y \leq y)$$

where $u = P(X \leq x)$ and $v = P(Y \leq y)$. Note that $C[0,0] = 0$ and $C[1,1] = 1$. This is because the copula takes the form of a *cumulative* distribution function.



This second diagram is a two-dimensional representation of the first diagram. It is generally referred to as a contour diagram, as the (contour) lines indicate points of equal height in the 3D surface shown in the first diagram. (Think of taking a bird's eye view of the first diagram – looking down upon it from above.)

Question

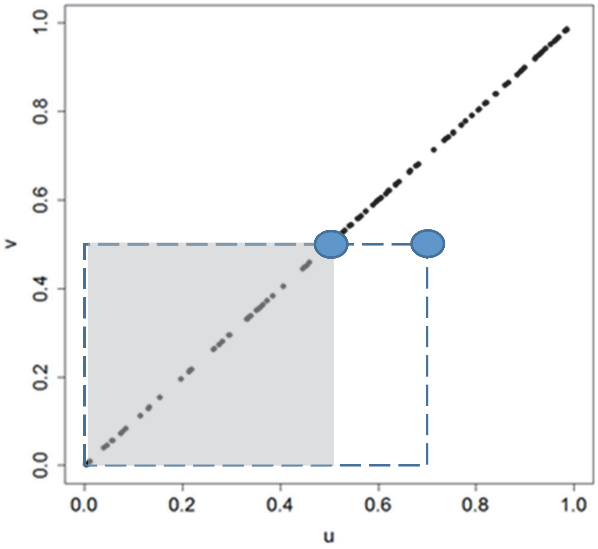
By considering the scatterplot for the co-monotonic copula that was shown earlier, and recognising that $C[u,v] = P(U \leq u, V \leq v)$, explain why the second diagram (shown above) represents the co-monotonic copula.

Solution

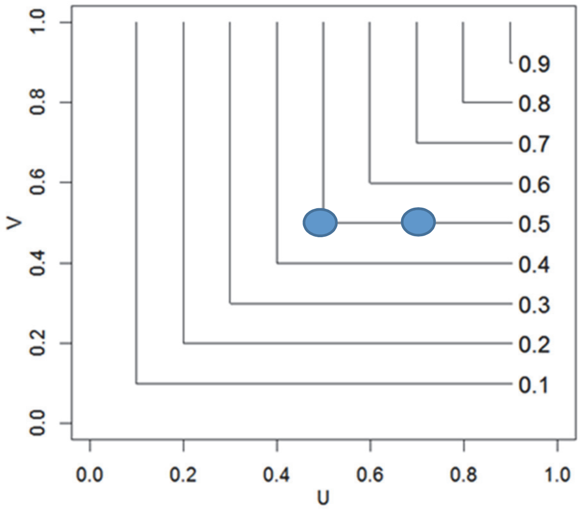
We can estimate $P(U \leq u, V \leq v)$ by looking at the number of simulations where $U \leq u$ and $V \leq v$ as a percentage of the total number of simulations. For example, in the simulation used to produce the scatterplot for the co-monotonic copula:

- Consider the condition $U \leq 0.5$ and $V \leq 0.5$, as indicated by the shaded square in the diagram below. About 50% of the simulations are where $U \leq 0.5$ and $V \leq 0.5$, *ie* the sample estimate of $P(U \leq 0.5, V \leq 0.5)$ is about 0.5.

- The same (number of) simulations meet the condition $U \leq 0.7$ and $V \leq 0.5$ (indicated by the dashed-line rectangle in the diagram below).

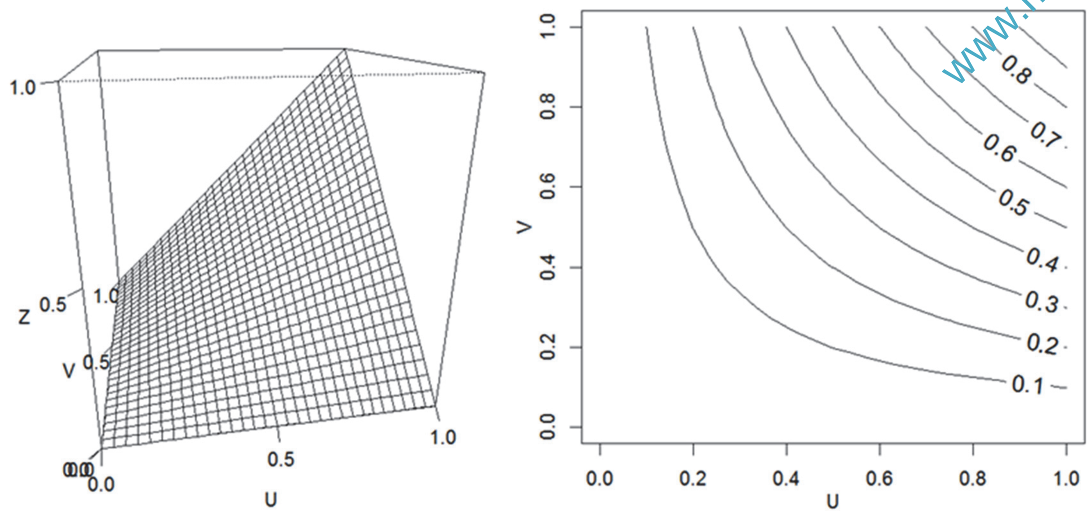


This is why (0.5,0.5) and (0.7,0.5) both lie on the same (0.5) contour line in the diagram below:

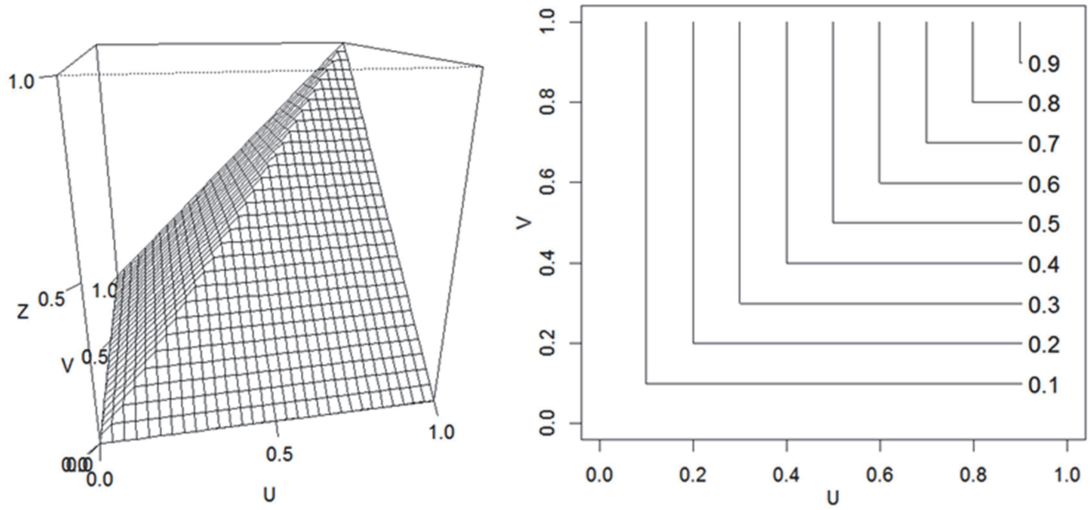


Similar perspective and contour diagrams (to those above) can be created for the other two fundamental copulas. All three sets of diagrams are shown on the next page for ease of comparison.

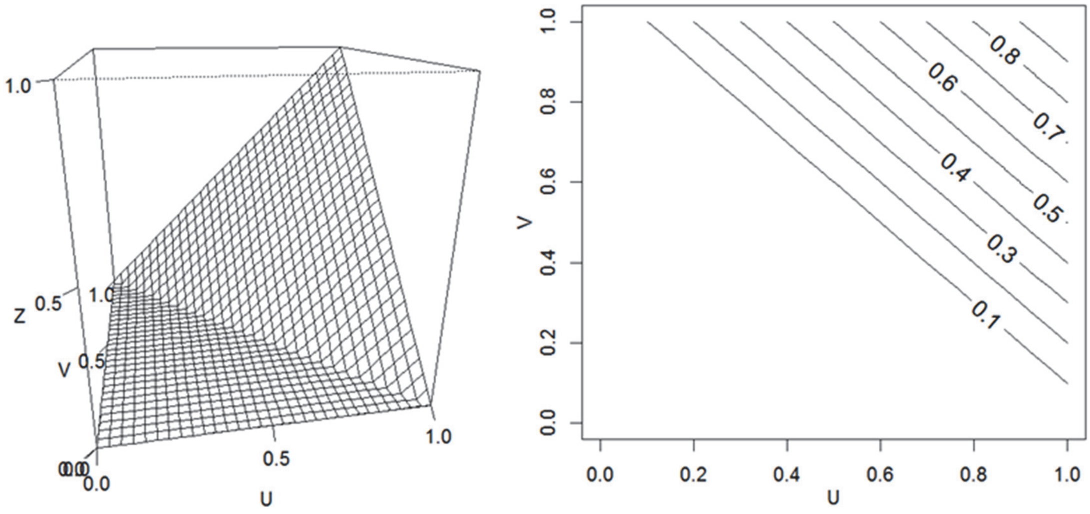
(i) independence (or product) copula:



(ii) co-monotonic (or minimum) copula



(ii) counter-monotonic (or maximum) copula:



5 Explicit copulas (including Archimedean copulas)

Explicit copulas have simple closed-form expressions. Below are three examples of commonly used explicit copulas.

As with statistical distributions such as the Poisson distribution or normal distribution, there are also a number of bespoke copulas that arise naturally in specific contexts. Examples of these are:

- **the Gumbel copula**
- **the Clayton copula**
- **the Frank copula.**

In these examples, α is a parameter whose value can be specified. We will see later that this can be used to adjust the strength of the dependence between the variables.

5.1 Gumbel copula

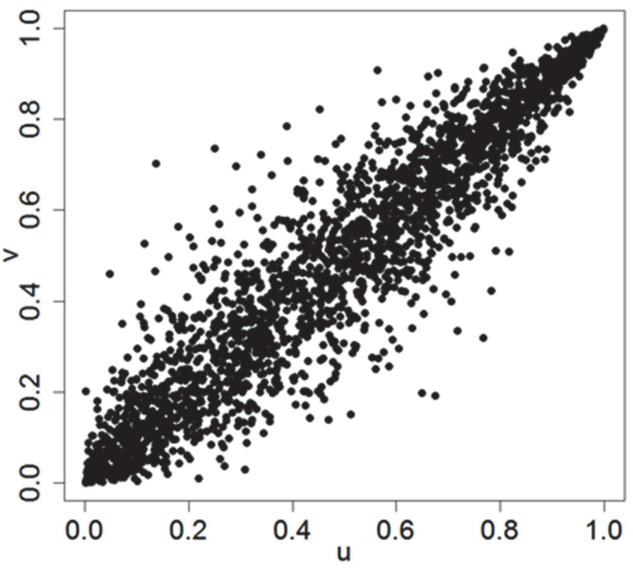
The *Gumbel copula* is defined in the bivariate case as:

$$C[u,v] = \exp\left\{-\left((-\ln u)^{\alpha} + (-\ln v)^{\alpha}\right)^{1/\alpha}\right\}$$

Note that the Gumbel copula is often referred to as the Gumbel-Hougaard copula.

The Gumbel copula describes an interdependence structure in which there is upper tail dependence (the level of which is determined by the parameter α), but there is no lower tail dependence.

A scatterplot of simulated values from the Gumbel copula with parameter value $\alpha = 5$ is as follows:



Question

(i) Derive the coefficient of upper tail dependence for the Gumbel copula.

(ii) Comment on how the value of the parameter α affects the degree of upper tail dependence in the case of the Gumbel copula.

Solution

(i) ***Tail dependence of the Gumbel copula***

For the Gumbel copula setting $u = v$ gives:

$$\begin{aligned}
C[u,u] &= \exp\left\{-\left((-\ln u)^{\alpha} + (-\ln u)^{\alpha}\right)^{1/\alpha}\right\} \\
&= \exp\left\{-\left(2(-\ln u)^{\alpha}\right)^{1/\alpha}\right\} \\
&= \exp\left\{-\left(2^{1/\alpha}(-\ln u)\right)\right\} \\
&= \exp\left\{\left(2^{1/\alpha}\ln u\right)\right\} \\
&= \exp\left\{\ln u^{2^{1/\alpha}}\right\} \\
&= u^{2^{1/\alpha}}
\end{aligned}$$

The coefficient of upper tail dependence is given by:

$$\lambda_U = \lim_{u \to 1^-} \frac{1 - 2u + C[u,u]}{1-u} = \lim_{u \to 1^-} \frac{1 - 2u + u^{2^{1/\alpha}}}{1-u}$$

In the limit this fraction has the form $\dfrac{0}{0}$, which is undefined. However, we can use L'Hôpital's rule, $\lim\limits_{x \to a} \dfrac{f(x)}{g(x)} = \lim\limits_{x \to a} \dfrac{f'(x)}{g'(x)}$, to find the value of the limit:

$$\lambda_U = \lim_{u \to 1^-} \frac{1 - 2u + u^{2^{1/\alpha}}}{1-u} = \lim_{u \to 1^-} \frac{-2 + 2^{1/\alpha} u^{2^{1/\alpha}-1}}{-1} = 2 - 2^{1/\alpha}$$

(ii) ***Comment***

As α increases, $2^{1/\alpha}$ reduces and hence $2 - 2^{1/\alpha}$ increases. So increasing the value of the parameter α increases the degree of upper tail dependence of the Gumbel copula.

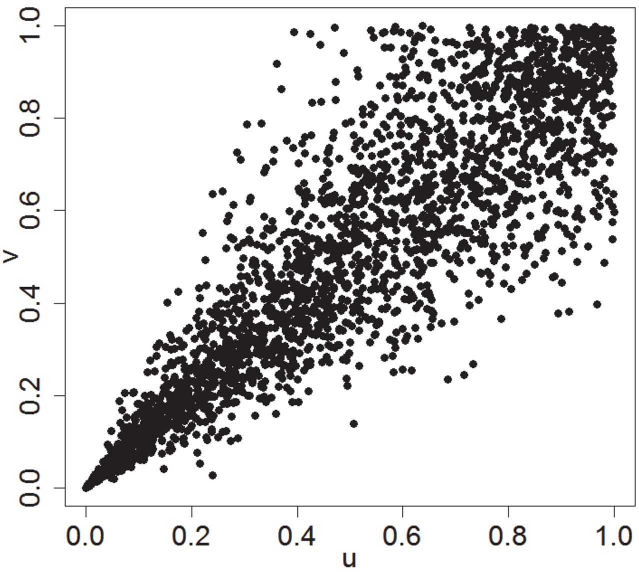
5.2 Clayton copula

The *Clayton copula* is defined in the bivariate case as:

$$C[u,v] = \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}$$

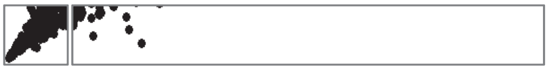
The Clayton copula describes an interdependence structure in which there is lower tail dependence (the level of which is determined by the parameter α), but there is no upper tail dependence.

A scatterplot of simulated values from the Clayton copula with parameter value $\alpha = 5$ is as follows:

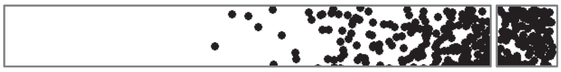


From the diagram:

- We might anticipate that there *may* be a degree of lower tail dependence because the simulations in a 'thin rectangle' at the bottom of the diagram lie mostly in the 'small square at the left-hand end' (*ie* if $V \le v$, for small values of v, then it is likely that $U \le v$).



- We might anticipate that there *may* be a lack of upper tail dependence because the simulations in a 'thin rectangle' at the top of the diagram do not lie mostly in the 'small square at the right-hand end' (*ie* if $V \ge v$, for large values of v, then we are not confident that $U \ge v$).



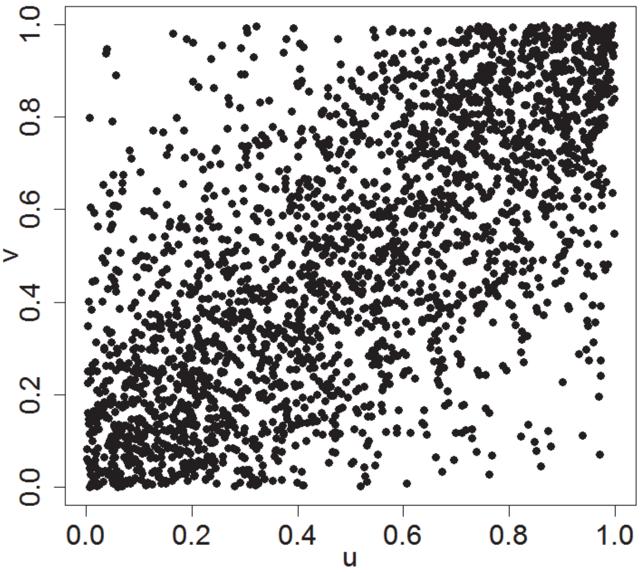
5.3 Frank copula

The *Frank copula* is defined in the bivariate case as:

$$C[u,v] = -\frac{1}{\alpha}\ln\left(1+\frac{\left(e^{-\alpha u}-1\right)\left(e^{-\alpha v}-1\right)}{\left(e^{-\alpha}-1\right)}\right)$$

The Frank copula describes an interdependence structure in which there is no upper or lower tail dependence.

A scatterplot of simulated values from the Frank copula with parameter value $\alpha = 5$ is as follows:



The lack of tail dependency is indicated in the scatterplot in that, for example, if V takes a value very close to 0 or 1, then we are uncertain whether or not U will also take a value similarly close to 0 or 1.

5.4 Archimedean copulas

A number of copulas can be specified by a special form of generator function that automatically captures the properties required for a copula. These are called *Archimedean copulas.*

Archimedean copulas are described by reference to a *generator function*. In the bivariate case, they take the form:

$$C[u,v] = \psi^{[-1]}\left(\psi(u)+\psi(v)\right)$$

where $\psi(x)$ is the generator function, and $\psi^{[-1]}$ is the *pseudo-inverse function* (explained below).

Archimedean copulas are a subset of explicit copulas. The Gumbel, Clayton and Frank copulas are all examples of Archimedean copulas.

For example, consider the Gumbel copula:

$$C[u,v]=\exp\left\{-\left((-\ln u)^{\alpha}+(-\ln v)^{\alpha}\right)^{1/\alpha}\right\}$$

We can see that the Gumbel copula is an example of an Archimedean copula where the generator function is $\psi(x)=(-\ln x)^{\alpha}$ and the inverse generator function is $\psi^{-1}(x)=\exp\left(-x^{1/\alpha}\right)$.

Before we go any further with Archimedean copulas, we need to introduce some maths relating to pseudo-inverse functions.

Pseudo-inverse functions

In order to define Archimedean copulas, we need to extend the familiar idea of an inverse function (*eg* the inverse of $f(x)=x^2$ is $f^{-1}(x)=\sqrt{x}$) to ensure that the inverse function is defined for all possible arguments. This is done by defining the *pseudo-inverse* function $\psi^{[-1]}(x)$ of a function $\psi(x)$ as:

$$\psi^{[-1]}(x)=\begin{cases}\psi^{-1}(x) & \text{if } 0\le x\le\psi(0)\\ 0 & \text{if } \psi(0)<x\le\infty\end{cases}$$

where $\psi^{-1}(x)$ denotes the ordinary inverse function obtained by inverting the equation $x=\psi(y)$ to express y in terms of x .

The pseudo-inverse function gives us a means of determining the inverse where the function ψ outputs values on a finite rather than infinite range.

If $\psi(0)=\infty$, the pseudo-inverse is always equal to the 'ordinary' inverse and the generator function is called a *strict generator function*.

General definition of an Archimedean copula

The definition of Archimedean copulas can be extended to more than 2 dimensions.

Copulas in the Archimedean family are of the form:

$$C[u_1,\ldots,u_d]=\psi^{[-1]}\left(\sum_{i=1}^{d}\psi(u_i)\right)$$

In order to be valid, **the generator function $\psi:[0,1]\to[0,\infty]$ must be a continuous, strictly decreasing, convex function with $\psi(1)=0$.**

Although this may all seem a little theoretical and complicated, the idea behind Archimedean copulas is, in fact, very intuitive. Remember that a copula function takes as inputs marginal CDFs and outputs a joint CDF. Using the generator function and its inverse, we are:

- taking probabilities between 0 and 1 (the u_i's or marginal CDFs)
- converting these to numbers greater than 0 using the generator function ψ
- summing the results
- converting the result back to a probability (*ie* the joint CDF) using the inverse function ψ^{-1}.

Note that the definition of the pseudo-inverse function ensures that, whatever the sum $\sum_{i=1}^{d}\psi(u_i)$, the value of the Archimedean copula will always be a valid probability.

In the bivariate case, we have:

$$C[u,v] = \psi^{[-1]}(\psi(u)+\psi(v))$$

We now look at how the Gumbel, Clayton and Frank copulas can be derived using the generator function approach.

Gumbel copula

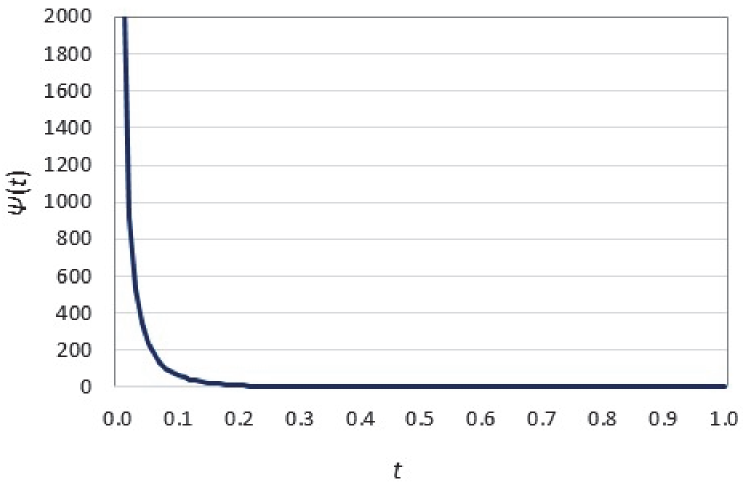
For example, the *Gumbel copula* can be defined by the generator function:

$$\psi(t) = (-\ln t)^{\alpha} \quad \text{where} \quad 1 \le \alpha < \infty$$

which we can use to deduce an explicit formula for the copula function.

In this case $\psi(0) = \lim_{t\to 0}(-\ln t)^{\alpha} = \infty$, so the pseudo-inverse function equals the normal inverse function.

The graph of the Gumbel generator function where $\alpha = 5$ is illustrated below:



We can see from this graph of the Gumbel generator function that it looks valid, *ie* it is a continuous, strictly decreasing, convex function, $\psi:[0,1]\to[0,\infty]$, with $\psi(1)=0$.

The inverse generator function can be found by inverting the relationship $x=(-\ln y)^{\alpha}$ to obtain $y=\exp\left(-x^{1/\alpha}\right)$, so that $\psi^{[-1]}(x)=\psi^{-1}(x)=\exp\left(-x^{1/\alpha}\right)$. We then have:

$$C[u,v]=\psi^{[-1]}(\psi(u)+\psi(v))=\exp\left\{-\left((-\ln u)^{\alpha}+(-\ln v)^{\alpha}\right)^{1/\alpha}\right\}$$

Question

Confirm algebraically that the Gumbel generator function, $\psi(t)=(-\ln t)^{\alpha}$, is valid.

Solution

The generator function is $\psi(t)=(-\ln t)^{\alpha}$.

- When $t=0$, $\psi(0)=\lim\limits_{t\to 0}(-\ln t)^{\alpha}=\infty$.
- When $t=1$, $\psi(1)=(-\ln 1)^{\alpha}=0$.
- In the range $0<t<1$, $\ln t$ takes increasingly negative values, so that $-\ln t$ takes decreasing positive values, and hence so does $\psi(t)=(-\ln t)^{\alpha}$.

Hence the generator function, $\psi(t)=(-\ln t)^{\alpha}$, for the Gumbel copula is valid.

Clayton copula

The *Clayton copula* is defined by the generator:

$$\psi(t)=\frac{1}{\alpha}\left(t^{-\alpha}-1\right)\quad \textbf{where}\quad -1\le\alpha<\infty$$

Frank copula

The *Frank copula* is defined by the generator:

$$\psi(t)=-\ln\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)\quad \textbf{where}\quad -\infty<\alpha<\infty$$

Independence (or product) copula

The independence (or product) copula is also Archimedean. Its generator is $\psi(t)=-\ln t$.

Question

Show that the generator function $\psi(t) = -\ln t$ generates the independence (or product) copula.

Solution

The inverse function is found by rearranging the equation $x = -\ln t$ to make t the subject of the formula:

$$x = \psi(t) = -\ln t \quad \Rightarrow \quad t = \psi^{-1}(x) = \exp(-x)$$

We substitute in the generator and inverse generator functions giving:

$$\begin{aligned} C[u,v] &= \psi^{-1}\big((-\ln u) + (-\ln v)\big) \\ &= \exp\left[-(-\ln uv)\right] \\ &= uv \end{aligned}$$

The generator function is valid because:

- $\psi(0) = \lim\limits_{t \to 0}(-\ln t) = \infty$.
- $\psi(1) = -\ln 1 = 0$.
- In the range $0 < t < 1$, $\ln t$ takes increasing values, so that $-\ln t$ takes decreasing values.

6 Implicit copulas

The final group of copulas that we consider are called implicit copulas. These copulas are based on (or implied by) well-known multivariate distributions, but no simple closed-form expression exists for them. We look at:

- the Gaussian copula (based on the multivariate normal distribution)
- the Student's t copula (based on the multivariate Student's t distribution).

6.1 Gaussian copula

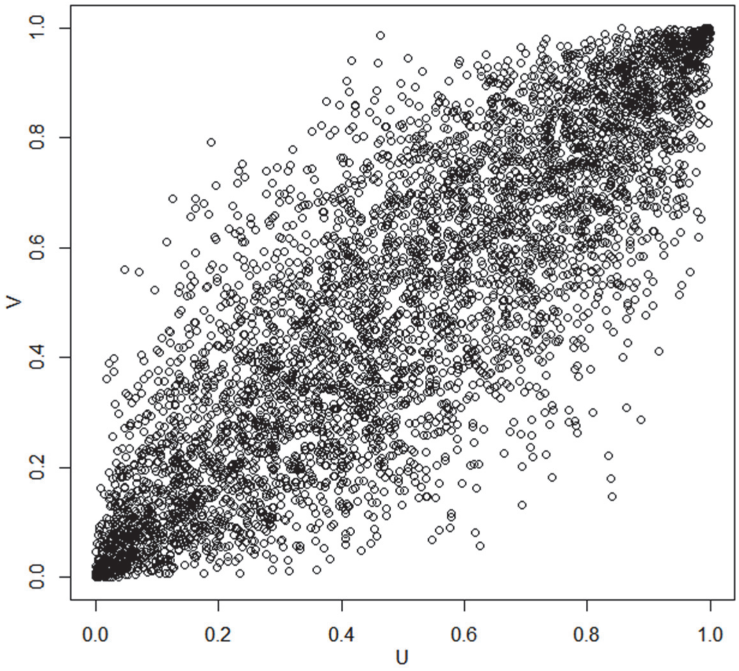
The bivariate *Gaussian copula* is defined by:

$$C[u,v] = \Phi_\rho\left[\Phi^{-1}(u), \Phi^{-1}(v)\right]$$

where Φ is the distribution function of the standard normal distribution and Φ_ρ is the distribution function of a bivariate normal distribution with correlation ρ.

Applying this Gaussian copula to normal marginal distributions will result in a bivariate normal distribution with correlation ρ.

An example scatterplot for the Gaussian copula, with a correlation parameter of $\rho = 0.85$ is:



In this example ($\rho = 0.85$):

- if $u = 1$, then it is extremely unlikely that we observe $v = 0$
- if $u = 0$, then it is extremely unlikely that we observe $v = 1$.

The independence, co-monotonic and counter-monotonic copulas are special cases of the Gaussian copula where $\rho = 0$, $\rho = +1$ and $\rho = -1$, respectively.

Unless the correlation parameter is equal to 1, *ie* unless the two random variables are perfectly positively interdependent, there is no tail dependence (upper or lower) exhibited by the Gaussian copula.

The bivariate Gaussian copula can be extended to the multivariate case incorporating the $d \times d$ correlation matrix of the individual random variables. So this is the unique copula that reproduces a joint normal distribution with a specified correlation matrix from the individual marginal distributions. Because it reproduces the joint distribution in this way, it is sometimes called an *implicit* copula.

The formula defining the bivariate Gaussian copula is mathematically equivalent to the following integral form:

$$C[u,v] = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(u)} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left(s^2 + t^2 - 2\rho st\right)\right\} ds\, dt$$

This can be simplified further to:

$$C[u,v] = \int_0^u \Phi\left(\frac{\Phi^{-1}(v) - \rho\Phi^{-1}(t)}{\sqrt{1-\rho^2}}\right) dt$$

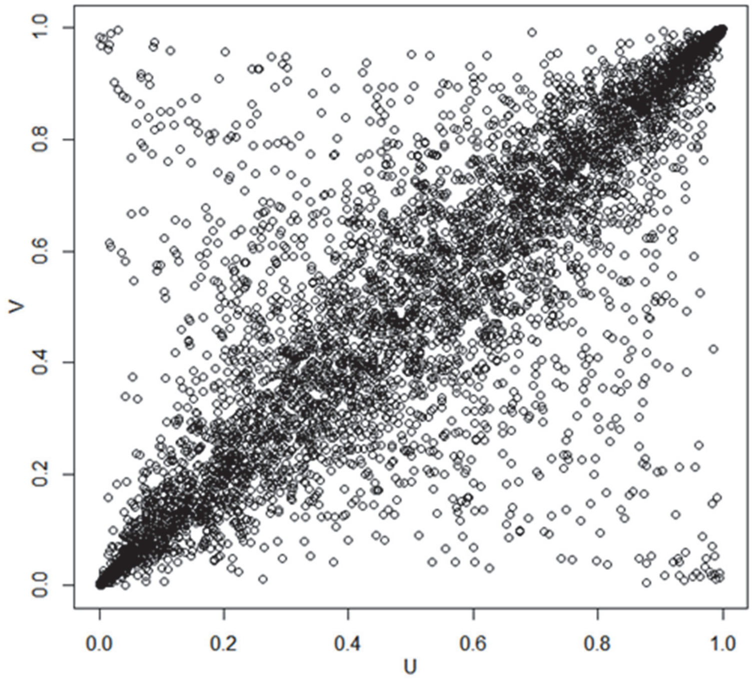
6.2 Student's *t* copula

The *Student's t copula* is defined by:

$$C[u,v] = t_{\gamma,\rho}\left[t_\gamma^{-1}(u), t_\gamma^{-1}(v)\right]$$

where t_γ is the distribution function of a random variable with a Student's t distribution with γ degrees of freedom and $t_{\gamma,\rho}$ is the distribution function of a bivariate Student's t distribution with correlation ρ.

An example scatterplot for the Student's t copula, with a correlation coefficient of $\rho = 0.85$ and $\gamma = 1$, is shown below.



There some simulations in the upper-left and lower-right corners of this scatterplot, which was not the case in the example scatterplot for the Gaussian copula that we saw earlier (and which was based on the same correlation parameter).

The Student's t copula allows the dependencies between the variables to be adjusted more finely than the corresponding Gaussian copula.

One of the disadvantages of the Gaussian copula, is that it has just a single parameter, ρ. The Student's t copula has an additional parameter, γ, the number of degrees of freedom, which can be used to vary the strength of the association between the two variables, *ie* the degree of tail dependence. The smaller the value of γ, the greater the level of tail dependence.

In the same way that the standard normal distribution is a limiting case of the Student's t distribution (as the number of degrees of freedom tends to infinity), the Gaussian copula is the limiting case of the *Student's t copula*.

Specifically, as the number of degrees of freedom tends to infinity, $\gamma \to \infty$, the Student's t copula tends approaches the Gaussian copula.

7 Choosing and fitting a suitable copula function

7.1 Choosing a suitable copula function

If we want to create a mathematical model to represent real-world phenomena then we might look at past data and:

- select and parameterise marginal distributions for each of the relevant variables, and
- describe and quantify the form and extent of the associations between the variables.

Examination of the form and levels of association between the variables of interest allows us to select a suitable candidate copula from the list of established copulas or to develop a new bespoke copula.

Different copulas result in different levels of tail dependence.

For example:

- **the Frank copula and the Gaussian copula have zero dependence in both tails, while the Student's t copula has equal positive dependence in both tails.**
- **the Gumbel copula has zero lower tail dependence but upper tail dependence of $2-2^{1/\alpha}$. The Clayton copula, on the other hand, has zero upper tail dependence but lower tail dependence of $2^{-1/\alpha}$.**

As we would expect, variables related by the independence (product) copula have a concordance of 0, whereas variables related by the co-monotonic (minimum) or counter-monotonic (maximum) copulas have a concordance of +1 and –1 respectively.

We can summarise the upper and lower tail dependence results in the table below:

Copula name	$_L\lambda$	$_U\lambda$
Independence	0	
Co-monotonic	1	
Counter-monotonic	0	
Gumbel	0	$2-2^{1/\alpha}$
Clayton	$2^{-1/\alpha}$ if $\alpha>0$ 0 if $\alpha\le 0$	0
Frank	0	
Gaussian	0 if $\rho<1$ 1 if $\rho=1$	
Student's t	>0 if $\gamma<\infty$, increasing as γ decreases 0 if $\gamma=\infty$ and $\rho\neq 1$ 1 for all γ when $\rho=1$	

So the Gumbel copula, with an appropriate value for the parameter α, might be a suitable copula to use when modelling large general insurance claims resulting from a common underlying cause.

Question

State an appropriate copula to use if the data exhibit the following features:

(a) independence

(b) high upper tail dependence, but no lower tail dependence

(c) a high degree of positive interdependence throughout

(d) high lower tail dependence, but no upper tail dependence

(e) a high degree of negative interdependence throughout

(f) no upper or lower tail dependence

(g) both upper and lower tail dependence.

Solution

(a) the independence (or product) copula

(b) the Gumbel copula

(c) the co-monotonic copula (perfect positive interdependence), or Frank copula (with a high positive parameter to give a strong positive association throughout)

(d) the Clayton copula

(e) the counter-monotonic copula, (perfect negative interdependence), or Frank copula (with a high negative parameter to give a strong negative association throughout)

(f) the Frank copula, Gaussian copula, or independence copula

(g) the Student's t copula

7.2 Fitting a copula

This short section is beyond the syllabus of Subject CS2.

We have seen above how the form and level of association between the variables (including the degree of lower and upper tail dependence) can help us choose a particular copula. The final step is how to parameterise that chosen copula. For example, if we have chosen to use the Gumbel copula in our model, we then have to choose a suitable value of the parameter α.

For Archimedean copulas, Kendall's tau is a function of the parameters of the copula. Therefore, we can fit Gumbel, Clayton and Frank copulas using a method of moments approach. We would calculate Kendall's tau for the observations and equate this to the formula for Kendall's tau for the copula, solving to obtain the fitted value of α .

For example, for the Gumbel copula, Kendall's tau is given by the formula:

$$\tau = 1 - \frac{1}{\alpha}$$

If the observed value of Kendall's tau is 0.75, then using the method of moments gives:

$$1 - \frac{1}{\hat{\alpha}} = 0.75 \Rightarrow \frac{1}{\hat{\alpha}} = 0.25$$

$$\Rightarrow \hat{\alpha} = 4$$

For the Clayton copula, Kendall's tau is given by the formula:

$$\tau = \frac{\alpha}{\alpha + 2}$$

If the observed value of Kendall's tau is 0.75, then using the method of moments gives:

$$\frac{\hat{\alpha}}{\hat{\alpha} + 2} = 0.75 \Rightarrow \hat{\alpha} = 0.75\hat{\alpha} + 1.5$$

$$\Rightarrow 0.25\hat{\alpha} = 1.5$$

$$\Rightarrow \hat{\alpha} = 6$$

A similar method can be used for the Frank copula, although the formula for Kendall's tau is more complicated.

It is also possible to use the method of maximum likelihood estimation to fit the copula parameters.

8 Calculating probabilities using copulas

Recall that a copula:

- is a function
- takes marginal probabilities of random variables as inputs, and outputs a corresponding joint probability
- for a bivariate distribution is defined by:

$$C_{XY}\left[F_X(x),F_Y(y)\right]=P(X\le x,Y\le y)=F_{X,Y}(x,y)$$

which is often written in the more compact form:

$$C[u,v]=F_{X,Y}(x,y) \quad \text{where } u=F_X(x) \text{ and } v=F_Y(y)$$

Question

Let $X=$ a person's height measured in *cm*, and $Y=$ weight measured in *kg*. Heights and weights are each assumed to be normally distributed, and:

$$P(X\le 180)=0.81594 \quad \text{and} \quad P(Y\le 70)=0.69146$$

(i) Calculate the joint probability that a person's height is less than or equal to 180*cm* and that their weight is less than or equal to 70*kg* using:

(a) the independence (or product) copula

(b) the Gaussian copula with $\rho=0$.

The following table is required for (i)(a). It shows an excerpt of values from the bivariate standard normal cumulative distribution function: $\Phi_{\rho=0}(x,y)$.

							X					
	Φ(x,y)	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
	0	0.2500	0.2699	0.2896	0.3090	0.3277	0.3457	0.3629	0.3790	0.3941	0.4080	0.4207
	0.1	0.2699	0.2914	0.3127	0.3336	0.3538	0.3733	0.3918	0.4092	0.4255	0.4405	0.4542
	0.2	0.2896	0.3127	0.3355	0.3579	0.3797	0.4005	0.4204	0.4391	0.4565	0.4726	0.4874
	0.3	0.3090	0.3336	0.3579	0.3818	0.4050	0.4273	0.4484	0.4684	0.4870	0.5042	0.5199
	0.4	0.3277	0.3538	0.3797	0.4050	0.4296	0.4532	0.4757	0.4968	0.5166	0.5348	0.5514
Y	**0.5**	0.3457	0.3733	0.4005	0.4273	0.4532	0.4781	0.5018	0.5242	0.5450	0.5642	0.5818
	0.6	0.3629	0.3918	0.4204	0.4484	0.4757	0.5018	0.5267	0.5501	0.5720	0.5922	0.6106
	0.7	0.3790	0.4092	0.4391	0.4684	0.4968	0.5242	0.5501	0.5746	0.5974	0.6185	0.6378
	0.8	0.3941	0.4255	0.4565	0.4870	0.5166	0.5450	0.5720	0.5974	0.6212	0.6431	0.6631
	0.9	0.4080	0.4405	0.4726	0.5042	0.5348	0.5642	0.5922	0.6185	0.6431	0.6658	0.6865
	1	0.4207	0.4542	0.4874	0.5199	0.5514	0.5818	0.6106	0.6378	0.6631	0.6865	0.7079

(ii) Compare the two results in (i).

Solution

We have $u = P(X \le 180) = 0.81594$ and $v = P(Y \le 70) = 0.69146$.

(i)(a) ***Independence (or product) copula***

The independence (or product) copula is given by:

$$C[u,v] = uv$$

The joint probability that a person's height is less than or equal to 180*cm* and their weight is less than or equal to 70*kg* is given by the independence (or product) copula as:

$$P(X \le 180, Y \le 70) = 0.81594 \times 0.69146 = 0.5642$$

(i)(b) ***Gaussian copula with perfect positive correlation***

The Gaussian copula is given by:

$$C[u,v] = \Phi_\rho\left[\Phi^{-1}(u), \Phi^{-1}(v)\right]$$

The joint probability that a person's height is less than or equal to 180*cm* and their weight is less than or equal to 70*kg* is given by the Gaussian copula with $\rho = 0$ as:

$$P(X \le 180, Y \le 70) = \Phi_\rho\left[\Phi^{-1}(0.81594), \Phi^{-1}(0.69146)\right]$$

From page 160 of the *Tables*, we have:

$$\Phi^{-1}(0.81594) = 0.90$$

and $\Phi^{-1}(0.69146) = 0.50$

Therefore:

$$P(X \le 180, Y \le 70) = \Phi_\rho[0.90, 0.50]$$

Using the excerpt provided in the question for the bivariate standard normal CDF:

$$P(X \le 180, Y \le 70) = \Phi_\rho[0.90, 0.50] = 0.5642$$

(ii) ***Compare the results***

The results of both calculations are the same. This is because, for the multivariate Gaussian distribution, a correlation coefficient of 0 implies independence.

The chapter summary starts on the next page so that you can keep all the chapter summaries together for revision purposes.

Chapter 17 Summary

Association and concordance

Variables are said to be associated if there is some form of statistical relationship between them, whether causal or not.

Coefficients of association are generally designed so that their values vary between −1 and +1. Their absolute values increase with the strength of the relationship. They take a value of +1 (or −1) when there is perfect positive (or negative) association. A positive association between two variables does not necessarily imply that one is *dependent* on the other.

Pearson's correlation coefficient measures the degree to which there is a linear relationship between the variables.

Concordance is another particular form of association. Broadly speaking, two random variables are concordant if small values of one are likely to be associated with small values of the other, and vice versa.

Spearman's rho and Kendall's tau are two examples of measures of concordance.

Definition of a copula and properties

A copula function takes as its inputs the marginal cumulative distribution functions, and outputs a joint cumulative distribution function. A copula in *d*-dimensions is expressed as:

$$C\left[u_1,u_2,\ldots,u_d\right]=F_{X_1,X_2,\ldots,X_d}\left(x_1,x_2,\ldots,x_d\right) \text{ where } u_i=F_{X_i}\left(x_i\right).$$

Copulas provide a way of deconstructing the joint distribution of a set of variables into components (the marginal distributions plus a copula). This means that we can see the nature of the interdependence between the variables.

A copula is determined by the relative order (ranking) of the observations rather than by the exact shape of the marginal distribution.

A copula has three properties:

1. It must be an increasing function of its inputs.
2. If the values of all but one of the marginal CDFs are equal to 1, then the copula is equal to the value of the remaining marginal CDF.
3. The copula must always return a valid probability.

Sklar's theorem

Sklar's theorem says that if F is a joint CDF and $F_1,...,F_d$ are marginal CDFs, then there exists a copula C, such that for all $x_1,\ldots,x_d \in [-\infty,\infty]$:

$$F(x_1,\ldots,x_d) = C\left[F_1(x_1),\ldots,F_d(x_d)\right]$$

Furthermore if the marginal distributions are continuous, then the copula is unique.

Fundamental copulas

Fundamental copulas represent the three basic (or fundamental) dependencies that a set of variables can display, namely independence, perfect positive interdependence, and perfect negative interdependence. The copulas are:

Independence (or product) copula

The independence (or product) copula is defined in the bivariate case as:

$$C[u,v] = uv$$

Co-monotonic (or minimum copula)

The co-monotonic copula is defined in the bivariate case as:

$$C[u,v] = \min(u,v)$$

Counter-monotonic (or maximum copula)

The counter-monotonic copula is defined in the bivariate case as:

$$C[u,v] = \max(u+v-1,0)$$

The minimum and maximum copulas form the upper and lower bounds for all copulas, called the Fréchet-Höffding bounds.

Explicit copulas

Explicit copulas have simple closed-form expressions. An important subclass is that of Archimedean copulas. Archimedean copulas take the form:

$$C[u,v] = \psi^{[-1]}\left(\psi(u)+\psi(v)\right)$$

where $\psi(t)$ is a generator function and $\psi^{[-1]}$ is a pseudo-inverse generator function.

Three examples of Archimedean copulas are: the Gumbel, Clayton and Frank copulas.

Gumbel copula

The Gumbel copula is defined in the bivariate case as:

$$C[u,v] = \exp\left\{-\left((-\ln u)^{\alpha} + (-\ln v)^{\alpha}\right)^{1/\alpha}\right\}$$

The generator function is $\psi(t) = (-\ln t)^{\alpha}$ where $1 \le \alpha < \infty$.

Clayton copula

The Clayton copula is defined in the bivariate case as:

$$C[u,v] = \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}$$

The generator function is $\psi(t) = \dfrac{1}{\alpha}\left(t^{-\alpha} - 1\right)$ where $-1 \le \alpha < \infty$.

Frank copula

The Frank copula is defined in the bivariate case as:

$$C[u,v] = -\frac{1}{\alpha}\ln\left(1 + \frac{\left(e^{-\alpha u} - 1\right)\left(e^{-\alpha v} - 1\right)}{\left(e^{-\alpha} - 1\right)}\right)$$

The generator function is $\psi(t) = -\ln\left(\dfrac{e^{-\alpha t} - 1}{e^{-\alpha} - 1}\right)$ where $-\infty < \alpha < \infty$.

Implicit copulas

Implicit copulas are based on well-known multivariate distributions, but no simple closed-form expression exists for them. Examples are the Gaussian copula and the Student's *t*-copula.

Gaussian copula

The Gaussian copula is defined in the bivariate case as:

$$C[u,v] = \Phi_{\rho}\left[\Phi^{-1}(u), \Phi^{-1}(v)\right]$$

where Φ is the distribution function of the standard normal distribution and Φ_{ρ} is the distribution function of a bivariate standard normal distribution with correlation ρ.

Combining marginal normal random variables using the Gaussian copula results in the multivariate normal distribution.

The independence, co-monotonic and counter-monotonic copulas are special cases of the Gaussian copula where $\rho=0$, $\rho=+1$ and $\rho=-1$ respectively.

Student's t copula

The Student's t copula defined by:

$$C\left[u,v\right]=t_{\gamma,\rho}\left[t_{\gamma}^{-1}(u),t_{\gamma}^{-1}(v)\right]$$

where t_{γ} is the distribution function of a random variable with a Student's t distribution with γ degrees of freedom and $t_{\gamma,\rho}$ is the distribution function of a bivariate Student's t distribution with correlation ρ.

The Student's t copula allows for more flexibility than the normal copula since it involves an extra parameter, namely the number of degrees of freedom, γ.

The Gaussian copula is the limiting case of the Student's t copula as the number of degrees of freedom tends to infinity.

Tail dependence of copula functions

Tail dependence looks at the relationship between two random variables at the extremes of the distributions, *ie* in the upper and lower tails. Tail dependence can be measured using the coefficients of lower and upper tail dependence, which can be calculated directly from the copula function.

The coefficient of lower tail dependence is defined as:

$$\lambda_L=\lim_{u\to0^+}P\left(X\le F_X{}^{-1}(u)\,\middle|\,Y\le F_Y{}^{-1}(u)\right)=\lim_{u\to0^+}\frac{C\left[u,u\right]}{u}$$

The coefficient of upper tail dependence is defined as:

$$\lambda_U=\lim_{u\to1^-}P\left(X>F_X{}^{-1}(u)\,\middle|\,Y>F_Y{}^{-1}(u)\right)=\lim_{u\to1^-}\frac{1-2u+C\left[u,u\right]}{1-u}$$

or as:

$$\lambda_U=\lim_{u\to1^-}\frac{\bar{C}\left[1-u,1-u\right]}{1-u}=\lim_{u\to0^+}\frac{\bar{C}\left[u,u\right]}{u}$$

where $\bar{C}$ is the survival copula, defined as:

$$\bar{C}\left[1-u,1-v\right]=1-u-v+C\left[u,v\right]$$

For the copulas discussed in this chapter, the table below summarises the coefficients of lower and upper tail dependency.

Copula name	$_L\lambda$	$_U\lambda$
Independence	0	
Co-monotonic	1	
Counter-monotonic	0	
Gumbel	0	$2-2^{1/\alpha}$
Clayton	$2^{-1/\alpha}$ if $\alpha>0$ 0 if $\alpha\le 0$	0
Frank	0	
Gaussian	0 if $\rho<1$ 1 if $\rho=1$	
Student's t	>0 if $\gamma<\infty$, increasing as γ decreases 0 if $\gamma=\infty$ and $\rho\neq 1$ 1 for all γ when $\rho=1$	

The degree of concordance and the level of tail dependencies exhibited by a particular set of data helps to indicate which copula(s) might be appropriate to consider using.

The practice questions start on the next page so that you can keep the chapter summaries together for revision purposes.

Chapter 17 Practice Questions

17.1 List, in words, the three technical properties which a copula function must satisfy to ensure that it correctly captures the properties expected of a joint distribution function.

17.2 Exam style An investor purchases three 5-year bonds from different companies within the same industry sector. The probability that an individual bond defaults within the first year is 10%.

(i) Using a Gumbel copula with parameter $\alpha = 2$, calculate the probability that all three bonds default within the first year. [3]

(ii) Discuss the suitability of the Gumbel copula in this situation. [3]

[Total 6]

17.3 For the Clayton copula:

Exam style (i) Determine whether the generator function $\psi(t) = \frac{1}{\alpha}\left(t^{-\alpha} - 1\right)$ is valid. [3]

(ii) Determine the inverse generator function. [1]

(iii) Derive the Clayton copula function in the bivariate case. [3]

[Total 7]

17.4 For the Frank copula:

Exam style (i) Determine whether the generator function $\psi(t) = -\ln\left(\frac{e^{-\alpha t} - 1}{e^{-\alpha} - 1}\right)$ is valid. [3]

(ii) Determine the inverse generator function. [1]

(iii) Derive the Frank copula function. [3]

[Total 7]

17.5 Exam style (i) Derive the coefficient of lower and upper tail dependence for the Clayton copula in the case where the parameter $\alpha > 0$. [4]

(ii) Comment on how the value of the parameter α affects the degree of lower tail dependence in the case of the Clayton copula. [1]

[Total 5]

17.6 Exam style Derive the coefficient of lower tail dependence for the Gumbel copula in the case where the parameter $\alpha > 0$. [4]

17.7 Let X and Y be two random variables representing the future lifetimes of two 40-year old individuals. The two lives are married.

Exam style

You are given that:

$$P(X \le 25) = 0.17831 \text{ and } P(Y \le 25) = 0.11086$$

(i) Calculate the joint probability that both lives will die by the age of 65 using:

(a) the Gumbel copula with $\alpha = 5$

(b) the Clayton copula with $\alpha = 5$

(c) the Frank copula with $\alpha = 5$. [6]

(ii) Comment on the results as well as on which copula you think is most appropriate to use for modelling joint life expectancy. [3]

[Total 9]

17.8 You are considering whether there is a link between heights and weights, and have gathered some pairs of data:

$$(172cm, 68kg),\ (182cm, 70kg),\ (158cm, 75kg),\ (150cm, 60kg),\ (174cm, 65kg)$$

Calculate Spearman's rho for this dataset.

Chapter 17 Solutions

17.1 Three technical properties a copula function must satisfy are:

1. A copula is an increasing function of its inputs.

2. If all the marginal CDFs are equal to 1 except for one of the marginal CDFs then the copula function is equal to the value of that one marginal CDF.

3. A copula function always returns a valid probability.

17.2 (i) ***Calculation of probability all three bonds default within the next year.***

Let T_i be the time until default of bond i where $i=1,2,3$. We want to calculate the joint probability:

$$P(T_1 \leq 1, T_2 \leq 1, T_3 \leq 1) = C[u_1, u_2, u_3]$$

where $u_i = P[T_i \leq 1] = 0.1$ for $i=1,2,3$. [1]

Using the Gumbel copula with parameter $\alpha = 2$, we have:

$$\begin{aligned} P(T_1 \leq 1, T_2 \leq 1, T_3 \leq 1) &= C[u_1, u_2, u_3] \\ &= \exp\left\{-\left((-\ln u_1)^2 + (-\ln u_2)^2 + (-\ln u_3)^2\right)^{1/2}\right\} \\ &= \exp\left\{-\left(3(-\ln 0.1)^2\right)^{1/2}\right\} \\ &= 0.0185 \text{ or } 1.85\% \end{aligned}$$

[2]

(ii) ***Suitability of the Gumbel copula***

The Gumbel copula exhibits (non-zero) upper-tail dependence, the degree of which can be varied by adjusting the single parameter. However, it exhibits no lower tail dependence. [1]

Hence, the Gumbel copula is appropriate if we believe that the three investments are likely to behave similarly as the term approaches five years but not at early durations. [½]

This is unlikely to be the case though. If one bond defaults early on, then it may be indicative of problems in the industry sector or the economy and so the other investments may also be likely to default early on. [½]

If we believe the performance of investments issued by companies within the industry are much more closely associated (*eg* subject to the same systemic and operational risk factors), then a copula that exhibits both lower and upper tail dependence, such as the Student's t copula, may be more appropriate. [1]

17.3 (i) ***Is the Clayton generator function valid?***

- $\psi(0) = \lim\limits_{t\to 0} \dfrac{1}{\alpha}\left(t^{-\alpha} - 1\right) = \dfrac{1}{\alpha}\lim\limits_{t\to 0}\left(\dfrac{1}{t^{-\alpha}} - 1\right) = \infty$ [1]

- $\psi(1) = \dfrac{1}{\alpha}\left(1^{-\alpha} - 1\right) = 0$ [1]

- In the range $0 < t < 1$, t^{α} takes increasing values (starting at 1), so that $t^{-\alpha}$ takes decreasing values, and hence so does $\psi(t) = \dfrac{1}{\alpha}\left(t^{-\alpha} - 1\right)$. [1]

(ii) ***Inverse generator function***

The inverse function is found by rearranging the equation $x = \dfrac{1}{\alpha}\left(t^{-\alpha} - 1\right)$:

$$x = \psi(t) = \frac{1}{\alpha}\left(t^{-\alpha} - 1\right) \quad \Rightarrow \quad t = \psi^{-1}(x) = (1 + \alpha x)^{-1/\alpha}$$ [1]

(iii) ***Derive the Clayton copula function***

$$C[u,v] = \psi^{-1}\left[\frac{1}{\alpha}\left(u^{-\alpha} - 1\right) + \frac{1}{\alpha}\left(v^{-\alpha} - 1\right)\right]$$

$$= \left[1 + \alpha\left(\frac{1}{\alpha}\left(u^{-\alpha} - 1\right) + \frac{1}{\alpha}\left(v^{-\alpha} - 1\right)\right)\right]^{-1/\alpha}$$

$$= \left[1 + \left(u^{-\alpha} - 1\right) + \left(v^{-\alpha} - 1\right)\right]^{-1/\alpha}$$

$$= \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}$$ [3]

17.4 (i) ***Is the Frank generator function valid?***

- $\psi(0) = \lim\limits_{t\to 0}\left[-\ln\left(\dfrac{e^{-\alpha t} - 1}{e^{-\alpha} - 1}\right)\right] = \infty$ [1]

- $\psi(1) = -\ln\left(\dfrac{e^{-\alpha} - 1}{e^{-\alpha} - 1}\right) = 0$ [1]

- In the range $0 < t < 1$, $\dfrac{e^{-\alpha t} - 1}{e^{-\alpha} - 1}$ takes increasing values, so that $\ln\left(\dfrac{e^{-\alpha t} - 1}{e^{-\alpha} - 1}\right)$ takes increasing values, and $\psi(t) = -\ln\left(\dfrac{e^{-\alpha t} - 1}{e^{-\alpha} - 1}\right)$ takes decreasing values. [1]

(ii) ***Inverse generator function***

The inverse function is found by rearranging the equation $x=-\ln\left(\dfrac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$:

$$x=\psi(t)=-\ln\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right) \Rightarrow t=\psi^{-1}(x)=-\frac{1}{\alpha}\ln\left[1+\left(e^{-\alpha}-1\right)e^{-x}\right] \qquad [1]$$

(iii) ***Derive the Frank copula function***

$$C[u,v]=\psi^{-1}\left[\left(-\ln\left(\frac{e^{-\alpha u}-1}{e^{-\alpha}-1}\right)\right)+\left(-\ln\left(\frac{e^{-\alpha v}-1}{e^{-\alpha}-1}\right)\right)\right]$$

$$=\psi^{-1}\left[-\ln\left(\left(\frac{e^{-\alpha u}-1}{e^{-\alpha}-1}\right)\left(\frac{e^{-\alpha v}-1}{e^{-\alpha}-1}\right)\right)\right]$$

$$=-\frac{1}{\alpha}\ln\left[1+\left(e^{-\alpha}-1\right)\left(\frac{e^{-\alpha u}-1}{e^{-\alpha}-1}\right)\left(\frac{e^{-\alpha v}-1}{e^{-\alpha}-1}\right)\right]$$

$$=-\frac{1}{\alpha}\ln\left[1+\frac{\left(e^{-\alpha u}-1\right)\left(e^{-\alpha v}-1\right)}{\left(e^{-\alpha}-1\right)}\right] \qquad [3]$$

17.5 (i)(a) ***Coefficients of lower and upper tail dependence – Clayton copula***

The coefficient of lower tail dependence is defined as:

$$\lambda_L=\lim_{u\to0^+}\frac{C[u,u]}{u} \qquad [½]$$

Substituting in for the Clayton copula formula:

$$\lambda_L=\lim_{u\to0^+}\left[\frac{\left(u^{-\alpha}+u^{-\alpha}-1\right)^{-1/\alpha}}{u}\right]$$

$$=\lim_{u\to0^+}\left[\frac{\left(2u^{-\alpha}-1\right)^{-1/\alpha}}{u}\right]$$

$$=\lim_{u\to0^+}\left[\left(2u^{-\alpha}-1\right)^{-1/\alpha}\left(u^{\alpha}\right)^{-1/\alpha}\right]$$

$$=\lim_{u\to0^+}\left[\left(2-u^{\alpha}\right)^{-1/\alpha}\right]=2^{-1/\alpha} \qquad [1½]$$

The coefficient of upper tail dependence is given by:

$$\lambda_U = \lim_{u \to 1^-} \frac{1-2u+C[u,u]}{1-u}$$

[½]

From the calculation of the coefficient of lower tail dependence above, we have:

$$C[u,u] = \left(2u^{-\alpha}-1\right)^{-1/\alpha}$$

Therefore:

$$\lambda_U = \lim_{u \to 1^-} \frac{1-2u+\left(2u^{-\alpha}-1\right)^{-1/\alpha}}{1-u}$$

[½]

In the limit this fraction has the form $\frac{0}{0}$, which is undefined. We can use L'Hôpital's rule, $\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}$, to find the value of the limit:

$$\begin{aligned}
\lambda_U &= \lim_{u \to 1^-} \frac{1-2u+\left(2u^{-\alpha}-1\right)^{-1/\alpha}}{1-u} \\
&= \lim_{u \to 1^-} \frac{-2-\frac{1}{\alpha}\left(2u^{-\alpha}-1\right)^{-\frac{1}{\alpha}-1}\left(-2\alpha u^{-\alpha-1}\right)}{-1} \\
&= \lim_{u \to 1^-} \frac{-2+\left(2u^{-\alpha}-1\right)^{-\frac{1}{\alpha}-1}\left(2u^{-\alpha-1}\right)}{-1} = 0
\end{aligned}$$

[1]

(ii) ***Comment***

As α increases, $-\frac{1}{\alpha}$ increases and hence $2^{-1/\alpha}$ increases. So the higher the value of the parameter α, the higher the degree of upper tail dependence. [1]

17.6 *Lower tail dependence of the Gumbel copula*

The Gumbel copula is expressed as:

$$C\left[u,v\right]=\exp\left\{-\left((-\ln u)^{\alpha}+(-\ln v)^{\alpha}\right)^{1/\alpha}\right\}$$ [1]

The coefficient of lower tail dependence is given by:

$$\begin{aligned}\lambda_L &= \lim_{u\to 0^+}\frac{C\left[u,u\right]}{u}\\ &= \lim_{u\to 0^+}\left[\frac{\exp\left\{-\left((-\ln u)^{\alpha}+(-\ln u)^{\alpha}\right)^{1/\alpha}\right\}}{u}\right]\\ &= \lim_{u\to 0^+}\left[\frac{\exp\left\{-\left(2^{1/\alpha}(-\ln u)\right)\right\}}{u}\right]\\ &= \lim_{u\to 0^+}\left[\frac{\exp\left\{\left(2^{1/\alpha}\ln u\right)\right\}}{u}\right]\\ &= \lim_{u\to 0^+}\left[\frac{u^{2^{1/\alpha}}}{u}\right]=\lim_{u\to 0^+}u^{2^{1/\alpha-1}}=0\end{aligned}$$ [3]

17.7 *Calculating probabilities using explicit copulas*

We have $u=P\left(X\le 25\right)=0.17831$ and $v=P\left(Y\le 25\right)=0.11086$.

(i)(a) ***Gumbel copula***

The joint probability that both lives die by age 65 is given by the Gumbel copula with $\alpha=5$ as:

$$\begin{aligned}P\left(X\le 25,Y\le 25\right)&=C\left[u,v\right]\\ &=\exp\left\{-\left((-\ln u)^{\alpha}+(-\ln v)^{\alpha}\right)^{1/\alpha}\right\}\\ &=\exp\left\{-\left((-\ln 0.17831)^{5}+(-\ln 0.11086)^{5}\right)^{1/5}\right\}\\ &=0.0986\end{aligned}$$ [2]

(i)(b) *Clayton copula*

The joint probability that both lives die by age 65 is given by the Clayton copula with $\alpha = 5$ as:

$$\begin{aligned} P(X \le 25, Y \le 25) &= C[u,v] \\ &= \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha} \\ &= \left(0.17831^{-5} + 0.11086^{-5} - 1\right)^{-1/5} \\ &= 0.1089 \end{aligned}$$

[2]

(i)(c) *Frank copula*

The joint probability that both lives die by age 65 is given by the Frank copula with $\alpha = 5$ as:

$$\begin{aligned} P(X \le 25, Y \le 25) &= C[u,v] \\ &= -\frac{1}{\alpha}\ln\left(1 + \frac{\left(e^{-\alpha u} - 1\right)\left(e^{-\alpha v} - 1\right)}{\left(e^{-\alpha} - 1\right)}\right) \\ &= -\frac{1}{5}\ln\left(1 + \frac{\left(e^{-5\times 0.17831} - 1\right)\left(e^{-5\times 0.11086} - 1\right)}{\left(e^{-5} - 1\right)}\right) \\ &= 0.0583 \end{aligned}$$

[2]

(ii) *Comment*

The Clayton copula gives the highest probability of both lives dying within 25 years. This is because the Clayton copula exhibits lower tail dependence. This means that if one life does not survive for long (*ie* dies early), there is a high probability that the other life will not survive for long (*ie* will also die early). [1]

The Gumbel copula gives the lowest probability of both lives dying within 25 years. This is because the Gumbel copula exhibits upper tail dependence. This means that if one life survives for a long time, there is a high probability that the other life will also survive for a long time. [1]

Studies also suggest that if one member of a married couple dies, this can precipitate the death of the other member ('broken heart syndrome'). On this basis, we might choose to use a copula function where there is a degree of positive interdependence throughout, *eg* the co-monotonic (or minimum) copula. [1]

Although we used the same parameter $\alpha = 5$ in each of the three copula functions, the effect of this parameter on the calculation will vary depending on the copula, and so the results are not directly comparable.

17.8 *Spearman's rho*

Height	*Rank*	*Weight*	*Rank*	d_i	d_i^2
150cm	1	60kg	1	0	0
158cm	2	75kg	5	-3	9
172cm	3	68kg	3	0	0
174cm	4	65kg	2	2	4
182cm	5	70kg	4	1	1

$$_S\rho = 1 - \frac{6}{T(T^2-1)}\sum_{i=1}^{T} d_i^2 = 1 - \frac{6}{5(5^2-1)} \times 14 = 0.3$$

18

Reinsurance

Syllabus objectives

1.1 Loss distributions, with and without risk sharing

1.1.2 Explain the concepts of excesses (deductibles), and retention limits.

1.1.3 Describe the operation of simple forms of proportional and excess of loss reinsurance.

1.1.4 Derive the distribution and corresponding moments of the claim amounts paid by the insurer and the reinsurer in the presence of excesses (deductibles) and reinsurance.

1.1.5 Estimate the parameters of a failure time or loss distribution when the data is complete, or when it is incomplete, using maximum likelihood and the method of moments.

0 Introduction

The claims on an insurance company must be met in full, but, to protect itself from large claims, the company itself may take out an insurance policy; such a policy is called a reinsurance policy. For the purposes of this chapter, it will be assumed that the reinsurance contract is one of two very simple types: individual excess of loss reinsurance or proportional reinsurance.

0.1 Proportional reinsurance

Under a proportional reinsurance arrangement, the *direct writer* (*ie* the original insurance company) and the reinsurer share the cost of all claims for each risk. For example, for a particular building insured against fire, the direct writer might retain 75% of the premium and will be liable to pay 75% of all claims, large or small. The direct writer must pay a premium to effect this reinsurance. The direct writer is sometimes referred to as the *direct insurer* or even just the *insurer*.

Proportional reinsurance operates in two forms:

1. With *quota share* reinsurance, the proportions are the same for all risks.

2. With *surplus* reinsurance, the proportions can vary from one risk to the next.

In this course we will focus on quota share reinsurance.

0.2 Non-proportional reinsurance

Under a non-proportional reinsurance arrangement, the direct writer pays a fixed premium to the reinsurer. The reinsurer will only be required to make payments where part of the claim amount falls in a particular reinsurance *layer* (*eg* between £1m and £5m). The layer will be defined by a lower limit, the *retention limit* (*eg* £1m), and an upper limit (*eg* £5m or infinity if the cover is *unlimited*). Usually, most claims are paid in full by the direct writer.

We will mention two forms of non-proportional reinsurance here:

1. With *individual excess of loss* (XOL) reinsurance, the reinsurer will be required to make a payment when the claim amount for any individual claim exceeds a specified *excess point* or *retention*. For example, the reinsurer might agree to pay the excess when any claim from a motor policy exceeds £50,000, but with an upper limit of £2 million.

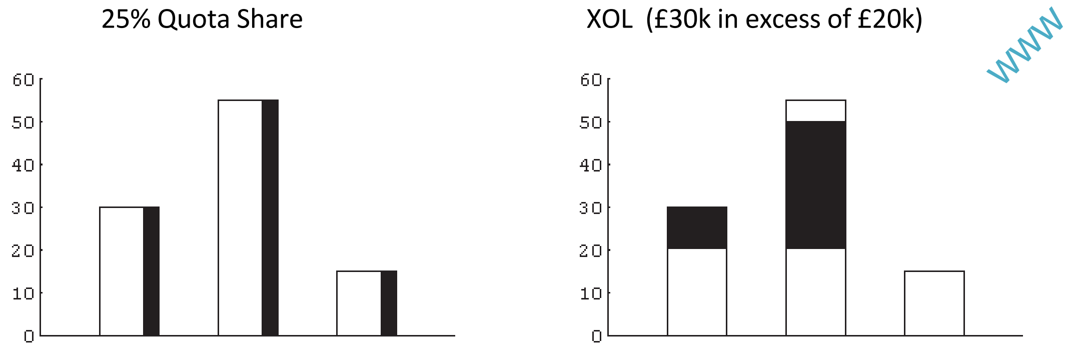
2. With *stop loss* reinsurance, the reinsurer will be required to make payments if the total claim amount for a specified group of policies exceeds a specified amount (which may be expressed as a percentage of the gross premium). We will look at this in Chapter 20.

The diagram below shows how much the direct writer and the reinsurer would pay when there are claims for £30,000, £55,000 and £15,000:

(a) under a 25% quota share arrangement, and

(b) under an individual XOL arrangement with a reinsurance layer of £30,000 in excess of £20,000.

The parts of each claim paid by the reinsurer are shown in black.


25% Quota Share

XOL (£30k in excess of £20k)

1 Reinsurance arrangements

The actual amount that the direct insurer ends up paying after allowing for payments under the reinsurance arrangements is called the *net claim amount*. The actual premium that the direct writer gets to keep after making any payments for reinsurance is the *insurer's net premium income*. The original amounts without adjustment for reinsurance are referred to as the *gross claim amount* and the *insurer's gross premium income*.

In this chapter we will use the following notation.

Notation

X is the gross claim amount random variable

Y is the net claim amount, *ie* the amount of the claim paid by the insurer in respect of a single claim (after receiving the reinsurance recovery)

Z is the amount paid by the reinsurer in respect of a single claim.

For a given reinsurance arrangement, we can express the random variables Y and Z in terms of X.

For example, suppose that a reinsurer has agreed to make the following payments in respect of individual claims incurred by a direct insurer:

- nothing, if the claim is less than £5,000
- the full amount reduced by £5,000, if the claim is between £5,000 and £10,000
- half the full amount, if the claim is between £10,000 and £20,000
- £10,000, if the claim exceeds £20,000.

Then:

$$Z = \begin{cases} 0 & \text{if } X \le £5,000 \\ X - 5,000 & \text{if } £5,000 < X \le £10,000 \\ X/2 & \text{if } £10,000 < X \le £20,000 \\ 10,000 & \text{if } X > £20,000 \end{cases}$$

and:

$$Y = \begin{cases} X & \text{if } X \le £5,000 \\ 5,000 & \text{if } £5,000 < X \le £10,000 \\ X/2 & \text{if } £10,000 < X \le £20,000 \\ X - 10,000 & \text{if } X > £20,000 \end{cases}$$

Note that $Y + Z = X$.

We are now in a position to consider the statistical calculations relating to reinsurance arrangements.

1.1 Excess of loss reinsurance

In excess of loss reinsurance, the insurer will pay any claim in full up to an amount M, the retention level; any amount above M will be borne by the reinsurer.

The excess of loss reinsurance arrangement can be written in the following way: if the claim is for amount X, then the insurer will pay Y where:

$$Y = X \quad \text{if } X \le M$$

$$Y = M \quad \text{if } X > M$$

The reinsurer pays the amount $Z = X - Y$.

Question

Write down an expression for Y if only a layer between M and $2M$ is reinsured.

Solution

$$Y = \begin{cases} X & \text{if } \ X \le M \\ M & \text{if } \ M < X \le 2M \\ X - M & \text{if } \ X > 2M \end{cases}$$

The insurer's liability is affected in two obvious ways by reinsurance:

(i) the mean amount paid is reduced;

(ii) the variance of the amount paid is reduced.

Both these conclusions are simple consequences of the fact that excess of loss reinsurance puts an upper limit on large claims.

The mean amounts paid by the insurer and the reinsurer under excess of loss reinsurance can now be obtained. Observe that the mean amount paid by the insurer without reinsurance is:

$$E(X) = \int_0^\infty x\, f(x)\, dx \tag{18.1}$$

where $f(x)$ is the PDF of the claim amount X. With a retention level of M the mean amount paid by the insurer becomes:

$$E(Y) = \int_0^M xf(x)dx + M\,P(X > M) \tag{18.2}$$

This is because:

$$E(Y)=\int_0^M x\,f(x)dx+\int_M^\infty M\,f(x)dx=\int_0^M x\,f(x)dx+M\int_M^\infty f(x)dx$$

and:

$$\int_M^\infty f(x)dx=P(X>M)$$

We can calculate $E(Y^2)$ in a similar way:

$$E(Y^2)=\int_0^M x^2\,f(x)dx+\int_M^\infty M^2\,f(x)dx=\int_0^M x^2\,f(x)dx+M^2\,P\left(X>M\right)$$

Then $\text{var}(Y)=E(Y^2)-\left[E(Y)\right]^2$.

More generally, the moment generating function of Y, the amount paid by the insurer, is:

$$\boldsymbol{M_Y(t)=E(e^{tY})=\int_0^M e^{tx}f(x)\,dx+e^{tM}\;P(X>M)}$$

Here we are using the formula for the expected value of a function of a continuous random variable:

$$E(h(X))=\int_x h(x)\,f(x)\,dx$$

with:

$$h(X)=\begin{cases}e^{tX} & \text{if } X\le M\\ e^{tM} & \text{if } X>M\end{cases}$$

Question

Suppose that claim amounts are uniformly distributed over the interval $(0,500)$. The insurer effects individual excess of loss reinsurance with a retention limit of 375.

Calculate the expected amounts paid by the insurer and the reinsurer in respect of a single claim.

Solution

Since $X\sim U(0,500)$, the expected gross claim amount is:

$$E(X)=\frac{500}{2}=250$$

The expected amount paid by the insurer is:

$$E(Y)=\int_0^{375} x\,f(x)\,dx+375P(X>375)$$

$$=\int_0^{375}\frac{x}{500}\,dx+375\left[1-F(375)\right]$$

$$=\left[\frac{x^2}{1{,}000}\right]_0^{375}+375\left[1-\frac{375}{500}\right]$$

$$=140.625+93.75$$

$$=234.375$$

Also, since $Y+Z=X$:

$$E(Z)=E(X)-E(Y)=250-234.375=15.625$$

Under excess of loss reinsurance, the reinsurer will pay Z where:

$$Z=\begin{cases}0 & \text{if } X\le M\\ X-M & \text{if } X>M\end{cases}$$

The mean amount paid by the reinsurer is:

$$\boldsymbol{E(Z)=\int_M^{\infty}(x-M)f(x)dx} \qquad \textbf{(18.3)}$$

Similarly, we can calculate $E(Z^2)$ using:

$$E(Z^2)=\int_M^{\infty}(x-M)^2\,f(x)\,dx$$

Then $\text{var}(Z)=E(Z^2)-\left[E(Z)\right]^2$.

More generally, the moment generating function of Z is:

$$M_Z(t)=E(e^{tZ})$$

$$=\int_0^M e^{t0}f(x)\,dx\;+\int_M^{\infty}e^{t(x-M)}f(x)\,dx$$

$$=\int_0^M f(x)\,dx\;+\int_M^{\infty}e^{t(x-M)}f(x)\,dx$$

$$=P(X\le M)+\int_M^{\infty}e^{t(x-M)}f(x)\,dx$$

We can use R to simulate gross claim amounts and hence the amounts paid by the insurer and reinsurer for any given retention limit.

Suppose claims (in £'s) have an exponential distribution with parameter $\lambda = 0.0005$. The R code for simulating 10,000 claims, *x*, is given by:

```
x <- rexp(10000,rate=0.0005)
```

We can then obtain the claims paid by the insurer, *y*, and the reinsurer, *z* with retention *M* using:

```
y <- pmin(x,M)
z <- pmax(0,x-M)
```

We can then obtain their means and variances using the R functions `mean` **and** `var`**.**

We can use these vectors to estimate probabilities. For example, to estimate the probability that the insurer pays less than £1,000 we would use:

```
length(y[y<1000])/length(y)
```

Similarly we could estimate the claim size for a given percentile. For example, to estimate the claim size corresponding to the 90th percentile of the insurer's claims we would use:

```
quantile(y,0.9)
```

1.2 The reinsurer's conditional claims distribution

Now consider reinsurance (once again) **from the point of view of the reinsurer. The reinsurer may have a record only of claims that are greater than M. If a claim is for less than M the reinsurer may not even know a claim has occurred. The reinsurer thus has the problem of estimating the underlying claims distribution when only those claims greater than M are observed. The statistical terminology is to say that the reinsurer observes claims from a truncated distribution.**

In this case the values observed by the reinsurer relate to a conditional distribution, since the numbers are conditional on the original claim amount exceeding the retention limit.

Let W be the random variable with this truncated distribution. Then:

$$W = X - M \mid X > M$$

This can also be expressed as follows:

$$W = Z \mid Z > 0$$

Suppose that the underlying claim amounts have PDF $f(x)$ and CDF $F(x)$. Suppose that the reinsurer is only informed of claims greater than the retention *M* and has a record of $w = x - M$. What is the PDF $g(w)$ of the amount, w, paid by the reinsurer?

The argument goes as follows:

$$\begin{aligned}
\boldsymbol{P(W < w) = P(X < w + M \mid X > M)} \\
&= \frac{\boldsymbol{P(X < w + M \text{ and } X > M)}}{\boldsymbol{P(X > M)}} \\
&= \frac{\boldsymbol{P(M < X < w + M)}}{\boldsymbol{P(X > M)}} \\
&= \int_M^{w+M} \frac{\boldsymbol{f(x)}}{\boldsymbol{1 - F(M)}}\, \boldsymbol{dx} \\
&= \frac{\boldsymbol{F(w + M) - F(M)}}{\boldsymbol{1 - F(M)}}
\end{aligned}$$

This derivation also uses the result:

$$P(a < X < b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

Differentiating with respect to w, the PDF of the reinsurer's claims is:

$$\boldsymbol{g(w) = \frac{f(w+M)}{1-F(M)}\,,\quad w > 0} \qquad \textbf{(18.4)}$$

This is just the original PDF evaluated at the gross amount $w+M$, divided by the probability that the claim exceeds M.

The PDF of W may be denoted by $f_W(w)$ rather than $g(w)$. With this notation, the result can be stated as follows:

PDF of the reinsurer's conditional claim amount random variable

If $W = X - M \mid X > M$, then:

$$f_W(w) = \frac{f_X(w+M)}{1-F_X(M)} = \frac{f_X(w+M)}{P(X > M)}$$

Question

Using the notation above, determine the distribution of W if:

(a) $X \sim Exp(\lambda)$

(b) $X \sim Pa(\alpha, \lambda)$

Solution

(a) ***Exponential***

If $X \sim Exp(\lambda)$, then $f_X(x) = \lambda e^{-\lambda x}$ and $F_X(x) = 1 - e^{-\lambda x}$. So:

$$f_W(w) = \frac{f_X(w+M)}{1 - F_X(M)} = \frac{\lambda e^{-\lambda(w+M)}}{e^{-\lambda M}} = \lambda e^{-\lambda w}, \quad w > 0$$

This is the PDF of $Exp(\lambda)$. So $W \sim Exp(\lambda)$, the same as the original claims distribution. This illustrates the memoryless property of the exponential distribution.

(b) ***Pareto***

If $X \sim Pa(\alpha, \lambda)$, then $f_X(x) = \dfrac{\alpha\lambda^\alpha}{(\lambda + x)^{\alpha+1}}$ and $F_X(x) = 1 - \left(\dfrac{\lambda}{\lambda + x}\right)^\alpha$. So:

$$f_W(w) = \frac{\alpha\lambda^\alpha \big/ (\lambda + w + M)^{\alpha+1}}{\lambda^\alpha \big/ (\lambda + M)^\alpha} = \frac{\alpha(\lambda + M)^\alpha}{(\lambda + M + w)^{\alpha+1}}, \quad w > 0$$

This is the PDF of $Pa(\alpha, \lambda + M)$. So $W \sim Pa(\alpha, \lambda + M)$.

We can now calculate the expected value of W. Using the PDF of $W = X - M \mid X > M$, we have:

$$E(W) = \int_0^\infty w\, f_W(w)\, dw = \frac{\int_0^\infty w\, f_X(w+M)\, dw}{1 - F_X(M)} = \frac{\int_M^\infty (x-M)\, f_X(x)\, dx}{1 - F_X(M)} = \frac{E(Z)}{P(X > M)} = \frac{E(Z)}{P(Z > 0)}$$

So the reinsurer’s expected claim payment on a claim in which it is involved is just the reinsurer’s expected claim payment (on all claims), $E(Z)$, divided by the probability that the claim involves the reinsurer.

If *z* is the vector of the reinsurer’s claims:

```
z <- pmax(0,x-M)
```

Then we can obtain the truncated distribution, *w*, using:

```
w <- z[z>0]
```

We can then calculate moments, probabilities and quantiles as before.

1.3 Proportional reinsurance

In proportional reinsurance the insurer pays a fixed proportion of the claim, whatever the size of the claim. Using the same notation as above, the proportional reinsurance arrangement can be written as follows: if the claim is for an amount X then the company will pay Y where:

$$Y = \alpha X \qquad 0 < \alpha < 1$$

The parameter α is known as the retained proportion or retention level; note that the term retention level is used in both excess of loss and proportional reinsurance though it means different things.

Since $Y + Z = X$, we must have $Z = (1-\alpha)X$. The mean and variance of Y and Z are calculated as follows:

$$E(Y) = \alpha E(X) \qquad E(Z) = (1-\alpha)E(X)$$

$$\text{var}(Y) = \alpha^2 \,\text{var}(X) \qquad \text{var}(Z) = (1-\alpha)^2 \,\text{var}(X)$$

Question

Claims from a particular portfolio have a generalised Pareto distribution with parameters $\alpha = 6$, $\lambda = 200$ and $k = 4$. A proportional reinsurance arrangement is in force with a retained proportion of 80%.

Calculate the mean and variance of the amount paid by the insurer and the reinsurer in respect of a single claim.

Solution

Using X to represent the individual claim amount random variable and the formulae for the mean and variance of a three-parameter Pareto random variable (from page 15 of the *Tables*), we have:

$$E(X) = \frac{k\lambda}{\alpha - 1} = \frac{4 \times 200}{6-1} = \frac{800}{5} = 160$$

and:

$$\text{var}(X) = \frac{k(k+\alpha-1)\lambda^2}{(\alpha-1)^2(\alpha-2)} = \frac{4(4+6-1)\times 200^2}{(6-1)^2(6-2)} = \frac{1{,}440{,}000}{100} = 14{,}400$$

The amount paid by the insurer is $Y = 0.8X$. So:

$$E(Y) = 0.8 \times 160 = 128$$

and:

$$\text{var}(Y) = 0.8^2 \times 14{,}400 = 9{,}216$$

The amount paid by the reinsurer is $Z = 0.2X$. So:

$$E(Z) = 0.2 \times 160 = 32$$

and:

$$\text{var}(Z) = 0.2^2 \times 14{,}400 = 576$$

As the amount paid by the insurer on a claim X is $Y = \alpha X$ and the amount paid by the reinsurer is $Z = (1-\alpha)X$, the distribution of both of these amounts can be found by a simple change of variable.

Question

Claims from a particular portfolio have an exponential distribution with mean 1,000. The insurer takes out proportional reinsurance with a retained proportion of 0.9.

Determine the distribution of the insurer's net claim amount random variable.

Solution

We know that X is exponential with mean 1,000, so the exponential parameter is $\dfrac{1}{1{,}000}$.

From page 11 of the *Tables*, the MGF of X is:

$$M_X(t) = (1 - 1{,}000t)^{-1} \ , \quad t < \frac{1}{1{,}000}$$

Since $Y = 0.9X$, the MGF of Y is:

$$M_Y(t) = E(e^{tY}) = E(e^{0.9tX}) = M_x(0.9t) = (1 - 1{,}000 \times 0.9t)^{-1} = (1 - 900t)^{-1} \ , \ t < \frac{1}{900}$$

This is the MGF of the exponential distribution with mean 900. By the uniqueness property of MGFs, it follows that the distribution of the insurer's net claim amount random variable is exponential with mean 900.

The payments of the insurer, y, and the reinsurer, z, with retained proportion a would be:

```
y <- a*x
z <- (1-a)*x
```

We can then calculate moments, probabilities and quantiles as before.

2 Normal and lognormal distributions

There are useful integral formulae that simplify reinsurance calculations when working with normal and lognormal distributions.

2.1 Normal distribution

Truncated mean of the normal distribution

If $X \sim N(\mu,\sigma^2)$, then:

$$\int_L^U x\, f_X(x)\, dx = \mu\,[\Phi(U') - \Phi(L')] - \sigma\,[\phi(U') - \phi(L')]$$

where:

$$L' = \frac{L-\mu}{\sigma}$$

$$U' = \frac{U-\mu}{\sigma}$$

$\phi(z)$ is the PDF of the standard normal distribution

$\Phi(z)$ is the CDF of the standard normal distribution.

This result is given on page 18 of the *Tables*. It is proved as follows.

Using the formula for $f_X(x)$ and the substitution $z = \dfrac{x-\mu}{\sigma}$:

$$\int_L^U x\, f_X(x)\, dx = \int_L^U x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \int_{L'}^{U'} (\mu + \sigma z) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

$$= \mu \int_{L'}^{U'} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz + \sigma \int_{L'}^{U'} z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Now, since $\dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ is the PDF of $N(0,1)$:

$$\int_{L'}^{U'} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = P\left(L' < N(0,1) < U'\right)$$

So:

$$\int_L^U x\,f_X(x)\,dx = \mu P\left(L' < N(0,1) < U'\right) + \sigma \int_{L'}^{U'} z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,dz$$

$$= \mu P\left(L' < N(0,1) < U'\right) + \sigma \left[-\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \right]_{L'}^{U'}$$

$$= \mu\left[\Phi(U') - \Phi(L')\right] - \sigma\left[\phi(U') - \phi(L')\right]$$

$$= \mu\left[\Phi\left(\frac{U-\mu}{\sigma}\right) - \Phi\left(\frac{L-\mu}{\sigma}\right)\right] - \sigma\left[\phi\left(\frac{U-\mu}{\sigma}\right) - \phi\left(\frac{L-\mu}{\sigma}\right)\right]$$

When $L = -\infty$ or $U = \infty$, these formulae can be simplified because:

$$\phi(-\infty) = \phi(\infty) = 0\,,\ \Phi(-\infty) = 0,\ \Phi(\infty) = 1$$

Question

Claims from a particular portfolio are normally distributed with mean 800 and standard deviation 100. An individual excess of loss arrangement with retention limit is 860 is in place.

Calculate the insurer's mean claim payment net of reinsurance.

Solution

The insurer's mean claim payment is:

$$E(Y) = \int_0^{860} x\,f_X(x)\,dx + 860\,P(X > 860)$$

where $X \sim N(800, 100^2)$.

Using the formula for the truncated mean of a normal random variable:

$$\int_0^{860} x\,f_X(x)\,dx = 800\left[\Phi\left(\frac{860-800}{100}\right) - \Phi\left(\frac{0-800}{100}\right)\right]$$

$$-100\left[\phi\left(\frac{860-800}{100}\right) - \phi\left(\frac{0-800}{100}\right)\right]$$

$$= 800\left[\Phi(0.6) - \Phi(-8)\right] - 100\left[\phi(0.6) - \phi(-8)\right]$$

From pages 160 and 161 of the *Tables*:

$$\Phi(0.6) = 0.72575$$

$$\Phi(-8) = 1 - \Phi(8) \approx 0$$

Also, using the formula for the PDF of the standard normal distribution from page 10 of the *Tables*:

$$\phi(0.6)=\frac{1}{\sqrt{2\pi}}e^{-½\times 0.6^2}=0.33322$$

$$\phi(-8)=\frac{1}{\sqrt{2\pi}}e^{-½\times(-8)^2}\approx 0$$

So:

$$\int_0^{860} x\,f_X(x)\,dx \approx 800\big[0.72575-0\big]-100\big[0.33322-0\big]=547.278$$

The second term in the expression for $E(Y)$ is:

$$\begin{aligned}860\,P(X>860)&=860\left[1-\Phi\left(\frac{860-800}{100}\right)\right]\\&=860\big[1-\Phi(0.6)\big]\\&=860\big[1-0.72575\big]\\&=235.855\end{aligned}$$

Hence:

$$E(Y)\approx 547.278+235.855=783.13$$

2.2 Lognormal distribution

Truncated moments of the lognormal distribution

If $X\sim\log N(\mu,\sigma^2)$, then:

$$\int_L^U x^k f_X(x)\,dx = e^{k\mu+½k^2\sigma^2}\big[\Phi(U_k)-\Phi(L_k)\big]$$

where:

$$L_k=\frac{\ln L-\mu}{\sigma}-k\sigma$$

$$U_k=\frac{\ln U-\mu}{\sigma}-k\sigma$$

$\Phi(z)$ is the CDF of the standard normal distribution.

This result is also given on page 18 of the *Tables*. It is proved as follows.

Using the formula for the PDF of the lognormal distribution from page 14 of the *Tables*, we have:

$$\int_L^U x^k f_X(x)\,dx = \int_L^U x^k \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x}e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2}\,dx$$

Making the substitution $t = \dfrac{\ln x - \mu}{\sigma} - k\sigma$ gives:

$$\int_L^U x^k f_X(x)\,dx = \int_{L_k}^{U_k} e^{k(\mu+\sigma t+k\sigma^2)}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(t+k\sigma)^2}\,dt$$

$$= \int_{L_k}^{U_k} \frac{1}{\sqrt{2\pi}}e^{k\mu+k\sigma t+k^2\sigma^2}e^{-\frac{1}{2}t^2-k\sigma t-\frac{1}{2}k^2\sigma^2}\,dt$$

$$= e^{k\mu+\frac{1}{2}k^2\sigma^2}\int_{L_k}^{U_k}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}\,dt$$

$$= e^{k\mu+\frac{1}{2}k^2\sigma^2}[\Phi(U_k)-\Phi(L_k)]$$

When $L=0$ or $U=\infty$, these formulae can be simplified using the facts that:

$$\Phi(-\infty)=0\,,\ \Phi(0)=\tfrac{1}{2}\,,\ \Phi(\infty)=1$$

By setting $k=1$ in the truncated moments formula, we can calculate the insurer's expected claim payment under excess of loss reinsurance when the original claims follow a lognormal distribution.

Question

An insurer is considering taking out one of the following reinsurance treaties:

Treaty 1: Proportional reinsurance with a retained proportion of 0.75

Treaty 2: Individual excess of loss cover with a retention limit of £25,000

The claims distribution is lognormal with parameters $\mu = 8.5$ and $\sigma^2 = 0.64$.

Calculate the insurer's expected net claim payments in the following cases:

(a) without either treaty

(b) with Treaty 1 only

(c) with Treaty 2 only.

Solution

(a) ***No reinsurance***

Without either treaty, the insurer pays the full amount of each loss. So:

$$E(Y)=E(X)=e^{\mu+½\sigma^2}=e^{8.5+½\times 0.64}=£6,768$$

(b) ***Treaty 1***

Under Treaty 1, the insurer pays 75% of each loss. So $Y=0.75X$ and:

$$E(Y)=0.75E(X)=0.75\times 6,768=£5,076$$

(c) ***Treaty 2***

Under Treaty 2, the insurer pays the first £25,000 of each loss. So:

$$Y=\begin{cases} X & \text{if } X\le 25,000 \\ 25,000 & \text{if } X>25,000 \end{cases}$$

and:

$$E(Y)=\int_0^{25,000} x\,f_X(x)\,dx+25,000P(X>25,000)$$

Using the truncated moments formula with $k=1$ gives:

$$\int_0^{25,000} x\,f_X(x)\,dx=e^{\mu+½\sigma^2}\left[\Phi(U_1)-\Phi(L_1)\right]$$

where:

$$U_1=\frac{\ln U-\mu}{\sigma}-\sigma=\frac{\ln 25,000-8.5}{\sqrt{0.64}}-\sqrt{0.64}=1.23329$$

$$\begin{aligned}\Phi(U_1)&\approx(1-0.329)\Phi(1.23)+0.329\,\Phi(1.24)\\&=0.671\times 0.89065+0.329\times 0.89251\\&=0.89126\end{aligned}$$

and:

$$\Phi(L_1)=\Phi\left(\frac{\ln L-\mu}{\sigma}-\sigma\right)=0 \;\text{ since } \ln L\to-\infty \text{ as } L\to 0$$

So:

$$\int_0^{25,000} x\, f_X(x)\, dx = e^{8.5+½\times 0.64} \times 0.89126 = 6,032.30$$

The second term in the expression for $E(Y)$ is:

$$25,000P(X > 25,000) = 25,000\left[1 - \Phi\left(\frac{\ln 25,000 - 8.5}{\sqrt{0.64}}\right)\right]$$

$$= 25,000\left[1 - \Phi(2.03329)\right]$$

Interpolating gives:

$$\Phi(2.03329) \approx (1-0.329)\Phi(2.03) + 0.329\,\Phi(2.04)$$

$$= 0.671 \times 0.97882 + 0.329 \times 0.97932$$

$$= 0.97898$$

So:

$$E(Y) \approx 6,032.30 + 25,000 \times (1 - 0.97898) = £6,558$$

Setting $k = 2$ in the truncated moments formula, we can calculate the second non-central moment of the insurer's claim payment. We now extend the previous question to calculate the standard deviation of the insurer's net claim amount.

Question

An insurer is considering taking out one of the following reinsurance treaties:

Treaty 1: Proportional reinsurance with a retained proportion of 0.75

Treaty 2: Individual excess of loss cover with a retention limit of £25,000

The claims distribution is lognormal with parameters $\mu = 8.5$ and $\sigma^2 = 0.64$.

Calculate the standard deviation of the insurer's net claim payments in the following cases:

(a) without either treaty

(b) with Treaty 1 only

(c) with Treaty 2 only.

Solution

(a) ***No reinsurance***

Without either treaty, the variance is:

$$\text{var}(Y) = \text{var}(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) = e^{2(8.5)+0.64}(e^{0.64} - 1) = 41{,}067{,}256.6$$

So the standard deviation is £6,408.

(b) ***Treaty 1***

Under Treaty 1:

$$\text{var}(Y) = 0.75^2 \text{var}(X)$$

So the standard deviation is $0.75 \times 6{,}408 = £4{,}806$.

(c) ***Treaty 2***

Under Treaty 2:

$$E(Y^2) = \int_0^M x^2 f_X(x)\,dx + \int_M^\infty M^2 f_X(x)\,dx$$

Using the truncated moments formula with $k = 2$ gives:

$$\int_0^{25{,}000} x^2\, f_X(x)\,dx = e^{2\mu+2\sigma^2}\left[\Phi(U_2) - \Phi(L_2)\right]$$

where:

$$U_2 = \frac{\ln U - \mu}{\sigma} - 2\sigma = \frac{\ln 25{,}000 - 8.5}{\sqrt{0.64}} - 2\sqrt{0.64} = 0.43329$$

$$\begin{aligned}\Phi(U_2) &\approx (1 - 0.329)\Phi(0.43) + 0.329\,\Phi(0.44)\\ &= 0.671 \times 0.66640 + 0.329 \times 0.67003\\ &= 0.66759\end{aligned}$$

and:

$$\Phi(L_2) = \Phi\left(\frac{\ln L - \mu}{\sigma} - 2\sigma\right) = 0 \quad \text{since } \ln L \to -\infty \text{ as } L \to 0$$

So:

$$\int_0^{25,000} x^2 \, f_X(x)\, dx = e^{2\times 8.5+2\times 0.64} \times 0.66759 = 57,998,646$$

$$E(Y^2) = 57,998,646 + 25,000^2 \times (1-0.97898) = 71,130,942$$

$$\text{var}(Y) = 71,130,942 - 6,558^2 = 5,304^2$$

and the standard deviation is £5,304. (The calculations are very sensitive to rounding, so we have used Excel to obtain accurate values.)

Our calculations have shown that reinsurance reduces both the mean and variance of the insurer's claim payments, as expected.

3 Inflation

The examples we have considered so far have assumed that claim distributions don't change over time (or at least that we are looking at a sufficiently short time period for us to be able to make this assumption).

In practice claims are likely to increase because of inflation, at least in the longer term. A claim distribution that is suitable for modelling claim amounts in one year may well not be suitable a year or two later. We need to adjust our claim distributions to allow for inflation.

In this section we will look at how claims inflation affects reinsurance arrangements. It is easy to deal with claims inflation in the proportional reinsurance situation.

Question

Claims from a portfolio of policies are believed to follow an $Exp(\lambda)$ distribution. A proportional reinsurance arrangement with a retained proportion α is in force.

(i) Give an expression for the insurer's expected claim payment.

(ii) Next year, claim amounts are expected to increase by an inflationary factor of k. Derive an expression for the insurer's expected claim payment next year.

Solution

(i) ***Expected claim payment this year***

We know that $X \sim Exp(\lambda)$ and $Y = \alpha X$. So:

$$E(Y) = E(\alpha X) = \alpha E(X) = \frac{\alpha}{\lambda}$$

(ii) ***Expected claim payment next year***

Next year, the gross claim amount random variable is kX and the insurer's net claim payment is αkX. So the insurer's expected claim payment is:

$$E(\alpha kX) = \alpha kE(X) = \frac{\alpha k}{\lambda}$$

With excess of loss reinsurance, inflation can cause a problem. Suppose that the claims X are inflated by a factor of k but the retention M remains fixed. What effect does this have on the arrangement?

The amount claimed is kX, and the amount paid by the insurer, Y, is:

$$\begin{aligned} Y &= kX && \text{if } kX \le M \\ Y &= M && \text{if } kX > M \end{aligned}$$

In other words:

$$Y = \begin{cases} kX & \text{if } X \le \frac{M}{k} \\ M & \text{if } X > \frac{M}{k} \end{cases}$$

The mean amount paid by the insurer is:

$$\boldsymbol{E(Y) = \int_0^{M/k} kx\, f(x)\, dx + M\, P(X > M/k)} \qquad \textbf{(18.5)}$$

For example, if $X \sim Exp(\lambda)$, then:

$$E(Y) = \int_0^{M/k} kx\,\lambda e^{-\lambda x}\, dx + M\,P(X > M/k)$$

Integrating by parts:

$$E(Y) = \left[-kx\,e^{-\lambda x}\right]_0^{M/k} + \int_0^{M/k} ke^{-\lambda x}\, dx + M\,P(X > M/k)$$

$$= -Me^{-\lambda M/k} + \int_0^{M/k} ke^{-\lambda x}\, dx + Me^{-\lambda M/k}$$

The first and last terms in the line above cancel to give:

$$E(Y) = \int_0^{M/k} ke^{-\lambda x}\, dx = \left[-\frac{k}{\lambda}e^{-\lambda x}\right]_0^{M/k} = \frac{k}{\lambda}\left(1 - e^{-\lambda M/k}\right)$$

One important general point that can be made is that the new mean claim amount paid by the insurer is not k times the mean claim amount paid by the insurer without inflation.

The insurer's mean claim amount will inflate by less than k. We can see this by considering different sizes of claim. From the insurer's point of view, the amount it has to pay out on small claims (those that are nowhere near the retention limit) will increase by k. However, the amount paid on claims that were already above the limit will not increase at all (and the amount paid on claims that didn't reach the limit before but now do will increase, but by less than k).

The reinsurer's mean claim amount will increase by a factor of more than k to compensate.

A similar approach can also be taken in situations where the retention limit is linked to some index of inflation.

Of course if the retention limit increases by a factor of k as well, both mean claim amounts (for the insurer and reinsurer) will increase by the same factor.

When examining the details of a reinsurance arrangement in real life it is very important to check whether the retention limits are fixed or are linked to an agreed inflation index. There are special published indices specifically for use in connection with general insurance claims.

The payments of the insurer, *y*, and the reinsurer, *z*, with retention *M* and inflation factor *k* would be:

```
y <- pmin(k*x,M)
z <- pmax(0,k*x-M)
```

We can then calculate moments, probabilities and quantiles as before.

4 Estimation

Consider the problem of estimation in the presence of excess of loss reinsurance. Suppose that the claims record shows only the net claims paid by the insurer. A typical claims record might be:

$$x_1, x_2, M, x_3, M, x_4, x_5, \ldots \qquad (18.6)$$

and an estimate of the underlying gross claims distribution is required.

As before, we wish to estimate the parameters for the distribution we have assumed for the claims.

The method of moments is not available since even the mean claim amount cannot be computed. On the other hand, it may be possible to use the method of percentiles without alteration; this would happen if the retention level M is high and only the higher sample percentiles were affected by the (few) reinsurance claims.

The statistical terminology for a sample of the form (18.6) is censored. In general, a censored sample occurs when some values are recorded exactly and the remaining values are known only to exceed a particular value, here the retention level M.

Maximum likelihood can be applied to censored samples. The likelihood function is made up of two parts. If the values of $x_1, x_2, \ldots, x_n$ are recorded exactly these contribute a factor of:

$$L_1(\underline{\theta}) = \prod_{i=1}^{n} f(x_i; \underline{\theta})$$

If a further m claims are referred to the reinsurer, then the insurer records a payment of M for each of these claims. These censored values then contribute a factor:

$$L_2(\underline{\theta}) = \prod_{j=1}^{m} P(X > M) \quad ie \left[P(X > M)\right]^m$$

The complete likelihood function is:

$$L(\underline{\theta}) = \prod_{i=1}^{n} f(x_i; \underline{\theta}) \times \left[1 - F(M; \underline{\theta})\right]^m$$

where $F(.;\underline{\theta})$ is the CDF of the claims distribution.

The reason for multiplying is that the likelihood reflects the probability of getting the n claims with known values *and* m claims exceeding M. Also, we are assuming that the claims are independent.

In R, we can define the censored log-likelihood function and use the function `nlm` on the negative value of this as before.

Question

Claims from a portfolio are believed to follow an $Exp(\lambda)$ distribution. The insurer has effected individual excess of loss reinsurance with a retention limit of 1,000.

The insurer observes a random sample of 100 claims, and finds that the average amount of the 90 claims that do not exceed 1,000 is 82.9. There are 10 claims that do exceed the retention limit.

Calculate the maximum likelihood estimate of the parameter λ.

Solution

Here $X \sim Exp(\lambda)$ and $P(X > 1{,}000) = e^{-1{,}000\lambda}$. So the likelihood function is:

$$L(\lambda) = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \ldots \lambda e^{-\lambda x_{90}} \times (e^{-1{,}000\lambda})^{10} = \lambda^{90} e^{-(10{,}000+\sum x_i)\lambda}$$

Taking logs:

$$\ln L = 90 \ln\lambda - (10{,}000 + \sum x_i)\lambda$$

Differentiating with respect to λ:

$$\frac{\partial}{\partial\lambda}\ln L = \frac{90}{\lambda} - (10{,}000 + \sum x_i)$$

This is equal to 0 when:

$$\lambda = \frac{90}{10{,}000 + \sum x_i} = \frac{90}{10{,}000 + (90 \times 82.9)} = 0.005154$$

Differentiating again:

$$\frac{\partial^2}{\partial\lambda^2}\ln L = -\frac{90}{\lambda^2}$$

This is negative when $\lambda = 0.005154$. (In fact it is always negative.) So we have a maximum turning point and hence $\hat{\lambda} = 0.005154$.

5 Policy excess

Insurance policies with an excess are common in motor insurance and many other kinds of property and accident insurance. Under this kind of policy, the insured agrees to carry the full burden of the loss up to a limit, L, called the excess. If the loss is an amount X, greater than L, then the policyholder will claim only $X-L$. If Y is the amount actually paid by the insurer, then:

$$\begin{aligned} Y &= 0 && \text{if } X \le L \\ Y &= X - L && \text{if } X > L \end{aligned}$$

Clearly, the premium due on any policy with an excess will be less than that on a policy without an excess.

This assumes that some of the saving is actually passed on to the policyholder. A policy excess may also be referred to as a *deductible*.

The position of the insurer for a policy with an excess is exactly the same as that of the reinsurer under excess of loss reinsurance. The position of the policyholder as far as losses are concerned is exactly the same as that of an insurer with an excess of loss reinsurance contract.

In practice, expenses form a significant part of the insurance cost. So the presence of an excess might not affect the premium as much as might be expected. A premium calculated ignoring expenses is called a 'risk premium'.

Question

An insurer believes that claims from a particular type of policy follow a Pareto distribution with parameters $\alpha = 2$ and $\lambda = 900$. The insurer wishes to introduce a policy excess so that 20% of losses result in no claim to the insurer.

Calculate the size of the excess.

Solution

Let L be the size of the excess. The insurer wants to set L so that $P(X<L)=0.2$. Using the given loss distribution, we have:

$$P(X<L) = 1 - \left(\frac{900}{900+L}\right)^2$$

So we require:

$$1 - \left(\frac{900}{900+L}\right)^2 = 0.2$$

Rearranging:

$$\left(\frac{900}{900+L}\right)^2 = 0.8$$

$$\Rightarrow \frac{900}{900+L} = \sqrt{0.8}$$

$$\Rightarrow 900+L = \frac{900}{\sqrt{0.8}}$$

$$\Rightarrow L = \frac{900}{\sqrt{0.8}} - 900 = 106.23$$

The chapter summary starts on the next page so that you can keep all the chapter summaries together for revision purposes.

Chapter 18 Summary

Reinsurance

Reinsurance is insurance for insurance companies. By using reinsurance, the insurer seeks to protect itself from large claims. The mean amount paid by the insurer is reduced, and the variance of the amount paid by the insurer is reduced.

Reinsurance may be proportional or non-proportional (*ie* excess of loss).

We use the following notation:

X is the gross claim amount random variable

Y is the net claim amount, *ie* the amount of the claim paid by the insurer

Z is the amount paid by the reinsurer

Proportional reinsurance

Under proportional reinsurance, the insurer and the reinsurer split the claim in pre-defined proportions. For a claim amount X, the amount paid by the insurer is $Y=\alpha X$ and the amount paid by the reinsurer is $Z=(1-\alpha)X$ where α is known as the retained proportion or retention level, $0<\alpha<1$.

$$E(Y)=\alpha E(X) \qquad E(Z)=(1-\alpha)E(X)$$

$$\text{var}(Y)=\alpha^2\,\text{var}(X) \qquad \text{var}(Z)=(1-\alpha)^2\,\text{var}(X)$$

Non-proportional reinsurance (individual excess of loss)

Under individual excess of loss, the insurer will pay any claim in full up to an amount M, the retention level. Any amount above M will be met by the reinsurer.

$$Y=\begin{cases} X & \text{if } X\le M \\ M & \text{if } X>M \end{cases} \qquad Z=\begin{cases} 0 & \text{if } X\le M \\ X-M & \text{if } X>M \end{cases}$$

$$E(Y)=\int_0^M x\,f_X(x)\,dx+\int_M^\infty M\,f_X(x)\,dx \qquad E(Z)=\int_M^\infty (x-M)\,f_X(x)\,dx$$

$$E(Y^2)=\int_0^M x^2\,f_X(x)\,dx+\int_M^\infty M^2\,f_X(x)\,dx \qquad E(Z^2)=\int_M^\infty (x-M)^2\,f_X(x)\,dx$$

Reinsurer's conditional claims distribution

It may be the case that the reinsurer is only informed of claims greater than the retention level M. In this case, the reinsurer observes claims from a truncated (or conditional) distribution. Let W be the random variable associated with this distribution, then:

$$W = Z \mid Z > 0 = X - M \mid X > M$$

$$f_W(w) = \frac{f_X(w+M)}{P(X>M)} = \frac{f_X(w+M)}{P(Z>0)}$$

$$E(W) = \frac{E(Z)}{P(Z>0)}$$

Excesses

When a policy excess applies, the policyholder pays for the first part of each loss up to an excess level L. Any amount greater than L will be met by the insurer. The positions of the policyholder and the insurer as far as losses are concerned are the same as those of the insurer and the reinsurer respectively under individual excess of loss reinsurance. When a policy excess applies, the insurer's conditional distribution takes the same form as that of the reinsurer's conditional distribution above.

Inflation and individual excess of loss reinsurance

If claims are inflated by a factor of k but the retention level remains fixed at M then the amount paid by the insurer is:

$$Y = \begin{cases} kX & \text{if } X \le \frac{M}{k} \\ M & \text{if } X > \frac{M}{k} \end{cases}$$

The amount paid by the reinsurer is:

$$Z = \begin{cases} 0 & \text{if } X \le \frac{M}{k} \\ kX - M & \text{if } X > \frac{M}{k} \end{cases}$$

Estimation of parameters from a censored sample

The likelihood function of a vector of parameters $\underline{\theta}$, based on a sample of n exact observations and m censored observations known to exceed M is:

$$L(\underline{\theta}) = \left[\prod_{i=1}^{n} f_X(x_i)\right]\left[P(X>M)\right]^m$$

assuming that the observations are realisations of $n+m$ IID random variables.

Chapter 18 Practice Questions

18.1 An insurer insures a risk for which individual claim sizes (in £000s) have mean 500 and standard deviation 250. The insurer arranges excess of loss reinsurance for this risk with a retention limit of £1,000,000.

Calculate the proportion of claims from this risk for which the insurer expects to receive a payment from the reinsurer if the loss distribution is:

(a) gamma

(b) lognormal.

18.2 Claims arising from a particular portfolio have a Pareto distribution with parameters $\alpha = 6$ and $\lambda = 200$. The insurer effects individual excess of loss reinsurance with a retention limit of 80.

(i) Calculate the insurer's expected claim amount before and after reinsurance.

(ii) Calculate the mean amount paid by the reinsurer on claims in which it is involved.

18.3 A sample of a reinsurer's payments made under a proportional reinsurance arrangement consists of the following values, in units of thousands of pounds:

4.6, 6.8, 22.9, 1.4, 3.8, 10.2, 19.4, 32.1

If the original claim amounts have a $Gamma(\alpha,\lambda)$ distribution, and the retained proportion is 80%, determine the distribution of the reinsurer's claim payments. Hence estimate the parameters α and λ using the method of moments.

18.4 If $X \sim \log N(7.5, 0.85^2)$, calculate:

(a) $$\int_{1,000}^{5,000} f(x)\,dx$$

(b) $$\int_{0}^{1,000} x\,f(x)\,dx$$

(c) $$\int_{5,000}^{\infty} x^2 f(x)\,dx$$

18.5 Claims from a portfolio are believed to have a Pareto distribution with parameters α and λ. In Year 0, $\alpha = 6$ and $\lambda = 1{,}000$. An excess of loss reinsurance arrangement is in force, with a retention limit of 500. Inflation is a constant 10% *pa*.

(i) Determine the distribution of the gross claim amounts in Years 1 and 2.

(ii) Calculate the reinsurer's mean claim payment on all claims in Years 0, 1 and 2.

18.6 Claim amounts from a portfolio follow a Weibull distribution with PDF:

$$f(x) = 2cxe^{-cx^2}, x \geq 0$$

An individual excess of loss reinsurance arrangement with retention limit $M = 3$ is in force. A sample of the *reinsurer's* non-zero payment amounts gives the following values:

$$n = 10 \qquad \sum w_i = 8.7 \qquad \sum w_i^2 = 92.3$$

where the units are millions of pounds. Calculate the maximum likelihood estimate of c.

18.7 Exam style

(i) A random variable X has the lognormal distribution with density function $f(x)$ and parameters μ and σ. Show that for $a > 0$:

$$\int_a^\infty x\, f(x)dx = \exp\left(\mu + \frac{\sigma^2}{2}\right)\left(1 - \Phi\left(\frac{\log a - \mu - \sigma^2}{\sigma}\right)\right)$$

where Φ is the cumulative distribution function of the standard normal distribution. [4]

(ii) Claims under a particular class of insurance follow a lognormal distribution with mean 9.070 and standard deviation of 10.132 (figures in £000s). In any one year 20% of policies are expected to give rise to a claim.

An insurance company has 200 policies on its books and wishes to take out individual excess of loss reinsurance to cover all the policies in the portfolio. The reinsurer has quoted premiums for two levels of reinsurance as follows (figures in £000s):

Retention limit	Premium
25	50
30	40

(a) Calculate the probability, under each reinsurance arrangement, that a claim arising will involve the reinsurer.

(b) By investigating the average amount of each claim ceded to the reinsurer, calculate which of the retention levels gives the best value for money for the insurer (ignoring the insurer's attitude to risk).

(c) The following year, assuming all other things equal, the insurer believes that inflation will increase the mean and standard deviation of the claims in its portfolio by 8%. If the reinsurer charges the same premiums as before, determine which of the retention levels will give best value for money next year. [18]

[Total 22]

18.8 Exam style

(i) Loss amounts from a particular type of insurance have a Pareto distribution with parameters α and λ. If the company applies a policy excess, *E*, derive the distribution function of claim amounts paid by the insurer. [3]

(ii) Assuming that $\alpha = 4$ and $\lambda = 15$, calculate the mean claim amount paid by the insurer:

(a) with no policy excess ($E = 0$),

(b) with an excess of 10 ($E = 10$). [2]

(iii) Using your answers to (ii), comment on the effect of introducing a policy excess. [2]

[Total 7]

18.9 Exam style

Losses from a group of travel insurance policies are assumed to follow a Pareto distribution with parameters $\alpha = 4.5$ and $\lambda = 3{,}000$.

Next year losses are expected to increase by 3%, and the insurer has decided to introduce a policy excess of 100 per claim.

Calculate the probability that a loss next year is borne entirely by the policyholder. [2]

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 18 Solutions

18.1 (a) ***Gamma distribution***

Let X denote the loss random variable and suppose that $X \sim Gamma(\alpha,\lambda)$. Then:

$$E(X) = \frac{\alpha}{\lambda} = 500 \qquad \text{and} \qquad \text{var}(X) = \frac{\alpha}{\lambda^2} = 250^2$$

Solving these equations simultaneously gives:

$$\alpha = 4 \qquad \text{and} \qquad \lambda = 0.008$$

The reinsurer will make a payment if the claim size exceeds £1m. Since we are working in £000s, we have to calculate:

$$P(X > 1{,}000)$$

To do this, we can use the relationship between the gamma and chi-squared distributions:

$$X \sim Gamma(\alpha,\lambda) \Rightarrow 2\lambda X \sim \chi^2_{2\alpha}$$

So:

$$P(X > 1{,}000) = P(2\lambda X > 2{,}000\lambda) = P(\chi^2_8 > 16) = 1 - 0.9576 = 0.0424$$

(b) ***Lognormal distribution***

If $X \sim \log N(\mu,\sigma^2)$, then:

$$E(X) = e^{\mu + ½\sigma^2} = 500 \qquad \text{and} \qquad \text{var}(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) = 250^2$$

Squaring the first equation and substituting this into the second gives:

$$e^{\sigma^2} - 1 = \frac{250^2}{500^2} = 0.25 \quad \Rightarrow \quad \sigma^2 = \ln 1.25 = 0.22314$$

Then, from the equation for $E(X)$ we have:

$$\mu = \ln 500 - ½\sigma^2 = 6.1030$$

and:

$$P(X > 1{,}000) = P\left(Z > \frac{\ln 1{,}000 - 6.103036}{\sqrt{0.2231436}}\right) = P(Z > 1.70354)$$

$$= 1 - 0.95576 = 0.04424$$

18.2 (i) ***Insurer's expected claim payments***

Let X denote the gross claim amount random variable. Then $X \sim Pa(6,200)$ and:

$$E(X)=\frac{200}{6-1}=40$$

ie the mean claim amount before reinsurance is 40.

The insurer's net claim amount is:

$$Y=\begin{cases} X & \text{if } X\le 80 \\ 80 & \text{if } X>80 \end{cases}$$

So:

$$E(Y)=\int_0^{80} x\frac{6(200)^6}{(200+x)^7}\,dx+80P(X>80)$$

The integral can be evaluated by substitution. Let $u=200+x$. Then:

$$\int_0^{80} x\frac{6(200)^6}{(200+x)^7}\,dx=\int_{200}^{280}(u-200)\frac{6(200)^6}{u^7}du$$

$$=6(200)^6\int_{200}^{280}(u^{-6}-200u^{-7})du$$

$$=6(200)^6\left[\frac{u^{-5}}{-5}-\frac{200u^{-6}}{-6}\right]_{200}^{280}$$

$$=6(200)^6\left[\left(-\frac{280^{-5}}{5}+\frac{200\times 280^{-6}}{6}\right)-\left(-\frac{200^{-5}}{5}+\frac{200\times 200^{-6}}{6}\right)\right]$$

$$=21.9378$$

Alternatively, we could integrate by parts.

Also:

$$P(X>80)=1-F(80)=\left(\frac{200}{200+80}\right)^6=0.1328$$

So the mean claim amount after reinsurance is:

$$E(Y)=21.9378+80\times 0.1328=32.5626$$

(ii) ***Mean amount paid by the reinsurer on claims in which it is involved***

The mean amount paid by the reinsurer on all claims (including those where the reinsurer makes no payment) is:

$$E(Z)=E(X)-E(Y)=40-32.5626=7.4374$$

and the mean amount paid by the reinsurer on claims in which it is involved is:

$$\frac{E(Z)}{P(X>80)}=\frac{7.4374}{0.1328}=56$$

Alternatively, we could say that W, the reinsurer's conditional claim payment random variable has a $Pa(6,280)$ distribution. The mean of this distribution is $\frac{280}{6-1}=56$.

18.3 Let X be the gross claim amount random variable and Z be the reinsurer's claim payment. Then $X \sim Gamma(\alpha,\lambda)$ and $Z=0.2X$. The moment generating function of Z is:

$$M_Z(t)=E(e^{tZ})=E(e^{0.2tX})=M_X(0.2t)=\left(1-\frac{0.2t}{\lambda}\right)^{-\alpha}=\left(1-\frac{t}{5\lambda}\right)^{-\alpha}\ , \quad t<5\lambda$$

This is the MGF of the $Gamma(\alpha,5\lambda)$ distribution. By the uniqueness property of MGFs, it follows that $Z \sim Gamma(\alpha,5\lambda)$. So:

$$E(Z)=\frac{\alpha}{5\lambda} \qquad \text{and} \qquad \text{var}(Z)=\frac{\alpha}{(5\lambda)^2}=\frac{\alpha}{25\lambda^2}$$

The sample mean and n-denominator variance are:

$$\bar{z}=\frac{\sum z_i}{8}=\frac{101.2}{8}=12.65$$

and:

$$s^2=\frac{\sum z_i^2}{8}-\bar{z}^2=\frac{2{,}119.02}{8}-12.65^2=104.855$$

The method of moments estimates of α and λ are the solutions of the equations:

$$\frac{\alpha}{5\lambda}=12.65 \qquad \text{and} \qquad \frac{\alpha}{25\lambda^2}=104.855$$

Solving these gives $\hat{\alpha}=1.526$ and $\hat{\lambda}=0.02413$.

18.4 (a) Using the truncated moments formula with $k=0$:

$$\int_{1,000}^{5,000} f(x)\,dx = \Phi\left(\frac{\ln 5,000 - 7.5}{0.85}\right) - \Phi\left(\frac{\ln 1,000 - 7.5}{0.85}\right)$$

$$= \Phi(1.19670) - \Phi(-0.69676)$$

$$= 0.88429 - 0.24298$$

$$= 0.6413$$

This is $P(1,000 < X < 5,000)$.

(b) Using the formula with $k=1$:

$$\int_{0}^{1,000} x\,f(x)\,dx = e^{7.5+½\times 0.85^2}\left[\Phi\left(\frac{\ln 1,000 - 7.5}{0.85} - 0.85\right) - \Phi(-\infty)\right]$$

$$= e^{7.86125}[\Phi(-1.54676) - 0]$$

$$= 2,594.76 \times 0.06096$$

$$= 158.2$$

(c) Using the formula with $k=2$:

$$\int_{5,000}^{\infty} x^2 f(x)\,dx = e^{2(7.5)+2\times 0.85^2}\left[\Phi(\infty) - \Phi\left(\frac{\ln 5,000 - 7.5}{0.85} - 2(0.85)\right)\right]$$

$$= e^{16.445}[1 - \Phi(-0.50330)]$$

$$= 13,866,688(1 - 0.30738)$$

$$= 9.604\text{m}$$

18.5 (i) ***Distribution of insurer's claim payments before reinsurance***

Let X_j be the gross claim amount random variable in Year j. Then $X_0 \sim Pa(6,1000)$ and $X_1 = 1.1X_0$. The Pareto distribution does not have a moment generating function, but we can determine the distribution of X_1 by considering its CDF. For $x > 0$:

$$F_{X_1}(x) = P(X_1 \le x) = P(1.1X_0 \le x) = P\left(X_0 \le \tfrac{x}{1.1}\right) = F_{X_0}\left(\tfrac{x}{1.1}\right) = 1 - \left(\frac{1,000}{1,000 + \frac{x}{1.1}}\right)^6$$

Multiplying the numerator and the denominator of the bracketed fraction by 1.1, we see that:

$$F_{X_1}(x) = 1 - \left(\frac{1,100}{1,100 + x}\right)^6 , \quad x > 0$$

This is the CDF of the $Pa(6,1100)$ distribution. So $X_1 \sim Pa(6,1100)$.

Inflation has no effect on the first parameter, but the second parameter has increased by 10%.

Similarly, $X_2 \sim Pa(6,1210)$.

(ii) ***Reinsurer's expected claim payments***

Let Z_j be the reinsurer's claim payment random variable in Year j. Then:

$$E(Z_0)=\int_{500}^{\infty}(x-500)\,f_{X_0}(x)\,dx=\int_{500}^{\infty}(x-500)\frac{6(1,000)^6}{(1,000+x)^7}\,dx$$

Substituting $t=x-500$, we see that:

$$E(Z_0)=\int_0^{\infty}t\frac{6(1,000)^6}{(1,500+t)^7}\,dt$$

We can rewrite this as follows:

$$E(Z_0)=\left(\frac{1,000}{1,500}\right)^6\int_0^{\infty}t\frac{6(1,500)^6}{(1,500+t)^7}\,dt$$

The integrand is of the form $t\,f_T(t)$, where $T\sim Pa(6,1500)$. So:

$$E(Z_0)=\left(\frac{1,000}{1,500}\right)^6 E(T)=\left(\frac{1,000}{1,500}\right)^6\times\frac{1,500}{5}=26.337$$

The only change from Year 0 to Year 1 is in the λ parameter. So, using the same approach:

$$E(Z_1)=\int_{500}^{\infty}(x-500)\frac{6(1,100)^6}{(1,100+x)^7}\,dx=\int_0^{\infty}t\frac{6(1,100)^6}{(1,600+t)^7}\,dt=\left(\frac{1,100}{1,600}\right)^6\int_0^{\infty}t\frac{6(1,600)^6}{(1,600+t)^7}\,dt$$

The final integral is the mean of the $Pa(6,1600)$ distribution. So:

$$E(Z_1)=\left(\frac{1,100}{1,600}\right)^6\times\frac{1,600}{5}=33.790$$

Similarly:

$$E(Z_2)=\int_{500}^{\infty}(x-500)\frac{6(1,210)^6}{(1,210+x)^7}\,dx=\int_0^{\infty}t\frac{6(1,210)^6}{(1,710+t)^7}\,dt=\left(\frac{1,210}{1,710}\right)^6\int_0^{\infty}t\frac{6(1,710)^6}{(1,710+t)^7}\,dt$$

$$=\left(\frac{1,210}{1,710}\right)^6\times\frac{1,710}{5}=42.930$$

The percentage increase from Year 0 to Year 1 is 28.3%, and the percentage increase from Year 1 to Year 2 is 27.1%. These figures are more than 10% as expected.

18.6 Since we are given information about claim payments made by the reinsurer, we need to consider the reinsurer's conditional claim amount random variable. This has PDF:

$$g(w)=\frac{f(w+M)}{1-F(M)}$$

The gross claim amount random variable has a Weibull distribution with parameters c and 2. So:

$$g(w)=\frac{2c(w+M)e^{-c(w+M)^2}}{e^{-cM^2}}=2c(w+3)e^{-c(w^2+6w)}$$

So the likelihood function based on a random sample of n payments made by the reinsurer is:

$$L(c)=2^n c^n\prod_{i=1}^{n}(w_i+3)\exp\left(-c\sum_{i=1}^{n}(w_i^2+6w_i)\right)$$

Taking logs:

$$\ln L=n\ln 2+n\ln c+\sum_{i=1}^{n}\ln(w_i+3)-c\sum_{i=1}^{n}(w_i^2+6w_i)$$

Differentiating with respect to c:

$$\frac{\partial}{\partial c}\ln L=\frac{n}{c}-\sum_{i=1}^{n}(w_i^2+6w_i)$$

This is 0 when:

$$c=\frac{n}{\sum_{i=1}^{n}(w_i^2+6w_i)}$$

Differentiating again:

$$\frac{\partial^2}{\partial c^2}\ln L=-\frac{n}{c^2}$$

This is negative, so we have a maximum.

Substituting in the given numerical values, we find that:

$$\hat{c}=\frac{10}{92.3+6\times 8.7}=0.0692$$

18.7 (i) ***Proof***

We want to simplify the integral:

$$\int_a^\infty x \frac{1}{x\sigma\sqrt{2\pi}} e^{-½\left(\frac{\log x-\mu}{\sigma}\right)^2} dx$$

Making the substitution $u = \dfrac{\log x - \mu}{\sigma} - \sigma$, the integral becomes:

$$\int_{\frac{\log a-\mu}{\sigma}-\sigma}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-½(u+\sigma)^2} \sigma e^{\mu+u\sigma+\sigma^2} du \qquad [1]$$

Multiplying out the brackets in the exponent and simplifying, we have:

$$e^{\mu+½\sigma^2} \int_{\frac{\log a-\mu}{\sigma}-\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-½u^2} du \qquad [1]$$

This integrand is the PDF of the standard normal distribution and the integral is:

$$P\left(N(0,1) > \frac{\log a - \mu}{\sigma} - \sigma\right) \qquad [1]$$

So we have:

$$\int_a^\infty x\, f(x)dx = e^{\mu+½\sigma^2}\left[1-\Phi\left(\frac{\log a - \mu}{\sigma} - \sigma\right)\right] \qquad [1]$$

This is the required result.

(ii)(a) ***Probability***

We first need the parameter values for the lognormal distribution. Using the formulae for the mean and variance of the lognormal distribution from page 14 of the *Tables* we have the following equations:

$$e^{\mu+½\sigma^2} = 9.070 \qquad \text{and} \qquad e^{2\mu+\sigma^2}(e^{\sigma^2}-1) = 10.132^2$$

Solving these simultaneous equations (by squaring the first equation and then substituting into the second equation), we obtain the values:

$$\sigma^2 = 0.80999 \qquad \text{and} \qquad \mu = 1.79998 \qquad [2]$$

The probability that a claim involves the reinsurer is the probability that it exceeds the retention limit. So if X represents the amount of a claim, we have, for the first reinsurance arrangement:

$$P(X>25)=P\left(\log N(\mu,\sigma^2)>25\right)=P\left(N(\mu,\sigma^2)>\log 25\right)$$

$$=P\left(N(0,1)>\frac{\log 25-\mu}{\sigma}\right)=1-\Phi\left(\frac{\log 25-\mu}{\sigma}\right) \qquad [1]$$

Substituting in the values for μ and σ^2, we get:

$$P(X>25)=1-\Phi(1.57656)=0.05745 \qquad [1]$$

So the probability that a claim will involve the reinsurer if the first arrangement is in force is 5.745%.

Using exactly the same argument for the second arrangement, we get:

$$P(X>30)=1-\Phi\left(\frac{\log 30-\mu}{\sigma}\right)=1-\Phi(1.77914)=0.03761 \qquad [1]$$

So the probability that a claim will involve the reinsurer if the second arrangement is in force is 3.761%.

(ii)(b) ***Better arrangement***

Consider the first arrangement. The amount ceded to the reinsurer (*ie* the amount paid by the reinsurer on a claim) is:

$$Z=\begin{cases}0 & \text{if } X\leq 25\\ X-25 & \text{if } X>25\end{cases}$$

So:

$$E(Z)=\int_{25}^{\infty}(x-25)\ f(x)\ dx=\int_{25}^{\infty}x\ f(x)\ dx-25\int_{25}^{\infty}f(x)\ dx \qquad [1]$$

where $f(x)$ is the PDF of the original lognormal distribution.

We can calculate the first of these integrals by using the result from the first part of the question:

$$\int_{25}^{\infty}x\ f(x)\ dx=e^{\mu+½\sigma^2}\left[1-\Phi\left(\frac{\log 25-\mu-\sigma^2}{\sigma}\right)\right]$$

$$=9.070\left[1-\Phi(0.67657)\right]$$

$$=9.070\times 0.24934$$

$$=2.2615 \qquad [1]$$

The second integral is just the probability that we worked out in part (ii)(a). So:

$$E(Z) = 2.2615 - 25 \times 0.05745 = 0.8253$$ [1]

Working in exactly the same way for the second arrangement (where Z is now the amount paid by the reinsurer in excess of 30), we have:

$$\begin{aligned} E(Z) &= 9.070\left[1 - \Phi(0.87915)\right] - 30 \times 0.03761 \\ &= 9.070 \times 0.18966 - 30 \times 0.03761 \\ &= 0.5919 \end{aligned}$$ [1]

So the expected amount paid out by the reinsurer per £1 of premium is (under the first arrangement):

$$\frac{200 \times 0.2 \times 0.8253}{50} = £0.660$$ [1]

Under the second arrangement, it is:

$$\frac{200 \times 0.2 \times 0.5919}{40} = £0.592$$ [1]

So, other things being equal, the first arrangement looks better value.

(ii)(c) ***Better arrangement under new circumstances***

The new mean and standard deviation are now 9.7956 and 10.94256 respectively. So we can calculate the new parameter values:

$$e^{\mu + \frac{1}{2}\sigma^2} = 9.7956 \qquad \text{and} \qquad e^{2\mu + \sigma^2}(e^{\sigma^2} - 1) = 10.94256^2$$

Solving these exactly as we did before, we find that $\mu = 1.87694$ and σ^2 is unchanged at 0.80999. [2]

So the value for $E(Z)$ is now (under the first arrangement):

$$\begin{aligned} E(Z) &= 9.7956\left[1 - \Phi\left(\frac{\log 25 - 1.87694 - 0.80999}{\sqrt{0.80999}}\right)\right] \\ &\qquad - 25 \times \left[1 - \Phi\left(\frac{\log 25 - 1.87694}{\sqrt{0.80999}}\right)\right] \\ &= 9.7956\left[1 - \Phi(0.59106)\right] - 25\left[1 - \Phi(1.49105)\right] \\ &= 9.7956 \times 0.27724 - 25 \times 0.06797 \\ &= 1.0164 \end{aligned}$$ [2]

Under the second arrangement:

$$E(Z) = 9.7956\left[1-\Phi\left(\frac{\log 30 - 1.87694 - 0.80999}{\sqrt{0.80999}}\right)\right]$$

$$-30\times\left[1-\Phi\left(\frac{\log 30 - 1.87694}{\sqrt{0.80999}}\right)\right]$$

$$= 9.7956\left[1-\Phi(0.79364)\right] - 30\left[1-\Phi(1.69363)\right]$$

$$= 9.7956\times 0.21370 - 30\times 0.04517$$

$$= 0.73830$$ [2]

So working exactly as before, the payment per £1 of premium under the first arrangement is now:

$$\frac{200\times 0.2\times 1.0164}{50} = £0.813$$

In the second arrangement the corresponding figure is:

$$\frac{200\times 0.2\times 0.73830}{40} = £0.738$$

So the first arrangement is still better value for the insurer. [1]

18.8 (i) ***Distribution function***

Let X denote the amount of the loss and Y denote the amount paid by the insurer in respect of the loss. With a policy excess of E in force, we have:

$$Y = X - E \mid X > E$$

The CDF of Y is given by:

$$F_Y(y) = P(Y \le y) = P(X - E \le y \mid X > E)$$

$$= \frac{P(X - E \le y \text{ and } X > E)}{P(X > E)}$$

$$= \frac{P(X \le y + E \text{ and } X > E)}{P(X > E)}$$

$$= \frac{P(E < X \le y + E)}{P(X > E)}$$

$$= \frac{F_X(y+E) - F_X(E)}{1 - F_X(E)}$$ [2]

Since $X \sim Pa(\alpha,\lambda)$:

$$F_Y(y) = \frac{\left[1-\left(\frac{\lambda}{\lambda+y+E}\right)^{\alpha}\right]-\left[1-\left(\frac{\lambda}{\lambda+E}\right)^{\alpha}\right]}{1-\left[1-\left(\frac{\lambda}{\lambda+E}\right)^{\alpha}\right]}$$

$$= \frac{\left(\frac{\lambda}{\lambda+E}\right)^{\alpha}-\left(\frac{\lambda}{\lambda+y+E}\right)^{\alpha}}{\left(\frac{\lambda}{\lambda+E}\right)^{\alpha}}$$

$$= 1-\left(\frac{\lambda+E}{\lambda+y+E}\right)^{\alpha}, \quad y>0$$ [1]

So $Y \sim Pa(\alpha,\lambda+E)$.

(ii) ***Mean values***

The mean of the $Pa(\alpha,\lambda+E)$ distribution is $\dfrac{\lambda+E}{\alpha-1}$.

If $E=0$ then $E(Y)=\dfrac{15}{3}=5$. [1]

If $E=10$ then $E(Y)=\dfrac{25}{3}=8\dfrac{1}{3}$. [1]

(iii) ***Effect of introducing a policy excess***

Introducing a policy excess of E increases the mean claim amount paid by the insurer by $\dfrac{E}{\alpha-1}$. This is because small losses are met in full by the policyholder. [2]

It may still be advantageous to the insurer to introduce a policy excess, since although the average claim payment will increase, fewer claim payments will be made.

18.9 Let X be the loss amount random variable for this year, and let X' be the loss amount random variable for next year. Then:

$$X \sim Pa(4.5, 3000) \quad \text{and} \quad X' = 1.03X$$

The probability that a loss next year is borne entirely by the policyholder is:

$$P(X' \le 100) = P(X \le \tfrac{100}{1.03}) = 1-\left(\frac{3,000}{3,000+\frac{100}{1.03}}\right)^{4.5} = 0.13353$$ [2]

19

Risk models 1

Syllabus objectives

1.2 Compound distributions and their application in risk modelling

1.2.1 Construct models appropriate for short-term insurance contracts in terms of the numbers of claims and the amounts of individual claims.

1.2.2 Describe the major simplifying assumptions underlying the models in 1.2.1.

1.2.3 Define a compound Poisson distribution and show that the sum of independent random variables each having a compound Poisson distribution also has a compound Poisson distribution.

1.2.4 Derive the mean, variance and coefficient of skewness for compound binomial, compound Poisson and compound negative binomial random variables.

0 Introduction

In the first section of this chapter, we describe the main features of general insurance policies. There is no mathematics in this section, and you should be able to read it through fairly quickly in order to obtain a good overview of the different types of product available.

In the remaining sections of the chapter we introduce the idea of a *compound distribution*. We will define and use the compound Poisson, compound geometric, compound negative binomial and compound binomial distributions.

We will also start to look at two models, the *individual risk model* and the *collective risk model*, which are used to describe *aggregate claims*, *ie* the total claims that arise during a period from a group of policies.

In the simplest case of a life assurance benefit (often referred to as 'long-term business'), each policy can result in at most one claim and claims will be for amounts specified in advance (*ie* the sum assured). The benefit level may be the same for all policies or it may vary between policies.

In general insurance (often referred to as 'short-term business'), policies can give rise to more than one claim and the amounts will not usually be known in advance.

In this chapter we look at the theory of risk models. In the next chapter we explain how to adapt these models when reinsurance is in place.

1 General features of a product

1.1 Insurable interest

Generally, for a risk to be insurable:

- **the policyholder must have an interest in the risk being insured, to distinguish between insurance and a wager**
- **a risk must be of a financial and reasonably quantifiable nature.**

1.2 Insurable risk

Ideally risk events also need to meet the following criteria if they are to be insurable:

- **Individual risk events should be independent of each other.**

  In practice we won't often get strict independence but a low correlation is desirable.

- **The probability of the event should be relatively small. In other words, an event that is nearly certain to occur is not conducive to insurance.**

  For example, a house would not be insured if it stood on the edge of a crumbling cliff.

- **Large numbers of potentially similar risks should be pooled in order to reduce the variance and hence achieve more certainty.**

  The similar risks should still be independent.

- **There should be an ultimate limit on the liability undertaken by the insurer.**

  This would help the risk event meet the above criteria that it must be of a reasonably quantifiable nature.

- **Moral hazards should be eliminated as far as possible because these are difficult to quantify, result in selection against the insurer and lead to unfairness in treatment between one policyholder and another.**

  Moral hazards occur when a person takes more risks because another party bears the cost of those risks.

However, the desire for income means that an insurer or reinsurer will usually be found to provide cover when these ideal criteria are not met.

Other characteristics that most general insurance products share are:

- **Cover is normally for a fixed period, most commonly one year, after which it has to be renegotiated. There is normally no obligation on the insurer or insured to continue the arrangement thereafter although in most cases a need for continuing cover may be assumed to exist.**
- **Claims are not of fixed amounts, and the amount of loss as well as the fact needs to be proved before a claim can be settled.**
- **A claim occurring does not bring the policy to an end.**

- **Claims may occur at any time during the policy period. Although there is normally a contractual obligation on the policyholder to report a claim to the insurer as quickly as possible, notification may take some time if the loss is not evident immediately. Settlement of the claim may take a long time if protracted legal proceedings are needed or if it is not straightforward to determine the extent of the loss. However, from the moment of the event giving rise to the claim the ultimate settlement amount is a liability of the insurer. Estimating the amounts of money that need to be reserved to settle these liabilities is one of the most important areas of actuarial involvement in general insurance.**

Classes of insurance in which claims tend to take a long time to settle are known as long-tail. Those which tend to take a short time to settle are known as short-tail, although the dividing line between the two categories is not always distinct.

2 Models for short-term insurance contracts

2.1 The basic model

Many forms of non-life insurance can be regarded as short-term contracts, for example motor insurance. Some forms of life insurance also fall into this category, for example group life and one-year term assurance policies.

A short-term insurance contract can be defined as having the following attributes:

- **The policy lasts for a fixed, and relatively short, period of time, typically one year.**
- **The insurance company receives from the policyholder(s) a premium.**
- **In return, the insurer pays claims that arise during the term of the policy.**

At the end of the policy's term the policyholder may or may not renew the policy. If it is renewed, the premium payable by the policyholder may or may not be the same as in the previous period.

The insurer may choose to pass part of the premium to a reinsurer. In return, the reinsurer will reimburse the insurer for part of the cost of the claims during the policy's term according to some agreed formula.

An important feature of a short-term insurance contract is that the premium is set at a level to cover claims arising during the (short) term of the policy only. This contrasts with life assurance policies where mortality rates increasing with age mean that the (level) annual premium in the early years would be more than sufficient to cover the expected claims in the early years. The excess amount would then be accumulated as a reserve to be used in the later years when the premium on its own would be insufficient to meet the expected cost of claims.

Now to be more specific, a short-term insurance contract covering a risk will be considered. A risk includes either a single policy or a specified group of policies. For ease of terminology the term of the contract is assumed to be one year, but it could equally well be any other short period, for example six months. The random variable S denotes the aggregate claims paid by the insurer in the year in respect of this risk. Models will be constructed for this random variable S. In Section 3 collective risk models will be studied. Later, in the next chapter, the idea of a collective risk model is extended to an individual risk model.

We will see shortly what these terms mean.

A first step in the construction of a collective risk model is to write S in terms of the number of claims arising in the year, denoted by the random variable N, and the amount of each individual claim. Let the random variable X_i denote the amount of the ith claim. Then:

$$S=\sum_{i=1}^{N} X_i \qquad \textbf{(19.1)}$$

where the summation is taken to be zero if N is zero.

This decomposition of S allows consideration of claim numbers and claim amounts separately. A practical advantage of this is that the factors affecting claim numbers and claim amounts may well be different. Take motor insurance as an example. A prolonged spell of bad weather may have a significant effect on claim numbers but little or no effect on the distribution of individual claim amounts. On the other hand, inflation may have a significant effect on the cost of repairing cars, and hence on the distribution of individual claim amounts, but little or no effect on claim numbers.

This approach is referred to as a *collective risk model* because it is considering the *claims* arising from a group of policies taken as a whole, rather than by considering the claims arising from each individual policy.

The random variable S is the sum of a random number of random quantities, and is said to have a *compound distribution*.

Because compound distributions arise commonly in general insurance examples, the random variable N is often referred to as the 'number of claims' and the distribution of the random variables $X_1, X_2, \ldots$ is referred to as the 'individual claim size distribution', even where the compound distribution arises in another context.

To define a compound distribution, we need to know:

- the distribution of N (which is a *discrete* distribution) and
- the distribution of the X_i's (which may be *any* distribution).

If the distribution of the X_i's is continuous, then S will have a *mixed distribution*, *ie* partly discrete and partly continuous. This is because of the possibility that $N=0$.

The problems that will be studied are the derivation of the moments and distribution of S in terms of the moments and distributions of N and the X_i's. Both will be studied with and without simple forms of reinsurance. The corresponding problems for the reinsurer will also be studied, *ie* the derivation of the moments and distribution of the aggregate claims paid in the year in respect of this risk by the reinsurer.

2.2 Discussion of the simplifications in the basic model

The model for short-term insurance described in the previous subsection contains a number of simplifications as compared to a real insurance operation. The first of these is that it is usually assumed that the moments, and sometimes the distributions, of N and the X_i's are known with certainty. In practice these would probably be estimated from some relevant data.

For example, we might assume that claim amounts have a $Gamma(500,4)$ distribution.

In practice it may not be possible to make such simple assumptions. For example:

- There may not be an appropriate theoretical distribution that models the distribution of claim amounts actually paid sufficiently well.
- Even if the shape of the distribution is satisfactory, appropriate parameter values may change over time, even in the short term.

- There may not be sufficient homogeneity in the portfolio. For example, different policies may produce claim amounts that have different sizes. This leads to the idea of a mixture distribution.

Another simplification is to assume, at least implicitly, that claims are settled more or less as soon as the incident causing the claim occurs, so that, for example, the insurer's profit is known at the end of the year. In practice, there will be at least a short delay in the settlement of claims and in some cases the delay can amount to many years. This will be especially true when the extent of the loss is difficult to determine, for example if it is to be decided in a court of law.

Delays will often lead to higher payments being made, owing to inflationary factors. (The relevant inflation rate may well be very different from that normally used to measure inflation.)

The model does not in general include any mention of expenses. The premium is assumed to pay the claims and include a loading for profit. In practice, the premium paid by the policyholder(s) will also include a loading for expenses. It is possible to include expenses in the model in a very simple way.

The simplest way to allow for expenses would be to use a claim size distribution that was artificially inflated to allow for some sort of claim expense amount (*eg* adding an extra 20%), although this might not give the right 'shape'. Alternatively we might express the random variable X as the sum of two other random variables, one to represent the actual claim amount and the other to represent the corresponding claim expense.

An important element in models for long-term insurance is a rate of interest since, as explained above, excess premium income would be invested to build up reserves. Interest is a relatively less important, but still important, feature of short-term insurance. It is possible to include interest in models for short-term insurance but it is more usual to ignore it, at least in elementary models.

We will ignore interest in the models used in this chapter.

There are a number of additional elements included when setting the premium to be charged to policyholders, including the policyholders' previous claims record and these are covered in Subject CP1 – Actuarial Practice. The allowance for policyholders' claim experience could be based on claim frequency or total claim amounts. This is beyond the scope of this subject.

In fact, we have already looked at models that make allowance for policyholders' claims experience in Chapter 2.

2.3 Notation and assumptions

Throughout this chapter the following two important assumptions will be made:

- **the random variables $\{X_i\}_{i=1}^{N}$ are independent and identically distributed**
- **the random variable N is independent of $\{X_i\}_{i=1}^{N}$.**

In words these assumptions mean that:

1. **the number of claims is not affected by the amount of individual claims**
2. **the amount of a given individual claim is not affected by the amount of any other individual claim**
3. **the distribution of the amounts of individual claims does not change over the (short) term of the policy.**

Point 1 follows from the second of the two assumptions above. Points 2 and 3 follow from the first.

Throughout this chapter it will be assumed that all claims are for non-negative amounts, so that $P(X_i \le x) = 0$ for $x < 0$. Many of the formulae in this chapter will be derived using the moment generating functions (from now on abbreviated to MGFs) of S, N and X_i. These MGFs will be denoted $M_S(t)$, $M_N(t)$ and $M_X(t)$, respectively, and will be assumed to exist for some positive values of the dummy variable t. The existence of the MGF of a non-negative random variable for positive values of t cannot generally be taken for granted; for example the MGFs of the Pareto and of the lognormal distributions do not exist for any positive value of t. However, all the formulae derived in this chapter with the help of MGFs can be derived, although less easily, without assuming the MGFs exist for positive values of t.

One method would be to use characteristic functions, $E(e^{itX})$, which don't have the same convergence problems as MGFs. However the Core Reading does not cover these.

$G(x)$ and $F(x)$ shall denote the distribution functions of S and X_i, respectively, so that:

$$G(x) = P(S \le x) \text{ and } F(x) = P(X_i \le x)$$

For convenience it will often be assumed that the density of $F(x)$ exists and it will be denoted $f(x)$. In cases where this density does not exist, so that X_i has a discrete or a mixed continuous/discrete distribution, expressions such as:

$$\int_0^\infty x\, f(x)\, dx$$

should be interpreted appropriately. The meaning should always be clear from the context.

The kth moment, ($k = 1,2,3...$) of X_i about zero will be denoted m_k.

Using this notation:

$$E(X_i) = m_1 \quad \text{and} \quad \text{var}(X_i) = m_2 - m_1^2$$

3 The collective risk model

3.1 The collective risk model

Recall from Section 2.1 that S is represented as the sum of N random variables X_i, where X_i denotes the amount of the ith claim. Thus:

$$S = X_1 + X_2 + \cdots + X_N$$

and $S = 0$ if $N = 0$.

S is said to have a *compound distribution*.

Note that it is the number of claims, N, from the risk as a collective (as opposed to counting the number of claims from individual policies) that is being considered and this gives the name 'collective risk model'. Within this framework, expressions in general terms for the distribution function, mean, variance and MGF of S can be developed.

3.2 Distribution functions and convolutions

An expression for $G(x)$, the distribution function of S, can be derived by considering the event $\{S \le x\}$. Note that if this event occurs, then one, and only one, of the following events must occur:

	$\{S \le x$ and $N = 0\}$	**(*ie* no claims)**
or	**$\{S \le x$ and $N = 1\}$**	**(*ie* one claim of amount $\le x$)**
or	**$\{S \le x$ and $N = 2\}$**	**(*ie* two claims which total $\le x$)**
	$\vdots$	
or	**$\{S \le x$ and $N = r\}$**	**(ie *r* claims which total $\le x$)**
	$\vdots$	

and so on. These events are mutually exclusive and exhaustive.

Thus:

$$\{S \le x\} = \bigcup_{n=0}^{\infty} \{S \le x \text{ and } N = n\}$$

and hence:

$$\begin{aligned} P(S \le x) &= \sum_{n=0}^{\infty} P(S \le x \text{ and } N = n) \\ &= \sum_{n=0}^{\infty} P(N = n) P(S \le x \mid N = n) \end{aligned}$$

Question

A group of policies can give rise to at most two claims in a year. The probability function for the number of claims is as follows:

Number of claims, n	0	1	2
$P(N=n)$	0.6	0.3	0.1

Each claim is either for an amount of 1 or an amount of 2, with equal probability. Claim amounts are independent of one another and are independent of the number of claims.

Determine the distribution function of the aggregate annual claim amount, S.

Solution

S can take the values 0, 1, 2, 3 or 4.

S will only equal 0 if $N=0$ and this has probability 0.6. So:

$$P(S\le 0)=P(S=0)=P(N=0)=0.6$$

S will equal 1 if there is one claim for amount 1. So:

$$\begin{aligned}P(S=1)&=P(N=1,X=1)\\&=P(N=1)P(X=1) \qquad \text{by independence}\\&=0.3\times 0.5\\&=0.15\end{aligned}$$

and:

$$P(S\le 1)=P(S=0)+P(S=1)=0.6+0.15=0.75$$

The other values of the CDF can be calculated in a similar way and are given below:

$$P(S\le 2)=P(S\le 1)+P(S=2)=0.75+0.3\times 0.5+0.1\times 0.5^2=0.925$$

$$P(S\le 3)=P(S\le 2)+P(S=3)=0.925+0.1\times 2\times 0.5^2=0.975$$

$$P(S\le 4)=P(S\le 3)+P(S=4)=1$$

The distribution of a sum of independent random variables can be found using *convolutions*.

If $Z=X+Y$, where X and Y are independent random variables with PDFs (or PFs) $f_X(x)$ and $f_Y(y)$, then $f_Z(z)$, the PDF (or PF) of Z, is called the *convolution* of X and Y.

This is written mathematically as $f_Z=f_X*f_Y$.

A formula for a convolution can be found by summing over all possible values of x and y that give a particular value z .

Finding a convolution

$$f_Z(z)=\sum_x f_X(x)f_Y(z-x) \qquad \text{for } \textit{discrete} \text{ random variables}$$

$$f_Z(z)=\int f_X\left(x\right)f_Y(z-x)dx \qquad \text{for } \textit{continuous} \text{ random variables}$$

'Sum or integrate over all values of x that could lead to a total of z .'

Similar formulae can be used to find the distribution function of a sum.

$$F_Z(z)=\sum_x f_X(x)F_Y(z-x) \qquad \text{or} \qquad \sum_x F_X(x)f_Y(z-x)$$

$$F_Z\left(z\right)=\int f_X(x)F_Y(z-x)dx \qquad \text{or} \qquad F_Z\left(z\right)=\int F_X(x)f_Y(z-x)dx$$

Question

Suppose that $N \sim Poisson(\lambda)$, $M \sim Poisson(\mu)$, and N and M are independent.

Use a convolution approach to derive the probability function of $N+M$.

Solution

Let $V=N+M$. Then, for $v=0,1,2,\ldots$:

$$\begin{aligned}
P(V=v) &= \sum_{n=0}^{v} P(N=n)P(M=v-n) \\
&= \sum_{n=0}^{v} \frac{\lambda^n e^{-\lambda}}{n!}\,\frac{\mu^{v-n} e^{-\mu}}{(v-n)!} \\
&= \frac{e^{-(\lambda+\mu)}}{v!}\sum_{n=0}^{v} \frac{v!}{n!(v-n)!}\lambda^n \mu^{v-n} \\
&= \frac{e^{-(\lambda+\mu)}}{v!}\sum_{n=0}^{v} \binom{v}{n}\lambda^n \mu^{v-n} \\
&= \frac{e^{-(\lambda+\mu)}}{v!}(\lambda+\mu)^v
\end{aligned}$$

This is the probability function of the $Poisson(\lambda+\mu)$ distribution. So $N+M \sim Poisson(\lambda+\mu)$.

This result can also be proved using MGFs.

We can now consider the distribution function for a compound distribution.

For convolutions of distribution functions, suppose that $\{X_i\}_{i=1}^n$ are independent and identically distributed (IID) random variables with common distribution function $F(x)$.

Then the distribution function of $\sum_{i=1}^{n} X_i$ is denoted by $F^{n^*}(x)$, so that:

$$F^{n^*}(x) = P(X_1 + X_2 + \cdots + X_n \leq x)$$

Now note that if $N = n$, then S is the sum of a fixed number n, of random variables, $\{X_i\}_{i=1}^n$, and hence:

$$P(S \leq x \mid N = n) = F^{n^*}(x)$$

where $F^{n^*}(x)$ is the n-fold convolution of the distribution $F(x)$.

In other words, $F^{3^*}(x)$ would be the convolution $F(x)*F(x)*F(x)$ *etc*.

(Note that $F^{1^*}(x)$ is just $F(x)$ and, for convenience, $F^{0^*}(x)$ is defined to equal 1 for all non-negative values of x. Otherwise $F^{0^*}(x) = 0$.) Thus:

$$G(x) = P(S \leq x) = \sum_{n=0}^{\infty} P(N = n) F^{n^*}(x) \qquad (19.2)$$

Formula (19.2) is a general expression for the distribution function of S. Neither the distribution of N nor of X_i has been specified.

Note that when X_i is distributed on the positive integers it is easy to calculate $P(S = x)$ for $x = 1,2,3,\ldots$ since:

$$P(S = x) = G(x) - G(x-1)$$

$$= \sum_{n=1}^{\infty} P(N = n)\left[F^{n^*}(x) - F^{n^*}(x-1)\right]$$

ie $$P(S = x) = \sum_{n=1}^{\infty} P(N = n) f_x^{n^*} \qquad (19.3)$$

where $f_x^{n^*} = F^{n^*}(x) - F^{n^*}(x-1)$ is the probability function of $\sum_{i=1}^{n} X_i$.

This is just saying that $f_x^{n^*} = P(\sum_{i=1}^{n} X_i = x)$.

As in the case when X_i is a continuous random variable, $P(S = 0) = P(N = 0)$.

When the number of claims is large, and the claim amount distribution is not too skewed, we can approximate the distribution of S by a normal distribution with mean $E(S)$ and variance $\text{var}(S)$. We explain how to calculate moments of S in the next section.

3.3 Moments of compound distributions

To calculate the moments of S, conditional expectation results are used, conditioning on the number of claims, N. To find $E[S]$, apply the identity:

$$\boldsymbol{E[S] = E[E[S|N]]}$$

Here we are using the conditional expectation formula, which is given on page 16 of the *Tables*.

Now $E[S \mid N=n] = \sum_{i=1}^{n} E[X_i] = nm_1$. Hence:

$$\boldsymbol{E[S \mid N] = Nm_1}$$

and:

$$\boldsymbol{E[S] = E[Nm_1] = E[N]m_1} \qquad \textbf{(19.4)}$$

This can also be written as follows:

$$E(S) = E(N)E(X)$$

where $E(X) = E(X_i)$, $i = 1,2,\ldots,N$.

Formula (19.4) has a very natural interpretation. It says that the expected aggregate claim amount is the product of the expected number of claims and the expected individual claim amount.

This formula is also given on page 16 of the *Tables.*

Question

If X has a Pareto distribution with parameters $\lambda = 400$ and $\alpha = 3$, and N has a *Poisson*(50) distribution, calculate the expected value of S.

Solution

The expected value is:

$$E(S) = E(N)E(X) = 50 \times \frac{400}{3-1} = 10{,}000$$

Variance

To find an expression for $\text{var}[S]$, apply the identity:

$$\textbf{var}[S] = E[\text{var}[S \mid N]] + \text{var}[E[S \mid N]]$$

Here we are using the conditional variance formula, which is given on page 16 of the *Tables*.

Since $E(S \mid N) = Nm_1$, we have:

$$\text{var}[S] = E[\text{var}[S \mid N]] + \text{var}[Nm_1]$$

$\text{var}[S \mid N]$ can be found by using the fact that individual claim amounts are independent.

Now:

$$\text{var}[S \mid N = n] = \text{var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{var}[X_i] = n(m_2 - m_1^2)$$

and so $\text{var}[S \mid N] = N(m_2 - m_1^2)$. Hence:

$$\text{var}[S] = E[N(m_2 - m_1^2)] + \text{var}[Nm_1]$$

***ie*:** $$\text{var}[S] = E[N](m_2 - m_1^2) + \text{var}[N]m_1^2 \qquad \textbf{(19.5)}$$

Alternatively, writing this solely in terms of means and variances:

$$\text{var}(S) = E(N)\text{var}(X) + \text{var}(N)[E(X)]^2$$

where $\text{var}(X) = \text{var}(X_i)$, $i = 1, 2, \ldots, N$.

This formula can also be found on page 16 of the *Tables*. We will use it to determine the variances of the various compound distributions in the next few sections.

Unlike the expression for $E[S]$, formula (19.5) does not have a natural interpretation. The variance of S is expressed in terms of the mean and variance of both N and X_i.

However, this formula shows that the variability of the overall aggregate claim distribution is a function of both the variability in the number of claims and the variability in the claim amounts.

Moment generating function

The MGF of S is also found using conditional expectation. By definition, $M_S(t) = E[\exp(tS)]$, so:

$$M_S(t) = E\big[E[\exp(tS) \mid N]\big] \qquad \textbf{(19.6)}$$

Again, we are conditioning on the number of claims, exactly as we did before.

Now $E[\exp(tS)|\,N=n]=E[\exp(tX_1+tX_2+\ldots+tX_n)]$, and as $\{X_i\}_{i=1}^n$ are independent random variables:

$$E[\exp(tX_1+tX_2+\ldots+tX_n)]=\prod_{i=1}^{n}E[\exp(tX_i)]$$

Also, since $\{X_i\}_{i=1}^n$ are identically distributed, they have common MGF, $M_X(t)$, so that:

$$\prod_{i=1}^{n}E[\exp(tX_i)]=\prod_{i=1}^{n}M_X(t)=[M_X(t)]^n$$

Hence:

$$E[\exp(tS)\,|\,N]=[M_X(t)]^N \tag{19.7}$$

These conditional expectations are random variables because they are functions of N.

Hence, inserting (19.7) in (19.6):

$$M_S(t)=E[M_X(t)^N]=E[\exp(N\log M_X(t))]=M_N(\log M_X(t)) \tag{19.8}$$

We can see this last step by observing that $E[\exp(N\log M_X(t))]$ is of the same form as $E(e^{Nt})$ but with t replaced by $\log M_X(t)$. So it is the MGF of N evaluated at $\log M_X(t)$.

Again, this is given on page 16 of the *Tables*.

Thus, the MGF of S is expressed in terms of the MGFs of N and of X_i. As with the previous results, the distributions of neither N nor of X_i have been specified.

A summary of the results for the mean, variance and MGF of S is given below.

Mean, variance and MGF of *S*

$$E(S)=E(N)E(X)$$

$$\text{var}(S)=E(N)\text{var}(X)+\text{var}(N)[E(X)]^2$$

$$M_S(t)=M_N\left(\log M_X(t)\right)$$

There is one special case that is of some interest. This is when all claims are for the same fixed amount.

Example

Consider a portfolio of one-year term assurances each with the same sum assured. Suppose that the amount of a claim is B with probability one (assuming that a claim occurs at all), ***ie* $P(X_i=B)=1$ so that $m_1=B$ and $m_2=B^2$.**

B is a constant here, not a random variable. So the expected value of an individual claim amount is B and its variance is 0.

Then S is distributed on $0, B, 2B, \ldots$. In fact, $S = BN$ so:

$$P(S \le Bx) = P(N \le x)$$

Formulae (19.4) and (19.5) give the mean and variance of S, but as $S = BN$ it is easier to note that $E[S] = E[N]B$ and $\text{var}[S] = \text{var}[N]B^2$.

The next three sections consider compound distributions using various models for the number of claims, N.

3.4 The compound Poisson distribution

First consider aggregate claims when N has a Poisson distribution with mean λ denoted $N \sim Poi(\lambda)$. S then has a compound Poisson distribution with parameter λ, and $F(x)$ is the CDF of the individual claim amount random variable.

S is sometimes referred to as a compound Poisson random variable.

The results required for this distribution for N are:

$$E[N] = \text{var}[N] = \lambda$$

$$M_N(t) = \exp\left[\lambda(e^t - 1)\right]$$

Note that these results are given in the *Tables*.

These results can be combined with those of Section 3.1 as follows.

From (19.4):

$$E(S) = E(N)E(X) = \lambda E(X)$$

ie:

$$E[S] = \lambda m_1 \qquad \textbf{(19.9)}$$

From (19.5):

$$\text{var}(S) = E(N)\,\text{var}(X) + \text{var}(N)[E(X)]^2 = \lambda\,\text{var}(X) + \lambda[E(X)]^2 = \lambda E(X^2)$$

ie:

$$\text{var}[S] = \lambda m_2 \qquad \textbf{(19.10)}$$

and from (19.8):

$$M_S(t) = \exp\left[\lambda(M_X(t) - 1)\right] \qquad \textbf{(19.11)}$$

The results for the mean and variance have a very simple form. Note that the variance of S is expressed in terms of the second moment of X_i about zero (and not in terms of the variance of X_i).

Note also that the formula for the skewness of S has a simple form when S is a compound Poisson random variable:

$$\boldsymbol{skew[S] = \lambda m_3} \qquad \textbf{(19.12)}$$

***ie*:**

$$skew(S) = \lambda E(X^3)$$

The easiest way to show that the third central moment of S is λm_3 is to use the cumulant generating function:

$$\boldsymbol{C_S(t) = \log M_S(t)}$$

To determine the skewness, we differentiate it three times with respect to t and set $t = 0$, *ie*:

$$\boldsymbol{E[(S - \lambda m_1)^3] = \frac{d^3}{dt^3} \log M_S(t) \Big|_{t=0}}$$

In other words:

$$skew(S) = C_S'''(0)$$

Recall also that:

$$E(S) = C_S'(0)$$

$$\operatorname{var}(S) = C_S''(0)$$

Since $M_S(t) = \exp\left[\lambda(M_X(t) - 1)\right]$, it follows that:

$$\boldsymbol{\log M_S(t) = \lambda\left(M_X(t) - 1\right)}$$

So:

$$\boldsymbol{\frac{d^3}{dt^3} \log M_S(t) = \lambda \left[\frac{d^3}{dt^3} M_X(t) - 1 \right]_{t=0} = \lambda m_3}$$

ie $\quad \boldsymbol{E[(S - \lambda m_1)^3] = \lambda m_3}$

This is because $M_X'''(0) = E(X^3)$ for any random variable.

The coefficient of skewness of S is given by:

$$\frac{skew(S)}{[\text{var}(S)]^{3/2}}$$

Hence the coefficient of skewness $= \lambda m_3 / (\lambda m_2)^{3/2}$.

This result shows that the distribution of S is positively skewed, since m_3 is the third moment about zero of X_i and hence is greater than zero because X_i is a non-negative valued random variable. Note that the distribution of S is positively skewed even if the distribution of X_i is negatively skewed. The coefficient of skewness of S is $\lambda m_3 / (\lambda m_2)^{3/2}$, and hence goes to 0 as $\lambda \to \infty$. Thus for large values of λ, the distribution of S is almost symmetric.

Mean, variance and skewness of a compound Poisson random variable

If $N \sim Poisson(\lambda)$, then S is a compound Poisson random variable and:

$$E(S) = \lambda E(X) = \lambda m_1$$

$$\text{var}(S) = \lambda E(X^2) = \lambda m_2$$

$$skew(S) = \lambda E(X^3) = \lambda m_3$$

These results are all given on page 16 of the *Tables*.

Sums of independent compound Poisson random variables

A very useful property of the compound Poisson distribution is that the sum of independent compound Poisson random variables is itself a compound Poisson random variable. A formal statement of this property is as follows.

Let $S_1, S_2, ..., S_n$ be independent random variables. Suppose that each S_i has a compound Poisson distribution with parameter λ_i, and that the CDF of the individual claim amount random variable for each S_i is $F_i(x)$.

Define $A = S_1 + S_2 + \cdots + S_n$. Then A has a compound Poisson distribution with parameter Λ, and $F(x)$ is the CDF of the individual claim amount random variable for A, where:

$$\Lambda = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad F(x) = \frac{1}{\Lambda} \sum_{i=1}^{n} \lambda_i F_i(x)$$

Recall that Λ is the capital form of the Greek letter λ.

This is a very important result.

To prove the result, first note that $F(x)$ is a weighted average of distribution functions and that these weights are all positive and sum to one. This means that $F(x)$ is a distribution function and this distribution has MGF:

$$M(t) = \int_0^\infty e^{tx} f(x)\,dx$$

where $f(x) = F'(x)$ is the PDF of the individual claim amount random variable for A.

So:

$$M(t) = \int_0^\infty \exp(tx)\, \frac{1}{\Lambda} \sum_{i=1}^{n} \lambda_i\, f_i(x)\,dx$$

where $f_i(x)$ is the density of $F_i(x)$. Hence:

$$M(t) = \frac{1}{\Lambda} \sum_{i=1}^{n} \lambda_i \int_0^\infty \exp\{tx\} f_i(x)\,dx = \frac{1}{\Lambda} \sum_{i=1}^{n} \lambda_i\, M_i(t) \qquad (19.13)$$

where $M_i(t)$ is the MGF for the distribution with CDF $F_i(x)$.

Let $M_A(t)$ denote the MGF of A. Then:

$$M_A(t) = E[\exp(tA)] = E[\exp(tS_1 + tS_2 + \cdots + tS_n)]$$

By independence of $\{S_i\}_{i=1}^n$:

$$M_A(t) = \prod_{i=1}^{n} E(\exp(tS_i))$$

As S_i is a compound Poisson random variable, its MGF is of the form given by formula (19.11), so:

$$E[\exp(tS_i)] = \exp\left[\lambda_i (M_i(t) - 1)\right]$$

Thus:

$$M_A(t) = \prod_{i=1}^{n} \exp\{\lambda_i (M_i(t) - 1)\} = \exp\left\{\sum_{i=1}^{n} \lambda_i (M_i(t) - 1)\right\}$$

***ie*:**

$$M_A(t) = \exp\{\Lambda (M(t) - 1)\} \qquad (19.14)$$

where:

$$\Lambda = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad M(t) = \frac{1}{\Lambda} \sum_{i=1}^{n} \lambda_i M_i(t)$$

By the one-to-one relationship between distributions and MGFs, formula (19.14) shows that A has a compound Poisson distribution with Poisson parameter Λ. By (19.13), the individual claim amount distribution has CDF *F*(*x*).

Question

The distributions of aggregate claims from two risks, denoted by S_1 and S_2, are as follows:

- S_1 has a compound Poisson distribution with parameter 100 and distribution function $F_1(x) = 1 - \exp(-x/\alpha)$, $x > 0$.
- S_2 has a compound Poisson distribution with parameter 200 and distribution function $F_2(x) = 1 - \exp(-x/\beta)$, $x > 0$.

Assuming that S_1 and S_2 are independent, determine the distribution of $S_1 + S_2$.

Solution

Let $S = S_1 + S_2$. Then S has a compound Poisson distribution with parameters $\Lambda = 300$ and $F(x)$, where:

$$F(x) = \tfrac{1}{3}F_1(x) + \tfrac{2}{3}F_2(x) = 1 - \tfrac{1}{3}\exp(-x/\alpha) - \tfrac{2}{3}\exp(-x/\beta)$$

We can use R to simulate values from a compound Poisson distribution.

The R code to simulate 10,000 values from a compound Poisson distribution with parameter 1,000 and a gamma claims distribution with $\alpha = 750$ and $\lambda = 0.25$ is:

```
set.seed(123)
n <- rpois(10000,1000)
s <- numeric(10000)
for(i in 1:10000)
{x <- rgamma(n[i],shape=750,rate=0.25)
s[i] <- sum(x)}
```

We can obtain a mean of 2,997,651, a standard deviation of 93,719.71 and a coefficient of skewness of 0.02655921 as follows:

```
mean(s)
sd(s)
skewness<-sum((s-mean(s))^3)/length(s)
coeff.of.skew<-skewness/var(s)^(3/2)
```

We can estimate P(S>3,000,000) to be 0.4881 as follows:

```
length(s[s>3000000])/length(s)
```

Finally we could estimate the 90th percentile to be 3,115,719 as follows::

```
quantile(s,0.9)
```

We can plot a histogram of the compound distribution using the `hist` function and an empirical density function using `density` in the `plot` function. We can then superimpose a normal or other distribution to see if they provide a good approximation.

However, a better way to check the fit with a normal distribution is to use the `qqnorm` function:

```
qqnorm(<simulated values>)
```

or the `qqplot` function to compare the sample data to simulated values from a fitted model distribution:

```
qqplot(<simulated theoretical values>,
<simulated compound distribution values>)
```

Note we have used `set.seed(123)` so you can obtain the same values as this example.

3.5 The compound binomial distribution

Under certain circumstances, the binomial distribution is a natural choice for N. For example, under a group life insurance policy covering n lives, the distribution of the number of deaths in a year is binomial if it is assumed that each insured life is subject to the same mortality rate, and that lives are independent with respect to mortality.

The notation $N \sim Bin(n,p)$ is used to denote the binomial distribution for N. The key results for this distribution are:

$$E[N] = np$$

$$\text{var}[N] = np(1-p)$$

$$M_N(t) = (pe^t + 1 - p)^n$$

Note that these results are given in the *Tables*.

However, the notation for the MGF is slightly different.

When N has a binomial distribution, S has a compound binomial distribution. One important point about choosing the binomial distribution for N is that there is an upper limit, n, to the number of claims.

Expressions for the mean, variance and MGF of S are now found in terms of n, p, m_1, m_2 and $M_X(t)$ when $N \sim Bin(n,p)$.

There is no need to memorise the formulae in this section. However, it is important to be able to derive them.

Formula (19.4) gives the mean:

$$E(S) = E(N)E(X)$$

$$\Rightarrow E[S] = npm_1 \qquad \textbf{(19.15)}$$

Formula (19.5) gives the variance:

$$\text{var}(S) = E(N)\,\text{var}(X) + \text{var}(N)[E(X)]^2$$

$$\Rightarrow \mathbf{var}[\boldsymbol{S}] = \boldsymbol{np(m_2 - m_1^2) + np(1-p)\, m_1^2}$$

$$\boldsymbol{= npm_2 - np^2m_1^2} \qquad \textbf{(19.16)}$$

Lastly, formula (19.8) gives the MGF:

$$M_S(t) = M_N\left(\log M_X(t)\right)$$

$$\Rightarrow \boldsymbol{M_S(t) = (pM_X(t) + 1 - p)^n}$$

We can also find expressions for the skewness and the coefficient of skewness.

The third central moment is found from the cumulant generating function:

$$C_S(t) = \ln M_S(t)$$

In the next few steps, liberal use is made of the chain rule $\left(\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx}\right)$ and the product rule for differentiation $\left(\frac{d}{dx}(uv) = \frac{du}{dx}v + u\frac{dv}{dx}\right)$. The third derivative of the cumulant generating function is:

$$\boldsymbol{\frac{d^3}{dt^3}\log M_S(t) = \frac{d^3}{dt^3}n\log\left(pM_X(t) + q\right)} \quad \textbf{where } \boldsymbol{q = 1-p}$$

$$\boldsymbol{= \frac{d^2}{dt^2}\left\{np\left(\frac{d}{dt}M_X(t)\right)(pM_X(t) + q)^{-1}\right\}}$$

$$\boldsymbol{= \frac{d}{dt}\left\{np\left(\frac{d^2}{dt^2}M_X(t)\right)(pM_X(t) + q)^{-1} - n\left(p\frac{d}{dt}M_X(t)\right)^2(pM_X(t) + q)^{-2}\right\}}$$

$$\boldsymbol{= np\left(\frac{d^3}{dt^3}M_X(t)\right)(pM_X(t) + q)^{-1}}$$

$$\boldsymbol{- 3np^2\left(\frac{d^2}{dt^2}M_X(t)\right)(pM_X(t) + q)^{-2}\left(\frac{d}{dt}M_X(t)\right)}$$

$$\boldsymbol{+ 2n\left(p\frac{d}{dt}M_X(t)\right)^3(pM_X(t) + q)^{-3}}$$

Setting *t* = 0 gives:

$$skew(S) = \boldsymbol{E[(S - npm_1)^3] = npm_3 - 3np^2m_2m_1 + 2np^3m_1^3} \qquad \textbf{(19.17)}$$

The coefficient of skewness is then given by:

$$\frac{skew(S)}{[\text{var}(S)]^{3/2}} = \boldsymbol{\frac{npm_3 - 3np^2m_2m_1 + 2np^3m_1^3}{(npm_2 - np^2m_1^2)^{3/2}}}$$

It can be deduced from formula (5.17) that it is possible for the compound binomial distribution to be negatively skewed. The simplest illustration of this fact is when all claims are of (a fixed) **amount B. Then $S = BN$ and:**

$$E\left[(S-E[S])^3\right] = B^3 E\left[(N-E[N])^3\right]$$

ie:

$$skew(S) = B^3 skew(N)$$

So the coefficient of skewness of S is a multiple of that for N.

In fact:

$$coeff\ of\ skew(S) = \frac{skew(S)}{[\text{var}(S)]^{3/2}} = \frac{B^3 skew(N)}{[B^2\ \text{var}(N)]^{3/2}} = coeff\ of\ skew(N)$$

If $p > 0.5$, then the binomial distribution for N is negatively skewed.

So the coefficient of skewness of S will also be negative in this case.

Question

Determine an expression for the MGF of the aggregate claim amount random variable if the number of claims has a $Bin(100, 0.01)$ distribution and individual claim sizes have a $Gamma(10, 0.2)$ distribution.

Solution

Since $N \sim Bin(100, 0.01)$ and $X \sim Gamma(10, 0.2)$, we have:

$$M_N(t) = (0.99 + 0.01e^t)^{100} \qquad \text{and} \qquad M_X(t) = \left(1 - \frac{t}{0.2}\right)^{-10} = (1-5t)^{-10}$$

So:

$$\begin{aligned} M_S(t) &= M_N[\log M_X(t)] \\ &= \left[0.99 + 0.01e^{\log M_X(t)}\right]^{100} \\ &= \left[0.99 + 0.01M_X(t)\right]^{100} \\ &= \left[0.99 + 0.01(1-5t)^{-10}\right]^{100} \end{aligned}$$

3.6 The compound negative binomial distribution

The final choice of distribution for N is the negative binomial distribution, which has probability function:

$$\boldsymbol{P(N=n)=\binom{k+n-1}{n}p^k q^n \quad \text{for } n=0,1,2,\ldots}$$

This is the Type 2 negative binomial distribution. See page 9 of the *Tables*.

The Type 1 negative binomial distribution has probability function:

$$P(N=n)=\binom{n-1}{k-1}p^k q^{n-k} \quad \text{for } n=k,\ k+1,\ k+2,\ \ldots$$

It is not likely to be appropriate here, unless there is a specific reason why the number of claims must be at least k.

The parameters of the distribution are k (>0) and p, where $p+q=1$ and $0<p<1$. This distribution is denoted by $NB(k,p)$. When $N \sim NB(k,p)$:

$$\boldsymbol{E[N]=\frac{kq}{p}}$$

$$\boldsymbol{\text{var}[N]=\frac{kq}{p^2}}$$

$$\boldsymbol{M_N(t)=p^k(1-qe^t)^{-k}}$$

The special case $k=1$ leads to the geometric distribution. Once again, note that these results are given in the *Tables*.

The negative binomial distribution is an alternative to the Poisson distribution for N.

This is because the negative binomial distribution can take any non-negative integer value, unlike the binomial distribution which has an upper limit.

One advantage that the negative binomial distribution has over the Poisson distribution is that its variance exceeds its mean. These two quantities are equal for the Poisson distribution. Thus, the negative binomial distribution may give a better fit to a data set which has a sample variance in excess of the sample mean. This is often the case in practice. In the next chapter a situation leading to the negative binomial distribution for N is discussed. When N has a negative binomial distribution, S has a compound negative binomial distribution.

Expressions for the mean, variance and MGF of S when $N \sim NB(k,p)$ come immediately from formulae (19.4), (19.5) and (19.8):

$$E(S)=E(N)E(X) \Rightarrow \mathbf{E[S]=\frac{kq}{p}m_1}$$

$$\text{var}(S)=E(N)\text{var}(X)+\text{var}(N)[E(X)]^2 \Rightarrow \mathbf{var[S]=\frac{kq}{p}(m_2-m_1^2)+\frac{kq}{p^2}m_1^2}$$

Multiplying out the brackets and regrouping the terms, we see that:

$$\begin{aligned}\frac{kq}{p}(m_2-m_1^2)+\frac{kq}{p^2}m_1^2&=\frac{kq}{p}m_2-\frac{kq}{p}m_1^2+\frac{kq}{p^2}m_1^2\\&=\frac{kq}{p}m_2-\frac{kpq}{p^2}m_1^2+\frac{kq}{p^2}m_1^2\\&=\frac{kq}{p}m_2+\frac{kq-kpq}{p^2}m_1^2\\&=\frac{kq}{p}m_2+\frac{kq(1-p)}{p^2}m_1^2\\&=\frac{kq}{p}m_2+\frac{kq^2}{p^2}m_1^2\end{aligned}$$

So:

$$\mathbf{var[S]=\frac{kq}{p}m_2+\frac{kq^2}{p^2}m_1^2}$$

and:

$$M_S(t)=M_N\left(\log M_X(t)\right) \Rightarrow \mathbf{M_S(t)=\frac{p^k}{(1-qM_X(t))^k}}$$

As before, the third central moment of S can be found from the cumulant generating function of S, as follows:

$$\begin{aligned}\frac{d}{dt}\log M_S(t)&=\frac{d}{dt}\left(k\log p-k\log\left[1-qM_X(t)\right]\right)\\&=\frac{kq}{1-qM_X(t)}\left(\frac{d}{dt}M_X(t)\right)\end{aligned}$$

Then:

$$\frac{d^2}{dt^2}\log M_S(t) = kq^2\left(\frac{d}{dt}M_X(t)\right)^2 \frac{1}{(1-qM_X(t))^2} + \frac{kq}{1-qM_X(t)}\left(\frac{d^2}{dt^2}M_X(t)\right)$$

and:

$$\frac{d^3}{dt^3}\log M_S(t) = 3kq^2\left(\frac{d}{dt}M_X(t)\right)\left(\frac{d^2}{dt^2}M_X(t)\right)\frac{1}{(1-qM_X(t))^2}$$

$$+\frac{2kq^3}{(1-qM_X(t))^3}\left(\frac{d}{dt}M_X(t)\right)^3 + \frac{kq}{1-qM_X(t)}\left(\frac{d^3}{dt^3}M_X(t)\right)$$

Setting $t=0$ in the third derivative gives:

$$skew(S) = E[(S-E[S])^3] = \frac{3kq^2 m_1\, m_2}{p^2} + \frac{2kq^3\, m_1^3}{p^3} + \frac{kqm_3}{p} \qquad \textbf{(19.18)}$$

The parameters k and p are positive, as are the moments of X. It therefore follows from formula (19.18) that the compound negative binomial distribution is positively skewed. The coefficient of skewness can be found from $E((S-E(S))^3)\,/\,(\text{var}(S))^{3/2}$.

Question

The distribution of the number of claims from a motor portfolio is negative binomial with parameters $k = 4{,}000$ and $p = 0.9$. The claim size distribution is Pareto with parameters $\alpha = 5$ and $\lambda = 1{,}200$. Calculate the mean and standard deviation of the aggregate claim distribution.

Solution

The first two moments of the Pareto distribution are:

$$m_1 = E(X) = \frac{\lambda}{\alpha - 1} = \frac{1{,}200}{4} = 300$$

$$m_2 = E(X^2) = \frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)} + m_1^2 = \frac{5 \times 1{,}200^2}{4^2 \times 3} + 300^2 = 240{,}000$$

So, using the formulae for the mean and variance of a compound negative binomial distribution:

$$E(S) = \frac{kq}{p} \times m_1 = \frac{4{,}000 \times 0.1}{0.9} \times 300 = 133{,}333$$

$$\text{var}(S) = \frac{kq}{p}m_2 + \frac{kq^2}{p^2}m_1^2 = \frac{4{,}000 \times 0.1}{0.9} \times 240{,}000 + \frac{4{,}000 \times 0.1^2}{0.9^2} \times 300^2 = 111{,}111{,}111$$

So the standard deviation is 10,541.

4 Appendix

There is some repetition in the Core Reading in Section 3.3. To improve the flow of the chapter, we have removed the repeated section from the main part of the text and placed it below.

Let $S = X_1 + X_2 + \cdots + X_N$ (and $S = 0$ if $N = 0$) where the X_i's are independent, identically distributed (as a variable X) and are also independent of N. S is said to have a *compound distribution*.

***Illustration*: N is the number of claims which arise in a portfolio of business and X_i is the amount of the ith claim. S is the total claim amount.**

The mean and variance of S are easily found:

$$E(S \mid N = n) = E(X_1 + \cdots + X_N \mid N = n) = E(X_1 + \cdots + X_n) = nE(X)$$

Similarly:

$$\text{var}(S \mid N = n) = n\,\text{var}(X)$$

Therefore:

$$E(S) = E\left[E(S \mid N)\right] = E\left[NE(X)\right] = E(N)E(X)$$

***ie*:**

$$\mu_S = \mu_N \mu_X$$

and:

$$\begin{aligned}\text{var}(S) &= E\left[\text{var}(S \mid N)\right] + \text{var}\left[E(S \mid N)\right] \\ &= E\left[N\,\text{var}(X)\right] + \text{var}\left[NE(X)\right] \\ &= E(N)\,\text{var}(X) + \text{var}(N)\left[E(X)\right]^2\end{aligned}$$

***ie*:**

$$\sigma_S^2 = \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2$$

The MGF of S is given by:

$$M_S(t) = E(e^{tS}) = E\left[E\left(e^{tS} \mid N\right)\right]$$

and:

$$\begin{aligned}E(e^{tS} \mid N = n) &= E\left[\exp\left(t(X_1 + X_2 + \cdots + X_N)\right) \mid N = n\right] \\ &= E\left[\exp\left(t(X_1 + X_2 + \cdots + X_n)\right)\right] \\ &= \prod E\left[\exp(tX_i)\right] = \left[M_X(t)\right]^n\end{aligned}$$

Therefore:

$$M_S(t) = E\left[\left(M_X(t)\right)^N\right] = E\left[\exp\left(N \log M_X(t)\right)\right] = M_N\left(\log M_X(t)\right)$$

Compound Poisson distribution

An important illustration is provided by the compound Poisson distribution, which is the case in which $N \sim Poisson(\lambda)$. In this case $\mu_N = \sigma_N^2 = \lambda$.

***Properties*:**

$$E(S) = \lambda E(X)$$

$$\text{var}(S) = \lambda\,\text{var}(X) + \lambda\left[E(X)\right]^2 = \lambda E(X^2)$$

$$M_N(t) = \exp\left[\lambda(e^t - 1)\right]$$

so:

$$M_S(t) = \exp\left[\lambda\left(M_X(t) - 1\right)\right]$$

from which the mean and variance can be obtained and the results above verified.

Chapter 19 Summary

Insurable risks

For a risk to be insurable the policyholder should have an interest in the risk being insured to distinguish between insurance and a wager, and it should be of a financial and reasonably quantifiable nature. Ideally, risk events should:

- be independent
- have low probability of occurring
- be pooled with similar risks
- have an ultimate liability
- avoid moral hazards.

Characteristics of general insurance products

Most general insurance contracts share the following characteristics:

- Cover is normally for a fixed period, typically a year, after which it needs to be renegotiated.
- There is usually no obligation to continue cover although in most cases a need for continuing cover may be assumed to exist.
- Claims are not of fixed amounts.
- The existence of a claim and its amount have to be proved before a claim can be settled.
- A claim occurring does not bring the policy to an end.
- Claims that take a long time to settle are known as long-tailed and those that take a short time to settle are known as short-tailed.

Features of short-term insurance contracts

A short-term insurance contract can be defined as having the following attributes:

- The policy lasts for a fixed, and relatively short time period, typically one year.
- The insurance company receives a premium from the policyholder.
- In return the insurer pays claims that arise during the term of the policy.
- At the end of the policy term, the policyholder may or may not renew the policy. If it is renewed, the premium may or may not be the same as in the previous period.
- The insurer may pass part of the premium to a reinsurer, who, in return, will reimburse the insurer for part of the claims cost.

Collective risk model

Aggregate claim amounts may be modelled using a compound distribution. The aggregate claim amount S is the sum of a random number of IID random variables:

$$S = X_1 + X_2 + \cdots + X_N$$

where S is taken to be zero if $N = 0$. We assume that the random variable N is independent of the random variables X_i so that the distributions of the claim numbers and the individual claim amounts can be analysed separately. The distribution of S is said to be a compound distribution.

Other simplifying assumptions include:

- The moments (and sometimes the distributions) of N and X_i are known.
- Claims are settled more or less as soon as the claims occur.
- Expenses and investment returns are ignored.

Specific types of compound distributions include the compound Poisson, compound binomial, compound negative binomial, and compound geometric. Formulae for the MGF and the moments of a compound random variable are given on page 16 of the *Tables*.

Convolutions

If $Z = X + Y$, and X and Y are independent, then:

$$P(Z = z) = \sum_x P(X = x)P(Y = z - x) \text{ if } X \text{ and } Y \text{ are discrete}$$

$$f_Z(z) = f_X * f_Y(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\,dx \text{ if } X \text{ and } Y \text{ are continuous}$$

Sums of independent compound Poisson random variables

Let $S_1, S_2, \ldots, S_n$ be a set of independent random variables where S_i has a compound Poisson distribution with parameter λ_i and $F_i(x)$ is the CDF of the individual claim amount random variable for S_i. Then $A = S_1 + \cdots + S_n$ is compound Poisson with parameter $\Lambda = \sum \lambda_i$. The CDF of the individual claim amount random variable for A is:

$$F(x) = \frac{1}{\Lambda} \sum \lambda_i \; F_i(x)$$

The MGF of the individual claim amounts for A is:

$$M(t) = \frac{1}{\Lambda} \sum \lambda_i M_i(t)$$

where $M_i(t)$ is the MGF of the individual claim amounts for S_i.

Chapter 19 Practice Questions

19.1 (i) State the two conditions that must hold for a risk to be insurable.

(ii) List five other risk criteria that would be considered desirable by a general insurer.

19.2 A group of policies can give rise to at most 2 claims. The probabilities of 0, 1 or 2 claims are ½, ¼ and ¼ respectively. Claim amounts are IID $U(0,10)$ random variables. Let S denote the aggregate claim amount random variable.

Sketch the frequency distribution of S.

19.3 The random variable S has a compound Poisson distribution with Poisson parameter 4. The individual claim amounts are either 1, with probability 0.3, or 3, with probability 0.7.

Calculate the probability that $S = 4$.

19.4 A compound random variable $S = X_1 + X_2 + \cdots + X_N$ has claim number distribution:

$$P(N = n) = 9(n+1)4^{-n-2}\ , \quad n = 0,1,2,\ldots$$

The individual claim size random variable, X, is exponentially distributed with mean 2.

Calculate $E(S)$ and $\text{var}(S)$.

19.5 Write down a formula for the MGF of a compound Poisson distribution with individual claim size distribution $Gamma(\alpha,\beta)$ and Poisson parameter λ.

19.6 S_1 and S_2 are independent random variables each with a compound Poisson distribution. The distribution of S_i, $i = 1,2$, has Poisson parameter λ_i and individual claim amount distribution $F_i(x)$.

Which one of the following statements about the distribution of $S_1 + S_2$ is correct?

A $S_1 + S_2$ has a compound Poisson distribution with Poisson parameter $\lambda_1\lambda_2$ and individual claim amount distribution $F_1(x) + F_2(x)$.

B $S_1 + S_2$ has a compound Poisson distribution with Poisson parameter $(\lambda_1 + \lambda_2)$ and individual claim amount distribution $\left(F_1(x) + F_2(x)\right)/2$.

C $S_1 + S_2$ has a compound Poisson distribution with Poisson parameter $(\lambda_1 + \lambda_2)$ and individual claim amount distribution $\left(\lambda_1 F_1(x) + \lambda_2 F_2(x)\right)/\left(\lambda_1 + \lambda_2\right)$.

D $S_1 + S_2$ does not have a compound Poisson distribution.

19.7 Exam style

Claims on a group of policies of a certain type arise as a Poisson process with parameter λ_1. Claims on a second, independent, group of policies arise as a Poisson process with parameter λ_2. The aggregate claim amounts on the respective groups are denoted S_1 and S_2.

Using moment generating functions (or otherwise), show that S (the sum of S_1 and S_2) also has a compound Poisson distribution and hence derive the Poisson parameter for S. [4]

19.8 Exam style

The aggregate claim amount from a portfolio has a compound negative binomial distribution.

(i) Show that if $S = X_1 + \cdots + X_N$, then:

$$M_S(t) = M_N\left[\log M_X(t)\right]$$ [3]

(ii) If N has Type 2 negative binomial distribution with $k=2$ and $p=0.9$, and X has a gamma distribution with $\alpha=10$ and $\lambda=0.1$, determine an expression for $M_S(t)$. [2]

(iii) (a) Calculate the mean and variance of S.

(b) Using a suitable approximation, estimate the aggregate amount which will be exceeded with probability 0.1%. [4]

(iv) The insurer in fact has 100 identical independent portfolios of this type. Let:

$$T = S_1 + \cdots + S_{100}$$

(a) Determine the moment generating function for T.

(b) Using a normal approximation, estimate the total aggregate claim amount from the whole business which will be exceeded with probability 0.1%.

(c) Comment on your answers to parts (iii)(b) and (iv)(b). [4]

[Total 13]

Chapter 19 Solutions

19.1 (i) ***Criteria for an insurable risk***

The two conditions are:

- The policyholder must have an interest in the risk being insured.
- The risk must be of a financial and reasonably quantifiable nature.

(ii) ***Other desirable features of a risk***

Other desirable features are:

- Individual risks should be independent of one another.
- The probability that the insured event will occur should be small.
- Large numbers of similar risks should be pooled in order to reduce the variance and achieve greater certainty.
- The insurer's liability should be limited.
- Moral hazards should be eliminated as far as possible since these are difficult to quantify, result in selection against the insurer and lead to unfairness in the treatment of some policyholders.

19.2 If $N=0$, *ie* there are no claims, then $S=0$. So there is a point mass (or a 'blob' of probability) at $S=0$.

If $N=1$, *ie* there is exactly one claim, then S has a $U(0,10)$ distribution. This will happen with probability ¼.

If $N=2$, *ie* there are exactly 2 claims, then $S=X_1+X_2$ where X_1 and X_2 are independent random variables with PDFs:

$$f_{X_1}(x)=f_{X_2}(x)=\begin{cases}0.1 & \text{if } 0\le x\le 10\\ 0 & \text{otherwise}\end{cases}$$

The PDF of *S* is:

$$f_S(s)=\int_{-\infty}^{\infty} f_{X_1}(s-x)f_{X_2}(x)\,dx=0.1\int_0^{10} f_{X_1}(s-x)\,dx$$

The integrand is 0.1 if $0\le s-x\le 10$ and 0 otherwise. So for $0\le s\le 10$:

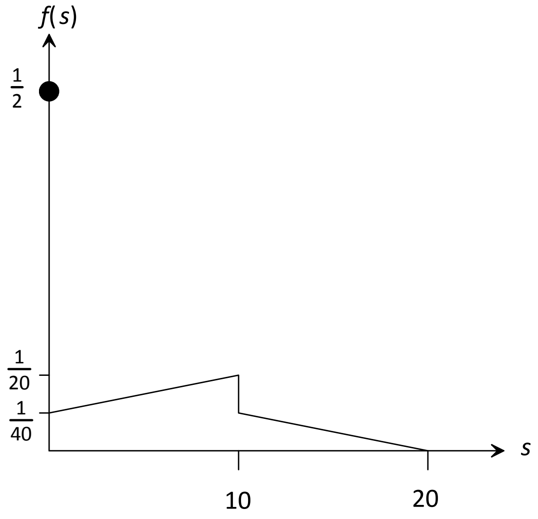
$$f_S(s)=0.1\int_0^s 0.1\,dx=0.01s$$

and for $10\le s\le 20$:

$$f_S(s)=0.1\int_{s-10}^{10} 0.1\,dx=0.2-0.01s$$

So for $N=2$, S has a symmetrical triangular shaped distribution on the interval (0, 20).

A graph of the distribution is shown below:



This graph is the combination of a blob at zero, a uniform distribution on (0,10) and a triangular distribution on (0,20).

19.3 We need to consider how we could get an aggregate claim amount of 4. This could happen in two ways:

(a) 2 claims, one for 1 and one for 3

(b) 4 claims, all for an amount of 1.

The probability of this happening is:

$$P(S=4)=P(N=2)P(X_1=1)P(X_2=3)+P(N=2)P(X_1=3)P(X_2=1)$$
$$+P(N=4)P(X_1=1)P(X_2=1)P(X_3=1)P(X_4=1)$$

Since the X_i's are identical, this simplifies to:

$$P(S=4)=2P(N=2)P(X=1)P(X=3)+P(N=4)[P(X=1)]^4$$
$$=2\times\frac{e^{-4}4^2}{2!}\times0.3\times0.7+\frac{e^{-4}4^4}{4!}\times0.3^4$$
$$=0.06312$$

19.4 The probability function of N can be written as:

$$P(N=n)=9(n+1)4^{-n-2}=\binom{n+1}{n}(3/4)^2(1/4)^n$$

We can see from this formula that N has a Type 2 negative binomial distribution with parameters $k=2$ and $p=3/4$.

Hence:

$$E(N)=\frac{kq}{p}=\frac{2\times 1/4}{3/4}=2/3$$

and: $$\text{var}(N)=\frac{kq}{p^2}=\frac{2\times 1/4}{(3/4)^2}=8/9$$

The individual claim amounts have an exponential distribution with $\lambda = ½$. So the mean and variance of the individual claims are:

$$E(X)=\frac{1}{\lambda}=2 \quad \text{and} \quad \text{var}(X)=\frac{1}{\lambda^2}=4$$

Hence:

$$E(S)=E(N)\,E(X)=\frac{2}{3}\times 2=\frac{4}{3}$$

and: $$\text{var}(S)=E(N)\,\text{var}(X)+\text{var}(N)[E(X)]^2=\frac{2}{3}\times 4+\frac{8}{9}\times 2^2=\frac{56}{9}$$

19.5 Since $N \sim Poisson(\lambda)$:

$$M_N(u)=\exp[\lambda(e^u-1)]$$

So:

$$M_S(t)=M_N[\log M_X(t)]=\exp\left[\lambda\left(e^{\log M_X(t)}-1\right)\right]=\exp\left[\lambda\left(M_X(t)-1\right)\right]$$

Since $X \sim Gamma(\alpha,\beta)$:

$$M_X(t)=\left(1-\frac{t}{\beta}\right)^{-\alpha}$$

So:

$$M_S(t)=\exp\left[\lambda\left(\left(1-\frac{t}{\beta}\right)^{-\alpha}-1\right)\right]$$

19.6 Option C is correct.

S_1+S_2 has a compound Poisson distribution with Poisson parameter $(\lambda_1+\lambda_2)$ and individual claim amount distribution $\left(\lambda_1F_1(x)+\lambda_2F_2(x)\right)/\left(\lambda_1+\lambda_2\right)$.

19.7 Let N_i denote the number of claims on policies of type i and let X_i denote the claim amount random variable for policies of type i, for $i=1,2$. Then:

$$M_{S_i}(t) = M_{N_i}\left(\ln M_{X_i}(t)\right) = \exp\left(\lambda_i\left(M_{X_i}(t)-1\right)\right)$$

By independence:

$$M_S(t) = E\left(e^{tS}\right) = E\left(e^{t(S_1+S_2)}\right) = E\left(e^{tS_1}e^{tS_2}\right) = E\left(e^{tS_1}\right)E\left(e^{tS_2}\right) = M_{S_1}(t)M_{S_2}(t) \qquad [1]$$

Hence:

$$\begin{aligned} M_S(t) &= \exp\left(\lambda_1\left(M_{X_1}(t)-1\right)\right)\exp\left(\lambda_2\left(M_{X_2}(t)-1\right)\right) \\ &= \exp\left(\lambda_1 M_{X_1}(t)+\lambda_2 M_{X_2}(t)-(\lambda_1+\lambda_2)\right) \\ &= \exp\left(\lambda(M_W(t)-1)\right) \qquad [1] \end{aligned}$$

where:

$$\lambda = \lambda_1+\lambda_2 \qquad [1]$$

and:

$$M_W(t) = \frac{\lambda_1 M_{X_1}(t)+\lambda_2 M_{X_2}(t)}{\lambda_1+\lambda_2} \qquad [1]$$

Hence S is a compound Poisson random variable with Poisson parameter $\lambda = \lambda_1+\lambda_2$.

19.8 (i) ***MGF of S***

The MGF of S is:

$$M_S(t) = E(e^{tS}) = E\left[E(e^{tS}\mid N)\right]$$

using the standard result for conditional means from page 16 of the *Tables*.

Looking at the inner expression, we have:

$$E(e^{tS}\mid N=n) = E\left(e^{t(X_1+\cdots+X_n)}\right) = E(e^{tX_1})\ldots E(e^{tX_n}) \qquad [1]$$

Now each of these terms is just the MGF of the random variable X.

So:

$$E(e^{tS} \mid N = n) = \left[M_X(t)\right]^n$$

and hence:

$$E(e^{tS} \mid N) = [M_X(t)]^N \qquad [1]$$

So:

$$M_S(t) = E\left[\left\{M_X(t)\right\}^N\right] = E\left(e^{N\log M_X(t)}\right) = M_N\left[\log M_X(t)\right] \qquad [1]$$

(ii) ***Compound negative binomial distribution***

We first need the individual MGFs. Using results from the *Tables*, we have:

$$M_X(t) = \left(1 - \frac{t}{0.1}\right)^{-10} = (1-10t)^{-10} \qquad [½]$$

and: $$M_N(t) = \left(\frac{0.9}{1-0.1e^t}\right)^2 \qquad [½]$$

Combining these, using the result from part (i):

$$M_S(t) = M_N\left[\log M_X(t)\right] = \left(\frac{0.9}{1-0.1(1-10t)^{-10}}\right)^2 \qquad [1]$$

(iii)(a) ***Mean and variance of S***

We could differentiate this expression to find the mean and variance of S. However, it is much easier to use the standard compound distribution formulae:

$$E(S) = E(X)E(N)$$

and: $$\text{var}(S) = \left[E(X)\right]^2 \text{var}(N) + \text{var}(X)E(N)$$

Using the results from the *Tables* for the individual distributions:

$$E(X) = \frac{\alpha}{\lambda} = \frac{10}{0.1} = 100$$

$$\text{var}(X) = \frac{\alpha}{\lambda^2} = \frac{10}{0.1^2} = 1{,}000 \qquad [1]$$

$$E(N) = \frac{kq}{p} = \frac{2\times 0.1}{0.9} = 0.22222$$

$$\text{var}(N) = \frac{kq}{p^2} = \frac{2\times 0.1}{0.9^2} = 0.24691 \qquad [1]$$

Using the formulae above:

$$E(S) = E(X)E(N) = 22.222$$

and: $\text{var}(S) = [E(X)]^2 \text{var}(N) + \text{var}(X)E(N) = 100^2 \times 0.24691 + 1{,}000 \times 0.22222 = 2{,}691.358$ [1]

(iii)(b) ***Aggregate amount***

We now assume that S has an approximate normal distribution with this mean and variance. So, standardising in the usual way, we have:

$$P(S > k) = 0.001 \quad \Rightarrow \quad P\left(N(0,1) > \frac{k - E(S)}{\sqrt{\text{var}(S)}}\right) = 0.001$$

Using the percentage points of the standard normal distribution, we find that:

$$\frac{k - E(S)}{\sqrt{\text{var}(S)}} = 3.0902 \quad \Rightarrow \quad k = 22.222 + 3.0902\sqrt{2{,}691.358} = 182.54$$ [1]

(iv)(a) ***MGF of T***

The MGF of T is:

$$M_T(t) = E(e^{tT}) = E[e^{t(S_1 + \dots + S_{100})}] = E(e^{tS_1}) \dots E(e^{tS_{100}}) = [M_S(t)]^{100}$$

$$= \left(\frac{0.9}{1 - 0.1(1-10t)^{-10}}\right)^{200}$$ [1]

The mean and variance of T are 100 times the corresponding results for S, *ie*:

$$E(T) = 100E(S) = 2{,}222$$

and: $\text{var}(T) = 100\,\text{var}(S) = 269{,}135.8$ [1]

(iv)(b) ***Total aggregate amount***

So the corresponding figure for the aggregate amount exceeded with probability 0.001 is:

$$2{,}222 + 3.0902\sqrt{269{,}135.8} = 3{,}825.37$$ [1]

(iv)(c) ***Comment***

This is substantially less than one hundred times the corresponding answer to part (iii)(b). The Central Limit Theorem tells us that as the number of portfolios increases, bad experience in some of the portfolios will be offset by better experience in others, leading to a situation where the overall variation is relatively smaller. Pooling of similar risks reduces the overall variance. We can see this happening here. [1]

20

Risk models 2

Syllabus objectives

1.2 Compound distributions and their application in risk modelling

1.2.1 Construct models appropriate for short-term insurance contracts in terms of the numbers of claims and the amounts of individual claims.

1.2.2 Describe the major simplifying assumptions underlying the models in 1.2.1.

1.2.4 Derive the mean, variance and coefficient of skewness for compound binomial, compound Poisson and compound negative binomial random variables.

1.2.5 Repeat 1.2.4 for both the insurer and the reinsurer after the operation of simple forms of proportional and excess of loss reinsurance.

0 Introduction

In this chapter we will look at some of the practical applications of risk models. We start by looking at how the models can be adapted for situations involving reinsurance. A section on the individual risk model is followed by some more complex examples of the use of risk models in practice.

1 Aggregate claim distributions under proportional and individual excess of loss reinsurance

In Chapter 19, we introduced the notation S to denote the aggregate claim amount random variable, *ie*:

$$S = X_1 + X_2 + \cdots + X_N$$

where N denotes the number of claims and X_i denotes the amount of the i th claim.

Here we extend this concept to consider the situation when reinsurance is in force. We will use the following notation:

- Y_i is the amount paid by the insurer in respect of the i th claim
- Z_i is the amount paid by the reinsurer in respect of the i th claim
- $S_I = Y_1 + Y_2 + \cdots + Y_N$ is the aggregate claim amount paid by the insurer
- $S_R = Z_1 + Z_2 + \cdots + Z_N$ is the aggregate claim amount paid by the reinsurer.

The formulae that we derived for the mean, variance and MGF of S can be adapted to cover the reinsurance situation by replacing X by Y or Z, as appropriate. For example:

$$E(S_I) = E(N)E(Y)$$

$$\text{var}(S_I) = E(N)\,\text{var}(Y) + \text{var}(N)[E(Y)]^2$$

$$M_{S_I}(t) = M_N(\log M_Y(t))$$

1.1 Proportional reinsurance

The distribution of the number of claims involving the reinsurer is the same as the distribution of the number of claims involving the insurer, as each pays a defined proportion of every claim.

For a retention level α ($0 \le \alpha \le 1$), the i th individual claim amount for the insurer is αX_i and for the reinsurer is $(1-\alpha)X_i$.

In other words:

$$Y_i = \alpha X_i$$

$$Z_i = (1-\alpha)X_i$$

So:

$$\begin{aligned} S_I &= Y_1 + Y_2 + \cdots + Y_N \\ &= \alpha X_1 + \alpha X_2 + \cdots + \alpha X_N \\ &= \alpha (X_1 + X_2 + \cdots + X_N) \\ &= \alpha S \end{aligned}$$

and:

$$\begin{aligned} S_R &= Z_1 + Z_2 + \cdots + Z_N \\ &= (1-\alpha) X_1 + (1-\alpha) X_2 + \cdots + (1-\alpha) X_N \\ &= (1-\alpha)(X_1 + X_2 + \cdots + X_N) \\ &= (1-\alpha) S \end{aligned}$$

***ie* the aggregate claims amounts are αS and $(1-\alpha)S$ respectively.**

Question

(i) Show that under a proportional reinsurance arrangement where the direct writer retains a proportion α, the MGF of the net individual claim amount Y paid by the direct insurer is $M_X(\alpha t)$.

(ii) Hence give a formula for $M_{S_I}(t)$ when the number of claims follows a Poisson distribution with mean 25 and individual claim amounts are exponentially distributed with mean 1,000.

Solution

(i) ***MGF of Y***

Under this arrangement, $Y = \alpha X$. So:

$$M_Y(t) = E(e^{tY}) = E(e^{t\alpha X}) = E[e^{(t\alpha)X}] = M_X(\alpha t)$$

(ii) ***MGF of insurer's aggregate claim amount***

Since $N \sim Poisson(25)$:

$$M_N(t) = e^{25(e^t - 1)}$$

and the MGF of S_I is:

$$M_{S_I}(t) = M_N(\log M_Y(t)) = e^{25(\exp(\log M_Y(t)) - 1)} = e^{25(M_Y(t) - 1)}$$

In addition, since X is exponentially distributed with mean 1,000 (*ie* with parameter $\frac{1}{1,000}$):

$$M_X(t) = (1 - 1{,}000t)^{-1}$$

and hence:

$$M_Y(t) = M_X(\alpha t) = (1 - 1{,}000\alpha t)^{-1}$$

So the MGF of S_I is:

$$M_{S_I}(t) = e^{25\left[(1-1{,}000\alpha t)^{-1} - 1\right]}$$

1.2 Individual excess of loss reinsurance

The amount that an insurer pays on the ith claim under individual excess of loss reinsurance with retention level M is $Y_i = \min\{X_i, M\}$.

Equivalently:

$$Y = \begin{cases} X & \text{if } X \le M \\ M & \text{if } X > M \end{cases}$$

The amount that the reinsurer pays is $Z_i = \max\{0, X_i - M\}$.

We can also write this as:

$$Z = \begin{cases} 0 & \text{if } X \le M \\ X - M & \text{if } X > M \end{cases}$$

As previously stated, **the insurer's aggregate claims net of reinsurance can be represented as:**

$$S_I = Y_1 + Y_2 + \cdots + Y_N$$

and the reinsurer's aggregate claims as:

$$S_R = Z_1 + Z_2 + \cdots + Z_N \qquad \textbf{(20.1)}$$

If, for example, $N \sim Poi(\lambda)$, S_I has a compound Poisson distribution with Poisson parameter λ, and the ith individual claim amount is Y_i. Similarly, S_R has a compound Poisson distribution with Poisson parameter λ, and the ith individual claim amount is Z_i.

Hence, for this compound Poisson distribution with reinsurance, we have:

$$E(S_I)=\lambda E(Y)$$

$$\text{var}(S_I)=\lambda E(Y^2)$$

$$skew(S_I)=\lambda E(Y^3)$$

Similar formulae can be obtained for S_R by replacing Y with Z.

Note, however that if $F(M)>0$, as will usually be the case, then Z_i may take the value 0.

Here, $F(x)$ denotes the distribution function of the original claim amount random variable, X. So:

$$F(M)=P(X\le M)$$

If this is greater than 0, then there is a non-zero probability that the reinsurer will not be involved in a claim.

In other words, 0 is counted as a possible claim amount for the reinsurer. From a practical point of view, this definition of S_R is rather artificial. The insurer will know the observed value of N, but the reinsurer will probably know only the number of claims above the retention level M, since the insurer may notify the reinsurer only of claims above the retention level.

Example

The annual aggregate claim amount from a risk has a compound Poisson distribution with Poisson parameter 10. Individual claim amounts are uniformly distributed on $(0,2000)$. The insurer of this risk has effected excess of loss reinsurance with retention level 1,600.

Calculate the mean, variance and coefficient of skewness of both the insurer's and reinsurer's aggregate claims under this reinsurance arrangement.

Solution

Let S_I and S_R be as above. To find $E[S_i]$ calculate $E[Y_i]$. Now:

$$E(Y_i)=\int_0^M x\,f(x)\,dx+M\,P(X_i>M)$$

where $f(x)=0.0005$ is the U(0,2000) density function and M = 1,600.

This gives:

$$E[Y_i]=\left[\frac{0.0005x^2}{2}\right]_0^M+0.2M=960$$

So:

$$E[S_I] = 10E[Y_i] = 9{,}600$$

To find $\text{var}[S_I]$, we must calculate $E[Y_i^2]$:

$$E[Y_i^2] = \int_0^M x^2 f(x)dx + M^2 P(X_i > M)$$

$$= \left[\frac{0.0005x^3}{3}\right]_0^M + 0.2M^2$$

$$= 1{,}194{,}666.7$$

So:

$$\text{var}[S_I] = 10E[Y_i^2] = 11{,}946{,}667$$

To find the coefficient of skewness of the insurer's claims, we must calculate $E[Y_i^3]$:

$$E[Y_i^3] = \int_0^M x^3 f(x)dx + M^3 P(X_i > M)$$

$$= \left[\frac{0.0005x^4}{4}\right]_0^M + 0.2M^3$$

$$= 1{,}638{,}400{,}000$$

So:

$$E\left[\left(S_I - E(S_I)\right)^3\right] = 10E[Y_i^3] = 16{,}384{,}000{,}000$$

and the coefficient of skewness of S_I is:

$$\frac{16{,}384{,}000{,}000}{11{,}946{,}667^{3/2}} = 0.397$$

To find $E[S_R]$, note that the expected annual aggregate claim amount from the risk is $E[S] = \lambda E[X] = 10 \times 1{,}000 = 10{,}000$. Then:

$$E[S_R] = 10{,}000 - E[S_I] = 400$$

To find $\text{var}[S_R]$, calculate $E[Z_i^2]$ from:

$$E[Z_i^2] = \int_M^{2,000} (x-M)^2 f(x)\,dx$$

$$= \int_0^{2,000-M} 0.0005y^2\,dy \quad \text{where } y = x-M$$

$$= \left[\frac{0.0005\,y^3}{3}\right]_0^{2,000-M}$$

$$= 10,666.7$$

So:

$$\text{var}[S_R] = 10E[Z_i^2] = 106,667$$

To find the coefficient of skewness of the reinsurer's claims, we need to calculate $E[Z_i^3]$:

$$E[Z_i^3] = \int_M^{2,000} (x-M)^3 f(x)\,dx$$

$$= \int_0^{2,000-M} 0.0005y^3\,dy \quad \text{where } y = x-M$$

$$= 3,200,000$$

So:

$$E\left[\left(S_R - E(S_R)\right)^3\right] = 10E[Z_i^3] = 32,000,000$$

and the coefficient of skewness of S_R is:

$$\frac{32,000,000}{106,667^{3/2}} = 0.92$$

Question

Calculate the variance of S, the aggregate claim amount before reinsurance for the example above and explain why:

$$\text{var}(S_I) + \text{var}(S_R) \neq \text{var}(S)$$

Solution

We have $\text{var}(S) = 10E(X^2)$, where:

$$E(X^2) = \int_0^{2000} \frac{x^2}{2000} dx = \frac{4{,}000{,}000}{3}$$

So $\text{var}(S) = 13{,}333{,}333$.

This is not equal to $11{,}946{,}667 + 106{,}667$ because S_I and S_R are not independent.

To simulate the collective risk model with *individual* reinsurance we can combine the R code from Chapters 15, 18 and 19.

For example, to simulate 10,000 values for a reinsurer where claims have a compound Poisson distribution with parameter 1,000 and a gamma claims distribution with $\alpha = 750$ and $\lambda = 0.25$ under *individual* excess of loss with retention 2,500 we would use:

```
set.seed(123)
M <-2500
n <- rpois(10000,1000)
sR <- numeric(10000)
for(i in 1:10000)
{x <- rgamma(n[i],shape=750,rate=0.25)
z <- pmax(0,x-M)
sR[i] <- sum(z)}
```

We can now find moments, the coefficient of skewness, probabilities and quantiles as before.

Earlier we mentioned that using $S_R = Z_1 + \cdots + Z_N$ is a bit artificial. We now look at an alternative way of modelling the reinsurer's compound claim amount distribution.

The reinsurer's aggregate claims can also be represented by:

$$\boldsymbol{S_R = W_1 + W_2 + \cdots + W_{NR}} \qquad \textbf{(20.2)}$$

where the random variable NR denotes the actual number of (non-zero) payments made by the reinsurer.

Here:

$$W_i = X_i - M \mid X_i > M$$

For example, suppose that the risk above gave rise to the following eight claim amounts in a particular year:

403 1,490 1,948 443 1,866 1,704 1,221 823

The retention limit is 1,600.

Then in formula (20.1) the observed value of N is 8, and the third, fifth and sixth claims require payments from the reinsurer of 348, 266 and 104 respectively. The reinsurer makes a 'payment' of 0 on the other five claims.

In formula (20.2), the observed value of NR is 3 and the observed values of W_1, W_2 and W_3 are 348, 266 and 104 respectively. Note that the observed value of S_R is the same (*ie* 718) under each definition.

W_i has density function:

$$g(w) = \frac{f(x+M)}{1-F(M)} \ , \quad w > 0$$

We saw this result in Section 1.2 of Chapter 18. It can also be written as:

$$f_W(w) = \frac{f_X(w+M)}{1-F_X(M)}$$

To specify the distribution for S_R as given in formula (20.2) the distribution of NR is needed.

In some contexts it may be obvious what this distribution is, but here is a general method for establishing the distribution.

This is found as follows. Define:

$$NR = I_1 + I_2 + \cdots + I_N$$

where N denotes the number of claims from the risk (as usual). I_j is an indicator random variable which takes the value 1 if the reinsurer makes a (non-zero) payment on the j th claim, and takes the value 0 otherwise. Thus NR gives the number of payments made by the reinsurer.

From its definition, we see that NR is a compound random variable. However, instead of being the sum of individual claims amounts, NR is a sum of indicator random variables.

Since I_j takes the value 1 only if $X_j > M$:

$$P(I_j = 1) = P(X_j > M) = \pi \text{, say}$$

and:

$$P(I_j = 0) = 1 - \pi$$

In other words, I_j has a $Binomial(1,\pi)$ distribution.

Further, I_j has MGF:

$$M_I(t) = \pi e^t + 1 - \pi$$

The formula for the MGF of a binomial distribution is given on page 6 of the *Tables*.

By formula (19.8) in Chapter 19 (the formula for the MGF of a compound random variable), ***NR*** **has MGF:**

$$\boldsymbol{M_{NR}(t) = M_N(\log M_I(t))}$$

Question

If N has a $Poisson(\lambda)$ distribution and $P(X > M) = ½$, show that NR has a $Poisson(½\lambda)$ distribution.

Solution

Here $\pi = ½$, so:

$$M_I(t) = ½e^t + ½$$

and:

$$M_{NR}(t) = M_N[\log M_I(t)] = \exp\{\lambda[M_I(t) - 1]\} = \exp\left\{\lambda(½e^t - ½)\right\} = \exp\left\{½\lambda(e^t - 1)\right\}$$

This is the MGF of the $Poisson(½\lambda)$ distribution. By the uniqueness property of MGFs, it follows that $NR \sim Poisson(½\lambda)$.

We now continue the above Core Reading example where the annual aggregate claim amount from a risk has a compound Poisson distribution with Poisson parameter 10, individual claim amounts are uniformly distributed on $(0, 2000)$, and the insurer of this risk has effected excess of loss reinsurance with retention level 1,600.

Example

Continuing the above example and using formula (20.2) as the model for S_R, it can be seen that S_R has a compound Poisson distribution with Poisson parameter $0.2 \times 10 = 2$. Individual claims, W_i, have density function:

$$\boldsymbol{g(w) = \frac{f(w+M)}{1-F(M)} = \frac{0.0005}{0.2} = 0.0025 \text{ , for } 0 < w < 400}$$

ie **W_i is uniformly distributed on (0,400).**

Using the formulae for the moments of a continuous uniform distribution from page 13 of the *Tables*, we have:

$$\boldsymbol{E[W_i] = 200 \text{ , } E[W_i^2] = 53{,}333.33 \text{ and } E[W_i^3] = 16{,}000{,}000}$$

giving the same results as before.

If we multiply these figures by 2 (the Poisson parameter of S_R), we get $E(S_R)=400$, $\text{var}(S_R)=106,667$ and $\text{skew}(S_R)=32,000,000$, which agree with the answers obtained previously.

Thus, there are two ways to specify and evaluate the distribution of S_R.

1.3 Aggregate excess of loss reinsurance

Under an aggregate excess of loss arrangement with retention limit M, the insurer pays all the claims if the total claim amount is less than or equal to the retention limit. The maximum payment made by the insurer is M. So the insurer's aggregate claim payment is:

$$S_I=\begin{cases} S & \text{if } S\le M \\ M & \text{if } S>M \end{cases}$$

The reinsurer's aggregate claim payment is:

$$S_R=\begin{cases} 0 & \text{if } S\le M \\ S-M & \text{if } S>M \end{cases}$$

Calculations involving aggregate excess of loss reinsurance are done using a first principles approach.

Question

The annual number of claims from a small group of policies has a Poisson distribution with a mean of 2. Individual claim amounts have the following distribution:

Amount	200	400
Probability	0.7	0.3

Individual claim amounts are independent of each other and are also independent of the number of claims. The insurer has purchased aggregate excess of loss reinsurance with a retention limit of 600.

Calculate the probability that the reinsurer is involved in paying the claims that arise in the next policy year.

Solution

The reinsurer will be involved if the total claim amount is more than 600. Since the total claim amount must be a multiple of 200, the probability is:

$$P(S>600)=1-P(S=0)-P(S=200)-P(S=400)-P(S=600)$$

The total claim amount will be 0 only if there are no claims. So:

$$P(S=0)=P(N=0)=\frac{e^{-2}2^0}{0!}=e^{-2}$$

Using the assumption that individual claim amounts are independent of the number of claims, we have:

$$P(S=200)=P(N=1,X_1=200)=P(N=1)P(X_1=200)=\frac{e^{-2}2^1}{1!}\times 0.7=1.4e^{-2}$$

$$P(S=400)=P(N=2,X_1=200,X_2=200)+P(N=1,X_1=400)$$

$$=\frac{e^{-2}2^2}{2!}\times 0.7^2+\frac{e^{-2}2^1}{1!}\times 0.3$$

$$=1.58e^{-2}$$

and:

$$P(S=600)=P(N=3,X_1=200,X_2=200,X_3=200)$$

$$+P(N=2,X_1=200,X_2=400)+P(N=2,X_1=400,X_2=200)$$

$$=\frac{e^{-2}2^3}{3!}\times 0.7^3+2\times\frac{e^{-2}2^2}{2!}\times 0.7\times 0.3$$

$$=1.29733e^{-2}$$

So:

$$P(S>600)=1-e^{-2}(1+1.4+1.58+1.29733)=0.28579$$

We can also simulate the collective risk model with *aggregate* reinsurance. For example to simulate 10,000 values for a reinsurer where claims have a compound Poisson distribution with parameter 1,000 and a gamma claims distribution with $\alpha=750$ and $\lambda=0.25$ under *aggregate* excess of loss with retention 3,000,000 we would take our S from Section 3.4 of Chapter 19 and then use:

```
sR <- pmax(0,s-3000000)
```

2 The individual risk model

Under this model a portfolio consisting of a fixed number of risks is considered. It will be assumed that:

- **these risks are independent**
- **claim amounts from these risks are not** (necessarily) **identically distributed random variables**
- **the number of risks does not change over the period of insurance cover.**

As before, aggregate claims from this portfolio are denoted by S. So:

$$S = Y_1 + Y_2 + \dots + Y_n$$

where Y_j denotes the claim amount under the j th risk and n denotes the number of risks. It is possible that some risks will not give rise to claims. Thus, some of the observed values of $\{Y_j\}_{j=1}^n$ may be 0.

In fact in most forms of insurance most policies would not give rise to any claims during a given year.

This approach is referred to as an *individual risk model* because it is considering the claims arising from each individual policy.

For each risk, the following assumptions are made:

- **the number of claims from the j th risk, N_j, is either 0 or 1 (20.3)**
- **the probability of a claim from the j th risk is q_j. (20.4)**

If a claim occurs under the j th risk, the claim amount is denoted by the random variable X_j. Let $F_j(x)$, μ_j and σ_j^2 denote the distribution function, mean and variance of X_j respectively.

Assumption (20.3) is very restrictive. It means that a maximum of one claim from each risk is allowed for in the model. This includes risks such as one-year term assurance (since a policyholder can only die once), **but excludes many types of general insurance policy. For example, there is no restriction on the number of claims that could be made in a policy year under household contents insurance.**

There are three important differences between this model and the collective risk model:

(1) The number of risks in the portfolio has been specified. In the collective risk model, there was no need to specify this number, nor to assume that it remained fixed over the period of insurance cover (not even when it was assumed that $N \sim Bin(n,q)$).

On the other hand we could argue that there was an implicit assumption of a constant number of risks in the very fact that we were using a binomial distribution to model the number of claims.

(2) The number of claims from each individual risk has been restricted. There was no such restriction in the collective risk model.

(3) It is assumed that individual risks are independent. In the collective risk model it was individual claim amounts that were independent.

The contrast here is between the occurrence of claims and the size of claims.

Assumptions (20.3) and (20.4) say that $N_j \sim Bin(1, q_j)$. Thus, the distribution of Y_j is compound binomial, with individual claim amount random variable X_j . From formulae (19.15) and (19.16) in Chapter 19 (for the mean and variance of a compound random variable) **it follows immediately that:**

$$E[Y_j] = q_j \mu_j \tag{20.5}$$

$$\text{var}[Y_j] = q_j \sigma_j^2 + q_j(1-q_j)\mu_j^2 \tag{20.6}$$

S is the sum of n independent compound binomial random variables. The distribution of S can be stated only when the compound binomial variables are identically distributed, as well as independent. It is possible, but complicated, to compute the distribution function of S under certain conditions.

However, it is easy to find the mean and variance of S :

$$E[S] = E\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} E[Y_j] = \sum_{j=1}^{n} q_j \mu_j \tag{20.7}$$

The assumption that individual risks are independent is needed to write:

$$\text{var}[S] = \text{var}\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} \text{var}[Y_j] = \sum_{j=1}^{n} (q_j \sigma_j^2 + q_j(1-q_j)\mu_j^2) \tag{20.8}$$

In the special case when $\{Y_j\}_{j=1}^{n}$ is a sequence of identically distributed, as well as independent, random variables, then for each policy the values of q_j , μ_j and σ_j^2 are identical, say q , μ and σ^2 . Since $F_j(x)$ is independent of j , we can refer to it simply as $F(x)$. Hence, S is compound binomial, with binomial parameters n and q , and individual claims have distribution function $F(x)$. In this special case, it reduces to the collective risk model, and it can be seen from (20.7) and (20.8) that:

$$E[S] = nq\mu$$

$$\text{var}[S] = nq\sigma^2 + nq(1-q)\mu^2$$

which correspond to (19.15) and (19.16) respectively in Chapter 19.

Question

The probability of a claim arising on any given policy in a portfolio of 1,000 one-year term assurance policies is 0.004. Claim amounts have a $Gamma(5, 0.002)$ distribution. Calculate the mean and variance of the aggregate claim amount.

Solution

We have:

$$\mu_j = \frac{5}{0.002} = 2{,}500 \qquad \text{and:} \qquad \sigma_j^2 = \frac{5}{0.002^2} = 1{,}250{,}000$$

So the mean and variance of the aggregate claim amount are:

$$E(S) = nq\,\mu = 1{,}000 \times 0.004 \times 2{,}500 = 10{,}000$$

and:

$$\begin{aligned}\text{var}(S) &= nq\sigma^2 + nq(1-q)\mu^2 \\ &= 1{,}000 \times \left[0.004 \times 1{,}250{,}000 + 0.004 \times 0.996 \times 2{,}500^2\right] \\ &= (£5{,}468)^2\end{aligned}$$

We can use R to simulate the total claim amount payable under the individual risk model.

Suppose we have n life policies, with the probabilities of death for each policy contained in the vector `q` **and simulated claim amounts for each policy contained in the vector** `claim`**. Then:**

```
S <- q*claim
```

We can now find moments, the coefficient of skewness, probabilities and quantiles as before, and also apply reinsurance if appropriate.

3 Parameter variability / uncertainty

This section forms part of the Core Reading, but does not address any specific syllabus objectives. However, the material here provides useful practice in applying the models we have studied.

3.1 Introduction

So far risk models have been studied assuming that the parameters, that is the moments and in some cases even the distributions, of the number of claims and of the amount of individual claims are known with certainty. In general, these parameters would not be known but would have to be estimated from appropriate sets of data. In this section it will be seen how the models introduced earlier can be extended to allow for parameter uncertainty / variability. This will be done by looking at a series of examples. Most, but not all, of these examples will consider uncertainty in the claim number distribution since this, rather than the individual claim amount distribution, has received more attention in the actuarial literature. All the examples will be based on claim numbers having a Poisson distribution.

3.2 Variability in a heterogeneous portfolio

Consider a portfolio consisting of n independent policies. The aggregate claims from the ith policy are denoted by the random variable S_i, where S_i has a compound Poisson distribution with parameters λ_i, and the CDF of the individual claim amounts distribution is $F(x)$. Notice that, for simplicity, the CDF of the distribution of individual claim amounts, $F(x)$, is assumed to be identical for all the policies.

In this example the CDF of individual claim amounts, *ie* $F(x)$, is assumed to be known but the values of the Poisson parameters, *ie* the λ_is, are not known. In this subsection the λ_is are assumed to be (sample values of) independent random variables with the same (known) distribution. In other words $\{\lambda_i\}_{i=1}^n$ is treated as a set of independent and identically distributed random variables with a known distribution. This means that if a policy is chosen at random from the portfolio it is assumed that the Poisson parameter for the policy is not known but that probability statements can be made about it. For example, 'there is a 50% chance that its Poisson parameter lies between 3 and 5'. It is important to understand that the Poisson parameter for a policy chosen from the portfolio is a fixed number; the problem is that this number is not known.

Example

Suppose that the Poisson parameters of policies in a portfolio are not known but are equally likely to be 0.1 or 0.3.

(i) Find the mean and variance (in terms of m_1 and m_2) of the aggregate claims from a policy chosen at random from the portfolio.

(ii) Find the mean and variance (in terms of m_1, m_2 and n) of the aggregate claims from the whole portfolio.

It may be helpful to think of this as a model of part of a motor insurance portfolio. The policies in the whole portfolio have been subdivided according to their values for rating factors such as 'age of driver', 'type of car' and even 'past claims experience'. The policies in the part of the portfolio being considered have identical values for these rating factors. However, there are some factors, such as 'driving ability', that cannot easily be measured and so they cannot be taken explicitly into account. It is supposed that some of the policyholders in this part of the portfolio are 'good' drivers and the remainder are 'bad' drivers. The individual claim amount distribution is the same for all drivers but 'good' drivers make fewer claims (0.1 *pa* on average) than 'bad' drivers (0.3 *pa* on average). It is assumed that it is known, possibly from national data, that a policyholder in this part of the portfolio is equally likely to be a 'good' driver or a 'bad' driver but that it cannot be known whether a particular policyholder is a 'good' driver or a 'bad' driver.

Solution

Let λ_i, $i = 1, 2, ..., n$ be the Poisson parameter of the ith policy in the portfolio. $\{\lambda_i\}_{i=1}^{\infty}$ is regarded as a set of independent and identically distributed random variables, each with the following distribution:

$$P(\lambda_i = 0.1) = 0.5$$

$$P(\lambda_i = 0.3) = 0.5$$

From this:

$$E[\lambda_i] = 0.2$$

$$\text{var}[\lambda_i] = 0.01$$

(i) The moments of S_i can be calculated by conditioning on the value of λ_i. Since $S_i \mid \lambda_i$ has a straightforward compound Poisson distribution, formulae (19.8) and (19.9) in Chapter 19 can be used to write:

$$E[S_i] = E[E[S_i \mid \lambda_i]] = E[\lambda_i m_1] = 0.2 m_1$$

$$\begin{aligned}\text{var}[S_i] &= E[\text{var}[S_i \mid \lambda_i]] + \text{var}[E[S_i \mid \lambda_i]]\\ &= E[\lambda_i m_2] + \text{var}[\lambda_i m_1]\\ &= 0.2 m_2 + 0.01 m_1^2\end{aligned}$$

(ii) The random variables $\{S_i\}_{i=1}^{n}$ are independent and identically distributed, each with the distribution of S_i given in part (i). Hence, the result in (i) above can be used to write:

$$E\left[\sum_{i=1}^{n} S_i\right] = nE[S_i] = 0.2\,n\,m_1$$

$$\text{var}\left[\sum_{i=1}^{n} S_i\right] = n\,\text{var}[S_i] = 0.2\,n\,m_2 + 0.01\,n\,m_1^2$$

Example

Suppose the Poisson parameters for individual policies are drawn from a gamma distribution with parameters α and δ. Find the distribution of the number of claims from a policy chosen at random from the portfolio.

Solution

Let N_i denote the number of claims from the *i* th policy in the portfolio and let λ_i be its Poisson parameter. Then N_i has a Poisson distribution with parameter λ_i but the problem is that (by assumption) the value of λ_i is not known. What *is* known is the distribution from which λ_i has been chosen.

The problem can be summarised as follows:

Given that:

$$N_i \mid \lambda_i \sim Poisson(\lambda_i) \text{ and } \lambda_i \sim Gamma(\alpha,\delta)$$

find the marginal distribution of N_i.

The marginal distribution of N_i is its unconditional distribution. In this example, N_i can only take whole number values, so it is a discrete random variable. To determine its marginal distribution, we need to derive a formula for the unconditional probability $P(N_i = x)$.

This problem can be solved by removing the conditioning in the usual way.

Recall that if X and Y are discrete random variables, then the unconditional probability $P(X = x)$ is given by:

$$P(X=x)=\sum_y P(X=x,Y=y)=\sum_y P(X=x \mid Y=y)P(Y=y)$$

In this example X is being replaced by N_i and Y is being replaced by λ_i. Since λ_i is a continuous random variable, we turn the sum into an integral and formula becomes:

$$P(N_i=x)=\int_{\lambda_i} P(N_i=x \mid \lambda_i) f(\lambda_i)\, d\lambda_i$$

For $x = 0,1,2,\ldots$:

$$P(N_i=x)=\int_0^\infty \exp\{-\lambda\}\frac{\lambda^x}{x!}\frac{\delta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp\{-\delta\lambda\}\,d\lambda$$

$$=\frac{\delta^\alpha}{\Gamma(\alpha)x!}\int_0^\infty \exp\{-\lambda(\delta+1)\}\,\lambda^{x+\alpha-1}\,d\lambda$$

Evaluate the integral by comparing the integrand with a gamma density.

We can make the integrand look like the PDF of the $Gamma(x+\alpha,\ \delta+1)$ distribution by inserting a factor of $\dfrac{(\delta+1)^{x+\alpha}}{\Gamma(x+\alpha)}$ inside the integral. We need to compensate for doing this by inserting a factor of $\dfrac{\Gamma(x+\alpha)}{(\delta+1)^{x+\alpha}}$ outside the integral. This gives:

$$P(N_i=x)=\frac{\delta^\alpha}{\Gamma(\alpha)x!}\frac{\Gamma(x+\alpha)}{(\delta+1)^{x+\alpha}}\int_0^\infty \frac{(\delta+1)^{x+\alpha}}{\Gamma(x+\alpha)}\lambda^{x+\alpha-1}e^{-\lambda(\delta+1)}\,d\lambda$$

$$=\frac{\delta^\alpha}{\Gamma(\alpha)x!}\frac{\Gamma(x+\alpha)}{(\delta+1)^{x+\alpha}}\int_0^\infty f(\lambda)d\lambda \quad \text{where } \lambda\sim Gamma(x+\alpha,\delta+1)$$

The integral in the line above is 1 (as we are integrating a PDF over all possible values of the random variable).

So:

$$\boldsymbol{P(N_i=x)=\frac{\delta^\alpha}{\Gamma(\alpha)x!}\frac{\Gamma(x+\alpha)}{(\delta+1)^{x+\alpha}}}$$

which shows that the marginal distribution of N_i is negative binomial with parameters α and $\dfrac{\delta}{\delta+1}$.

3.3 Variability in a homogeneous portfolio

Now a different example is considered. Suppose, as before, there is a portfolio of n policies. The aggregate claims from a single policy have a compound Poisson distribution with parameters λ, and the CDF of the individual claim amounts random variable is $F(x)$. The Poisson parameters are the same for all policies in the portfolio. If the value of λ were known, the aggregate claims from different policies would be independent of each other. It is assumed that the value of λ is not known, possibly because it changes from year to year, but that there *is* some indication of the probability that λ will be in any given range of values. As in the previous example, it is assumed for simplicity that there is no uncertainty about the moments or distribution of the individual claim amounts, *ie* about $F(x)$. The uncertainty about the value of λ can be modelled by regarding λ as a random variable (with a known distribution).

Example

Suppose that the Poisson parameter, λ, will be equal to 0.1 or to 0.3 with equal probability.

(i) Calculate the mean and variance (in terms of m_1 and m_2) of the aggregate claims from a policy chosen at random from the portfolio.

(ii) Calculate the mean and variance (in terms of m_1, m_2 and n) of the aggregate claims from the whole portfolio.

Solution

Using the same notation as before let S_i denote the aggregate claims from the i th policy in the portfolio. The situation can be summarised as follows:

The random variables $\{S_i \mid \lambda\}_{i=1}^{n}$ are independent and identically distributed, each with a compound Poisson distribution with parameters λ and $F(x)$. The random variable λ has the following distribution:

$$P(\lambda = 0.1) = 0.5$$

$$P(\lambda = 0.3) = 0.5$$

(i) Conditioning on the value of λ:

$$E[S_i] = E[E(S_i \mid \lambda)] = E[\lambda m_1] = 0.2 m_1$$

$$\text{var}[S_i] = E[\text{var}(S_i \mid \lambda)] + \text{var}[E(S_i \mid \lambda)] = E[\lambda m_2] + \text{var}[\lambda m_1]$$

$$= 0.2 m_2 + 0.01 m_1^2$$

(ii) $$E\left[\sum_{i=1}^{n} S_i\right] = n E[S_1] = 0.2 n m_1$$

(since $\{S_i\}_{i=1}^{n}$ are identically distributed)

$$\text{var}\left[\sum_{i=1}^{n} S_i\right] = E\left[\text{var}\left(\sum_{i=1}^{n} S_i \mid \lambda\right)\right] + \text{var}\left[E\left(\sum_{i=1}^{n} S_i \mid \lambda\right)\right]$$

$$= E[n\lambda m_2] + \text{var}[n\lambda m_1]$$

$$= 0.2 n m_2 + 0.01 n^2 m_1^2$$

Note that $S_1 \mid \lambda, \ldots, S_n \mid \lambda$ are independent but $S_1, \ldots, S_n$ are not unconditionally independent (since they all depend on the value of λ), so $\text{var}(\sum_{i=1}^{n} S_i) \neq \sum_{i=1}^{n} \text{var}(S_i)$.

It is useful to compare the answers to the above example with those to the first example in the Section 3.2. The values of the mean are in all cases the same, as are the variances when a single policy is considered (part (i)). The difference occurs when variances for more than one policy are considered (part (ii)), in which case the second example gives the greater variance. It is important to understand the differences (and the similarities) between the two examples. A practical situation where the second example could be appropriate would be a portfolio of policies insuring buildings in a certain area. The number of claims could depend on, among other factors, the weather during the year; an unusually high number of storms resulting in a high expected number of claims (*ie* a high value of λ) and vice versa for all the policies together.

3.4 Variability in claim numbers and claim amounts and parameter uncertainty

This section contains two more examples. The first is a rather complicated example involving uncertainty over claim amounts as well as claim numbers.

Example

An insurance company models windstorm claims under household insurance policies using the following assumptions.

The number of storms arising each year, K, is assumed to have a Poisson distribution with parameter λ.

The number of claims arising from the ith storm, N_i, $i = 1, 2, ..., K$, is assumed to have a Poisson distribution with parameter Θ_i.

The parameters Θ_i, $i = 1, 2, ..., K$, are assumed to be independent and identically distributed random variables, with $E(\Theta_i) = n$ and $\text{var}(\Theta_i) = s_1^2$.

The amount of the jth claim arising from the ith storm, X_{ij}, $j = 1, 2, ..., N_i$, has a lognormal distribution with parameters μ_i and σ^2, where σ^2 is assumed to be known. The mean claim amounts, $\Lambda_i = \exp(\mu_i + \sigma^2 / 2)$ are assumed to be independent and identically distributed random variables with mean p and variance s_2^2.

It is also assumed that Θ_i and Λ_i are independent.

(i) Show that $E[X_{ij}] = p$ and $\text{var}[X_{ij}] = \exp\{\sigma^2\}(p^2 + s_2^2) - p^2$.

(ii) Let S_i denote aggregate claims outgo from the i-th storm, so that $S_i \mid \{\Theta_i, \Lambda_i\}$ is a compound Poisson random variable. Show that:

$$E[S_i] = np$$

and:

$$\text{var}[S_i] = (p^2 + s_2^2)(n^2 + s_1^2 + n\exp\{\sigma^2\}) - n^2 p^2$$

(iii) Find expressions for the mean and variance of the annual aggregate claims outgo from all storms.

Solution

(i) $E[X_{ij}] = E[E(X_{ij} \mid \Lambda_i)] = E[\Lambda_i] = p$

$$\text{var}[X_{ij}] = E[\text{var}(X_{ij} \mid \Lambda_i)] + \text{var}[E(X_{ij} \mid \Lambda_i)]$$

$$= E[\Lambda_i^2(\exp\{\sigma^2\} - 1)] + \text{var}(\Lambda_i)$$

$$= (p^2 + s_2^2)(\exp\{\sigma^2\} - 1) + s_2^2$$

$$= (p^2 + s_2^2)\exp\{\sigma^2\} - p^2$$

(ii) $E[S_i] = E[E(S_i \mid \Theta_i, \Lambda_i)] = E[\Theta_i \Lambda_i] = np$ **since** Θ_i **and** Λ_i **are independent.**

Now, since $S_i \mid \{\Theta_i, \Lambda_i\}$ **has a compound Poisson distribution:**

$$\text{var}[S_i \mid \Theta_i, \Lambda_i] = \Theta_i E[X_{ij}^2 \mid \Lambda_i] = \Theta_i (\Lambda_i^2 \exp\{\sigma^2\})$$

and so:

$$E[\text{var}(S_i \mid \Theta_i, \Lambda_i)] = n(p^2 + s_2^2)\exp\{\sigma^2\}$$

Also:

$$\begin{aligned}\text{var}[E(S_i \mid \Theta_i, \Lambda_i)] &= \text{var}[\Theta_i \Lambda_i] = E[\Theta_i^2 \; \Lambda_i^2] - n^2 p^2 \\ &= (n^2 + s_1^2)(p^2 + s_2^2) - n^2 p^2\end{aligned}$$

Putting these last two results together:

$$\text{var}[S_i] = (n^2 + s_1^2)(p^2 + s_2^2) - n^2 p^2 + n(p^2 + s_2^2)\exp\{\sigma^2\}$$

(iii) **Let** R **be a random variable denoting the annual aggregate claims outgo from all storms. Then** R **can be written:**

$$R = \sum_{i=1}^{K} S_i$$

where K **has a Poisson distribution and the random variables** S_i **are IID (independent and identically distributed).**

Hence, R **has a compound Poisson distribution and so:**

$$E[R] = \lambda E(S_i) = \lambda n p$$

$$\begin{aligned}\text{var}[R] &= \lambda E(S_i^2) = \lambda(\text{var}[S_i] + E[S_i]^2) \\ &= \lambda(p^2 + s_2^2)(n^2 + s_1^2 + n\exp\{\sigma^2\})\end{aligned}$$

Example

Each year an insurance company issues a number of household contents insurance policies, for each of which the annual premium is £80. The aggregate annual claims from a single policy have a compound Poisson distribution; the Poisson parameter is 0.4 and individual claim amounts have a gamma distribution with parameters α **and** λ**. The expense involved in settling a claim is a random variable uniformly distributed between £50 and £**b **(>£50). The amount of the expense is independent of the amount of the associated claim. The random variable** S **represents the total aggregate claims and expenses in one year from this portfolio. It may be assumed that** S **has approximately a normal distribution.**

(i) Suppose that:

$$\alpha = 1\,;\ \lambda = 0.01\,;\ b = 100$$

Show that the company must sell at least 884 policies in a year to be at least 99% sure that the premium income will exceed the claims and expenses outgo.

(ii) Now suppose that the values of α , λ and b are not known with certainty but could be anywhere in the following ranges:

$$0.95 \le \alpha \le 1.05\,;\ 0.009 \le \lambda \le 0.011\,;\ 90 \le b \le 110$$

By considering what, for the insurance company, would be the worst possible combination of values for α , λ and b, calculate the number of policies the company must sell to be at least 99% sure that the premium income will exceed the claims and expenses outgo.

Solution

Let X_i be the amount of the ith claim and Y_i be the amount of the associated expense. Let N be the total number of claims from the portfolio and let n be the number of policies in the portfolio. Then N has a Poisson distribution with parameter 0.4n and S can be written:

$$S = \sum_{i=1}^{N} (X_i + Y_i)$$

where $\{X_i + Y_i\}_{i=1}^{\infty}$ is a sequence of independent and identically distributed random variables, independent of N. From this it can be seen that S has a compound Poisson distribution with $X_i + Y_i$ representing the 'amount of the ith individual claim'. Standard results can now be used to write down the following formulae for the moments of S:

$$E[S] = 0.4n\, E[X_i + Y_i]$$

$$\text{var}[S] = 0.4n\, E[(X_i + Y_i)^2\,] = 0.4n(E[X_i^2] + 2E[X_iY_i] + E[Y_i^2]\,)$$

In terms of α , λ and b , the moments of X_i and Y_i are as follows:

$$E[X_i] = \alpha / \lambda \qquad E[Y_i] = (b+50)/2$$

$$E[X_i^2] = \alpha(\alpha+1)/\lambda^2 \qquad E[Y_i^2] = (b^2 + 50b + 2{,}500)/3$$

$$E[X_iY_i] = E[X_i]\,E[Y_i]$$

where the final relationship follows from the independence of X_i and Y_i.

(i) **Now put:**

$$\alpha = 1; \ \lambda = 0.01; \ b = 100$$

into these formulae to show that:

$$E[S] = 70n \text{ and } \operatorname{var}[S] = 127.80^2 n$$

Hence, S has approximately a normal distribution with mean $70n$ and standard deviation $127.80\sqrt{n}$. The premium income is $80n$ and the smallest value of n is required such that:

$$P(S < 80n) \geq 0.99$$

Standardising S in the usual way for a normal distribution, this becomes:

$$P\left[\frac{S - 70n}{127.80\sqrt{n}} < \frac{80n - 70n}{127.80\sqrt{n}}\right] \geq 0.99$$

The upper 99% point of a standard normal distribution is 2.326 and so the condition for n is:

$$\frac{80n - 70n}{127.80\sqrt{n}} \geq 2.326$$

which gives:

$$n \geq 883.7$$

(or $n \geq 884$ to the next higher integer).

(ii) **For the insurance company, the worst possible combination of values for α, λ and b is the combination which gives the highest possible values for $E[S]$ and var[S]. To see this, let μ and σ denote the mean and the standard deviation of aggregate claims and expenses from a single policy. Both μ and σ will be functions of α, λ and b and:**

$$E[S] = n\mu \text{ and } \operatorname{var}[S] = n\sigma^2$$

Following the same steps as in part (i), the condition for n is:

$$\frac{(80 - \mu)\sqrt{n}}{\sigma} \geq 2.326$$

which becomes:

$$n \geq [2.326\sigma / (80 - \mu)]^2$$

Hence, the highest value of n results from the highest values for μ and σ (provided the highest value for μ is less than 80). Now note that:

$$\mu = 0.4E[X_i + Y_i] \quad \text{and} \quad \sigma^2 = 0.4E[(X_i + Y_i)^2]$$

From the formulae for the moments of X_i and Y_i given above, μ and σ are maximised when α and b are as large as possible and λ is as small as possible, *ie* when:

$$\alpha = 1.05\,; \quad \lambda = 0.009\,; \quad b = 110$$

This combination of values gives:

$$\mu = 78.67 \quad \text{and} \quad \sigma = 144.14$$

so that n must be at least 63,546 for the insurance company to be at least 99% sure that premium income will exceed claims and expenses outgo.

Chapter 20 Summary

Collective risk model with reinsurance

In the collective risk model, individual claims can be subject to a reinsurance agreement, either proportional or excess of loss.

Under the collective risk model, the aggregate claim amount S is given by:

$$S = X_1 + X_2 + \cdots + X_N$$

where X_i is the amount of the i th claim and N is the total number of claims.

If reinsurance is in place, the insurer's aggregate claims net of reinsurance can be represented as:

$$S_I = Y_1 + Y_2 + \cdots + Y_N$$

where Y_i is the amount of the i th claim paid by the insurer and N is defined as above. S_I is a compound random variable and:

$$E(S_I) = E(N)E(Y)$$

$$\text{var}(S_I) = E(N)\text{var}(Y) + \text{var}(N)[E(Y)]^2$$

$$M_{S_I}(t) = M_N[\ln M_Y(t)]$$

The reinsurer's aggregate claims can be represented as:

$$S_R = Z_1 + Z_2 + \cdots + Z_N$$

where Z_i is the amount of the i th claim paid by the reinsurer and N is defined as above. S_R is a compound random variable and:

$$E(S_R) = E(N)E(Z)$$

$$\text{var}(S_R) = E(N)\text{var}(Z) + \text{var}(N)[E(Z)]^2$$

$$M_{S_R}(t) = M_N[\ln M_Z(t)]$$

Under individual excess of loss reinsurance, some of the claims may fall below the retention level M. If this is the case, then some of the Z_i will be zero. An alternative way of expressing the reinsurer's aggregate claims is as:

$$S_R = W_1 + W_2 + \cdots + W_{NR}$$

where $W_i = Z_i \mid Z_i > 0$ and NR is the number of non-zero claims, *ie* the number of claims in which the reinsurer is involved.

Under an aggregate excess of loss arrangement with retention limit M, the maximum payment made by the insurer is M. The insurer's aggregate claim payment is:

$$S_I = \begin{cases} S & \text{if } S \le M \\ M & \text{if } S > M \end{cases}$$

The reinsurer's aggregate claim payment is:

$$S_R = \begin{cases} 0 & \text{if } S \le M \\ S - M & \text{if } S > M \end{cases}$$

Individual risk model

The individual risk model considers the payments made under each risk (*eg* policy) separately. The model assumes that:

- the number of risks is fixed
- the risks are independent
- claim amounts from these risks are not necessarily IID
- N_j, the number of claims from the j th risk is either 0 or 1.

For a portfolio containing n risks, the aggregate claim amount is given by:

$$S = Y_1 + Y_2 + \cdots + Y_n$$

where Y_j denotes the aggregate claims from risk j. Since each Y_j is the sum of a random number (0 or 1) of random claim amounts, each Y_j has a compound binomial distribution. Suppose that q_j is the probability of a claim from the j th risk. If a claim arises from the j th risk, suppose that the claim amount random variable is X_j. Then:

$$E(S) = \sum_{j=1}^{n} q_j \mu_j$$

$$\text{var}(S) = \sum_{j=1}^{n} \left[q_j \sigma_j^2 + q_j (1 - q_j) \mu_j^2 \right]$$

$$M_S(t) = \prod_{j=1}^{n} \left[q_j M_{X_j}(t) + (1 - q_j) \right]$$

where $\mu_j = E(X_j)$ and $\sigma_j^2 = \text{var}(X_j)$.

If, for a group of n risks, the probability of a claim is fixed and the claim amounts are IID random variables, then the individual risk model is equivalent to a collective risk model where S has a compound binomial distribution with $N \sim Bin(n,q)$.

Chapter 20 Practice Questions

20.1 The annual aggregate claims from a risk have a compound Poisson distribution with parameter 250. Individual claim amounts have a Pareto distribution with parameters $\alpha = 4$ and $\lambda = 900$. The insurer effects proportional reinsurance with a retained proportion of 75%.

Calculate the variances of the total amounts paid by the insurer and by the reinsurer.

20.2 The aggregate claims from a risk have a compound Poisson distribution with parameter μ. Individual claim amounts (in £) have a Pareto distribution with parameters $\alpha = 3$ and $\lambda = 1{,}000$.

The insurer of this risk calculates the premium using a premium loading factor of 0.2 (*ie* it charges 20% in excess of the risk premium).

The insurer is considering effecting individual excess of loss reinsurance with retention limit £1,000. The reinsurance premium would be calculated using a premium loading factor of 0.3.

The insurer's profit is defined to be the premium charged by the insurer less the reinsurance premium and less the claims paid by the insurer, net of reinsurance.

(i) Show that the insurer's expected profit before reinsurance is 100μ.

(ii) Calculate the insurer's expected profit after effecting the reinsurance, and hence find the percentage reduction in the insurer's expected profit.

(iii) Calculate the percentage reduction in the standard deviation of the insurer's profit as a result of effecting the reinsurance.

20.3 Aggregate annual claims from a portfolio of general insurance policies have a compound Poisson distribution with Poisson parameter 20. Individual claim amounts have a uniform distribution over the interval $(0,200)$. Excess of loss reinsurance is arranged so that the expected amount paid by the insurer on any claim is 50.

Calculate the variance of the aggregate annual claims paid by the insurer.

20.4 A portfolio of policies consists of one-year term assurances on 100 lives aged exactly 30 and 200 lives aged exactly 40. The probability of a claim during the year on any one of the lives is 0.0004 for the 30 year olds and 0.001 for the 40 year olds.

If the sum assured on a life aged x is uniformly distributed between $1{,}000(x-10)$ and $1{,}000(x+10)$, calculate the variance of the aggregate claims from this portfolio during the year (assuming that policies are independent with regard to claims).

20.5 The number of claims from a given portfolio has a Poisson distribution with a mean of 1.5 per month. Individual claim amounts have the following distribution:

Amount	200	300
Probability	0.65	0.35

An aggregate reinsurance contract has been arranged so that the insurer pays no more than 400 per month in total.

Assuming that the individual claim amounts are independent of each other and are also independent of the number of claims, calculate the expected aggregate monthly claim amounts for the insurer and the reinsurer.

20.6 A portfolio consists of 500 independent risks. For the i th risk, with probability $1-q_i$ there are no claims in one year, and with probability q_i there is exactly one claim ($0<q_i<1$). For all risks, if there is a claim, it has mean μ, variance σ^2 and moment generating function $M(t)$. Let T be the total amount claimed on the whole portfolio in one year.

Exam style

(i) Determine the mean and variance of T. [4]

The amount claimed in one year on risk i is approximated by a compound Poisson random variable with Poisson parameter q_i and claims with the same mean μ, the same variance σ^2, and the same moment generating function $M(t)$ as above. Let $\tilde{T}$ denote the total amount claimed on the whole portfolio in one year in this approximate model.

(ii) Determine the mean and variance of $\tilde{T}$, and compare your answers to those in part (i). [4]

Assume that $q_i = 0.02$ for all i, and if a claim occurs, it is of size μ with probability one.

(iii) Derive the moment generating function of T, and show that T has a compound binomial distribution. [2]

(iv) Determine the moment generating function of the approximating $\tilde{T}$, and show that $\tilde{T}$ has a compound Poisson distribution. [2]

[Total 12]

20.7 Exam style A company is analysing the number of accidents that occur each year on the factory floor. It believes that the number of accidents per year N has a geometric distribution with parameter 0.8, so that:

$$P(N=n)=0.8\times 0.2^{n}, \quad n=0,1,2,...$$

For each accident, the number of employees injured is Y, where $Y=X+1$, and X is believed to have a Poisson distribution with parameter 2.2.

The company has taken out an insurance policy, which provides a benefit of £1,000 to each injured employee, up to a maximum of three employees per accident, irrespective of the level of injury. There is no limit on the number of accidents that may be claimed for in a year.

(i) Show that $E(S)=0.634$ and $\text{var}(S)=2.125$, where S is the total number of employees claiming benefit in a year under this policy. [7]

(ii) Hence find the mean and variance of the aggregate amount paid out under this policy in a year. [1]

[Total 8]

20.8 Exam style An insurance company offers accident insurance for employees. A total of 650 policies have been issued split between two categories of employees. The first category contains 400 policies, and claims occur on each policy according to a Poisson process at a rate of one claim per 20 years, on average. In this category all claim amounts are £3,000. In the second category, claims occur on each policy according to a Poisson process at a rate of one claim per 10 years, on average. In this category, the claim amount is either £2,000 or £3,000 with probabilities 0.4 and 0.6, respectively. All policies are assumed to be independent. Let S denote the aggregate annual claims from the portfolio.

(i) Calculate the mean, variance and coefficient of skewness of S. [4]

(ii) Using the normal distribution as an approximation to the distribution of S, calculate Y such that the probability of S exceeding Y is 10%. [3]

(iii) The insurance company decides to effect reinsurance cover with aggregate retention £100,000, so that the insurance company then pays out no more than this amount in claims each year. In the year following the inception of this reinsurance, the numbers of policies in each of the two groups remains the same but, because of changes in the employment conditions of which the company was unaware, the probability of a claim in group 2 falls to zero. Using the normal distribution as an approximation to the distribution of S, calculate the probability of a claim being made on the reinsurance treaty. [3]

[Total 10]

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 20 Solutions

20.1 The mean and variance of the gross claim amounts are:

$$E(X)=\frac{\lambda}{\alpha-1}=\frac{900}{3}=300$$

$$\text{var}(X)=\frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)}=\frac{4\times 900^2}{3^2\times 2}=180,000$$

So the mean and variance of the net claims for the direct insurer and the reinsurer are:

$$E(Y)=0.75\times 300=225$$

$$\text{var}(Y)=0.75^2\times 180,000=101,250$$

$$E(Z)=0.25\times 300=75$$

$$\text{var}(Z)=0.25^2\times 180,000=11,250$$

Using the formula for the variance of a compound Poisson random variable, the variances of the aggregate claim payments made by the insurer and the reinsurer are:

$$\text{var}(S_I)=\lambda E[Y^2]=\lambda\left[\text{var}(Y)+[E(Y)]^2\right]=250[101,250+225^2]=37,968,750$$

$$\text{var}(S_R)=\lambda E[Z^2]=\lambda\left[\text{var}(Z)+[E(Z)]^2\right]=250[11,250+75^2]=4,218,750$$

20.2 (i) ***Expected profit before reinsurance***

We have:

$$E(X)=\frac{\lambda}{\alpha-1}=500$$

So the expected aggregate claim amount is:

$$E(S)=500\mu$$

The insurer's premium income is:

$$1.2E(S)=1.2\times 500\mu=600\mu$$

So the expected profit before reinsurance is:

$$600\mu-500\mu=100\mu$$

(ii) *Expected profit after reinsurance*

The reinsurance premium is given by $1.3E(S_R)$, where:

$$E(S_R) = E(Z)E(N) = \mu E(Z)$$

Now:

$$E(Z) = \int_{1,000}^{\infty} (x - 1,000) \frac{3 \times 1,000^3}{(1,000 + x)^4} dx$$

Setting $u = x - 1,000$:

$$E(Z) = \int_0^{\infty} u \frac{3 \times 1,000^3}{(2,000 + u)^4} du = \left(\frac{1,000}{2,000}\right)^3 \int_0^{\infty} u \frac{3 \times 2,000^3}{(2,000 + u)^4} du$$

Recognising this integral as the mean of the $Pareto(3, 2\,000)$ distribution, we see that:

$$E(Z) = \left(\frac{1}{2}\right)^3 \times \frac{2,000}{3-1} = 125$$

and:

$$E(S_R) = 125\mu$$

So the reinsurance premium is $1.3 \times 125\mu = 162.5\mu$.

Alternatively we could evaluate this integral using the substitution $t = 1000 + x$ *or using integration by parts.*

The insurer's expected aggregate claim payment is:

$$E(S_I) = E(S) - E(S_R) = 500\mu - 125\mu = 375\mu$$

So the insurer's expected profit after reinsurance is:

$$600\mu - 162.5\mu - 375\mu = 62.5\mu$$

This is the insurer's premium income, minus the premium paid by the insurer to the reinsurer, minus the insurer's expected aggregate claim payment.

The percentage reduction in the expected profit (which was 100μ without reinsurance) is 37.5%.

(iii) *Percentage reduction in standard deviation*

In the absence of reinsurance, the insurer's profit is equal to its premium income minus the aggregate claim amount. Since the premium income is a fixed amount and only the cost of claims is random, the variance of the profit is:

$$\text{var}(S) = \mu E(X^2)$$

We have:

$$E(X)=\frac{\lambda}{\alpha-1}=500 \qquad \text{and} \qquad \text{var}(X)=\frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)}=750,000$$

So:

$$E(X^2)=\text{var}(X)+[E(X)]^2=750,000+500^2=1,000,000$$

and:

$$\text{var}(S)=1,000,000\mu$$

Hence the standard deviation of the profit is $1,000\sqrt{\mu}$.

With reinsurance, the insurer's profit is equal to premiums charged less the reinsurance premium less the net claims paid. Since the premiums are fixed amounts, the variance of the insurer's profit is:

$$\text{var}(S_I)=\mu E(Y^2)$$

where:

$$E(Y^2)=\int_0^{1,000} x^2\frac{3\times1,000^3}{(1,000+x)^4}\,dx+\int_{1,000}^{\infty}1,000^2\frac{3\times1,000^3}{(1,000+x)^4}\,dx$$

$$=3\times1,000^3\int_0^{1,000}\frac{x^2}{(1,000+x)^4}\,dx+3\times1,000^5\int_{1,000}^{\infty}\frac{1}{(1,000+x)^4}\,dx$$

The second integral is:

$$\int_{1,000}^{\infty}\frac{1}{(1,000+x)^4}\,dx=\left[\frac{(1,000+x)^{-3}}{-3}\right]_{1,000}^{\infty}=\frac{1}{3\times2,000^3}$$

For the first integral, we can set $u=1,000+x$ to give:

$$\int_0^{1,000}\frac{x^2}{(1,000+x)^4}\,dx=\int_{1,000}^{2,000}\frac{(u-1,000)^2}{u^4}\,du=\left[-\frac{1}{u}+\frac{1,000}{u^2}-\frac{1,000,000}{3u^3}\right]_{1,000}^{2,000}=\frac{1}{24,000}$$

Alternatively, we could integrate by parts (twice).

So:

$$E(Y^2)=\frac{3\times1,000^3}{24,000}+\frac{3\times1,000^5}{3\times2,000^3}=250,000 \quad \text{and} \quad \text{var}(S_I)=250,000\mu$$

Hence the standard deviation of the insurer's profit is now $500\sqrt{\mu}$, which is a reduction of 50%.

The standard deviation is reduced by a greater percentage than the mean. This is very often the case for excess of loss reinsurance.

20.3 We have $X \sim U(0,200)$ and:

$$Y = \begin{cases} X & \text{if } X \le M \\ M & \text{if } X > M \end{cases} \qquad Z = \begin{cases} 0 & \text{if } X \le M \\ X-M & \text{if } X > M \end{cases}$$

The expected amount paid by the insurer on any claim is:

$$E(Y) = \int_0^M x\frac{1}{200}dx + \int_M^{200} M\frac{1}{200}dx = 50$$

Solving this:

$$\left[\frac{x^2}{400}\right]_0^M + \left[\frac{Mx}{200}\right]_M^{200} = 50$$

$$\Rightarrow \quad \frac{M^2}{400} + M - \frac{M^2}{200} = 50$$

$$\Rightarrow \quad M^2 - 400M + 20{,}000 = 0$$

$$\Rightarrow \quad M = \frac{-(-400) \pm \sqrt{(-400)^2 - 4\times1\times20{,}000}}{2\times1} = 58.579 \text{ or } 341.42$$

Since claims are a maximum of 200, M must be 58.579.

The variance of the aggregate annual claims paid by the insurer is:

$$\text{var}(S) = \lambda E(Y^2) = 20E(Y^2)$$

where:

$$E(Y^2) = \int_0^M x^2\frac{1}{200}dx + \int_M^{200} M^2\frac{1}{200}dx = \left[\frac{x^3}{600}\right]_0^M + \left[\frac{M^2x}{200}\right]_M^{200}$$

$$= \frac{M^3}{600} + M^2 - \frac{M^3}{200}$$

$$= \frac{58.579^3}{600} + 58.579^2 - \frac{58.579^3}{200}$$

$$= 2{,}761.42$$

Hence:

$$\text{var}(S) = 20\times2{,}761.42 = 55{,}228$$

20.4 For each age group, the individual claim amounts have a uniform distribution. So the mean and variance of the individual claim distributions are:

$$E(X) = \tfrac{1}{2}(b+a) = 1{,}000x$$

and: $\operatorname{var}(X) = \tfrac{1}{12}(b-a)^2 = \dfrac{20{,}000^2}{12}$

Using the individual risk model, the variance of the aggregate claim amount is:

$$\operatorname{var}(S) = \sum_{i=1}^{n} \left\{ q_i \sigma_i^2 + q_i(1-q_i)\mu_i^2 \right\}$$

$$= 100\left[(0.0004) \times \frac{20{,}000^2}{12} + (0.0004)(0.9996) \times 30{,}000^2\right]$$

$$+ 200\left[(0.001) \times \frac{20{,}000^2}{12} + (0.001)(0.999) \times 40{,}000^2\right]$$

$$= 37.32\text{m} + 326.35\text{m} = 363.67\text{m}$$

Alternatively, we could model the aggregate claim amount from each group as a compound binomial random variable. For example, $N \sim Bin(100, 0.0004)$ *for the 100 lives aged exactly 30. We could then use the formula for* $\operatorname{var}(S)$ *from the collective risk model.*

20.5 Under this reinsurance arrangement, we have:

$$S_I = \begin{cases} S & \text{if } S \le 400 \\ 400 & \text{if } S > 400 \end{cases} \qquad S_R = \begin{cases} 0 & \text{if } S \le 400 \\ S-400 & \text{if } S > 400 \end{cases}$$

where S is the total monthly claim amount.

Since individual claim amounts must be either 200 or 300, the possible values of S_I are 0, 200, 300, and 400 and:

$$E(S_I) = 0 \times P(S_I = 0) + 200 \times P(S_I = 200) + 300 \times P(S_I = 300) + 400 \times P(S_I = 400)$$

The insurer's aggregate claim amount is 0 if there are no claims. So:

$$P(S_I = 0) = P(N = 0) = \frac{e^{-1.5} \times 1.5^0}{0!} = e^{-1.5}$$

The insurer's aggregate claim amount is 200 if there is one claim and the amount of the claim is 200. So:

$$P(S_I = 200) = P(N = 1, X_1 = 200)$$

Since N and X_1 are independent, we have:

$$P(S_I = 200) = P(N = 1)P(X_1 = 200) = \frac{e^{-1.5} \times 1.5^1}{1!} \times 0.65 = 0.975e^{-1.5}$$

Similarly:

$$P(S_I=300)=P(N=1,X_1=300)=\frac{e^{-1.5}\times 1.5^1}{1!}\times 0.35=0.525e^{-1.5}$$

Finally, the insurer's aggregate claim amount is 400 if the total claim amount is 400 or more. This probability can be calculated by subtraction as follows:

$$\begin{aligned}P(S_I=400)&=1-P(S_I=0)-P(S_I=200)-P(S_I=300)\\&=1-e^{-1.5}-0.975e^{-1.5}-0.525e^{-1.5}\\&=1-2.5e^{-1.5}\end{aligned}$$

So:

$$E(S_I)=0\times e^{-1.5}+200\times 0.975e^{-1.5}+300\times 0.525e^{-1.5}+400(1-2.5e^{-1.5})=255.52$$

We can now calculate $E(S_R)$ using the result:

$$E(S_R)=E(S)-E(S_I)$$

We have:

$$E(S)=\lambda E(X)=1.5\left[200\times 0.65+300\times 0.35\right]=1.5\times 235=352.50$$

Hence:

$$E(S_R)=352.50-255.52=96.98$$

20.6 *This is part of Subject 106, April 2003, Question 9*.

(i) ***Mean and variance of T***

Let $T=Y_1+Y_2+\cdots+Y_{500}$, where Y_i is the total claim on the i th policy. Then:

$$E[T]=E[Y_1]+E[Y_2]+\cdots+E[Y_{500}]$$

$$\text{var}[T]=\text{var}[Y_1]+\text{var}[Y_2]+\cdots+\text{var}[Y_{500}]$$

Since $E[Y_i]=q_i\mu$ and $\text{var}[Y_i]=q_i\sigma^2+q_i(1-q_i)\mu^2$, we have:

$$E[T]=\mu\sum_{i=1}^{500}q_i$$ [2]

$$\text{var}[T]=\sigma^2\sum_{i=1}^{500}q_i+\mu^2\sum_{i=1}^{500}q_i(1-q_i)$$ [2]

since the risks are independent.

(ii) ***Mean and variance of $\tilde{T}$***

Let C be the amount claimed in one year on a single risk. Then, according to the approximation:

$$C = X_1 + X_2 + \cdots + X_N$$

where $N \sim Poi(q_i)$, $E[X] = \mu$ and $\text{var}[X] = \sigma^2$.

Also:

$$\tilde{T} = C_1 + C_2 + \cdots + C_{500}$$

where C_i is the total amount claimed on the i th risk.

Using the formulae for the mean and variance of compound Poisson random variable:

$$E[C_i] = \mu q_i \qquad \text{var}[C_i] = q_i(\mu^2 + \sigma^2)$$ [2]

Since $\tilde{T}$ is the sum of claims for the whole portfolio, we have:

$$E[\tilde{T}] = \mu \sum_{i=1}^{500} q_i \qquad \text{var}[\tilde{T}] = (\sigma^2 + \mu^2) \sum_{i=1}^{500} q_i$$ [2]

The mean is the same but the variance is larger than that obtained in part (i).

(iii) ***MGF***

By definition we have:

$$M_T(t) = E[e^{tT}] = E[e^{t(Y_1+Y_2+\cdots+Y_{500})}] = M_{Y_1}(t)M_{Y_2}(t)\ldots M_{Y_{500}}(t)$$ [½]

From the information given in the question, Y_i is either 0 with probability 0.98 or μ with probability 0.02. We can therefore work out the moment generating function of Y_i:

$$M_Y(t) = E[e^{tY}] = e^{t\times 0} \times 0.98 + e^{t\mu} \times 0.02 = 0.98 + 0.02e^{t\mu}$$ [½]

Substituting this into the expression for the moment generating function for T, we get:

$$M_T(t) = (0.98 + 0.02e^{t\mu})^{500}$$ [½]

This is of the form of the moment generating function for a compound binomial distribution with parameters 500 and 0.02, and claim size distribution that is constant. [½]

(iv) ***Compound Poisson***

By definition we have:

$$M_{\tilde{T}}(t) = E[e^{t\tilde{T}}] = E[e^{t(C_1+C_2+\cdots+C_{500})}] = M_{C_1}(t)M_{C_2}(t)\ldots M_{C_{500}}(t)$$

From the information given in the question, since C_i has a compound Poisson distribution it has moment generating function:

$$M_{C_i}(t) = \exp\left[q_i(M_X(t)-1)\right] = \exp\left[0.02(M_X(t)-1)\right] \qquad [½]$$

The random variable X takes the value μ with probability 1, so:

$$M_X(t) = e^{t\mu}$$

and:

$$M_{C_i}(t) = \exp\left[0.02(e^{t\mu}-1)\right] \qquad [½]$$

Substituting this into the expression for the moment generating function for $\tilde{T}$, we get:

$$M_{\tilde{T}}(t) = \exp\left[0.02(e^{t\mu}-1)\right]\exp\left[0.02(e^{t\mu}-1)\right]\ldots\exp\left[0.02(e^{t\mu}-1)\right]$$

$$= \exp\left[10(e^{t\mu}-1)\right] \qquad [½]$$

This is of the form of the moment generating function for a compound Poisson distribution with parameter 10, and claim size distribution that is constant. [½]

20.7 (i) ***Total number of claimants***

The aggregate amount paid out by the company is $1{,}000S$, where:

$S = Z_1 + \cdots + Z_N$ is the total number of employees claiming benefit in a year

$$Z = \begin{cases} Y & Y < 3 \\ 3 & Y \geq 3 \end{cases}$$

$$Y = X + 1$$

and:

$$X \sim Poisson(2.2)$$

Now:

$$P(Y=1) = P(X=0) = e^{-2.2} \qquad [½]$$

$$P(Y=2) = P(X=1) = 2.2e^{-2.2} \qquad [½]$$

$$P(Y \geq 3) = 1 - e^{-2.2} - 2.2e^{-2.2} = 1 - 3.2e^{-2.2} \qquad [½]$$

So:

$$E(Z) = 1P(Y=1) + 2P(Y=2) + 3P(Y \geq 3)$$

$$= e^{-2.2} + 4.4e^{-2.2} + 3\left(1 - 3.2e^{-2.2}\right) = 3 - 4.2e^{-2.2} = 2.53463 \qquad [1]$$

and:

$$E(Z^2) = 1^2 P(Y=1) + 2^2 P(Y=2) + 3^2 P(Y \geq 3)$$

$$= e^{-2.2} + 8.8e^{-2.2} + 9\left(1 - 3.2e^{-2.2}\right) = 9 - 19e^{-2.2} = 6.89474 \qquad [1]$$

Hence the variance of Z is:

$$\text{var}(Z) = E\left(Z^2\right) - \left(E(Z)\right)^2 = 6.89474 - 2.53463^2 = 0.47041 \qquad [½]$$

To find $E(N)$ and $\text{var}(N)$, we use the fact that N has a Type 2 negative binomial distribution with parameters $p = 0.8$, $q = 0.2$ and $k = 1$. Using the formulae for the moments given on page 9 the *Tables*, we have:

$$E(N) = \frac{kq}{p} = \frac{0.2}{0.8} = 0.25 \qquad [½]$$

and: $$\text{var}(N) = \frac{kq}{p^2} = \frac{0.2}{0.8^2} = 0.3125 \qquad [½]$$

Alternatively, we could derive the moment generating function of N, and then use MGF formulae to derive the mean and variance of N.

The mean and variance of S are:

$$E(S) = E(Z)E(N) = 2.53463 \times 0.25 = 0.63366 \qquad [1]$$

and:

$$\text{var}(S) = \left(E(Z)\right)^2 \text{var}(N) + \text{var}(Z)E(N) = 2.53463^2 \times 0.3125 + \ 0.47041 \times 0.25 = 2.12521 \qquad [1]$$

(ii) ***Mean and variance of the aggregate amount***

So the mean and variance of $1{,}000S$ are:

$$E(1{,}000S) = 1{,}000E(S) = 634 \qquad [½]$$

and:

$$\text{var}(1{,}000S) = 1{,}000^2 \, \text{var}(S) = 2{,}125{,}000 \qquad [½]$$

Alternatively, we could define $S = Z_1 + Z_2 + \cdots + Z_N$*, where:*

$$Z = \begin{cases} 1{,}000Y & \text{if } Y \le 3 \\ 3{,}000 & \text{if } Y > 3 \end{cases}$$

This would give us the aggregate claim amount directly.

20.8 (i) ***Mean, variance and coefficient of skewness***

We have $N_1 \sim Poisson(400 \times \frac{1}{20}) \equiv Poisson(20)$, $X_1 = £3{,}000$, $N_2 \sim Poisson(250 \times \frac{1}{10}) \equiv Poisson(25)$ and:

$$X_2 = \begin{cases} £2{,}000 & \text{with probability } 0.4 \\ £3{,}000 & \text{with probability } 0.6 \end{cases}$$

Working in £000s, we find that:

$$E(X_2) = (2 \times 0.4) + (3 \times 0.6) = 2.6$$

$$E\left(X_2^2\right) = \left(2^2 \times 0.4\right) + \left(3^2 \times 0.6\right) = 7$$

$$E\left(X_2^3\right) = \left(2^3 \times 0.4\right) + \left(3^3 \times 0.6\right) = 19.4 \qquad [1]$$

Let S_i denote the annual aggregate claims from category i. Using the assumption that the policies are independent and the result that, for a compound Poisson random variable T, the k th central moment of T is given by $\lambda E(X^k)$, we obtain:

$$E(S) = E(S_1) + E(S_2) = (20 \times 3) + (25 \times 2.6) = 125 = £125{,}000 \qquad [1]$$

$$\begin{aligned} \text{var}(S) &= \text{var}(S_1) + \text{var}(S_2) \\ &= \lambda_1 E(X_1^2) + \lambda_2 E(X_2^2) \\ &= (20 \times 9) + (25 \times 7) \\ &= 355 \\ &= £^2 355{,}000{,}000 \qquad [1] \end{aligned}$$

$$\begin{aligned} \text{skew}(S) &= \text{skew}(S_1) + \text{skew}(S_2) \\ &= \lambda_1 E(X_1^3) + \lambda_2 E(X_2^3) \\ &= (20 \times 27) + (25 \times 19.4) \\ &= 1{,}025 \\ &= £^3 (1{,}025 \times 10^9) \end{aligned}$$

So the coefficient of skewness is:

$$\frac{\text{skew}(S)}{\left(\text{var}(S)\right)^{3/2}}=\frac{1,025}{355^{3/2}}=0.15324$$ [1]

(ii) ***Calculate Y using a normal approximation***

Assuming that $S \sim N(125, 355)$, we have:

$$P(S>Y)=0.1 \Rightarrow P\left(N(0,1)>\frac{Y-125}{\sqrt{355}}\right)=0.1$$

$$\Rightarrow \frac{Y-125}{\sqrt{355}}=1.2816 \qquad \text{from page 162 } \textit{Tables}$$

$$\Rightarrow Y=149.147$$

So S exceeds £149,000 with a probability of approximately 0.1. [3]

(iii) ***Probability that reinsurer is involved***

The expected value and variance of S are now the same as those of S_1. Working in £000s and assuming that $S \sim N(60, 180)$, we obtain:

$$P(S>100)=P\left(N\left(0,1\right)>\frac{100-60}{\sqrt{180}}\right)\approx 1-\Phi(2.98)=1-0.9986=0.0014$$ [3]

21

Machine learning

Syllabus objectives

5.1 Explain and apply elementary principles of machine learning.

5.1.1 Explain the main branches of machine learning and describe examples of the types of problems typically addressed by machine learning.

5.1.2 Explain and apply high-level concepts relevant to learning from data.

5.1.3 Describe and give examples of key supervised and unsupervised machine learning techniques, explaining the difference between regression and classification and between generative and discriminative models.

5.1.4 Explain in detail and use appropriate software to apply machine learning techniques (*eg* penalised regression and decision trees) to simple problems.

5.1.5 Demonstrate an understanding of the perspective of statisticians, data scientists, and other quantitative researchers from non-actuarial backgrounds.

0 Introduction

The aim of this chapter is to provide an insight into the topic of machine learning. Machine learning is a vast topic and this chapter will only provide a high-level introduction. Specifically, the chapter has the following aims:

- **To provide a high-level knowledge of the various branches of machine learning and examples of their applications, both within general industry and within the specific sectors that actuarial work involves. The level of knowledge targeted is such as will allow you to identify whether any branch of machine learning would be useful in addressing any problem you face.**
- **To provide you with sufficient background information that you can participate in high-level conversations related to projects involving machine learning analyses and their results.**
- **To describe some of the most common machine learning techniques.**
- **To discuss the relationship between machine learning and other branches of data science and statistical analysis, so that you are able to communicate effectively with other quantitative researchers, and to understand the similarities and differences between machine learning and other approaches.**

There are many resources available to students to gain an insight into the key elements of machine learning. One excellent resource is a series of lectures given at Caltech by Yaser Abu-Mostafa which is freely available online at https://work.caltech.edu/telecourse.html.

Another is A. Chalk and C. McMurtrie 'A practical introduction to Machine Learning concepts for actuaries', *Casualty Actuarial Society E-forum,* Spring 2016.

1 What is machine learning?

Machine learning describes a set of methods by which computer algorithms are developed and applied to data to generate information. This information can consist simply of hidden patterns in the data, but often the information is applied to solve a specific problem.

Machine learning methods have become popular in recent years with the advent of increasing quantities of data and the concomitant rapid increase in computing power.

When describing the new age of 'big data' researchers often talk about the three V's: volume, velocity and variety. There are now very large *volumes* of data in use, which computers can process very rapidly (*velocity*) and the data can take many different forms (*variety*).

In order for machine learning to be useful in tackling a problem we need the following to apply:

- **A pattern should exist. If there is no pattern, there is no information to be had, and machine learning will not help (indeed, it might be counterproductive by 'discovering' patterns that do not exist).**
- **The pattern cannot be practically pinned down mathematically by classical methods. If it could be pinned down, we could proceed to describe it mathematically.**
- **We have data relevant to the pattern.**

Examples of problems which are commonly solved in this way include:

- **targeting of advertising at consumers using web sites**
- **location of stock within supermarkets to maximise turnover**
- **forecasting of election results**
- **prediction of which borrowers are most likely to default on a loan.**

Question

Give some other examples where machine learning is used:

(a) in everyday life

(b) in an actuarial / insurance / financial context.

Solution

(a) Other everyday examples include:

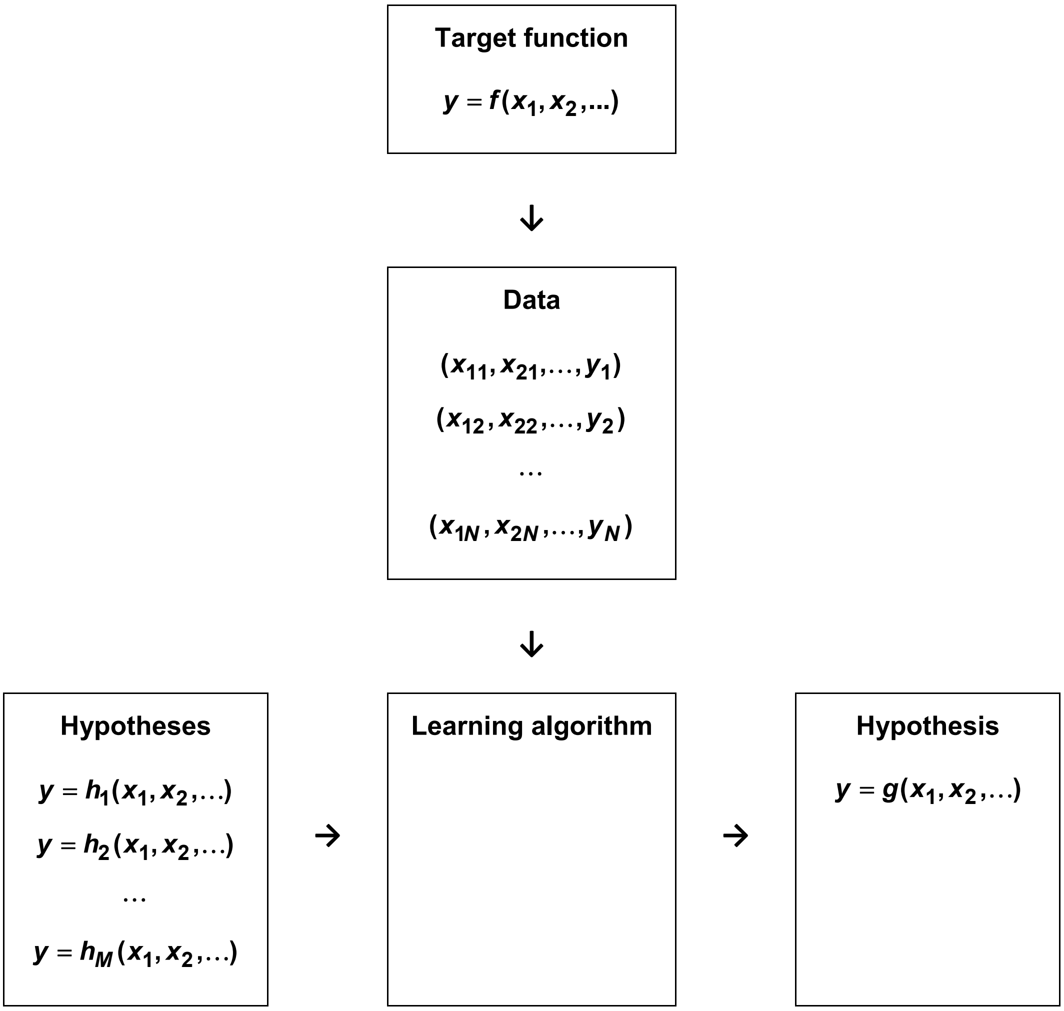
- using an internet search engine such as Google to find relevant web pages
- using a spam filter to identify and remove unwanted emails
- using face recognition to identify known criminals on CCTV footage
- identifying criminals using fingerprints or DNA
- recommending items to purchase on online shopping sites
- matching job applicants to available positions
- suggesting suitable matches on a dating app
- finding relevant historical records on family tree websites
- classifying people in your photos automatically by name
- recognising voice commands, eg in apps such as Siri, Cortana and Alexa
- converting handwriting to text
- translating text from one language to another.

(b) Other examples in an actuarial / insurance / financial context include:

- classifying the risk for motor insurance policyholders using in-car monitoring devices
- identifying marker genes that are associated with particular medical conditions
- identifying insurance claims that might be fraudulent
- identifying fraudulent benefit claims
- identifying fraudulent tax declarations.

2 An overview of machine learning

The diagram below (due to Yaser Abu-Mostafa) provides an overview of the machine learning process.


Target function
$y = f(x_1, x_2, ...)$
Data
$(x_{11}, x_{21}, \ldots, y_1)$
$(x_{12}, x_{22}, \ldots, y_2)$
$\ldots$
$(x_{1N}, x_{2N}, \ldots, y_N)$
Hypotheses
$y = h_1(x_1, x_2, \ldots)$
$y = h_2(x_1, x_2, \ldots)$
$\ldots$
$y = h_M(x_1, x_2, \ldots)$
Learning algorithm
Hypothesis
$y = g(x_1, x_2, \ldots)$

First, there is some target function f, which maps a set of variables, or *features*, that we can measure, onto some output y. (What we term 'variables' or 'covariates' in statistical modelling, machine learning terms 'features').

We will assume that we have identified J features that we consider relevant.

Let the variables, or features, be $x_1, x_2, \ldots, x_j, \ldots, x_J$.

Then we have:

$$y = f(x_1, x_2, \ldots, x_J)$$

The target function is unknown and it is this which we are trying to approximate. The target function might, for example, map life insurance data such as smoking behaviour, lifestyle factors and parental survival to life expectancy.

Second, we have data on y and $x_1, x_2, \ldots, x_J$ for a sample of N individuals.

We use the data to develop a hypothesis which relates the data to the output. Let the hypothesis be:

$$y = g(x_1, x_2, \ldots, x_J)$$

The idea is that $g(x_1, x_2, \ldots)$ should be close to the unknown function $f(x_1, x_2, \ldots)$.

The way the hypothesis $y = g(x_1, x_2, \ldots)$ is chosen is by trying out a large number, say M, of hypotheses $y = h_1(x_1, x_2, \ldots)$, $y = h_2(x_1, x_2, \ldots)$, …, $y = h_M(x_1, x_2, \ldots)$ on the data and using a learning algorithm to choose among them. The hypotheses are usually drawn from a *hypothesis set*, which has a general form.

So, for example, in classical linear modelling the hypothesis set might be the set of linear relationships:

$$y = w_{10} + w_{11}x_1 + w_{12}x_2 + \ldots + w_{1j}x_j + \ldots + w_{1J}x_J$$

$$y = w_{20} + w_{21}x_1 + w_{22}x_2 + \ldots + w_{2j}x_j + \ldots + w_{2J}x_J$$

$$\ldots$$

$$y = w_{m0} + w_{m1}x_1 + w_{m2}x_2 + \ldots + w_{mj}x_j + \ldots + w_{mJ}x_J$$

$$\ldots$$

$$y = w_{M0} + w_{M1}x_1 + w_{M2}x_2 + \ldots + w_{Mj}x_j + \ldots + w_{MJ}x_J$$

where the w_{mj} are weights to be applied to the features. There are M hypotheses, each with a different set of weights.

Linear regression (which is covered in Subject CS1) can be viewed within this framework. The weights are equivalent to regression coefficients and the final hypothesis $y = g(x_1, x_2, \ldots)$ is the set of weights which 'best fits' the data according to some criterion, such as minimising the squared distance between the values of y predicted by the model and the values of y observed in reality. Of course, the linear regression problem is typically solved in 'one step', whereas many machine learning problems are solved iteratively, or in many steps.

3 Concepts in machine learning

An important difference between machine learning and many statistical applications is that the goal of machine learning is to find an algorithm that can predict the outcome y in previously unseen cases.

In the example studied by Chalk and McMurtrie, the task was to predict the cause codes of aviation accidents from the words in brief narratives of the accidents.

This refers to a case study in the paper mentioned in the Introduction.

The cause codes in their example were 'aircraft', 'personnel issues', 'environmental issues' and 'organisational issues'. The idea of classifying insurance claims in this way has wide actuarial applications – for example, in the construction of different pricing models for different types of claim. But if an insurer uses such cause codes, a change in the IT system or in the staff that handle claims could result in claims not being coded or being coded inaccurately.

This is because the existing claims team will have established a 'rule book' with conventions that tell them which is the best code to record the accidents under when the cause is not clear-cut.

Doctors face a similar problem when recording the cause of death on a patient's death certificate. If an elderly patient dies during an operation to treat a cancer they were suffering from, this could be recorded as 'old age', 'cancer' or 'complications during surgery'.

It would be useful to develop a way of using narrative descriptions of claims to add cause codes to those for which codes are not available, so that continuity of coding could be maintained. We might do this by creating an algorithm which uses the claims narratives from data that were cause-coded to work out the cause codes that were given, and then apply this algorithm to claims *that were not cause-coded*.

Question

Suppose that we wish to use a similar system to identify cause codes for motor claims based on the descriptions policyholders write on their claim forms.

(i) List some cause codes that we might wish the system to identify.

(ii) List some keywords that the final algorithm might search for in the inputs.

Solution

(i) The cause codes would probably include the familiar third party damage, fire and theft, as well as others such as own vehicle damage and personal injury.

(ii) There would be a large number of relevant keywords in the policyholders' descriptions, *eg* 'crashed', 'hit', 'stolen', 'ran over', 'brakes', 'suddenly', 'tree', 'dog', 'ice' and 'exploded'.

Some of these keywords map in an obvious way to a cause (*eg* 'stolen' to 'theft') but others (*eg* 'suddenly') are much more open-ended.

A key element of this scenario is that we are going to apply the results of the exercise to data that were not used to develop the algorithm. This means that we are interested in the performance of the algorithm not just in the sample of N cases in our data, but 'out of sample'.

'Out of sample' refers to new data outside our original sample.

This is not always the case in statistical modelling (where we are often content with the model which 'fits' our data best).

For example, when we are graduating a set of crude mortality rates, we usually restrict our results to the same range of ages for which we have data.

3.1 The loss function

One way to evaluate a hypothesis is to calculate the predictions it makes and to penalise each incorrect prediction by some loss. For example, if the prediction involves the classification of something into categories, we could say that each incorrectly classified case incurs a loss of one. We then choose the hypothesis $y = g(x_1, x_2, \ldots)$ by minimising the loss function.

In this example the loss function we're using to penalise errors is the zero-one loss function, which assumes that all errors are equally bad. (The zero-one loss function is covered in the Bayesian statistics chapter in Subject CS1.)

It can be shown that for some common algorithms (such as logistic regression) maximising the likelihood is equivalent to minimising the loss function.

3.2 Model evaluation

When we fit statistical models to data, we have a range of criteria to allow us to choose the 'best' model from among a set of models (as we saw in Subject CS1).

For example, we saw that the parameters of a model can be estimated using the method of maximum likelihood or the method of moments, and that we can test the goodness of fit of a model using a chi-square test.

But evaluating a predictive model involves more than this. Even the 'best' model may not be a very good predictive model. And even if it is good, it might take a very long time to find the correct parameters, or it might be very difficult to interpret (and explain to clients).

Model evaluation therefore involves more than just applying some statistical criteria of 'fit'. We illustrate some possible measures using a model designed for classification.

***Accuracy*. This is the proportion of predictions that the model gets right. Usually we compare this proportion with the proportion predicted by a *naïve classifier* (*eg* a classifier that puts every case into the same category).**

***Precision and recall*. Consider a diagnostic test for a medical condition. The patients who take the test either have the condition or they do not. The test will classify (predict) patients as having the condition or not according to whether the outcome of the test fulfils certain criteria.**

An example of a naïve classifier in this situation is one that automatically says that the patient does not have the condition. This classifier has no ability to distinguish correctly between patients with and without the condition.

We now consider the possible outcomes of the model.

There are four possibilities, shown in the table below. This table is known as a *confusion matrix*.

The terminology is based on the idea that the matrix of possible outcomes quantifies the extent to which the test 'confuses' patients who do / do not have the condition.

		Test result classified / predicts patient as having condition	
		YES	**NO**
Patient actually has condition	**YES**	**True positive (TP)**	**False negative (FN)**
	NO	**False positive (FP)**	**True negative (TN)**

Question

A country is introducing a new screening programme for early identification of people with a particular type of cancer.

(i) Explain what 'false positive' and 'false negative' results would be in this context.

(ii) Discuss the impact of false positives and false negatives from the point of view of a patient.

(iii) State an additional concern regarding false negatives if this had been a test for an infectious disease.

Solution

(i) A false positive is a patient that the test flags as having the disease, but in fact does not.

A false negative is a patient that the test indicates as not having the disease, when in fact they do have it.

(ii) A false positive outcome is undesirable because the patient may be caused unnecessary worry or required to undergo further tests or treatment before it is established that they do not actually have the disease.

A false negative outcome is also undesirable because the patient may not now be identified early enough to receive effective treatment for the disease.

(iii) With an infectious disease, there is the additional concern that false negative patients may unknowingly spread the disease to other people.

There are several useful measures we can calculate from the confusion matrix to gauge the effectiveness of the test.

***Precision* is the percentage of cases classified as positive that are, in fact, positive. Using the abbreviations in the table this is:**

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP}+\textbf{FP}}$$

Ideally we would like to have $FP = 0$, which would result in a precision of 1, *ie* 100%.

***Recall* is the percentage of positives that we managed to identify** (correctly):

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP}+\textbf{FN}}$$

This measure is also called the *sensitivity*. If $FN = 0$, *ie* if the test has not missed anyone who has the condition, it will equal 100%.

These can be combined in a single measure known as the F_1 *score*:

$$F_1 \textbf{ score} = \frac{2\times\textbf{Precision}\times\textbf{Recall}}{\textbf{Precision}+\textbf{Recall}}$$

The '*F*' here arose historically and doesn't actually stand for anything. The '1' subscript just identifies this measure out of several similar measures that have also been proposed and could be used instead.

Question

(i) Derive and simplify a formula for the harmonic mean of the precision and the recall.

Hint: The harmonic mean of a set of values is the reciprocal of the arithmetic mean of their reciprocals.

(ii) Comment on the answer in (i).

Solution

(i) Using the hint given, we can see that the harmonic mean H of the precision and the recall can be found from the equation:

$$\frac{1}{H} = \frac{1}{2}\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)$$

So: $$\frac{1}{H} = \frac{1}{2}\left(\frac{\text{Recall}+\text{Precision}}{\text{Precision}\times\text{Recall}}\right) \Rightarrow H = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Recall}+\text{Precision}} = F_1$$

(ii) F_1 is the harmonic mean of the precision and the recall. This is different from the more familiar *arithmetic* mean, but it also gives an average value of the two measures taken together and results in a value in the same range, *ie* 0 to 1.

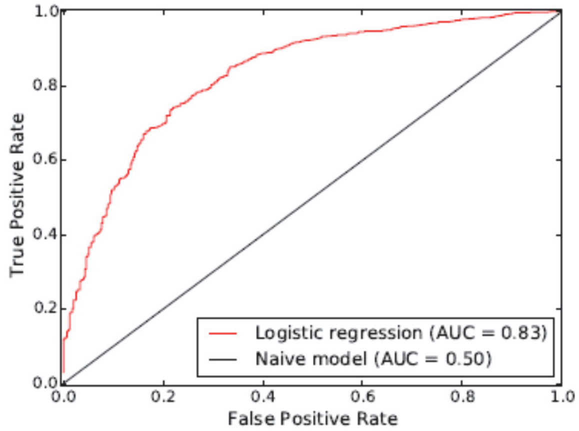
Another measure is the *false positive rate*.

There is a trade-off between the recall (the true positive rate) and the false positive rate (the percentage of cases which are not positives, but which are classified as such). The false positive rate is:

$$\textbf{False positive rate} = \frac{\mathbf{FP}}{\mathbf{TN}+\mathbf{FP}}$$

This is not the same as 1 – the true positive (*ie* the recall) rate.

***Receiver operating characteristic curve*. The trade-off between recall and the false positive rate can be illustrated using a *receiver operating characteristic* (ROC) curve. An example is shown below, taken from Alan Chalk and Conan McMurtrie 'A practical introduction to Machine Learning concepts for actuaries' *Casualty Actuarial Society E-forum,* Spring 2016.**



This figure compares the ROC curves for a logistic regression model fitted to the cause codes for aircraft accidents with a naïve model based on random guesswork. The area under the ROC provides another single-figure measure of the efficacy of the model. The further away from the diagonal is the ROC, the greater the area under the curve and the better the model is at correctly classifying the cases.

This type of graph is most useful when the test involves a threshold of some kind. For example, a medical test may involve measuring the concentration of a particular chemical in the patient's blood and labelling the patient as positive if this exceeds a particular level. We can then calculate the false positive rate (defined above) and the true positive rate (*ie* the recall rate) for different levels of the threshold and plot these on a graph.

Points near the top left of the graph correspond to a good test where the true positive rate is high and the false positive rate is low. The diagonal line corresponds to a neutral 'zero-sum' test where there is a simple trade-off with any improvement in the true positive rate being matched by an equal deterioration in the false positive rate. The area of the triangle below the diagonal is 0.5 and the area of the whole rectangle is 1 (the maximum possible score for the ROC).

Question

Two different tests have been applied to a sample of 100 individuals to identify whether or not a particular feature is present. These resulted in the following confusion matrices:

(a) **Test 1**

PREDICTED ACTUAL	YES	NO	TOTAL
YES	TP = 76	FN = 4	80
NO	FP = 4	TN = 16	20
TOTAL	80	20	100

(b) **Test 2**

PREDICTED ACTUAL	YES	NO	TOTAL
YES	TP = 5	FN = 5	10
NO	FP = 15	TN = 75	90
TOTAL	20	80	100

Calculate the precision, recall, F_1 score and false positive rate for each matrix and comment on the answers.

Solution

(a)

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{76}{76+4} = 95\%, \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{76}{76+4} = 95\%$$

$$F_1 \text{ score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2\times 95\%\times 95\%}{95\%+95\%} = 95\%$$

$$\text{False positive rate} = \frac{\text{FP}}{\text{TN}+\text{FP}} = \frac{4}{16+4} = 20\%$$

(b)

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{5}{5+15} = 25\%, \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{5}{5+5} = 50\%$$

$$F_1 \text{ score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2\times 25\%\times 50\%}{25\%+50\%} = 33\%$$

$$\text{False positive rate} = \frac{\text{FP}}{\text{TN}+\text{FP}} = \frac{15}{75+15} = 17\%$$

The precision values show that Test 1 is much more effective at correctly identifying individuals who do have the feature.

The recall values also show that Test 1 is much better at identifying individuals who do have the feature.

The F_1 scores show that the overall performance of Test 1 is much better than Test 2.

The false positive rates are quite low for both tests, indicating that only a small proportion of individuals who do not have the feature are incorrectly flagged as having it.

3.3 Generalisation error and model validation

The methods described in the previous section allow the assessment of model performance on existing data. But how can we assess the likely predictive performance of the model? Can we be sure that we can use machine learning to test numerous hypotheses and eventually pick one which will generalise acceptably to new data? The answer is that we can in theory (see Lectures 4-6 of Yaser Abu-Mostafa's course for a demonstration and proof of this). Specifically, we can show that if the in-sample error is $E_{in}(g)$ and the out-of-sample error is $E_{out}(g)$, then:

$$P\left[\left|E_{\text{in}}(g)-E_{\text{out}}(g)\right|>\varepsilon\right]\leq 4\left[H(N)\right]e^{-\frac{1}{8}\varepsilon^2 N}$$

where:

- **N is the sample size**
- **ε is some specified tolerance**
- **$H(N)$ is a polynomial in N which depends on the hypothesis set.**

This equation, called the Vapnik-Chervonenkis inequality, shows that, for large enough N, it will always be possible to use learning to choose a hypothesis g which will make the tolerance as small as we like.

In this inequality, $E_{in}(g)$ and $E_{out}(g)$ are some suitable measure of the error in the results, *eg* the average difference between the predicted and true values, or the proportion of records that are classified incorrectly. In a good model the difference between these two quantities will be small.

For large values of N the exponential factor $e^{-\frac{1}{8}\varepsilon^2 N}$ on the RHS of the inequality will dominate the polynomial $H(N)$. So the RHS provides an upper bound that tends to zero as we increase the sample size N. This means that we can make the probability of an error of a given size as small as we like by using a big enough sample size for the training set.

This may be true in theory, but how do we test the performance of our model out-of-sample?

3.4 Train-validation-test

The conventional approach in machine learning is to divide the data into two. One part of the data (usually the majority) is used to *train* the algorithm to choose the 'best' hypothesis from among the M competing ones. The other is used to *test* the chosen hypothesis g on data that the algorithm has not seen before.

In practice, the 'training' data is often split into a part used to estimate the parameters of the model, and a part used to validate the model.

This approach is often called the train-validation-test approach. It involves three data sets:

- **a *training data set*: the sample of data used to fit the model**
- **a *validation data set*: the sample of data used to provide an unbiased evaluation of model fit on the training dataset while tuning model hyper-parameters (see below)**
- **a *test data set*: the sample of data used to provide an unbiased evaluation of the final model fit on the training data set.**

3.5 Parameters and hyper-parameters

In statistical analysis, we often fit models to data, for example regression models such as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_J x_{Ji} + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim Normal(0, \sigma^2)$$

Here the β's and σ^2 are the *parameters* of the model. Most supervised machine learning algorithms involve models with similar parameters. The 'best' values for these parameters are estimated from the data.

We will explain the difference between supervised and unsupervised learning when we look at the different branches of machine learning in Section 4.

***Parameters* are required by the model when making predictions. They define the skill of the model when applied to your problem and they are estimated or learned from the data. They form an integral part of the learned model.**

We can say that the *parameters* of a model are variables internal to the model whose values are estimated from the data and are used to calculate predictions using the model.

Question

The number of road accidents each day is to be modelled using a linear model based on the average number of cars on the road each day. Identify the parameters in this model and describe the role they play.

Solution

If we let x denote the average number of cars on the road on a given day and y denote the number of road accidents the same day, then a linear model would take the form:

$$y = \alpha + \beta x + \varepsilon$$

In this equation, ε is an error term with mean 0, and α and β are the parameters. The values of α and β would be estimated from the data in the usual way, *ie* $\hat{\beta} = s_{xy} / s_{xx}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ (see page 24 of the *Tables*). The model incorporating these estimated values of the parameters would then be used to estimate values of y in the future, based on the value of x on that day, using the formula $\hat{y} = \hat{\alpha} + \hat{\beta}x$.

Machine learning algorithms, both supervised and unsupervised, however, also have higher-level attributes which must also be estimated or (in some sense) optimised. These might include:

- **the number of covariates J to include in a regression model**
- **the number of categories in a classification exercise**
- **the rate at which the model should learn from the data.**

We can speed up the learning process by allowing the machine to make quite big changes at each stage, based on the data it has just processed. However, this may result in the final model 'overshooting' the optimal solution or overfitting the model to the particular training set used.

A slower rate of learning will be more likely to come up with a good solution, but this will take longer to achieve.

These attributes are caller *hyper-parameters*. They cannot be estimated from the data – indeed they must often be defined before an algorithm can be implemented. Hyper-parameters are external to the model and their values cannot be estimated from the data. They are typically specified by the practitioner and may be set using heuristic guidelines. Nevertheless, they are critical to the predictive success of a model.

Hyper-parameters are variables external to the model whose values are set in advance by the user. They are chosen based on the user's knowledge and experience in order to produce a model that works well.

'Heuristic' means that there are no hard and fast rules for these. They are determined using rough guidelines and past experience of what works well, combined with experimentation.

Question

Give some other examples of hyper-parameters from other models that may be relevant to actuaries.

Solution

Life insurance

If we are calculating premiums for life insurance policies, we need to decide exactly how to define smoker categories such as Non-smoker, Light smoker and Heavy smoker, *eg* a person who smokes more than 20 cigarettes in an average day may be classed as a heavy smoker.

Graduation

If we are graduating mortality data using a Gompertz-Makeham formula $\mu_x = p_1(t) + \exp[p_2(t)]$, where $p_1(t) = \sum_{k=0}^{r-1} a_k t^k$ and $p_2(t) = \sum_{k=0}^{s-1} b_k t^k$ are polynomial functions (see page 32 of the *Tables*), we need to decide on the values to use for r and s, which determine the order of the two polynomials.

Time series

If we are fitting a linear time series using an ARIMA model, we need to decide on the values of d, p and q, which determine the number of levels of differencing to apply and the number of moving average and autoregressive terms to include.

GLM

If we are applying a generalised linear model, we need to decide on the form of the link function to use.

Reinsurance

If we are modelling large claims in general insurance, we need to specify the cut-off point for a claim to count as 'large'.

Motor insurance

If we are using geographical area as a rating factor in motor insurance, we need to decide on how many areas to use and which locations these cover.

Health

If we are using a patient's body mass index (BMI) as a predictor for the outcome of a medical procedure, we may need to specify the dividing lines between weight bands such as underweight (<20), normal (20–25), overweight (25–30) and obese (30+).

3.6 Validation and over-fitting

In a normal linear regression model, as we include more variables, the proportion of the variance in the dependent variable that is explained cannot decrease. A model with more variables will, in that sense, 'fit' the data better than one with fewer variables. The same is true with machine learning models, but the number of parameters in machine learning models can be very large.

There is a risk that, if the number of parameters / features is large, the estimates of the parameters in the model g that is chosen will reflect idiosyncratic characteristics of the specific data set we have used to 'train' the model, rather than the underlying relationships between the output, y, and the features $x_1, x_2, \dots, x_J$. This is known as *over-fitting* and is one of the biggest dangers faced by machine learning. Over-fitting leads to the identification of patterns that are not really there. More precisely, it leads to the identification of patterns that are specific to the training data and do not generalise to other data sets.

We saw this same issue when we considered the graduation of mortality rates by parametric formula in Chapter 11. If we use a formula with too many parameters, the graduated rates will be undergraduated. They will follow the crude rates too closely, reflecting a lot of the random 'noise' present in the data, rather than just capturing the underlying pattern of the rates.

On the other hand, if the number of parameters / features is small, we might miss important underlying relationships.

We saw this issue too with graduation by parametric formula. If we use a formula with too few parameters, the graduated rates will be overgraduated. They will be smoothed too much and will not follow the underlying pattern of the rates closely enough, *eg* 'smoothing over' genuine features such as the teenage accident hump.

So there is a trade-off here, between bias – the lack of fit of the model to the training data – and *variance* – the tendency for the estimated parameters to reflect the specific data we use for training the model.

One way to assess how the predictive ability of the model changes as the number of parameters / features increases is to withhold a portion of the 'training' data and use it to *validate* models with different numbers of parameters / features J. One approach is to divide the training data into, say, s slices, and to 'train' the model s times, using a different slice for validation each time. This is called *s*-fold *cross-validation*.

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In s-fold cross-validation, the original sample is randomly partitioned into s equal size subsamples.

One of the s subsamples is retained as the validation data for *testing* the model. The remaining $s-1$ subsamples are used as *training* data. The cross-validation process is then repeated s times (these are the 'folds'), with each of the s subsamples used exactly once as the validation data. The s results from the folds can then be compared (or averaged to produce a single prediction).

The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Typically, the error on the training data used to estimate the parameters decreases as J increases. But the prediction error on the validation data often decreases as J increases for small J, and reaches a minimum before increasing again as J gets larger. This suggests that models with a number of parameters / features close to the minimum might be most suitable and perform best out-of-sample.

So we might be able to get a better fit by adding extra parameters, but this won't necessarily make the predictions from the model any better when we apply it to new data.

3.7 Regularisation

How can we achieve a good balance between bias and variance? Put another way, is there a method that can use all the features to choose the final hypothesis g, but will prevent it becoming too complex so that generalisation is poor? There is, and it is called *regularisation* or *penalisation*. This approach exacts a penalty for having too many parameters. Recall that finding the 'best' values of the parameters, or feature weights, w_j in a machine learning problem involves minimising a loss function. Let the loss function be $L^*(w_1,w_2,\ldots,w_J)$. Then the hypothesis g will be chosen to be the hypothesis with a set of weights which minimises $L^*(w_1,w_2,\ldots,w_J)$.

The idea of regularisation, or penalisation, is to add to L^* a cost for model complexity. One possibility is to add a term equal to $\lambda\sum_{j=1}^{J}w_j^2$, so that we now minimise the expression:

$$L^*(w_1,w_2,\ldots,w_J)+\lambda\sum_{j=1}^{J}w_j^2$$

As noted earlier, since minimising the loss function is, in some models, equivalent to maximising the likelihood, minimising this expression is equivalent to maximising a penalised likelihood.

4 Branches of machine learning

Machine learning techniques can be divided into several branches, which we can refer to as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The difference between these lies not (as one might think) in the level of involvement of the human researcher in the development of the algorithm, or in the supervision of the machine. Instead, it lies in the extent to which the machine is given an instruction as to the end-point (or target) of the analysis.

The targets are sometimes referred to as *labels*.

Supervised and unsupervised learning

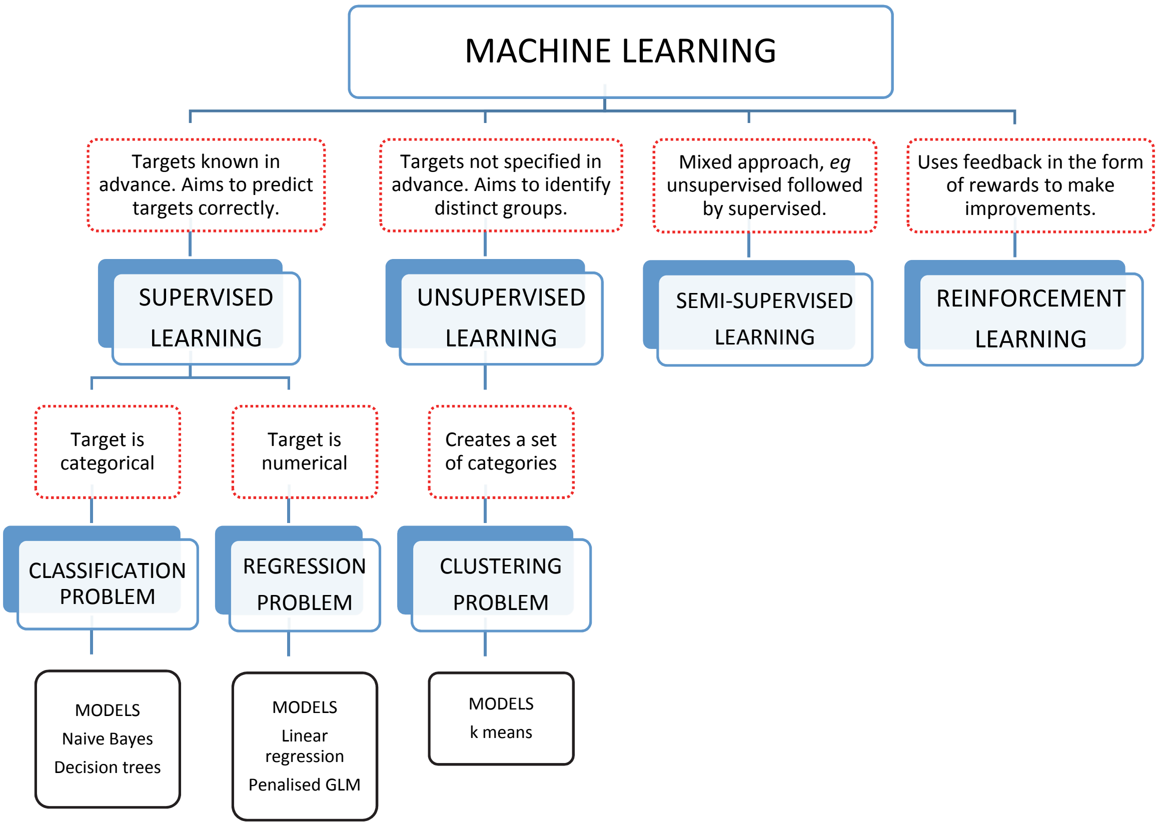
With *supervised learning*, the algorithm is given a set of specific targets to aim for.

With *unsupervised learning*, the algorithm aims to produce a set of suitable labels (*ie* targets).

With *semi-supervised learning*, the algorithm uses a combination of supervised and unsupervised methods.

With *reinforcement learning*, the algorithm aims to improve its performance through trial and error, using a rewards (or penalties) approach.

The diagram below shows the differences between the main branches of machine learning and the models we will consider here.


MACHINE LEARNING
Targets known in advance. Aims to predict targets correctly.
Targets not specified in advance. Aims to identify distinct groups.
Mixed approach, *eg* unsupervised followed by supervised.
Uses feedback in the form of rewards to make improvements.
SUPERVISED LEARNING
UNSUPERVISED LEARNING
SEMI-SUPERVISED LEARNING
REINFORCEMENT LEARNING
Target is categorical
Target is numerical
Creates a set of categories
CLASSIFICATION PROBLEM
REGRESSION PROBLEM
CLUSTERING PROBLEM
MODELS Naive Bayes Decision trees
MODELS Linear regression Penalised GLM
MODELS k means

Question

(i) Give examples of problems that would come under the headings of classification, regression and clustering.

(ii) Give examples of problems that could be solved using semi-supervised and reinforcement learning.

Solution

(i) ***Classification, regression and clustering***

An example of a classification problem is a spam filter that classifies emails into the two categories ‘Safe’ or ‘Suspicious’.

An example of a regression problem is a health awareness app that predicts the user’s life expectancy.

An example of a clustering problem is a system that groups together postcode areas that tend to have a similar experience of insurance claims.

(ii) ***Semi-supervised and reinforcement learning***

An example of semi-supervised learning is a photo app that groups photos featuring people with a similar appearance and then allows the user to name the people in order to add their names automatically to new photos

An example of reinforcement learning is a voice recognition app that adapts over time to the user’s voice.

4.1 Supervised learning

Supervised learning is associated with predictive models in which the output is specified. Here the machine is given a specific aim (*eg* to use the variables in the data to develop a model to predict whether a person will default on a loan), and the algorithm will try to converge on the parameters of the model which provide the ‘best’ prediction.

Examples relevant to the actuarial profession might be:

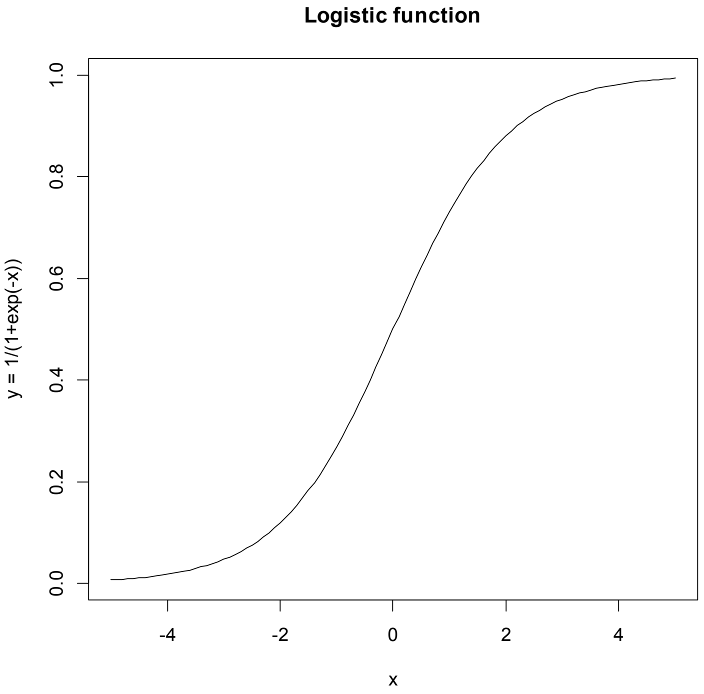
- **the prediction of future lifetime at age x, T_x, or survival probabilities from age x, $P(T_x > t)$**
- **the prediction of the risk of claims being made on certain classes of insurance.**

A distinction can be made between supervised learning that involves the prediction of a numerical value (such as future lifetime) and prediction of which category a case falls into (will a person default on a loan – yes or no?). For predicting numerical values, regression models are the normal approach, whereas predicting which category a case falls into is essentially a classification problem, and different algorithms, such as decision trees, are used. However, this distinction between regression and classification is somewhat fuzzy, as there are regression models, such as logistic regression or probit models, where the dependent variable is categorical. (These are examples of generalised linear models, which were covered in Subject CS1.)

Probit models (short for 'probability unit') produce outputs that can only take one of two values, *eg* Yes / No or 0 / 1.

Logistic regression is based on the logistic function $f(x)=\dfrac{1}{1+e^{-x}}$, shown in the graph below.

This function converts an input value, which can be anywhere in the entire range $-\infty<x<\infty$, to an output value on a continuous scale between 0 and 1. If we interpret the output value as a probability, we can convert it to a categorical output by saying that values exceeding a specified value p (*eg* $p=0.5$) correspond to Yes, while smaller values correspond to No.



This graph was plotted using the R code:

```
logistic<-function(x)1/(1+exp(-x))
plot(logistic,xlim=c(-5,5),main="Logistic function",
ylab="y = 1/(1+exp(-x))")
```

Question

Give another example where classification is achieved by converting a numerical output into a categorical value.

Solution

A familiar example is where a student's numerical score on an exam, *eg* 65%, is converted into an exam grade, *eg* Pass, Fail, *etc*.

Within classification algorithms, a distinction can be made between models that generate classifications and those that discriminate between classes.

For example, the first type of model would include an algorithm that identifies people in a user's photo albums in a photo app, while the second type would include one that 'tags' a user's 'friends' in a social media app. In the first case, the algorithm groups together faces that look similar without knowing in advance the names of the people or how many there will be, whereas in the second case, it just has to match the faces to a list of other users whose names are already known.

Consider the case where we have a categorical output value y and data (covariates), $x_1, x_2, \dots$. The aim is to predict into which category of y case i will fall given the values of the covariates for case i, $x_{1i}, x_{2i}, \dots$.

For example, we might have a set of historical texts written by different authors (y) and we wish to identify the most likely author based on the frequency of certain words in the text ($x_1, x_2, \dots$).

One approach is to model the joint probabilities $P(x_1, x_2, \dots, y)$. This generates a classification scheme. It is then possible to evaluate the conditional probability of being in category y, given $x_1, x_2, \dots$ as:

$$P(y \mid x_1, x_2, \ldots) = \frac{P(x_1, x_2, \ldots, y)}{P(x_1, x_2, \ldots)}$$

One problem with this approach is that the number of separate probabilities $P(x_1, x_2, \dots, y)$ to be computed increases exponentially with the number of covariates x_j.

For example, if we were considering ten words, which can each be recorded as frequent (=1) or infrequent (=0), we would already have $2^{10} = 1{,}024$ different combinations of the form $x_1 = 1, x_2 = 0, \dots, x_{10} = 1$.

This, however, can be overcome by assuming that, given the classes y, the covariates x_j $(j = 1, \dots, J)$ are independent.

We would then only have to consider the ten different probabilities for $x_1 = 1$, $x_2 = 0$, …, $x_{10} = 1$ individually.

With this assumption, we have:

$$P(x_1, x_2, ..., y) = P(y)\prod_{j=1}^{J} P(x_j \mid y)$$

This is called the *naïve Bayes classifier*.

An alternative method is to model the conditional probability $P(y \mid x_1, x_2, ...)$ directly, and to find, say, a linear combination of the x_k that best discriminates between the categories of y. This is the aim of a method known as *discriminant analysis*, which is effectively the same as *logistic regression*.

Other supervised learning techniques described in machine learning textbooks include the perceptron, neural networks and support vector machines.

Perceptrons and *neural networks* use interconnected layers of artificial neurons that can be activated or deactivated in a way that mimics the behaviour of the neurons in animal brains.

A *support vector machine* is a classification algorithm that considers the input data values as a vector defining a point in space and tries to place hyperplanes in a way that segregates the points most effectively.

4.2 Unsupervised learning

Other machine learning techniques operate without a target for the algorithm to aim at. We might, for example, set the machine the task of identifying clusters within the data.

Given a set of covariates, the idea is that the machine should try to find groups of cases which are similar to one another but different from cases in other groups. In the language we used in the exposed to risk chapter, we try to divide the data into homogeneous classes. However, we may not tell the machine in advance what the characteristics of each of these classes should be, or even how many such classes there should be. We allow the machine to determine these given a set of rules which form part of the algorithm. Machine learning where the output is not specified in advance is called unsupervised learning.

Examples of unsupervised learning techniques include *cluster analysis*, and the use of association rules such as the *apriori algorithm*.

The *apriori algorithm* is a machine learning technique that identifies combinations of data values that frequently occur together in a data set, *eg* where users of a music website will tend to download items by the same artist or items of the same genre. It can be used by online retailers as the basis for the 'Other customers also bought …' recommendations or for promoting bundles of items that are frequently bought together.

Apart from their use to divide data into homogeneous classes, unsupervised learning techniques are commonly used with very large data sets. Example would be market basket analysis, which uses data generated from retail transactions to identify items which are commonly purchased together, and text analysis.

For example, the predictive text feature on a mobile phone looks for combinations of words that are commonly used together so that it can auto-complete phrases such as 'Have a happy …' with the word 'birthday'.

4.3 Semi-supervised learning

It is possible to perform machine learning analysis by using a mixture of supervised and unsupervised learning. For example, cluster analysis could be used to identify clusters. These clusters could then be labelled using a variable y, and a supervised classification algorithm such as naïve Bayes or logistic regression used to develop predictions of the class into which each case would fall.

For example, a system that aims to identify pickpockets operating in a busy shopping street might first identify people who appear several times throughout the day with the same clothing. These people's faces could then be matched against a database of known offenders.

This makes obvious sense if the clusters identified by the unsupervised learner make substantive sense for the problem at hand. But even if your clusters do not make sense to you (a human), you will have constructed a machine called an *autoencoder* – which can considerably speed up any future modelling analysis.

An *autoencoder* compresses the raw data by focusing on features that appear to be significant, *eg* it might identify the different types of object that appear in a photo, even though it doesn't know what they actually are.

4.4 Reinforcement learning

In *reinforcement learning* the learner is not given a target output in the same way as with *supervised* learning. The learner uses the input data to choose some output, and is then told how well it is doing, or how close the chosen output is to the desired output. The learner can then use this information as well as the input data to choose another hypothesis.

Example

Imagine a world that can be modelled as a finite-state discrete-time stochastic process with state space S. An agent in this world who is in state u at time t can take many possible actions, A_l, and each of these actions will result in a probability that the agent is in state v at time $t+1$. We can define two functions:

- **the state transition function, $P(X_{t+1}=v \mid X_t=u, A_l)$, and**
- **the observation, or output function $P(Y \mid X_t=u, A_l)$.**

An example to have in mind here would be a fund manager who is investing the assets of a pension fund. Here state u might be the value of the assets at the time (t, $t+1$, *etc*) of an annual valuation, A_l might be a particular investment strategy that the manager could follow and Y might be some measure of the fund's solvency (which will be affected by unknown factors on the liability side).

Some values of Y are more desirable than others, and we want the agent to take the actions which will lead to desirable outcomes of Y. How do we achieve this? The agent does not know the future, and cannot necessarily see how the actions taken at time t will enhance or reduce the probabilities of Y.

One possibility is to define a reward function $E(R_t \mid X_t = i, A_l)$, in which the reward, R_t depends on the probability that the action A_l will lead to desirable values of Y. The agent then tries to maximise its overall rewards (discounted as appropriate). Clearly, if the agent had full information about the model, we could treat this is a standard maximisation problem. But the agent does not know this: all the agent knows is the rewards it received for particular actions at specific time points up to the present. Reinforcement learning is the process by which the agent updates the probabilities of taking particular actions on the basis of past rewards received.

This process is reminiscent of the idea of the *actuarial control cycle* where adjustments are made periodically based on feedback from past experience to ensure that a particular strategy remains on track.

Another popular machine learning technique that involves reinforcement is *genetic algorithms*, which are based on the idea of selective breeding from biology. These create successive generations of possible solutions. Small random variations ('mutations') are introduced into each solution to create the next generation. The solutions are then tested and the ones that perform best are selected to continue to the next generation. After a few generations, a good solution may have been discovered amongst the surviving solutions.

5 Stages of analysis in machine learning

In this section we will look at the stages involved in applying machine learning and discuss some of the issues these raise.

Machine learning tasks can be broken down into a series of steps.

5.1 Collecting data

The data must be assembled in a form suitable for analysis using computers. Several different tools are useful for achieving this: a spreadsheet may be used, or a database such as Microsoft Access.

Data may come from a variety of sources, including sample surveys, population censuses, company administration systems, databases constructed for specific purposes (such as the Human Mortality Database, www.mortality.org).

During the last 20 to 30 years the size of datasets available for analysis by actuaries and other researchers has increased enormously. Datasets, such as those on purchasing behaviour collected by supermarkets, relate to millions of transactions.

5.2 Types of data

There are many different types of data we might need to deal with. The table below illustrates the 'traditional' types of data that have been used by actuaries and statisticians.

DATA TYPES				
NUMERICAL (*ie* numbers)		CATEGORICAL (*ie* not numbers)		
DISCRETE	CONTINUOUS	ATTRIBUTE (DICHOTOMOUS)	NOMINAL	ORDINAL
↓	↓	↓	↓	↓
Age last birthday	Exact age	Alive / Dead	Customer name	Date of birth (DD/MM/YY)
Number of children	Salary	Male / Female	Type of claim	Month (Jan, Feb, Mar, …)
Number of claims	Claim amount	Claim / No claim	Occupation	Exam grade (A, B, C, …)
		Pass / Fail	Marital status	Size (S, M, L, XL)
			Country	Agree/Don't know/Disagree
			Colour of car	

Attribute (or *dichotomous*) data refers to variables whose values consist of just two categories.

Ordinal variables take values that can be ordered in a natural way, whereas the values for *nominal* variables cannot.

Some variables can be classed in several ways, *eg* the number of claims could be treated as a continuous variable, rather than discrete, if the values were large. We've seen this idea before when we used a normal approximation to a binomial or Poisson distribution in Subject CS1.

Similarly, colour of car could be recorded as red, orange, yellow, *etc* (*ie* nominal data) or (if it came from a video image) it could be measured on the RGB scale (*ie* a vector of three discrete numerical values). It's also quite common to record attribute data as 1's and 0's (*ie* with discrete numerical values) to make it easier to count subsets and to calculate proportions and averages.

Nowadays, however, there are many other types of data that don't correspond to these familiar types. For example, a motor insurer might be provided with a memory card containing footage of an accident recorded on a vehicle's dash cam (dashboard camera). This will typically be a very large file containing a mixture of video and audio information, as well as structural information (*eg* markers for the start and end of each frame of the video), header information (*eg* the date it was captured, the serial number of the camera and the software version) and other embedded information such as time markers for the footage and satellite coordinates.

5.3 Exploring and preparing the data

Some forms of 'raw' data or complex data structures first need to be converted to one of the types in the table. For example, a Word .docx file, although actually just a long sequence of 1's and 0's, has a specific structure from which the page layouts, the fonts and the text itself would need to be extracted before we could analyse the content. Similarly, we would need to prepare a raw audio or video file before we could use the data it contains for speech recognition or face recognition.

This stage can be divided into several elements:

- **The data need to be prepared in such a way that a computer is able to access it and apply a range of algorithms. If the data are already in a spreadsheet, this may be a simple matter of importing the data into whatever computer package is being used to develop the algorithms. If the data are stored in complex file formats, it will be useful to convert the data to rectangular format, with one line per case and one column per variable. It is also important here to recognise the nature of the variables being analysed: are they nominal, ordinal or continuous?**
- **Cleaning the data, replacing missing values, and checking the data for obvious errors is an important stage of any analysis, including machine learning.**
- **Exploratory data analysis (EDA). In machine learning applications it is probably not a good idea to do extensive EDA, as the outcome might influence your choice of model and hypothesis set.**

Question

Suggest how the raw data could be converted to a usable 'rectangular' format in each of the following cases:

(a) identifying people in a photograph stored as an image file

(b) identifying vehicles from CCTV video footage.

Solution

(a) We could create a separate record for each identified face in the picture and then record columns with the coordinates (or relative positions) of key features such as the centres of the eyes or the corners of the mouth. If the images are in colour, we could also include columns for eye colour, hair colour and complexion.

(b) We could create a separate record for each vehicle with a group of columns for each character in the vehicle number plate to record distinctive features such as whether the character contains a vertical line and which corners contain white space. If the video footage is in colour, we could also record the colour of the vehicle, which we could use as a check.

These data processing tasks are themselves non-trivial and form an important part of the classification process. What we have described here is just one possible approach.

5.4 Feature scaling

Some machine learning techniques will only work effectively if the variables are of similar *scale*. We can see this by recalling that, in a linear regression model (which we covered in Subject CS1) the parameter, β_j, associated with covariate x_j, measures the impact on *y* of a one-unit change in x_j. If x_j is measured in, say, metres, the value of β_j will be 100 times larger than it would be with the same data if x_j were measured in centimetres.

In machine learning the weights w_j play the role of the β's in the linear regression model. Consider the expression in Section 3.7:

$$L^*(w_1, w_2, \ldots, w_J) + \lambda \sum_{j=1}^{J} w_j^2$$

The penalty imposed for model complexity is $\lambda \sum_{j=1}^{J} w_j^2$ which clearly depends on the weights and hence on the scale at which the features are measured.

The scaling of the variables is particularly important with nearest neighbour algorithms, which we will discuss later. The relative distances between the points will be heavily dependent on the units of measurement. To ensure that all the factors are taken into account, we would want the ranges of values for each variable to be comparable.

Descriptive statistics such as frequency distributions, measures of central tendency and of dispersion might be useful here to establish an appropriate scale for each feature, as will cross tabulations of nominal or ordinal data, or correlation coefficients between continuous variables. Pictorial representations, such as histograms and boxplots, are invaluable.

5.5 Splitting the table into the training, validation and training data sets

We discussed the 'train-validate-test' approach in Section 3.4, which involves splitting the data into three parts.

A typical split might be to use 60% of the data for training, 20% for validation and 20% for testing. However it depends on the problem and not on the data. A guide might be to select enough data for the validation data set and the testing data set so that the validation and testing processes can function, and to allocate the rest of the data to the training data set. In practice, this often leads to around a 60% / 20% / 20% split.

5.6 Training a model on the data

This involves choosing a suitable machine learning algorithm using a subset of the data. The algorithm will typically represent the data as a model and the model will have parameters which need to be estimated from the data. This stage is analogous to the process of fitting a model to data as described in the chapters on regression and generalised linear models in Subject CS1.

5.7 Validation and testing

The model should then be validated using the 20% of the data set aside for this purpose. This should indicate, for example, whether we are at risk of over-fitting our data. The results of the validation exercise may mean that further training is required.

Once the model has been trained on a set of data, its performance should be evaluated. How this is done may depend on the purpose of the analysis. If the aim is prediction, then one obvious approach is to test the model on a set of data different from the one used for development. If the aim is to identify hidden patterns in the data, other measures of performance may be needed.

5.8 Improving model performance

We can measure the performance of the model by testing it on the 20% of the data we have reserved for this purpose. The hope is that the performance of the final hypothesis g on the 'test' data set is similar to that achieved by the same hypothesis on the training data set. This amounts to stating that the difference between the in-sample error and the out-of-sample error $|E_{\text{in}}(g) - E_{\text{out}}(g)|$ will be generally small, or that:

$$P\left[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \varepsilon\right] \leq Z$$

where Z is some threshold which may depend on the precise task to hand (the greater the value at risk, the smaller Z).

This is the same type of inequality that we saw in Section 3.3. 'Value at risk' here just means the amount of money at stake, *eg* the cost of the damage that would be caused by a wrong decision. However, the idea is similar to the value at risk (VaR) from investment theory (covered in Subject CM2), where we aim to keep the probability of a large loss below a specified probability level.

If the performance of the model is not sufficient for the task at hand, it may be possible to improve its performance. Sometimes the combination of several different algorithms applied to the same data set will produce a performance which is substantially better than any individual model. In other cases, the use of more data might provide a boost to performance. However, except when considering very simple combinations of models, care should be taken not to overfit the evaluation set.

5.9 The reproducibility of research

It is important that data analysis be *reproducible.* This means that someone else can take the same data, analyse it in the same way, and obtain the same results. In order that an analysis be reproducible the following criteria are necessary:

- **The data used should be fully described and available to other researchers.**
- **Any modifications to the data (*eg* recoding or transformation of variables, or computation of new variables) should be clearly described, ideally with the computer code used. In machine learning this is often called 'features engineering', whereby combinations of features are used to create something more meaningful.**
- **The selection of the algorithm and the development of the model should be described, again with computer code being made available. This should include the parameters of the model and how and why they were chosen.**

There is an inherent problem with reproducing stochastic models (which are studied in Subject CM1), in that those of necessity have a random element. Of course, details of the random number generator seeds chosen, and the precise command and package used to generate any randomness, could be presented. However, since stochastic models are typically run many times to produce a distribution of results, it normally suffices that the distribution of the results is reproducible.

To ensure reproducibility in stochastic models in R, use the same numerical seed in the function `set.seed()`.

6 Applications: supervised learning

In this section we describe some commonly-used supervised machine learning techniques. Some of these methods are direct extensions of methods covered elsewhere in the syllabus, notably the linear regression and generalised linear models in Subject CS1 and the proportional hazards models in Chapter 8 of this course.

6.1 Penalised generalised linear models

Suppose we have a set of data with J covariates $(x_1,\dots,x_J)$ and N cases (the sample size)**. To fit a generalised linear model, we normally specify the link function and the parameters of the model, $\beta_0,\dots,\beta_J$ (where β_0 is the intercept) and maximise the likelihood $L(\beta_0,\dots,\beta_J \mid x_1,\dots,x_J)$. However, this might not work well in certain situations.**

For example, if there are many covariates, including all of them might make the model unstable (the estimators of the parameters $\beta_1,\dots,\beta_J$ will have large variances) because of correlations among the $x_1,\dots,x_J$. If we wish to use the model for prediction on new data, this is very undesirable. We want to be able to trust that the estimated values of the parameters linking the covariates to the outcome variable are solidly grounded and unlikely to shift greatly when the model is applied to a new data set. Another way of saying this is that we only want to include features which really do have a general effect on output.

Question

Give an example of this problem of correlations between covariates in an actuarial setting.

Solution

Suppose that two of the variables that a motor insurer uses to determine premiums for its policyholders are ‘postcode’ and ‘type of car’, and that the FX99 postcode area and Nano cars are usually considered to be low-risk and therefore cheap to insure. However, last year there was a single policyholder in the FX99 area who drove a Nano car and this policyholder was involved in a freak accident that led to a claim that cost £1 million.

Next year we would expect the insurer’s pricing model to increase premiums for any policyholder living in the FX99 area or driving a Nano car to reflect the past experience of these two categories. Any other policyholders who happened to live in the FX99 area *and* drive a Nano car would have their premiums increased on both counts.

If the insurer also uses the different *combinations* of postcode and type of car in its pricing model (*ie* separate premiums are determined for each pairing of postcode and type of car), it would now charge a prohibitively high premium to any other policyholder who happened to live in the FX99 area *and* drive a Nano car. This is purely because the values of their covariates match those of the policyholder who made the large claim and the pricing model has put them in the same category as this policyholder and has ‘tarred them with the same brush’.

So it is important for us to avoid over-parameterising our model.

One way to solve this is to choose a subset of the J covariates to include in the model. But how do we choose this? We could look at all possible subsets of J and use criteria such as the Akaike Information Criterion or the Bayesian Information Criterion.

These both exact a penalty for additional parameters. If the number of parameters is J and the sample size in the data is N:

$$\text{AIC} = \text{deviance} + 2J$$

$$\text{BIC} = \text{deviance} + [\log_e(N)]J$$

The deviance here is a measure of the average error of the model's outputs and measures the goodness of fit. The extra terms reflect the number of parameters in the model. If we aim to minimise the AIC or the BIC, this will allow us to find a good trade-off between the two objectives of obtaining a good fit to the data and minimising the number of parameters in the model.

However as J increases the number of possible subsets rises. In many machine learning applications, J is large, and the number of cases is also large, so that comparing all possible subsets is computationally infeasible.

Penalised regression involves exacting a penalty for unrealistic or extreme values of the parameters, or just having too many parameters. The penalty may be written $\lambda P(\beta_1,...,\beta_J)$, so that we maximise:

$$\log_e L(\beta_0,\dots,\beta_J \mid x_1,\dots,x_J) - \lambda P(\beta_1,...,\beta_J)$$

By maximising this expression we are aiming for a trade-off where we try to maximise the log-likelihood but, at the same time, try to minimise the penalty applied (since this is subtracted).

Two common examples of penalties are:

- **ridge regression, where $P(\beta_1,...,\beta_J) = \sum_{j=1}^{J} \beta_j^2$,**
- **the LASSO (Least Absolute Shrinkage and Selection Operator), where $P(\beta_1,...,\beta_J) = \sum_{j=1}^{J} \left|\beta_j\right|$.**

The names are based on the geometrical interpretations of these measures, which you are not expected to know.

These penalty functions assume that the distribution has been parameterised in such a way that we would expect the true values of the parameters $\beta_1,\dots,\beta_J$ to be close to zero. If this is not the case, we would subtract the target value from each β in the penalty functions shown above.

The parameter λ is called the regularisation parameter, and its choice is important. Too small a value of λ leads to over-fitting the data and the problems associated with using just the likelihood. Too large a value of λ means only gross effects of the covariates will be included in the final model, and we may miss many important effects of the covariates on the outcome.

The parameter λ is an example of a hyper-parameter.

Question

A random sample $x_1,\dots,x_n$ of n values has been taken from a Poisson distribution with unknown mean μ. The value of μ is to be estimated using the penalty function $\lambda(\mu-5)^2$.

(i) Explain why this particular penalty function might have been chosen.

(ii) Write down an expression for the penalised log-likelihood function.

(iii) Show that $\hat{\mu}$, the estimated value of μ, satisfies the equation:

$$n(\hat{\mu}-\bar{x})+2\lambda\hat{\mu}(\hat{\mu}-5)=0$$

(iv) Calculate the value of $\hat{\mu}$ based on the sample of values 5.7, 5.4, 4.6, 5.0 and 4.9 when $\lambda=0.2$.

(v) Use the equation in part (iii) to show algebraically what happens to the value of $\hat{\mu}$:

(a) when λ is equal to zero and

(b) when λ is very large.

(vi) Comment on your answers in part (v).

Solution

In this question we have just one parameter (*ie* $J=1$), which is called μ (corresponding to β_1).

(i) We might choose this penalty function if we believe the true value of μ is close to 5, as the penalty when $\mu=5$ would be zero.

(ii) The likelihood function for the sample is:

$$L=\prod_{i=1}^{n}\frac{e^{-\mu}\mu^{x_i}}{x_i\,!}=e^{-n\mu}\mu^{\sum x_i}\times\text{constant}=e^{-n\mu}\mu^{n\bar{x}}\times\text{constant}$$

So the log-likelihood is:

$$\log L=-n\mu+n\bar{x}\log\mu+\text{constant}$$

and the penalised log-likelihood is:

$$(\log L)^*=-n\mu+n\bar{x}\log\mu+\text{constant}-\lambda(\mu-5)^2$$

(iii) To maximise this, we equate the derivative with respect to the parameter μ to zero:

$$\frac{\partial(\log L)^*}{\partial\mu}=-n+\frac{n\bar{x}}{\mu}-2\lambda(\mu-5)=0$$

Multiplying by μ and rearranging gives:

$$n(\mu - \bar{x}) + 2\lambda\mu(\mu - 5) = 0$$

So $\hat{\mu}$ satisfies the equation:

$$n(\hat{\mu} - \bar{x}) + 2\lambda\hat{\mu}(\hat{\mu} - 5) = 0$$

(iv) For this sample, we have:

$$n = 5 \quad \text{and} \quad \bar{x} = \frac{1}{5}(5.7 + 5.4 + 4.6 + 5.0 + 4.9) = \frac{25.6}{5} = 5.12$$

So with $\lambda = 0.2$, the equation in part (iii) becomes:

$$5(\hat{\mu} - 5.12) + 2(0.2)\hat{\mu}(\hat{\mu} - 5) = 0$$

This can be rearranged to give:

$$0.4\hat{\mu}^2 + 3\hat{\mu} - 25.6 = 0$$

Using the quadratic formula:

$$\hat{\mu} = \frac{-3 \pm \sqrt{3^2 - 4(0.4)(-25.6)}}{2(0.4)} = \frac{-3 \pm \sqrt{49.96}}{0.8} = -12.585 \text{ or } 5.085$$

Since μ must be positive, the required estimate is $\hat{\mu} = 5.085$.

(v)(a) If $\lambda = 0$, the equation in part (iii) becomes:

$$n(\hat{\mu} - \bar{x}) = 0 \quad \Rightarrow \hat{\mu} = \bar{x}$$

(v)(b) If we divide the equation in part (iii) by λ, we get:

$$\frac{n(\hat{\mu} - \bar{x})}{\lambda} + 2\hat{\mu}(\hat{\mu} - 5) = 0$$

As $\lambda \to \infty$, this becomes:

$$2\hat{\mu}(\hat{\mu} - 5) = 0 \quad \Rightarrow \hat{\mu} = 0 \quad \text{or} \quad \hat{\mu} = 5$$

Since the value of μ must be strictly positive, the required estimate would be $\hat{\mu} = 5$.

(vi) If we apply no penalty, as in part (v)(a), the method reduces to maximum likelihood estimation and we get the usual estimate for μ, *ie* the sample mean of 5.12.

If we apply a high penalty to values that are not close to the target value of 5, as in part (v)(b), the method will produce a value close to 5, irrespective of the actual values in the sample.

As expected, the estimated value of 5.085 lies between these two values.

6.2 Naïve Bayes classification

Recall from Subject CS1 that if $B_1, B_2, ..., B_R$ constitute a partition of a sample space S and $P(B_i) \neq 0$ for $i = 1,2,\dots,R$, then for any event A in S such that $P(A) \neq 0$:

$$\boldsymbol{P(B_r \mid A) = \frac{P(A|B_r)P(B_r)}{P(A)} \quad \text{for} \quad r = 1,2,\dots,R}$$

where:

$$\boldsymbol{P(A) = \sum_{i=1}^{R} P(A|B_i)P(B_i)}$$

This is *Bayes' Theorem*, which allows us to 'invert' conditional probabilities, *ie* to work out the values of the probabilities $P(B_r \mid A)$ when we know the probabilities $P(A \mid B_r)$.

Question

Derive Bayes' Theorem.

Solution

The proof uses the definition of conditional probabilities, $P(X \mid Y) = \dfrac{P(X,Y)}{P(Y)}$, and the equivalent identity $P(X,Y) = P(X \mid Y)P(Y)$.

Using the definition of conditional probabilities, the probability we want to find is:

$$P(B_r \mid A) = \frac{P(B_r, A)}{P(A)}$$

Using the identity above, the numerator on the right-hand side can be written as:

$$P(B_r, A) = P(A \mid B_r)P(B_r)$$

If we condition the denominator on the different possible values of B_r, we can write it as:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_R)P(B_R) = \sum_{r=1}^{R} P(A \mid B_r)P(B_r)$$

If we substitute these in, we then get Bayes' formula:

$$P(B_r \mid A) = \frac{P(A \mid B_r)P(B_r)}{P(A)} = \frac{P(A \mid B_r)P(B_r)}{\sum_{r=1}^{R} P(A \mid B_r)P(B_r)}$$

Naïve Bayes classification uses this formula to classify cases into mutually exclusive categories on some outcome y, on the basis of a set of covariates $x_1,\ldots,x_J$. The events A are equivalent to the covariates taking some set of values, and the partition $B_1, B_2, \ldots, B_R$ is the set of values that the outcome can take.

Here B_1 corresponds to the *vector* of covariates $(x_{11}, x_{21}, \ldots, x_{J1})$ for the first individual.

Suppose the outcome is whether or not a person will survive for 10 years. Let $y_i = 1$ denote the outcome that person i survives, and $y_i = 0$ denote the outcome that person i dies. Then, if we have J covariates, we can write:

$$P(y_i = 1 | x_{1i}, \ldots, x_{Ji}) = \frac{P(x_{1i}, \ldots, x_{Ji} | y_i = 1) P(y_i = 1)}{P(x_{1i}, \ldots, x_{Ji})}$$

This is difficult to estimate because all possible combinations of the $x_1,\ldots,x_J$ need to be estimated, and all combinations are unlikely to be in your data set.

This is a particular problem here because in practice many of the combinations of values will not be present in the sample, so the corresponding probabilities cannot be estimated. For example, for a motor insurer that uses several rating factors (*eg* age of policyholder, occupation, make of car, age of car, postcode area), many of the subsets will be empty.

The naïve Bayes algorithm assumes that the values of the x_i are independent, conditional on the value of y_i.

So we are assuming that:

$$P(x_{11}, x_{21}, \ldots, x_{J1} | y_i = 1) = P(x_{11} | y_i = 1) \times P(x_{21} | y_i = 1) \times \cdots \times P(x_{J1} | y_i = 1)$$

Question

Illustrate why this assumption might not be accurate for a motor insurer using the rating factors age of policyholder, occupation, make of car, age of car, postcode area.

Solution

As an example, 25% of policyholders may be under 25, 20% may drive high performance cars and 40% may drive cars less than 3 years old.

However, it is unlikely that 2% (*ie* $25\% \times 20\% \times 40\%$) of policyholders would be under 25 with high performance cars under 3 years old, as young drivers are unlikely to be able to afford such vehicles – or to be able to pay to insure them.

This allows the formula to be re-written:

$$P(y_i = 1 | x_{1i},\dots,x_{Ji}) = \frac{P(x_{1i} | y_i = 1)P(x_{2i} | y_i = 1)\dots P(x_{Ji} | y_i = 1)\ P(y_i = 1)}{P(x_{1i},\dots,x_{Ji})}$$

so that:

$$P(y_i = 1 | x_{1i},\dots,x_{Ji}) \propto P(y_i = 1) \prod_{j=1}^{J} P(x_{ji} | y_i = 1)$$

For a given set of values of $x_{1i},\dots,x_{J1}$, the denominator $P(x_{1i},\dots,x_{J1})$ doesn't change. So we can treat this as a constant in the calculations and just look at the relative values – hence the proportional sign. To find the actual values of the individual probabilities, we can just divide by the total, to produce a set of probabilities that add up to 1.

Question

A motor insurer has analysed a sample of 1,000 claims for three different geographical regions split by the size of claim (Small / Medium / Large). It has then classified them according to whether they proved to be fraudulent or genuine claims. The results are shown in the tables below:

FRAUDULENT	Region 1	Region 2	Region 3	Total
Small	3	0	6	9
Medium	10	5	20	35
Large	3	1	2	6
Total	16	6	28	50

GENUINE	Region 1	Region 2	Region 3	Total
Small	57	38	70	165
Medium	250	95	180	525
Large	176	46	38	260
Total	483	179	288	950

(i) Give a formula that could be used to estimate the probability that a new claim from Region 3 for a Medium amount will prove to be fraudulent.

(ii) Estimate the probability that each of the following types of new claim will be fraudulent:

(a) a claim from Region 3 for a Medium amount

(b) a claim from Region 1 for a Large amount

(c) a claim from Region 2 for a Small amount.

Solution

(i) Using Bayes' Theorem (and obvious abbreviations for the events), we have:

$$P(F\,|\,R3,M)=\frac{P(F,R3,M)}{P(R3,M)}=\frac{P(R3,M\,|\,F)P(F)}{P(R3,M\,|\,F)P(F)+P(R3,M\,|\,G)P(G)}$$

(ii)(a) From the tables above, we have:

$$P(R3,M\,|\,F)=\frac{20}{50}=0.4\,,\quad P(R3,M\,|\,G)=\frac{180}{950}\,,$$

$$P(F)=\frac{50}{50+950}=0.05\,,\quad P(G)=\frac{950}{50+950}=0.95$$

So: $$P(F\,|\,R3,M)=\frac{0.4\times 0.05}{0.4\times 0.05+180/950\times 0.95}=0.1$$

So the estimated probability that a claim from Region 3 for a Medium amount is fraudulent is 10%.

In fact, we can do this calculation directly from the table without introducing probabilities. For Region 3 and Medium amounts there were 20 fraudulent claims and 180 genuine claims. So the estimated probability that a claim from Region 3 for a Medium amount is fraudulent is $\frac{20}{20+180}=0.1$, *ie* 10%.

(ii)(b) Similarly, the estimated probability that a claim from Region 1 for a Large amount is fraudulent is $\frac{3}{3+176}=0.168$, *ie* 1.68%.

(ii)(c) Since there were no fraudulent claims from Region 2 for a Small amount in the sample, the estimated probability that a claim from Region 2 for a Small amount is fraudulent is 0.

6.3 Decision trees (classification and regression tree algorithm)

Classification and regression trees (CART) is a term introduced by Leo Breiman to refer to decision tree algorithms that can be used for classification or regression in predictive modelling problems.

Classically, this algorithm is referred to as 'decision trees' but on some platforms like R they are referred to by the more modern term CART.

The CART algorithm provides a foundation for important algorithms like:

- **bagged decision trees**
- **random forest**
- **boosted decision trees.**

We will not discuss these three types of decision trees in detail in this course, but we will provide a very brief insight into these approaches.

In *bagged decision trees*, we create random sub-samples of our data with replacement, train a CART model on each sample, and (given new data) calculate the average prediction from each model.

Random forests apply a method based on averaging a number of randomly generated decision trees.

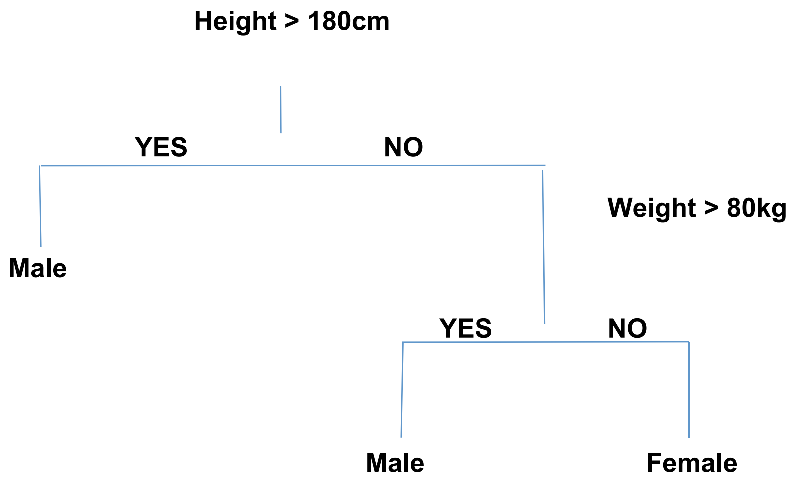
Boosting here refers to a method of repeatedly making small adjustments to improve the effectiveness of a model by reducing the residual error.

The representation for the CART model is a binary tree.

Each root node on a tree represents a single input variable x and a split point on that variable (assuming the variable is numeric).

The leaf nodes of the tree contain an output variable y which is used to make a prediction.

Given a dataset with two inputs of height in centimetres and weight in kilograms, the output of gender as male or female, below, is a crude example of a binary decision tree.



Example

Given an input of [height = 60cm, weight = 65kg] the above tree would be traversed as follows:

Node 1: Height > 180cm? No

Node 2: Weight > 80kg? No

Therefore, my result is: Female

Question

(i) Comment on the usefulness of this tree when applied to an ActEd tutor whose height is 175cm and whose weight is approximately 80kg.

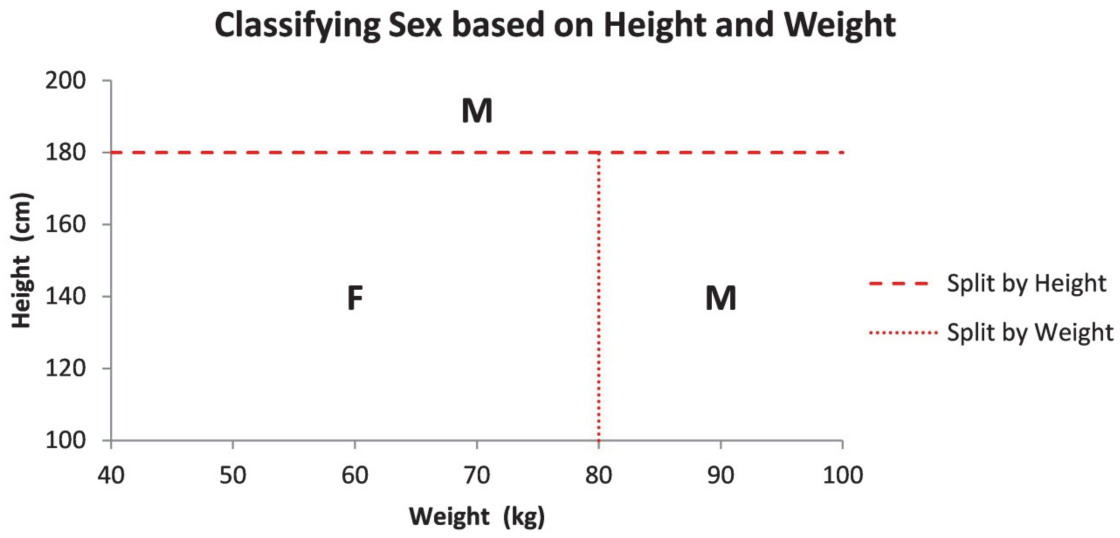
(ii) Explain the rationale behind it.

Solution

(i) Using this tree, this tutor would be on the cusp between Male and Female. The prediction of the tutor's gender would not be reliable, as it would depend on whether their weight was above or below 80kg on a particular day.

(ii) The rationale is that tall people (Height > 180cm) tend to be male, so we can separate them out at the first stage. Of the remaining people, males tend to be heavier than females, so we can split these at a level that is likely to a reasonable boundary between males and females (80kg, say).

We can represent this tree using the rectangle below.



With the binary tree representation of the CART model described above, making predictions is relatively straightforward.

Given a new input, the tree is traversed by evaluating the specific input at the root node of the tree.

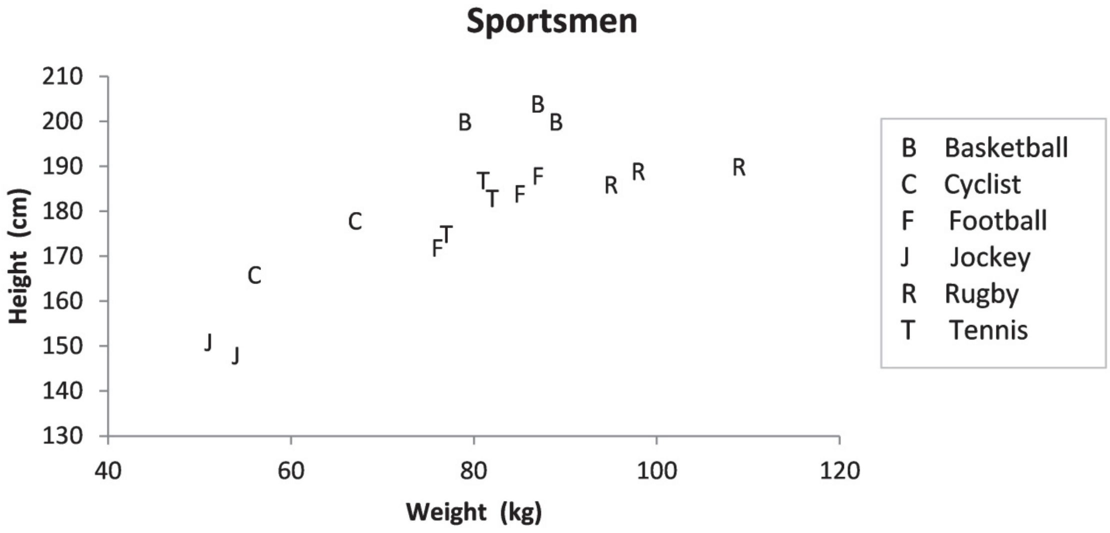
A learned binary tree is a partitioning of the input space. You can think of each input variable as a dimension on a p-dimension space. The decision tree splits this up into rectangles (when $p = 2$ input variables) or hyper-rectangles with more inputs.

In the model above we have $p = 2$ dimensions, *ie* there are two input variables, Height and Weight.

New data is filtered through the tree and lands in one of the rectangles and the output value for that rectangle is the prediction made by the model. This gives an intuition for the type of decisions that a CART model can make, *eg* boxy decision boundaries.

Question

The graph below shows the heights and weights of a sample of 16 sportsmen, which were measured at the start of a competition. Devise a binary decision tree that could be used to identify the most likely sport for other sportsmen who are competing, based only on their height and weight. Explain the rationale behind your algorithm.



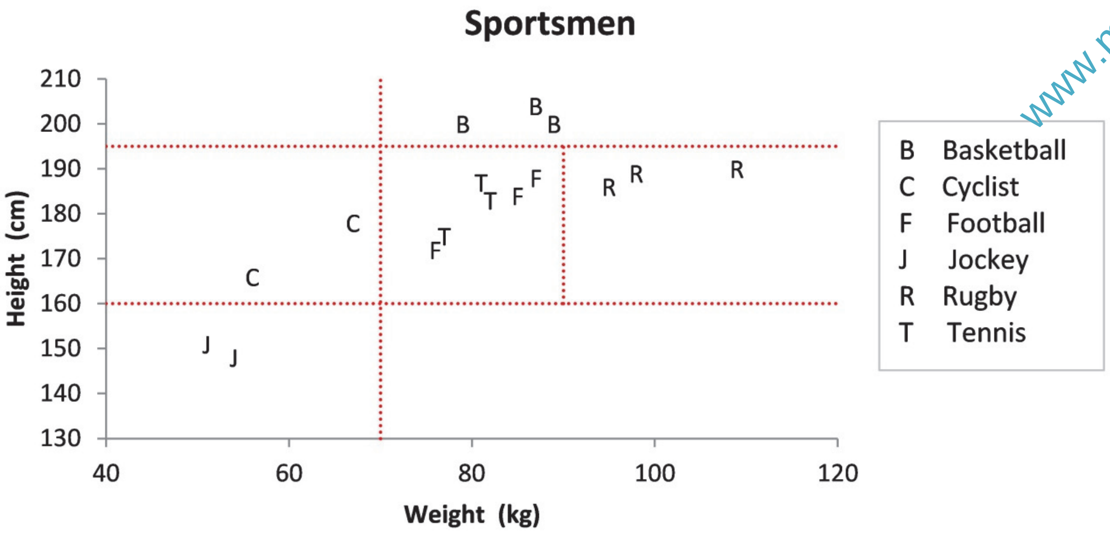
Solution

We can easily section off the basketball players, who tend to be tall, and the jockeys, who tend to be short.

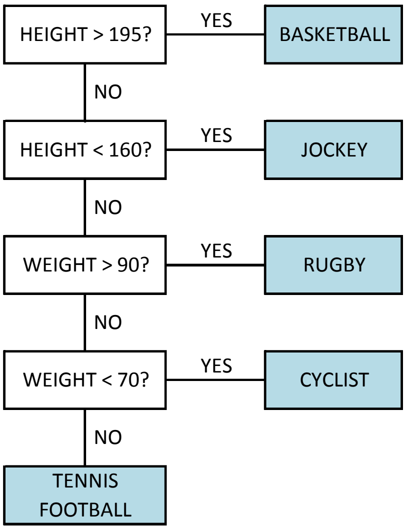
We can then section off the rugby players, who tend to be heavier, and the cyclists, who tend to be lighter.

However, the tennis players and footballers are intermixed, so we can't distinguish effectively between these two sports.

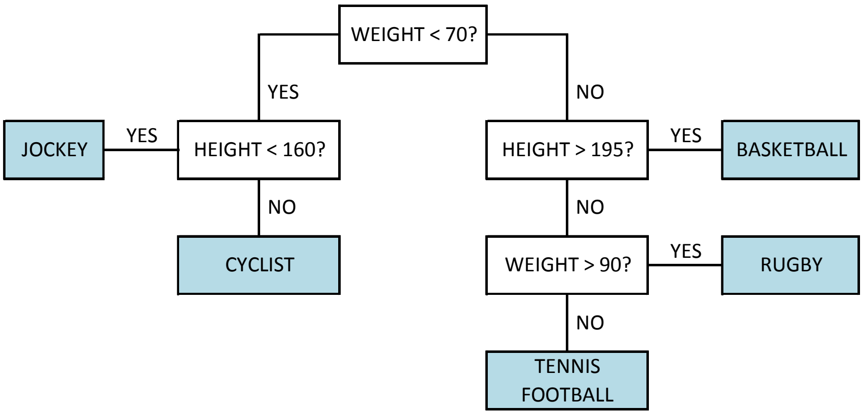
This leads to the partitions shown in the graph below.



Here is one possible algorithm based on these partitions.



This is not the only way to achieve the same result. For example, we could instead start with the test 'Weight < 70?' This alternative version has the slight advantage that the maximum number of tests that might be required is now 3, rather than 4.



Greedy splitting

Creating a binary decision tree is a process of dividing up the input space. A 'greedy' approach is used to divide the space, called recursive binary splitting.

This is a numerical procedure where all the values are lined up and different split points are tried and tested using a cost function. The split with the best cost (lowest cost because we minimize cost) is selected.

All input variables and all possible split points are evaluated and chosen in a greedy manner (*ie* the very best split point is chosen each time).

This is the approach we used in the previous example for sportsmen. In this particular example, most of the split points were clear-cut. However, to separate the tennis players and footballers, we would need to try different possible splits and then choose the one that minimises the number of incorrect classifications based on the training sample. So we would be applying a zero-one cost function where the cost is 1 unit if the answer is wrong and 0 if it is right.

The word 'greedy' here means that, at each stage, we just choose the split that appears to be the most effective at separating the remaining elements, without thinking ahead to the consequences this might have on the later splits. So we just 'bite off as much as we can' at each stage.

With just two variables, it is easy to visualise the best splits on a 2D graph. However, in higher dimensions, this is not so easy and we will need to calculate the effectiveness of different possible splits.

For regression problems, the cost function that is minimized to choose split points is the sum squared error across all training samples that fall within the rectangle:

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i is the output for the training sample and $\hat{y}_i$ is the predicted output for the rectangle.

For classification, the Gini index function is used, which provides an indication of how 'pure' the leaf nodes are (*ie* how mixed the training data assigned to each node is):

$$G = \sum_{k} p_k(1 - p_k)$$

This formula gives the Gini index for one of the final nodes in the tree or for a rectangle in the diagram.

Here p_k is the proportion of training instances with class k in the rectangle of interest. A node that has all classes of the same type (perfect class purity) will have $G = 0$, whereas a node that has a 50-50 split of classes for a binary classification problem (worst purity) will have $G = 0.5$.

The Gini index is a measure of inequality of a distribution that was introduced by the Italian statistician Corrado Gini. It was originally used to measure inequality in the distribution of income levels within a population.

It is calculated as the probability that, if two items are selected at random (with replacement), they will be of different types (*ie* if the first item is of type k, the second item will not be of type k). So, if all the items are the same, the probability will be 0, whereas, if they are all different, it will be close to 1.

For a binary classification problem, this can be re-written as:

$$\boldsymbol{G = 2p_1p_2}$$

or $$\boldsymbol{G = 1-\left(p_1^2+p_2^2\right)}$$

This follows since $p_1+p_2=1$, so that $(p_1+p_2)^2=p_1^2+p_2^2+2p_1p_2=1$.

The Gini index calculation for each node is weighted by the total number of instances in the parent node. The Gini score for a chosen split point in a binary classification problem is therefore calculated as follows:

$$\boldsymbol{G=\left[1-\left(g_{1,1}^2+g_{1,2}^2\right)\right]\times\frac{n_{g_1}}{n}+\left[1-\left(g_{2,1}^2+g_{2,2}^2\right)\right]\times\frac{n_{g_2}}{n}}$$

This formula gives the Gini index for the split point, *ie* it is a combined value that covers both of the nodes that meet at that point.

Here:

- **$g_{1,1}$ is the proportion of instances in group 1 for class 1, $g_{1,2}$ for group1 and class 2**
- **$g_{2,1}$ is the proportion of instances in group 2 for class 1, $g_{2,2}$ for group 2 and class 2**
- **n_{g_1} and n_{g_2} are the total number of instances in groups 1 and 2**
- **n is the total number of instances we are trying to group from the parent node.**

We can write the general formula for the Gini index for the whole tree as:

$$G=\sum_{nodes}\frac{n_{node}}{n}\left(1-\sum_{k=1}^{m}p_k^2\right)$$

where the sum is taken over all the nodes and n_{node} is the number of items at the node we are currently evaluating.

Question

Calculate the Gini index for each of the final nodes in the decision tree for the sportsmen and hence calculate the overall Gini index for the tree. Comment on your answer.

Solution

In this example the classes are the different sports.

We need to consider each of the shaded nodes (which are the same for both of the trees given).

For the node containing 3 tennis players and 3 footballers we have $p_1 = p_2 = \frac{3}{6} = \frac{1}{2}$.

So: $$G = 1 - \left(p_1^2 + p_2^2\right) = 1 - \left[\left(\tfrac{1}{2}\right)^2 + \left(\tfrac{1}{2}\right)^2\right] = \tfrac{1}{2}$$

Each of the other nodes is 'pure', with $p_1 = 1$. So these nodes have a Gini index of 0.

The overall Gini index for the tree is obtained by weighting these values by the number of data points:

$$G = \frac{1}{2} \times \frac{6}{16} + 0 \times \frac{10}{16} = \frac{3}{16} = 0.1875$$

This value is quite low, indicating that the tree is reasonably effective at classifying the sportsmen.

Stopping criterion

The recursive binary splitting procedure described above needs to know when to stop splitting as it works its way down the tree with the training data.

The most common stopping procedure is to use a minimum count on the number of training instances (*ie* the data items in our sample) **assigned to each leaf node. If the count is less than some minimum then the split is not accepted and the node is taken as a final leaf node.**

So, for example, we could say that if we try to split a node and it would result in fewer than, say, 5 data items in our sample on one of the new nodes, then we stop and leave the node as it is.

The minimum number we choose here (*ie* 5) is another example of a hyper-parameter.

The count of training members (*ie* the data items in our sample) **is tuned to the dataset, *eg* 5 or 10. It defines how specific to the training data the tree will be. Too specific (*eg* a count of 1) and the tree will overfit the training data and likely have poor performance on the test set.**

If we keep splitting until there is just one item on each of the final leaves in the tree, we will end up adding decision tests that are very specific to the particular data in the training set. For example, if there happened to be a female in the training data whose height was 183cm, we might include a test 'Height = 183cm?' to allocate her correctly. This would give a perfect result for this particular training set, but would not work for other data sets. So this would be an example of overfitting.

Pruning the tree

The stopping criterion is important as it strongly influences the performance of the tree. Pruning may be used after learning to further enhance the tree's performance.

The complexity of a decision tree is defined as the number of splits in the tree. Simpler trees are preferred. They are easy to understand (you can print them out and show them to subject matter experts), and they are less likely to overfit your data.

The fastest and simplest pruning method is to work through each leaf node in the tree and evaluate the effect of removing it using a hold-out test set. Leaf nodes are removed only if it results in a drop in the overall cost function on the entire test set. You stop removing nodes when no further improvements can be made.

The 'hold-out test set' just refers to another data set that wasn't used in training the model.

More sophisticated pruning methods can be used such as cost complexity pruning (also called 'weakest link pruning') where a learning parameter (alpha) is used to weigh whether nodes can be removed based on the size of the sub-tree.

7 Applications: unsupervised learning

7.1 *K*-means clustering

Suppose we have a set of data consisting of several variables (or features) measured for a group of individuals. These might relate to demographic characteristics, such as age, occupation, gender. Alternatively, they might relate to life insurance policies for which we have information such as sales channel, policy size, postcode, level of underwriting, *etc*.

We might ask whether we can identify groups (clusters) of policies which have similar characteristics. We may not know in advance what these clusters are likely to be, or even how many there are in our data.

Here we would be particularly interested in how we could group policies by occupation or postcode, as it would not be obvious at the outset how to do this in a logical way so that ones with similar outcomes are grouped together.

There are a range of clustering algorithms available, but many are based on the K *-means algorithm*. This is an iterative algorithm which starts with an initial division of the data into K clusters, and adjusts that division in a series of steps designed to increase the homogeneity within each cluster and to increase the heterogeneity between clusters.

The K -means algorithm proceeds as follows. Let us suppose we have data on J variables.

1. **Choose a number of clusters, K , into which the data are to be divided. This could be done on the basis of prior knowledge of the problem. Alternatively, the algorithm could be run several times with different numbers of clusters to see which produces the most satisfactory and interpretable result. There are various measures of within- and between-group heterogeneity, often based on within-groups sums of squares. Comparing within-groups sums of squares for different numbers of clusters might identify a value of K beyond which no great increase in within-group homogeneity is obtained.**

   These measures are based on the same identity for partitioning the sum of squares into 'between groups' and 'within groups' that forms the basis for the technique of analysis of variance (ANOVA), which is mentioned in the regression chapter in Subject CS1:

$$\underbrace{\sum\sum(y_{ij}-\bar{y}_{\bullet\bullet})^2}_{\text{Total sum of squares}} = \underbrace{\sum\sum(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2}_{\text{Between groups}} + \underbrace{\sum\sum(y_{ij}-\bar{y}_{i\bullet})^2}_{\text{Within groups}}$$

2. **Identify (perhaps arbitrarily) cluster centres in the J -dimensional space occupied by the data. This initial location of the centres could be done on the basis of prior knowledge of the problem to hand, or by random assignment of cases.**

3. **Assign cases to the cluster centre which is nearest to them, using some measure of distance. One common measure is Euclidean distance:**

$$\textbf{dist}(\boldsymbol{x},\boldsymbol{k}) = \sqrt{\sum_{j=1}^{J}(\boldsymbol{x}_j-\boldsymbol{k}_j)^2}$$

Here x_j is the standardised value of covariate j for case x, and k_j is the value of covariate j at the centre of cluster k ($k = 1,\ldots,K$). Note that it is often necessary to standardise the data before calculating any distance measure, for example by assuming a normal distribution using *z*-scores or by assuming a uniform distribution on (x_a, x_b), where x_a and x_b are the lowest and highest observed values of covariate x.

This is because the measure of distance is based purely on the numerical values of the variables. If some of the variables take large values because of the units of measurement that have been adopted, these variables will dominate the calculations and the other variables will effectively be ignored.

4. **Calculate the centroid of each cluster, using the mean values of the data points assigned to that cluster. This centroid becomes the new centre of each cluster.**

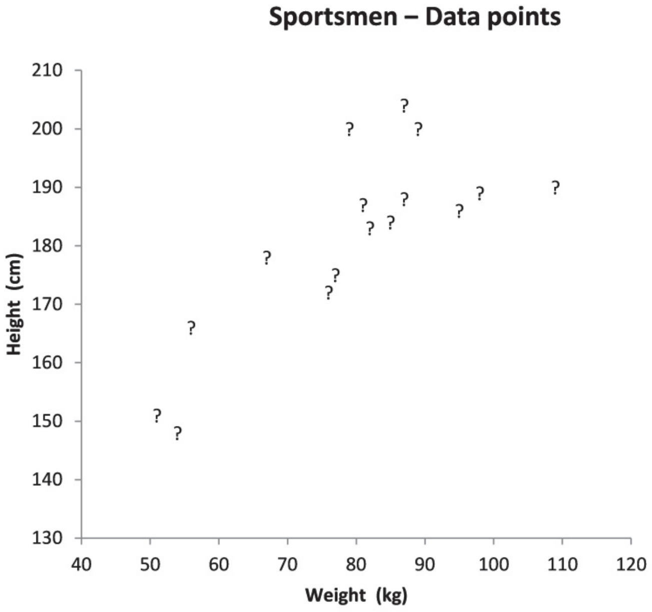
   The centroid is just the centre of gravity of the data points when each has the same weight. To calculate it, we find the average of each 'coordinate' in the data set.

5. **Re-assign cases to the nearest cluster centre using the new cluster centres.**

Iterate steps 4 and 5 until no re-assignment of cases takes place.

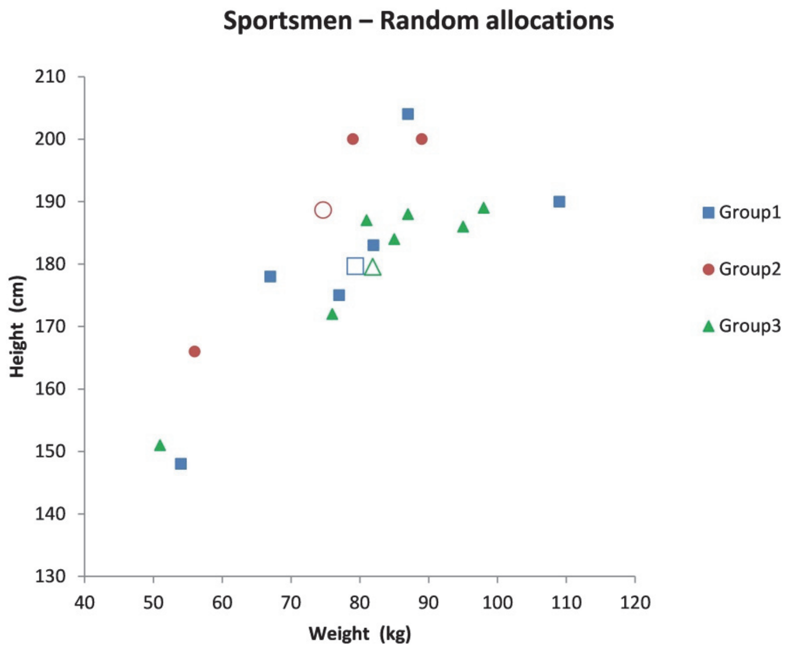
Example

The graph below shows the heights and weights of the same sample of 16 sportsmen from the previous example, but without showing the sports they were competing in. We have rescaled the vertical axis so that the spacing is the same on each axis. This ensures that the distances between the points appear correctly.



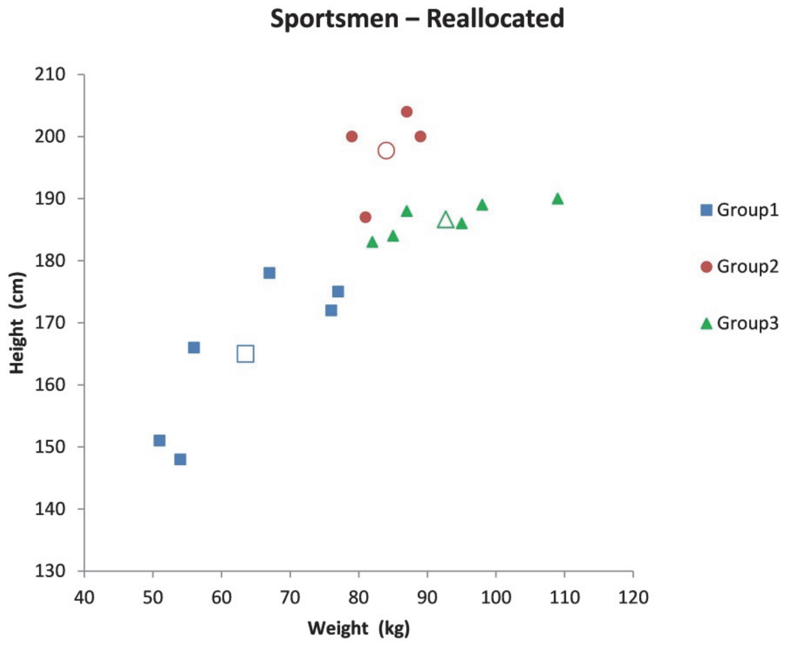
We can use the k-means method with $k = 3$ to try to identify any natural clusters in the data.

We start by allocating each of the data points to one of the 3 clusters at random. We've used squares, circles and triangles to distinguish the three groups. We then need to calculate the position of the centroid of each cluster by averaging the x and y coordinates of the data points within each cluster. The centroids are indicated by the larger hollow shapes.

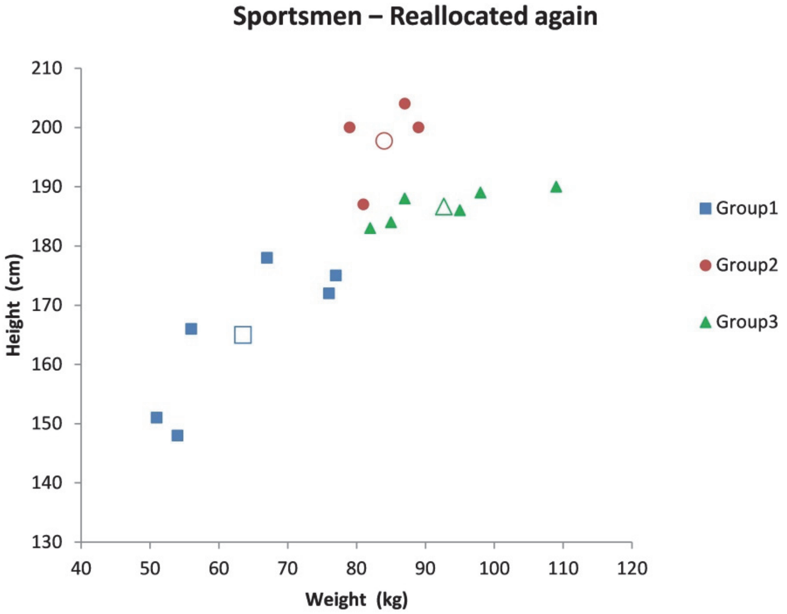


Since the initial allocation was random, the three centroids are all quite close together at this stage.

We now reallocate each data point to the cluster whose centroid it is nearest to and then recalculate the positions of the centroids of the new groups.



Again, we reallocate each data point to the cluster whose centre it is nearest to.



We can see that there has been no change to the allocations from the previous diagram. So the algorithm has now converged, with the points allocated to the three groups indicated by the different shapes.

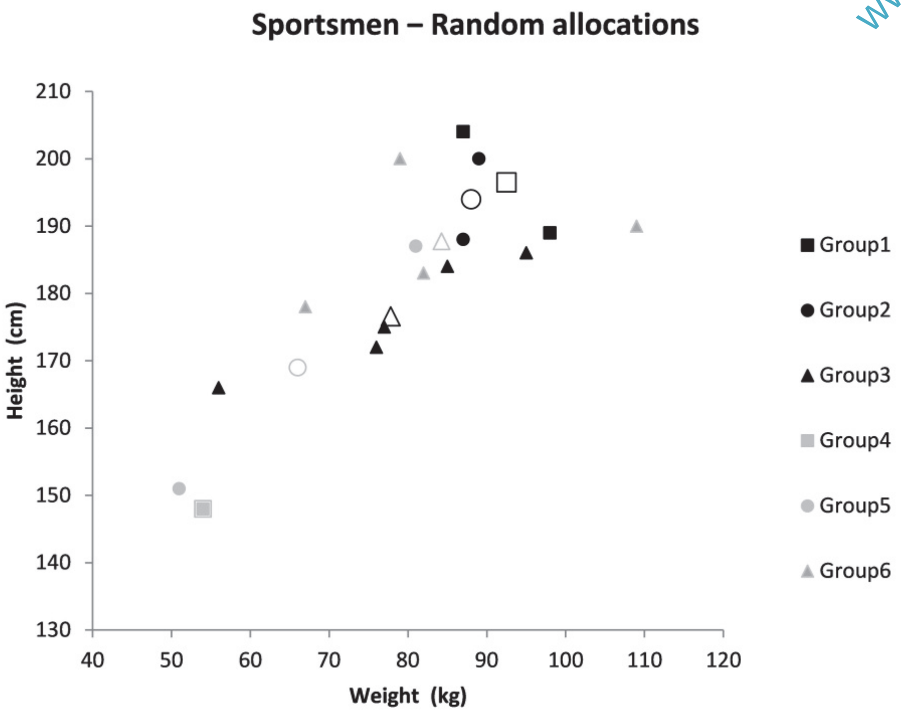
Comment

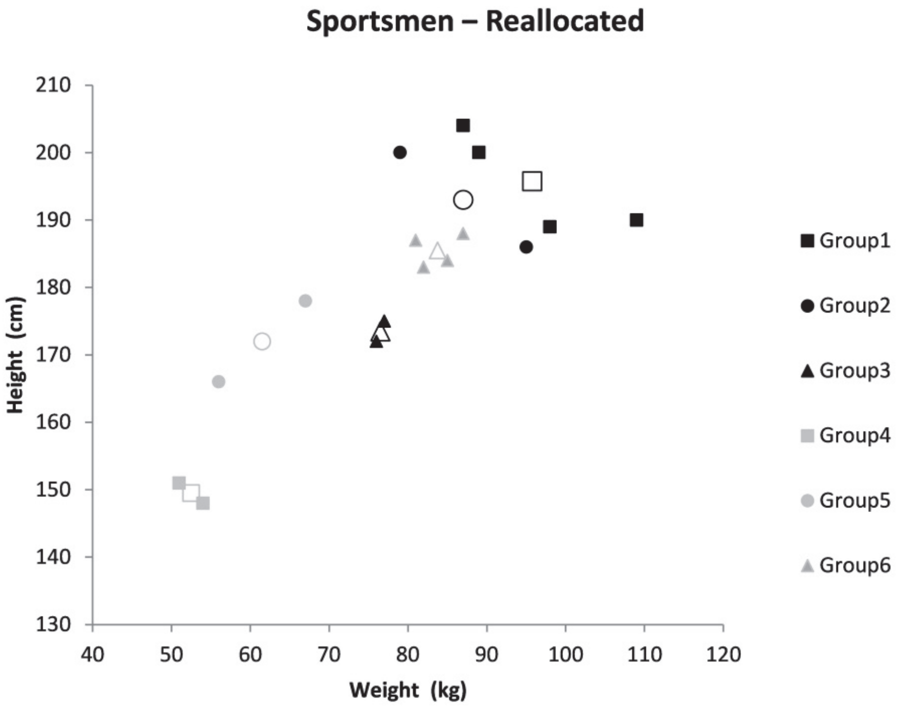
If we compare this grouping with the actual sports, we can see that:

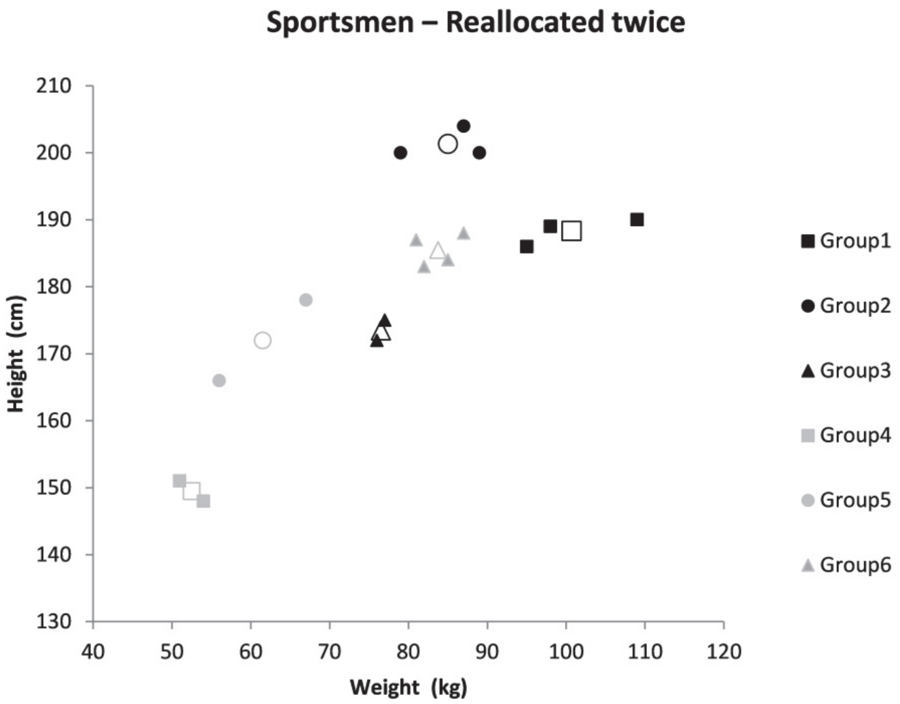
- the jockeys and cyclists have all been allocated to Group 1 (squares)
- the basketball players have all been allocated to Group 2 (circles)
- the rugby players have all been allocated to Group 3 (triangles)
- the footballers and tennis players have been spread amongst the three groups.

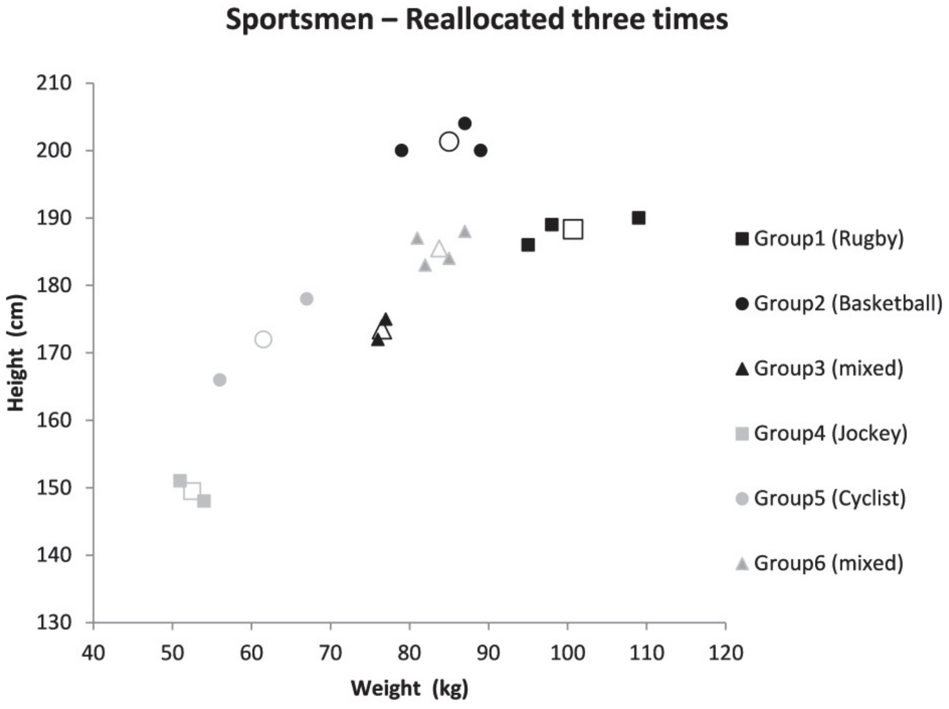
So with $k = 3$ the algorithm has correctly distinguished the sportsmen with more extreme characteristics (*eg* the very tall basketball players), but was not able to distinguish between the footballers and tennis players, who have quite similar characteristics.

We can try repeating the same steps with $k = 6$ groups (since there were six sports). We've used lighter coloured markers for the three extra groups.









There have been no changes, so the algorithm has converged at this stage.

We can see that:

- the jockeys (corresponding to Group 4), cyclists (Group 5), basketball players (Group 2) and rugby players (Group 1) have now all been allocated correctly to separate groups
- the footballers and tennis players have been allocated to Groups 3 and 6, but the split between the two sports is not correct.

So with $k=6$ the algorithm has performed better, correctly distinguishing all the sportsmen apart from the footballers and tennis players who have quite similar characteristics.

We could now use this final diagram to identify the most likely sport of another sportsman in the competition by checking which of the group centroids his weight and height are closest to.

For example, for a sportsman with weight 65kg and height 185cm, this would be a cyclist.

The table below shows the strengths and weaknesses of the K-means algorithm.

Strengths	Weaknesses
Uses simple principles for identifying clusters which can be explained in non-statistical terms	**Less sophisticated than more recent clustering algorithms**
Highly flexible and can be adapted to address nearly all its shortcomings with simple adjustments	**Not guaranteed to find the optimal set of clusters because it incorporates a random element**
Fairly efficient and performs well	**Requires a reasonable guess as to how many clusters naturally exist in the data**

Source: B. Lantz, *Machine Learning with R* (Birmingham, Packt Publishing, 2013), p. 271

The interpretation and evaluation of the results of K-means clustering can be somewhat subjective. If the K-means exercise has been useful, the characteristics of the clusters will be interpretable within the context of the problem being studied, and will either confirm that the pre-existing opinion about the existence of homogeneous groups has an evidential base in the data, or provide new insights into the existence of groups that were not seen before. One objective criterion that can be examined is the size of each of the clusters. If one cluster contains the vast majority of the cases, or there are clusters with only a few cases, this may indicate that meaningful groups do not exist.

Another way to judge the effectiveness of the algorithm is to repeat the process several times with different random allocations at the start. If similar groupings are obtained each time, it is likely that there is a valid basis underlying the clusters.

R has several machine learning packages that will achieve K-means clustering. One simple one is `kmeans`**.**

7.2 Principal components analysis

This is covered in Subject CS1.

Principal components analysis is a technique that identifies the dominant combinations of factors that are present in a dataset. The number of such combinations can be chosen so that the model is sufficient to give reasonably accurate results, but not so large that it reflects a lot of the random noise contained in the data.

8 Perspectives of statisticians, data scientists and other quantitative researchers

Machine learning involves the application to data of a range of methods aimed at using data to solve real-world problems. However, many other quantitative researchers would claim to be doing the same thing.

It is certainly true that practitioners from other backgrounds often do work that overlaps with machine learning. Statisticians, for example, do data mining, data reduction using principal components analysis, and routinely estimate logistic regression models. There are differences between the perspectives of many statisticians and that normally adopted by the users of machine learning techniques.

8.1 Terminology

Some of the challenges of communicating with other quantitative researchers are straightforward differences of terminology. In machine learning we talk of 'training' a model, or 'training' hyper-parameters, whereas statisticians might talk of 'fitting' a model or 'choosing' higher-level parameters. These are really different words being used for the same activity.

8.2 What is the aim of the analysis?

Some of the differences in perspectives of different groups of researchers are related to the aims of their analyses. This results in interest focusing on different aspects of the models.

This may be illustrated using logistic regression, or discriminant analysis. The logistic regression model may be written:

$$\log\left(\frac{P(y_i=1)}{P(y_i=0)}\right)=\beta_0+\beta_1 x_{1i}+\ldots+\beta_J x_{Ji}$$

where y is a binary variable dividing the data into two categories, coded 1 and 0, $x_{1i},\ldots,x_{Ji}$ are the values of the J covariates for case i, and the β's are parameters to be estimated from the data.

Statisticians will tend to be most interested in the values and significance of the β's, that is in the effect of the covariates on the probability that a case is in either group. They will often present these in tables of *odds ratios*, showing the effect of a difference in the value of a covariate on $\dfrac{P(y_i=1)}{P(y_i=0)}$. Often the purpose of their analyses is to test hypotheses about the effect of a covariate on the odds of y_i being 1.

For example, in a clinical trial, $y_i=1$ might denote recovery from an illness and $y_i=0$ denote death, x_1 might be a new drug treatment and $x_2,\ldots,x_J$ might be controls. The statistician's interest is mainly in the size and significance of the parameter β_1, and especially whether or not β_1 suggests that the new treatment leads to an increase in the odds of recovery. How good the model is at predicting who will recover and who will die is less of an issue.

In machine learning applications, however, the actual values of the β's are less important than the success of the model in predicting who will recover and who will die, or at discriminating between the two groups (those who recover and those who die). A useful model will be one that makes successful predictions of recovery / death when tested on new data.

Other criteria for assessing the usefulness of models are explicability, and persuading regulators and other supervisory bodies that you have not introduced a classification or discrimination which is perceived as undesirable (for example one based on gender).

In R, there are a wide range of packages which will perform machine learning techniques. This range changes over time. See for example https://www.r-bloggers.com/what-are-the-best-machine-learning-packages-in-r/ for an overview.

Chapter 21 Summary

What is machine learning?

Machine learning describes a set of methods by which computer algorithms are developed and applied to data to generate information.

Examples can be found in many areas of everyday life, *eg* in targeting online advertising.

Machine learning is increasingly being used in the areas of finance and insurance, *eg* for predicting which borrowers are most likely to default on a loan.

In machine learning, we have a set of features $x_1, x_2, \dots$, each associated with a corresponding output y. The aim is to find a hypothesis, *ie* a function $g(x_1, x_2, \dots)$ that provides a good approximation to the unknown underlying function $y = f(x_1, x_2, \dots)$.

Branches of machine learning

Machine learning algorithms can be divided into:

- supervised learning, where the algorithm is given a set of specific targets to aim for
- unsupervised learning, where the algorithm aims to produce a set of suitable labels
- semi-supervised learning, which involves a combination of supervised and unsupervised learning
- reinforcement learning, where the algorithm aims to improve its performance through repeated use.

Data types

Traditional data analysis in actuarial and statistical applications involves variables of the following types:

- numerical, which may be discrete or continuous
- categorical (ie non-numerical), which may be attribute (dichotomous), nominal or ordinal.

However, machine learning often involves data of much more complicated types, *eg* video footage, and the file sizes can be very large.

Concepts and terminology

Loss function

A key element of machine learning is that we need to apply the results to cases not seen in the data used to develop the algorithm. One way to evaluate a hypothesis is to calculate the predictions it makes and to penalise each incorrect prediction by applying a loss function.

Model evaluation

There are various criteria we can use to determine the best model.

Confusion matrices

Classification models that result in Yes or No outputs can be assessed using a confusion matrix for the test set, which shows:

True Positives (TP)	False Negatives (FN)
False Positives (FP)	True Negatives (TN)

We can use this to calculate:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall (sensitivity)} = \frac{TP}{TP+FN} \qquad F_1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

These all take values in the range 0% (worst) to 100% (best).

A receiver operating characteristic (ROC) curve can be used to compare models. This plots the true positive rate against the false positive rate.

Generalisation error

An upper bound can be determined for the magnitude of out-of-sample errors. This shows that, with a large enough training set, the out-of-sample error can be made as small as desired.

Train-validate-test

A common approach is to split the available data into three parts:

- a training data set (*eg* 60%)
- a validation data set (*eg* 20%)
- a test data set (*eg* 20%).

Parameters and hyper-parameters

The parameters of a model are variables internal to the model whose values are estimated from the data and are used to calculate predictions using the model.

Hyper-parameters are variables external to the model whose values are set in advance by the user and are chosen based on the user's knowledge and experience to produce a model that works well.

Validation and over-fitting

Machine learning involves a trade-off between bias and variance.

The model should have a sufficient number of parameters to produce accurate results, but not so many that it reflects specific features of the training set that we would not expect to be present in new data.

One way of striking a balance between these two is to apply regularisation or penalisation. This involves working with a loss function of the form:

$$L^*(w_1,w_2,\dots,w_J)+\lambda\sum_{j=1}^{J}w_j^2$$

Stages of analysis in machine learning

Machine learning tasks can be broken down into the following stages:

- collecting data
- exploring and preparing the data
- feature scaling (to ensure that the input variables have similar ranges of values)
- splitting the data
- training the model
- validation and testing
- improving model performance.

There is a trade-off here between:

- *bias, ie* the lack of fit of the model to the training data, and
- *variance*, *ie* the tendency for the estimated parameters to reflect the specific data we use for training the model.

It is also important for data analysis to be reproducible and well-documented.

Supervised learning techniques

Penalised generalised linear models

Penalised regression is an adaptation of the method of maximum likelihood where a penalty is applied to constrain the estimated values of the parameters to improve their reliability for making predictions. The method involves maximising the penalised likelihood:

$$\log_e L\left(\beta_0,\ldots,\beta_J \mid x_1,\ldots,x_J\right)-\lambda P(\beta_1,\ldots,\beta_J)$$

If the model is parameterised so that the parameter values are expected to be close to zero, we could use $\sum_{i=1}^{J}\beta_i^2$ (*ridge*) or $\sum_{i=1}^{J}|\beta_i|$ (*LASSO*) for the penalty function $P(\beta_1,\ldots,\beta_J)$.

To select an appropriate number of parameters J (with sample size N), we can minimise:

$$\text{AIC}=\text{Deviance}+2J \quad \text{(Akaike information criterion)}$$

or $$\text{BIC}=\text{Deviance}+\log N\times J \quad \text{(Bayesian information criterion)}$$

Naïve Bayes algorithm

The *naïve Bayes* algorithm uses Bayes' formula to classify items by determining the relative likelihood of each of the possible values for the covariates $x_{1i},\ldots,x_{Ji}$. For example, if the observed outcome is $y_i=1$:

$$P(y_i=1\mid x_{1i},\ldots,x_{Ji})\propto P(y_i=1)\prod_{j=1}^{J}P(x_{ji}\mid y_i=1)$$

This method can only be used with categorical covariates and it assumes that the conditional probabilities are independent.

Decision trees (CART analysis)

Decision trees, also known as classification and regression techniques (CART), classify items by asking a series of questions that home in on the most likely classification. It is important to avoid overfitting. Overfitting can be avoided by applying a stopping criterion or by pruning the decision tree. The simplest method of construction is to use greedy splitting.

Gini index for a decision tree

The 'purity' of one of the final nodes in a decision tree can be assessed by calculating the Gini index:

$$G=\sum_{k=1}^{m}p_k(1-p_k)=1-\sum_{k=1}^{m}p_k^2$$

where p_k is the proportion of sample items of class k present at that node. This gives a value between 0 ('pure') and 1 ('mixed'). So we would aim to minimise this.

The Gini index for a node where there is a split, or for the whole tree, is:

$$G=\sum_{nodes}\frac{n_{node}}{n}\left(1-\sum_{k=1}^{m}p_k^2\right)$$

where the sum is taken over all the nodes involved and n_{node} is the number of items at the node we are currently evaluating.

Unsupervised learning techniques

K-means clustering

The k-means clustering algorithm involves modelling the data values as points in space. Starting from an initial cluster allocation (usually random), the method repeatedly finds the centroid of the data points that have been allocated to each cluster and then reallocates the points to the cluster whose centroid they are nearest to. When this process reaches a stage where no further changes are made, the algorithm has converged to the solution.

This method has the advantages that:

- it uses a simple principle that can easily be explained
- it is highly flexible and can easily be adapted to address any shortcomings
- it is efficient and performs well.

However, it also has the disadvantages that:

- it is less sophisticated than more recent clustering algorithms
- it is not guaranteed to find the optimal set of clusters (because of the random element)
- it requires a reasonable guess as to how many clusters naturally exist in the data.

Perspectives

Machine learning techniques are used by statisticians, data scientists and other quantitative researchers, as well as actuaries. However, the different fields often have different aims and also use different terminology.

Chapter 21 Practice Questions

This is a new topic on the actuarial exam syllabus, so as yet the number of past exam questions is limited.

21.1 (i) Explain the distinction between supervised and unsupervised learning in the context of machine learning.

(ii) State whether each of the following applications involves supervised or unsupervised learning or a mixture of the two, and suggest a suitable machine learning algorithm that could be used in each case:

(a) predicting a university graduate's salary at age 40 based on the subject they studied, the grade they obtained in their degree and their sex

(b) grouping car insurance policyholders into different geographical areas that have similar experience

(c) recommending films for subscribers to watch on a subscription movie channel.

(iii) Explain the following terms relating to machine learning:

(a) hyper-parameters

(b) CART

(c) greedy splitting

(d) clustering.

21.2 A random sample $x_1,\ldots,x_n$ of values has been taken from a $N(\mu,\sigma_0^2)$ distribution, where the value of σ_0 is known but the value of μ is unknown. However, it is believed that the value of μ is close to 100. It has been suggested that μ could be estimated using a penalised log-likelihood function.

(i) Explain the rationale behind this method.

(ii) Suggest why the penalty function $\lambda(\mu-100)^2$ would be suitable to use in this case.

(iii) Show that the estimate of μ derived using this method with the penalty function in part (ii) is:

$$\hat{\mu}=\left(\frac{n}{\sigma_0^2}\bar{x}+200\lambda\right)\Big/\left(\frac{n}{\sigma_0^2}+2\lambda\right)$$

(iv) Comment on how the value of $\hat{\mu}$ calculated using the formula in part (iii) is influenced by the value chosen for the regularisation parameter.

(v) Explain why this method might be preferable in some circumstances to the basic method of maximum likelihood.

21.3 In a sample of size 100, 10% of individuals have a particular feature.

(a) Draw up a confusion matrix for a test that can identify this feature perfectly.

(b) Calculate four measures for the effectiveness of the test, based on the numbers in your matrix.

21.4 A warehouse stores four types of single malt whisky in identical bottles in an underground storeroom, before they are labelled and distributed. Recently the warehouse experienced a major flood in which the stock records were destroyed and the handwritten descriptions on some of the cases were washed off, so that they could no longer be identified.

The warehouse manager has asked you to supervise the process of identifying the type of whisky each of these cases belongs to, so that they can be correctly labelled. A professional whisky taster has provided a report based on a single bottle from each case, which he has classified based on three criteria: Smoky, Fruity, Colour (each on a scale of 1 to 3).

The standard descriptions of the four types are shown in the table below. These can be considered to have a probability of 80% of being correct, while any other description has a probability of 10%. The warehouse manager has also indicated the proportions of each type she believes were in stock at the time of the flood.

Case	Smoky	Fruity	Colour	Proportion in stock
Mactavish	1	3	1	40%
Western Isle	2	1	1	30%
Glenragh	2	2	3	10%
Dogavulin	3	2	2	20%

(i) Show that, under the assumptions of the naïve Bayes model:

$$P(y=A\,|\,x_1,x_2,x_3) \propto P(y=A)\times P(x_1\,|\,y=A)P(x_2\,|\,y=A)P(x_3\,|\,y=A)$$

The taster has described the sample bottle from one of the cases as:

Smoky = 2, Fruity = 2, Colour = 2

(ii) Use the formula from part (i) to estimate how likely this case is to be of each of the four types, and hence recommend how it should be labelled.

(iii) State two advantages and one disadvantage of the naïve Bayes classification method as a machine learning technique.

21.5 A doctor is using k-means clustering with $k=5$ to classify her patients by height and weight. The raw data shows the patients' statistics in m and kg, but she has converted the heights to cm. She used a method based on Euclidean distance, which converged after 3 iterations, giving the following results:

Group	1	2	3	4	5
Height (cm)	165	160	175	150	185
Weight (kg)	55	65	80	90	100

(i) Explain what the value of k represents.

(ii) Explain the reason for the doctor's choice of units.

(iii) Explain what convergence means in this context.

(iv) Three new patients have the following data values:

Mr Blobby:	(1.64m, 91kg)
Miss Twiggy:	(1.87m, 54kg)
Mrs Average:	(1.66m, 64kg)

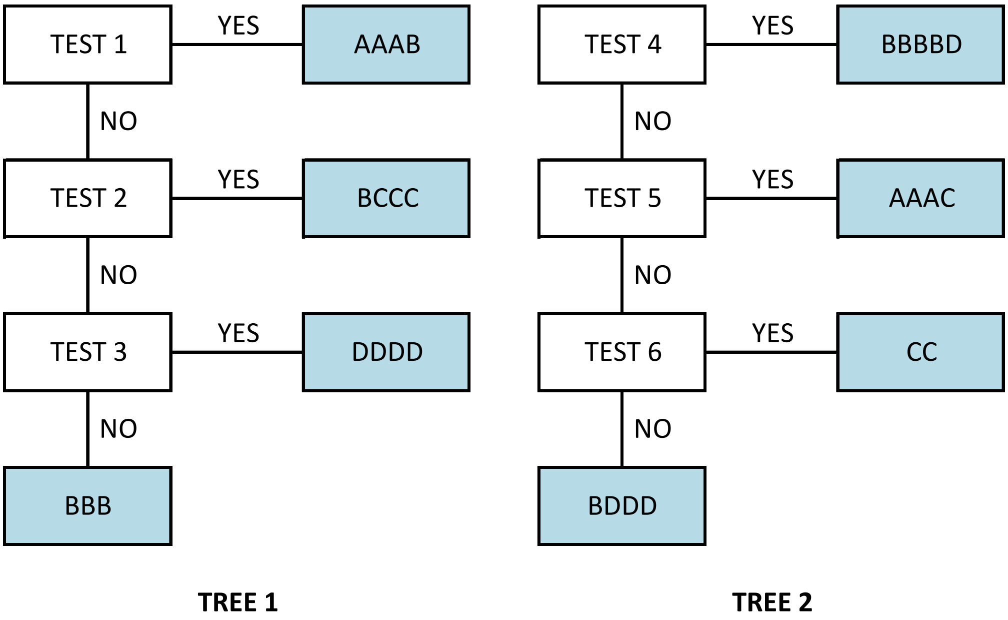
By using a graph, or otherwise, identify the groups to which these patients should be allocated based on their heights and weights.

(v) State whether the results in part (iv) would have differed if the clusters had been obtained using the absolute distance metric $D_{abs}(x,k)=\sum_{j=1}^{J}\left|x_j-k_j\right|$.

21.6 (i) (a) If $p_1 + p_2 + \cdots + p_m = 1$, prove the identity $\sum_{k=1}^{m} p_k \sum_{j=1, j \neq i}^{m} p_j = 1 - \sum_{k=1}^{m} p_k^2$.

(b) Explain how this identity can be used to calculate a measure of the effectiveness of a decision tree.

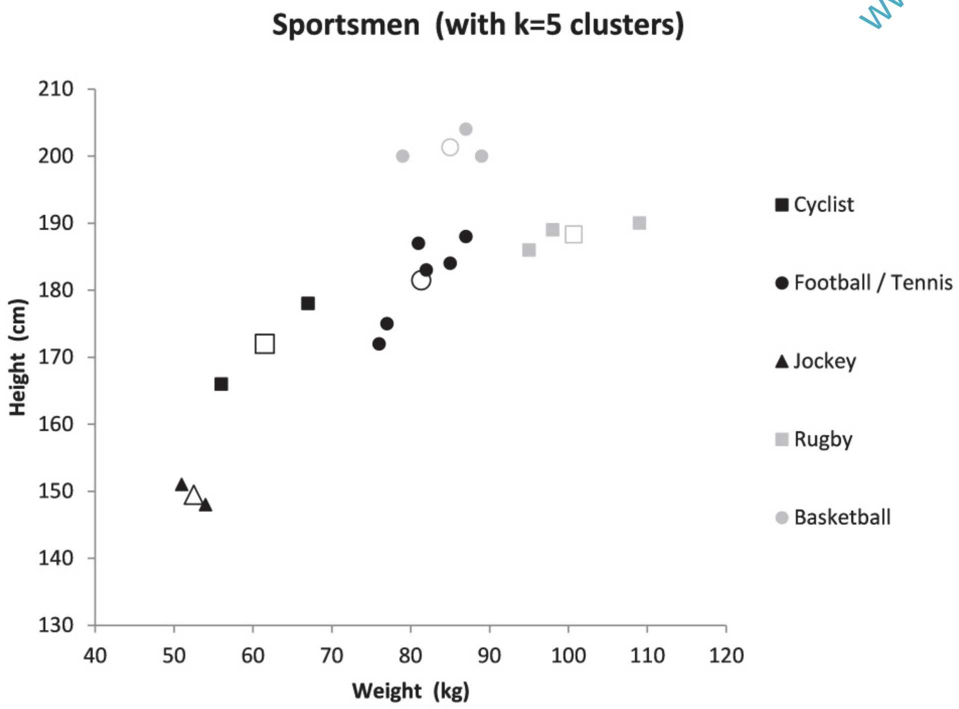
A researcher is considering two possible decision trees to classify items of four different types labelled A, B, C and D. A sample of 15 items classified using each of the trees gave the results shown below.



(ii) (a) Calculate the Gini index for each tree.

(b) Comment on your answers to part (ii)(a).

21.7 It has been decided to repeat the k-means method for the sportsmen with $k=5$, resulting in the following graph:



(i) Explain why the value $k=5$ might have been chosen.

(ii) Use this diagram to identify the most likely sports for three other sportsmen:

Thor (105kg, 200cm), Mo (85kg, 195cm) and Claude (70kg, 165cm)

The solutions start on the next page so that you can separate the questions and solutions.

Chapter 21 Solutions

21.1 (i) ***Supervised and unsupervised learning***

With supervised learning, the desired outcomes for the model are specified in advance and the algorithm aims to reproduce these as closely as possible.

With unsupervised learning, the desired outcomes for the model are not specified in advance and the algorithm aims to assign labels or classify each example in a logical way.

(ii) ***Applications***

(a) Here we would aim to reproduce the salaries for a sample of graduates as closely as possible based on the three variables. So this would be supervised learning. We could use a multiple regression model here.

(b) Here we would hope that the algorithm can find suitable homogeneous groups that we don't know in advance. So this would be unsupervised learning. We could use the k-means clustering algorithm here.

(c) Here we would initially make recommendations based on past data from other subscribers but this would be fine-tuned over time based on whether the user accepted or rejected the recommendations. So this would involve a mixture of supervised and unsupervised learning, or a reinforcement algorithm.

(iii) ***Terminology***

(a) As well as the 'internal' parameters that a model estimates from the data and uses to calculate predictions, machine learning methods often also require hyper-parameters, which are external to the model and whose values are set in advance based on the user's knowledge and experience in order to produce a model that works well. An example would be the number of clusters to aim for with the k-means algorithm.

(b) CART is an abbreviation for classification and regression techniques, which is another names for decision trees. These classify items by asking a series of questions that home in on the most likely classification.

(c) Greedy splitting is a method of constructing a decision tree. At each stage, we just choose the split that appears to be the most effective at separating the remaining elements, without thinking ahead to the consequences this might have for the later splits.

(d) Clustering refers to classifying data into a set of homogeneous groups or clusters, which can be done using methods such as the k-means algorithm.

21.2 (i) ***Rationale***

This method is based on the method of maximum likelihood where we choose the parameter values to maximise the log-likelihood of the data available. This gives the parameter values that best explain the values in the data.

However, we can also apply a penalty function, which is chosen to make the method more likely to produce parameter estimates close to a set of target values that we are expecting. The penalty is subtracted from the log-likelihood and we maximise this adjusted function instead.

(ii) ***Penalty function***

We expect the value of μ to be close to 100. The function $\lambda(\mu-100)^2$ would be suitable to use here because it takes a large positive value if μ is a long way from 100 (in either direction).

(iii) ***Formula***

The likelihood function for the sample is:

$$L=\prod_{i=1}^{n}\frac{1}{\sigma_0\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma_0}\right)^2\right]=\sigma_0^{-n}\exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]\times\text{constant}$$

So the log-likelihood is:

$$\log L=-n\log\sigma_0-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2+\text{constant}$$

and the penalised log-likelihood is:

$$(\log L)^*=-n\log\sigma_0-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2+\text{constant}-\lambda(\mu-100)^2$$

To maximise this, we equate the derivative with respect to the parameter μ to zero:

$$\frac{\partial(\log L)^*}{\partial\mu}=+\frac{2}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)-2\lambda(\mu-100)=\frac{n}{\sigma_0^2}(\bar{x}-\mu)-2\lambda(\mu-100)=0$$

Rearranging this gives:

$$\left(\frac{n}{\sigma_0^2}+2\lambda\right)\mu=\frac{n}{\sigma_0^2}\bar{x}+200\lambda$$

So the estimate of μ is:

$$\hat{\mu}=\left(\frac{n}{\sigma_0^2}\bar{x}+200\lambda\right)\Bigg/\left(\frac{n}{\sigma_0^2}+2\lambda\right)$$

(iv) ***Influence of the regularisation parameter***

The regularity parameter λ will be assigned a non-negative value. (Otherwise, it would correspond to a *reward* rather than a penalty.)

If λ was set equal to zero, there would be no penalty and the method would reduce to maximum likelihood estimation. As expected, the formula in part (iii) then gives the usual formula for the MLE of a normal distribution:

$$\hat{\mu} = \left(\frac{n}{\sigma_0^2}\bar{x} + 0\right) \bigg/ \left(\frac{n}{\sigma_0^2} + 0\right) = \frac{n}{\sigma_0^2}\bar{x} \bigg/ \frac{n}{\sigma_0^2} = \bar{x}$$

If λ is given a very high value, the penalty dominates the calculations and, in the limit, we have:

$$\hat{\mu} = \lim_{\lambda\to\infty}\left(\frac{n}{\sigma_0^2}\bar{x} + 200\lambda\right) \bigg/ \left(\frac{n}{\sigma_0^2} + 2\lambda\right) = \lim_{\lambda\to\infty}\left(\frac{n}{\lambda\sigma_0^2}\bar{x} + 200\right) \bigg/ \left(\frac{n}{\lambda\sigma_0^2} + 2\right) = \frac{200}{2} = 100$$

So now the estimate is equal to the target value of 100.

(v) ***Why this method might be preferable***

The basic (unpenalised) method of maximum likelihood can sometimes lead to unreliable results. The estimated values of the parameters can be very sensitive to the sample data and can vary wildly.

This is most likely to happen when the sample size is small or the likelihood function is very flat so that changes in the parameter values make very little difference to the log-likelihood.

Applying a penalty function encourages the method to produce parameter estimates that are close to the values that would be expected from prior expectations.

21.3 (a) The confusion matrix looks like this:

PREDICTED ACTUAL	YES	NO	TOTAL
YES	TP = 10	FN = 0	10
NO	FP = 0	TN = 90	90
TOTAL	10	90	100

(b) The four measures are:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{10}{10+0} = 100\%, \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{10}{10+0} = 100\%$$

$$F_1 \text{ score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2\times 100\%\times 100\%}{100\%+100\%} = 100\%$$

$$\text{False positive rate} = \frac{\text{FP}}{\text{TN}+\text{FP}} = \frac{0}{90+0} = 0\%$$

21.4 (i) ***Formula for naïve Bayes method***

$$P(y=A\,|\,x_1,x_2,x_3) = \frac{P(y=A,x_1,x_2,x_3)}{P(x_1,x_2,x_3)} \qquad \text{(definition of conditional probability)}$$

$$= \frac{P(x_1,x_2,x_3\,|\,y=A)P(y=A)}{P(x_1,x_2,x_3)} \qquad \text{(definition of conditional probability in reverse)}$$

$$= \frac{P(x_1\,|\,y=A)P(x_2\,|\,y=A)P(x_3\,|\,y=A)P(y=A)}{P(x_1,x_2,x_3)} \qquad \text{(independence assumption)}$$

$$\propto P(y=A)\times P(x_1\,|\,y=A)P(x_2\,|\,y=A)P(x_3\,|\,y=A) \quad \text{(ignore constant factor)}$$

(ii) ***Probabilities for each type***

Let y denote the type and let x_1,x_2,x_3 denote the three descriptions (Smoky, Fruity, Colour).

We can then apply the result from part (i) to calculate the probability that this case is a Mactavish whisky $(y=M)$, given that it has been described as Smoky $(S=2)$, Fruity $(F=2)$ and is medium Colour $(C=2)$:

$$P(y=M\,|\,S=2,F=2,C=2) \propto P(y=M)\times P(S=2\,|\,y=M)P(F=2\,|\,y=M)P(C=2\,|\,y=M)$$

$$= 0.4\times 0.1\times 0.1\times 0.1 = 0.0004$$

Similarly for the other three types:

$$P(y=W\,|\,S=2,F=2,C=2) \propto P(y=W)\times P(S=2\,|\,y=W)P(F=2\,|\,y=W)P(C=2\,|\,y=W)$$

$$= 0.3\times 0.8\times 0.1\times 0.1 = 0.0024$$

$$P(y=G\,|\,S=2,F=2,C=2) \propto P(y=G)\times P(S=2\,|\,y=G)P(F=2\,|\,y=G)P(C=2\,|\,y=G)$$

$$= 0.1\times 0.8\times 0.8\times 0.1 = 0.0064$$

$$P(y=D\,|\,S=2,F=2,C=2) \propto P(y=D)\times P(S=2\,|\,y=D)P(F=2\,|\,y=D)P(C=2\,|\,y=D)$$

$$= 0.2\times 0.1\times 0.8\times 0.8 = 0.0128$$

So the probabilities for each type are in the ratio:

$$\text{M}:\text{W}:\text{G}:D \;=\; 4:24:64:128 = \frac{4}{220}:\frac{24}{220}:\frac{64}{220}:\frac{128}{220} = 0.0182:0.1091:0.2909:0.5818$$

So the recommendation would be that this case is a Dogavulin whisky, as this has a far higher probability (58%) than the other three types.

(iii) ***Advantages and disadvantages***

Advantages of the naïve Bayes method include:

- it is easy to apply
- it requires very little data.

The main disadvantage is that it assumes that the conditional probabilities are independent (which can be a poor approximation when the variables are correlated).

21.5 (i) ***Meaning of k***

k is a hyper-parameter specifying the number of clusters the algorithm should aim to produce – in this case, 5.

(ii) ***Choice of units***

The weights in the original data provided were given in units of kilograms. These cover a range of values of about 50kg.

The heights in the original data provided were given in units of metres. These cover a range of values of about 0.50m.

So with units of (kg, m) the range for the weights is about 100 times greater, which would mean that the weights would totally dominate the calculations and the heights would effectively be ignored.

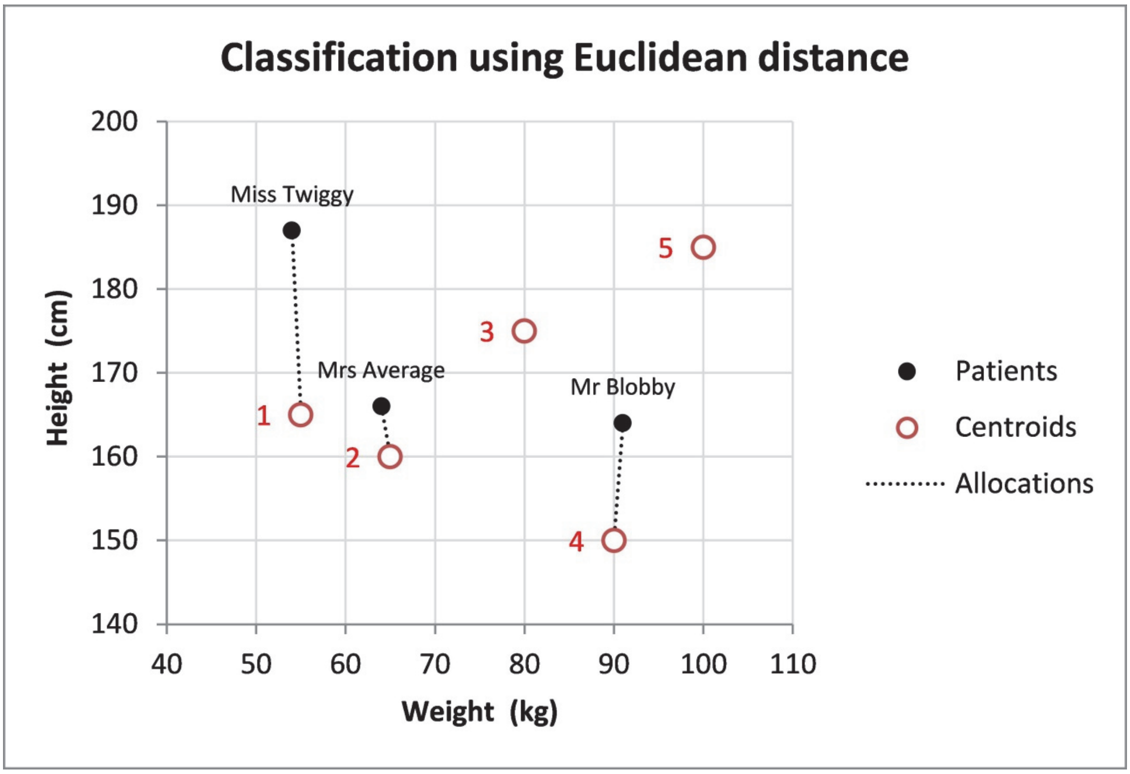
However, when the doctor converts the heights to centimetres, the range of values is then about 50cm, which is numerically very similar to the range for the weights. This gives the two variables a similar weighting in the calculations.

(iii) ***Convergence***

The algorithm involves repeatedly finding the centroid of the data points that have been allocated to each cluster and then reallocating the points to the cluster whose centroid they are nearest to. When this process reaches a stage where no further changes are made, the algorithm has converged to the solution.

(iv) ***Classification of new patients***

If we plot the patients and the centroids for the clusters on a graph, we can easily see which centroid each patient is nearest to, and hence classify these patients.



If we do the calculations, we find that the shortest Euclidean distances D_i (for the centroids $i = 1,2,\ldots,5$) are:

Mr Blobby: (1.64m, 91kg) $\rightarrow D_4 = \sqrt{(164-150)^2 + (91-90)^2} = 14.04$

Miss Twiggy: (1.87m, 54kg) $\rightarrow D_1 = \sqrt{(187-165)^2 + (54-55)^2} = 22.02$

Mrs Average: (1.66m, 64kg) $\rightarrow D_2 = \sqrt{(166-160)^2 + (64-65)^2} = 6.08$

(v) ***Absolute distance***

The absolute distance measures the distance between points assuming that we can only move horizontally or vertically.

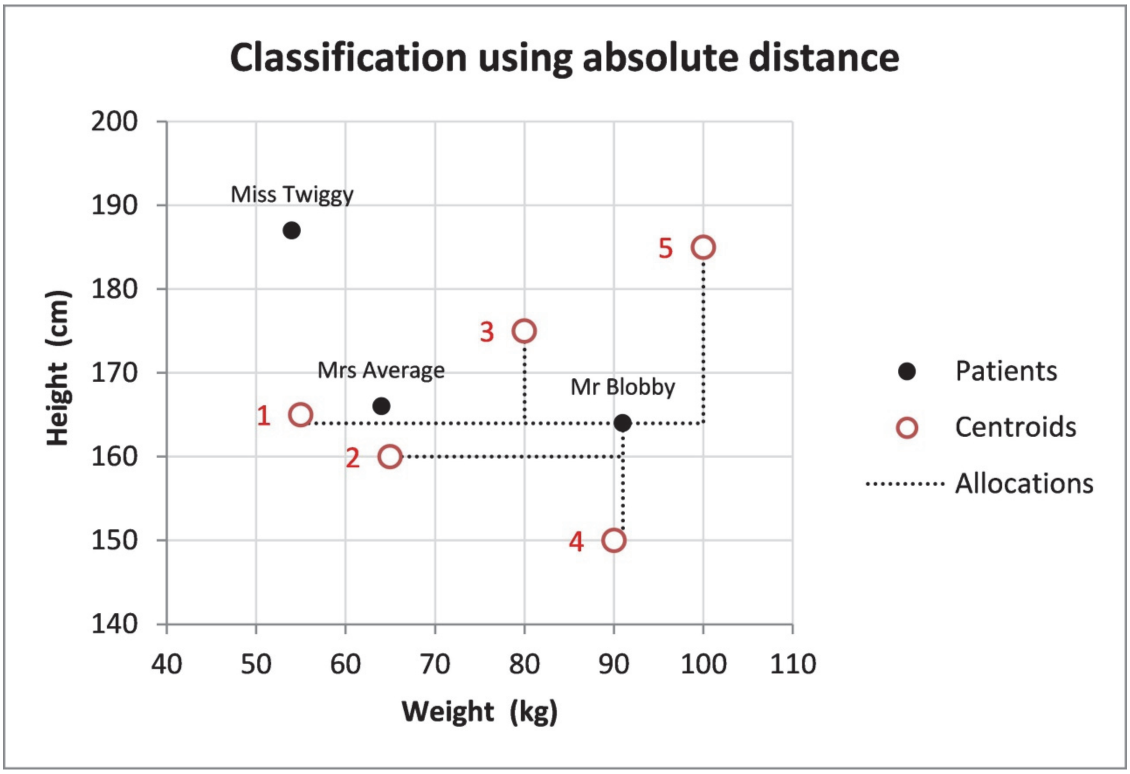
Here we have two dimensions (the two variables Weight and Height), so $J = 2$. With this metric the distance between the two points (x_1, x_2) and (k_1, k_2) is:

$$|x_1 - k_1| + |x_2 - k_2|$$

With this metric, Mr Blobby's distance from the centroid for cluster 4 is:

Mr Blobby: (1.64m, 91kg) $\rightarrow D_4 = |164 - 150| + |91 - 90| = 14 + 1 = 15$

The diagram below shows the distance to each centroid for Mr Blobby.



We can see that using absolute distances would give the same answers as Euclidean distance for all three of these patients.

21.6 (i)(a) ***Prove the identity***

$$\sum_{k=1}^{m} p_k \sum_{j=1, j\neq i}^{m} p_j = \sum_{k=1}^{m} p_k \left(1 - p_k\right) = \sum_{k=1}^{m} \left(p_k - p_k^2\right) = \sum_{k=1}^{m} p_k - \sum_{k=1}^{m} p_k^2 = 1 - \sum_{k=1}^{m} p_k^2$$

(i)(b) ***Measuring the effectiveness of a decision tree***

We can measure the effectiveness of a decision tree by examining the 'purity' of the groupings produced for a training data set.

This can be done by multiplying the proportion of items of type k at each node by the proportions for each other type $j \neq k$ and summing. These values are then weighted by the number of items at that node to calculate an overall measure called the Gini index. Using the identity above leads to the following formula:

$$G = \sum_{nodes} \frac{n_{node}}{n}\left(1 - \sum_{k=1}^{m} p_k^2\right)$$

where the sum is taken over all the nodes and n_{node} is the number of items at the node we are currently evaluating.

(ii)(a) ***Calculate the Gini index***

In Tree 1 the top final node contains AAAB, *ie* 3 A's and 1 B. So the proportions are $p_1 = \frac{3}{4}$ and $p_2 = \frac{1}{4}$. So the Gini index for this node is:

$$G = 1 - \left(\tfrac{3}{4}\right)^2 - \left(\tfrac{1}{4}\right)^2 = \tfrac{3}{8}.$$

The next node contains BCCC, *ie* 3 C's and 1 B. So it also has a Gini index of $\frac{3}{8}$.

The other two final nodes are 'pure', as each contains a single label. So they have a Gini index of 0.

To find the Gini index for the whole of Tree 1, we need to work out the weighted average of the Gini index at each of the final nodes weighted by the number of elements they contain, *ie* 4, 4, 4, and 3 (making a total of 15). So this gives:

$$G = \left(\frac{4}{15} \times \frac{3}{8}\right) + \left(\frac{4}{15} \times \frac{3}{8}\right) + \left(\frac{4}{15} \times 0\right) + \left(\frac{3}{15} \times 0\right) = \frac{1}{5} = 0.2$$

Similarly, for Tree 2 the Gini index is:

$$G = \frac{5}{15}\left[1 - \left(\tfrac{4}{5}\right)^2 - \left(\tfrac{1}{5}\right)^2\right] + \frac{4}{15}\left[1 - \left(\tfrac{3}{4}\right)^2 - \left(\tfrac{1}{4}\right)^2\right] + \frac{2}{15}\left[1 - 1^2\right] + \frac{4}{15}\left[1 - \left(\tfrac{1}{4}\right)^2 - \left(\tfrac{3}{4}\right)^2\right]$$

$$= \left(\frac{5}{15} \times \frac{8}{25}\right) + \left(\frac{4}{15} \times \frac{3}{8}\right) + 0 + \left(\frac{4}{15} \times \frac{3}{8}\right) = \frac{23}{75} = 0.307$$

(ii)(b) ***Comment***

The Gini index for Tree 1 is lower than for Tree 2. So, using this criterion, Tree 1 would be preferred.

21.7 (i) ***Choice of value of k***

The graph with $k = 6$ correctly identified the jockeys, cyclists, basketball players and rugby players into logical groups. However, the footballers and tennis players' characteristics are too similar for us to be able to distinguish them effectively using this approach. So a sensible approach is to combine the footballers and tennis players into a single group and look for 5 logical clusters.

(ii) ***Identify the sports***

In each case we allocate each sportsman to the group whose centroid he is closest to. This gives:

Thor = Rugby, Mo = Basketball, Claude = Cyclist

End of Part 5

What next?

1. Briefly **review** the key areas of Part 5 and/or re-read the **summaries** at the end of Chapters 17 to 21.

2. Ensure you have attempted some of the **Practice Questions** at the end of each chapter in Part 5. If you don't have time to do them all, you could save the remainder for use as part of your revision.

3. Attempt **Assignment X5**.

And finally …

Good luck!

Subject CS2: Assignment X1

2019 Examinations

Time allowed: 2¾ hours

Instructions to the candidate

1. *Please:*

   - *attempt all of the questions, as far as possible under exam conditions*
   - ***begin your answer to each question on a new page***
   - ***leave at least 2cm margin on all borders***
   - *write in black ink using a medium-sized nib because we will be unable to mark illegible scripts*
   - *note that assignment marking is not included in the price of the course materials. Please purchase Series Marking or a Marking Voucher before submitting your script.*
   - *note that we only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2019 exams.*

2. *Please* ***do not****:*

   - *use headed paper*
   - *use highlighting in your script.*

At the end of the assignment

If your script is being marked by ActEd, please follow
the instructions on the reverse of this page.

In addition to this paper, you should have available actuarial tables and an electronic calculator.

Submission for marking

You should aim to submit this script for marking by the recommended submission date. The recommended and deadline dates for submission of this assignment are listed on the summary page at the back of this pack and on our website at **www.ActEd.co.uk**.

Scripts received after the deadline date will not be marked, unless you are using a Marking Voucher. *It is your responsibility to ensure that scripts reach ActEd in good time*. If you are using Marking Vouchers, then please make sure that your script reaches us by the Marking Voucher deadline date to give us enough time to mark and return the script before the exam.

When submitting your script, please:

- complete the cover sheet, including the checklist
- scan your script, cover sheet (and Marking Voucher if applicable) and save as a pdf document, then email it to: ActEdMarking@bpp.com
- **do not submit a photograph of your script**
- **do not include the question paper in the scan.**

In addition, please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* 'CS2 Assignment X1 No. 12345', inserting your ActEd Student Number for 12345.
- The assignment should be scanned the **right way up** (so that it can be read normally without rotation) and as a single document. We cannot accept individual files for each page.
- Please set the resolution so that the script is legible and the resulting PDF **is less than 4 MB** in size.
- Do not protect the PDF in any way (otherwise the marker cannot return the script to ActEd, which causes delays).
- Please include the 'feedback from marker' sheet when scanning.
- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

Subject CS2: Assignment X1

2019 Examinations

Please complete the following information:

Name:

ActEd Student Number (see Note below):

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting it will help us to process your scripts quickly. If you do not know your ActEd Student Number, please email us at ActEd@bpp.com.

Your ActEd Student Number is not the same as your IFoA Actuarial Reference Number or ARN.

Number of following pages: _______

Please put a tick in this box if you have solutions and a cross if you do not: □

Please tick here if you are allowed extra time or other special conditions in the profession's exams (if you wish to share this information): □

Time to do assignment
(see Note below): _____ hrs _____ mins

Under exam conditions
(delete as applicable): yes / nearly / no

Note: If you take more than 2¾ hours, you should indicate how much you completed within this exam time so that the marker can provide useful feedback on your progress.

Score and grade for this assignment (to be completed by marker):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q9	**Total**	
__ 2	__ 5	__ 5	__ 5	__ 6	__ 7	__ 7	__ 14	__ 14	__ 15	__ 80	=_____%

Grade: A B C D E **Marker's initials:** _________

Please tick the following checklist so that your script can be marked quickly. Have you:

[ ] Checked that you are using the latest version of the assignments, *ie* 2019 for the sessions leading to the 2019 exams?

[ ] Written your full name in the box above?

[ ] Completed your ActEd Student Number in the box above?

[ ] Recorded your attempt conditions?

[ ] Numbered all pages of your script (excluding this cover sheet)?

[ ] Written the total number of pages (excluding the cover sheet) in the space above?

[ ] Included your Marking Voucher or ordered Series X Marking?

Please follow the instructions on the previous page when submitting your script for marking.

Feedback from marker

Notes on marker's section

The main objective of marking is to provide specific advice on how to improve your chances of success in the exam. The most useful aspect of the marking is the comments the marker makes throughout the script, however you will also be given a percentage score and the band into which that score falls. Each assignment tests only part of the course and hence does not give a complete indication of your likely overall success in the exam. However it provides a good indicator of your understanding of the material tested and the progress you are making with your studies:

A = Excellent progress B = Good progress C = Average progress
D = Below average progress E = Well below average progress

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

X1.1 Consider the following four stochastic processes:

counting process, time series, compound Poisson process, simple random walk

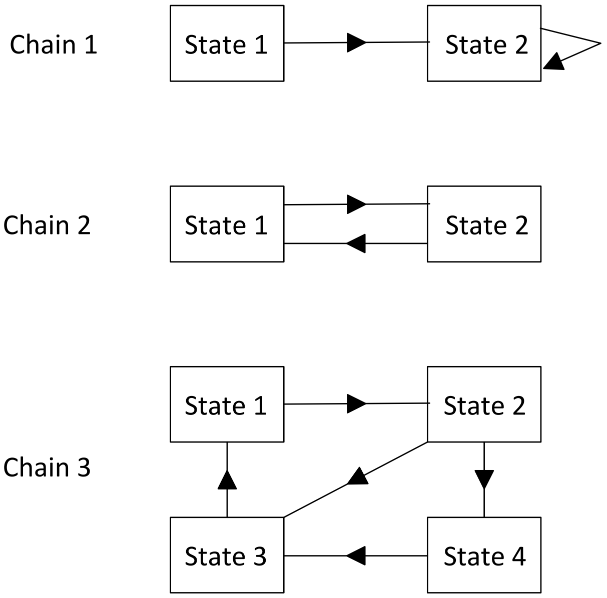
Place each process in a separate cell of the following table, so that each cell correctly describes the state space and the time set of the process placed in it.

		Time set	
		Discrete	*Continuous*
State Space	*Discrete*		
	Continuous		

[2]

X1.2 (i) Explain what it means for a Markov chain to be periodic with period d . [2]

(ii) The diagrams below show three Markov chains, where arrows indicate a non-zero transition probability.



Explain whether each chain is periodic or aperiodic, giving the period where relevant. [3]

[Total 5]

X1.3 (i) Give a mathematical definition of the Markov property. [2]

(ii) A stochastic process $X(t)$ has independent increments. Prove that it also has the Markov property. [3]

[Total 5]

X1.4 The weather in a particular city during the summer months is very variable. A research team has recorded the weather each day during the first three weeks of July. They use the notation S to denote a sunny day, C to denote a cloudy day, and R to denote a rainy day. Their results are as follows:

Week 1: SSRCSCC

Week 2: SCRRCSS

Week 3: RCCSCCS

One of the team suggests that the weather each day depends only on the weather for the previous day and decides to fit a Markov chain to the data.

(i) Estimate the transition probabilities for the Markov chain. [3]

(ii) The team plans to hold its summer barbecue on 23 July. Estimate the probability that this will be a sunny day. [2]

[Total 5]

X1.5 The time, in years, until a boiler breaks down is exponentially distributed with parameter λ, where:

$$\lambda = \begin{cases} \frac{1}{4} & \text{if the boiler has not previously broken down} \\ \frac{1}{2} & \text{if the boiler has broken down once previously} \\ 1 & \text{if the boiler has broken down more than once previously} \end{cases}$$

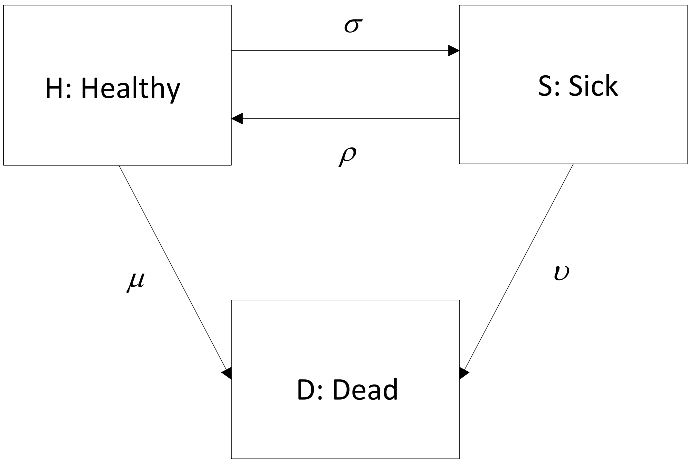
Once a boiler has broken down 10 times, it is scrapped. If a boiler has broken down fewer than 10 times, it is immediately repaired.

(i) By solving a suitable differential equation or otherwise, calculate the probability that a new boiler will break down more than once in the next 5 years. [5]

(ii) Calculate the expected lifetime of a new boiler. [1]

[Total 6]

X1.6 A life insurance company prices its long-term sickness policies using the following time-homogeneous Markov model:



For a group of policyholders over a 1-year period there are:

- 34 transitions from State H to State S
- 26 transitions from State S to State H
- 2 deaths from State H
- 7 deaths from State S.

The total time spent in State H is 904 years and the total time spent in State S is 112 years.

(i) Write down the likelihood function for these data values. [2]

(ii) Show that the maximum likelihood estimate of ρ is 0.23214. (You may assume that this gives a maximum.) [2]

(iii) Construct an approximate 95% confidence interval for ρ. [3]

[Total 7]

X1.7 Derive from first principles the Kolmogorov forward differential equation:

$$\frac{\partial}{\partial t}\,{}_t p_x = -{}_t p_x\,\mu_{x+t}$$

Hence show that:

$${}_t p_x = \exp\left(-\int_0^t \mu_{x+s}\,ds\right)$$ [7]

X1.8 A company assesses the credit-worthiness of various firms every quarter; the ratings are, in order of decreasing merit, *A, B, C* and *D* (default). Historical data support the view that the credit rating of a typical firm evolves as a Markov chain with transition matrix:

$$P=\begin{pmatrix} 1-\alpha-\alpha^2 & \alpha & \alpha^2 & 0 \\ \alpha & 1-2\alpha-\alpha^2 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1-2\alpha-\alpha^2 & \alpha \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

for some parameter α.

(i) Draw the transition graph of the chain. [2]

(ii) Determine the range of values of α for which the matrix P is a valid transition matrix. [4]

(iii) Explain whether the chain is irreducible and/or aperiodic. [3]

Company XYZ has a rating of *B* in Quarter 1.

(iv) Assuming that $\alpha = 0.1$, calculate:

(a) the expected number of quarters for which the company will hold a *B* rating before the rating changes.

(b) the probability that the first rating change for Company XYZ is an upgrade.

[5]

[Total 14]

X1.9 A no-claims discount system operated by an insurer selling private medical insurance has four levels of discount:

Level 1: 0% discount

Level 2: 10% discount

Level 3: 20% discount

Level 4: 25% discount

The insurer operates an accelerated discount scheme with the following rules:

- New policyholders start on Level 1.
- Following a year with one or more claims, move to the next lower level, or remain at Level 1.
- Following a claim-free year:
  - move up one level, or remain at Level 4, if, in the year before the most recent year, there were one or more claims or no insurance was in force
  - move up two levels, or move to Level 4 or remain at Level 4 if, in the year before the most recent year, there were no claims.

For any policyholder the probability of a claim-free year is 0.8.

(i) A stochastic process $X(t)$ is to be used to model the NCD system. $X(t)$ will denote the policyholder's discount Level (1, 2, 3 or 4) in year t. Explain why $X(t)$ does not possess the Markov property. [1]

(ii) Explain how the number of states that $X(t)$ can take can be increased to produce a new process $Y(t)$ that is Markov. [2]

(iii) Draw and label the transition graph for $Y(t)$. [2]

(iv) Write down the transition matrix of $Y(t)$. [1]

(v) Show that the conditions sufficient for $Y(t)$ to have a unique stationary distribution that will be reached are satisfied. [3]

(vi) Calculate the long-run probability that the policyholder is at discount level 2. [5]

[Total 14]

X1.10 Patients arriving at the Accident and Emergency department of a hospital (State *A*) wait for an average of one hour before being classified by a junior doctor as requiring in-patient treatment (State *I*), out-patient treatment (State *O*) or further investigation (State *F*). Only one new arrival in ten is classified as an in-patient, five in ten as out-patients.

If needed, further investigation takes an average of 3 hours, after which 50% of cases are discharged (State *D*), 25% are sent to receive out-patient treatment and 25% admitted as in-patients.

Out-patient treatment takes an average of 2 hours to complete, in-patient treatment an average of 60 hours. Both result in discharge.

It is suggested that a time-homogeneous Markov jump process with states *A*, *F*, *I*, *O* and *D* could be used to model the progress of patients through the system, with the ultimate aim of reducing the average time spent in the hospital.

(i) Sketch the transition diagram for this model, giving numerical values for the transition rates, and write down the generator matrix. [5]

(ii) Calculate the proportion of patients who eventually receive in-patient treatment. [1]

(iii) Determine an expression in terms of t for each of the following:

(a) the probability that a newly-arrived patient is yet to be classified by a junior doctor after t hours

(b) the probability that a newly-arrived patient is undergoing further investigation after t hours. [4]

(iv) Calculate the expected total time until discharge for a newly-arrived patient. [3]

(v) Explain whether a time-homogeneous process is appropriate for modelling this situation. [2]

[Total 15]

END OF PAPER

Subject CS2: Assignment X2

2019 Examinations

Time allowed: 2¾ hours

Instructions to the candidate

1. *Please:*

   - *attempt all of the questions, as far as possible under exam conditions*
   - ***begin your answer to each question on a new page***
   - ***leave at least 2cm margin on all borders***
   - *write in black ink using a medium-sized nib because we will be unable to mark illegible scripts*
   - *note that assignment marking is not included in the price of the course materials. Please purchase Series Marking or a Marking Voucher before submitting your script.*
   - *note that we only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2019 exams.*

2. *Please **do not**:*

   - *use headed paper*
   - *use highlighting in your script.*

At the end of the assignment

If your script is being marked by ActEd, please follow the instructions on the reverse of this page.

In addition to this paper, you should have available actuarial tables and an electronic calculator.

Submission for marking

You should aim to submit this script for marking by the recommended submission date. The recommended and deadline dates for submission of this assignment are listed on the summary page at the back of this pack and on our website at **www.ActEd.co.uk**.

Scripts received after the deadline date will not be marked, unless you are using a Marking Voucher. *It is your responsibility to ensure that scripts reach ActEd in good time*. If you are using Marking Vouchers, then please make sure that your script reaches us by the Marking Voucher deadline date to give us enough time to mark and return the script before the exam.

When submitting your script, please:

- complete the cover sheet, including the checklist
- scan your script, cover sheet (and Marking Voucher if applicable) and save as a pdf document, then email it to: ActEdMarking@bpp.com
- **do not submit a photograph of your script**
- **do not include the question paper in the scan.**

In addition, please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* ‘CS2 Assignment X2 No. 12345’, inserting your ActEd Student Number for 12345.
- The assignment should be scanned the **right way up** (so that it can be read normally without rotation) and as a single document. We cannot accept individual files for each page.
- Please set the resolution so that the script is legible and the resulting PDF **is less than 4 MB** in size.
- Do not protect the PDF in any way (otherwise the marker cannot return the script to ActEd, which causes delays).
- Please include the ‘feedback from marker’ sheet when scanning.
- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

Subject CS2: Assignment X2

2019 Examinations

Please complete the following information:

Name:

ActEd Student Number (see Note below):

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting it will help us to process your scripts quickly. If you do not know your ActEd Student Number, please email us at ActEd@bpp.com.

Your ActEd Student Number is not the same as your IFoA Actuarial Reference Number or ARN.

Number of following pages: _______

Please put a tick in this box if you have solutions and a cross if you do not: □

Please tick here if you are allowed extra time or other special conditions in the profession's exams (if you wish to share this information): □

Time to do assignment
(see Note below): _____ hrs _____ mins

Under exam conditions
(delete as applicable): yes / nearly / no

Note: If you take more than 2¾ hours, you should indicate how much you completed within this exam time so that the marker can provide useful feedback on your progress.

Score and grade for this assignment (to be completed by marker):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	**Total**	
__ 2	__ 3	__ 5	__ 7	__ 7	__ 7	__ 8	__ 8	__ 8	__ 12	__ 13	__ 80	=_____%

Grade: A B C D E **Marker's initials:** _________

Please tick the following checklist so that your script can be marked quickly. Have you:

[ ] Checked that you are using the latest version of the assignments, *ie* 2019 for the sessions leading to the 2019 exams?

[ ] Written your full name in the box above?

[ ] Completed your ActEd Student Number in the box above?

[ ] Recorded your attempt conditions?

[ ] Numbered all pages of your script (excluding this cover sheet)?

[ ] Written the total number of pages (excluding the cover sheet) in the space above?

[ ] Included your Marking Voucher or ordered Series X Marking?

[ ] Rated your X1 marker at **www.ActEd.co.uk/marking**?

Please follow the instructions on the previous page when submitting your script for marking.

Feedback from marker

Notes on marker's section

The main objective of marking is to provide specific advice on how to improve your chances of success in the exam. The most useful aspect of the marking is the comments the marker makes throughout the script, however you will also be given a percentage score and the band into which that score falls. Each assignment tests only part of the course and hence does not give a complete indication of your likely overall success in the exam. However it provides a good indicator of your understanding of the material tested and the progress you are making with your studies:

A = Excellent progress B = Good progress C = Average progress
D = Below average progress E = Well below average progress

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

X2.1 A certain species of insect is subject to a constant force of mortality of λ per day. Determine an exact expression in terms of λ for the curtate expectation of life of a new-born insect. [2]

X2.2 In a certain population, the force of mortality at age x is given by:

$$\mu_x = \begin{cases} 0.02 & 70 < x \le 75 \\ 0.04 & 75 < x \le 80 \\ 0.07 & 80 < x \le 85 \end{cases}$$

Calculate the probability that a life now aged exactly 73 will die between exact age 79 and exact age 82. [3]

X2.3 Let T_x denote the complete future lifetime of a life now aged exactly x.

(i) Define in terms of probabilities involving T_x :

(a) the survival function, $S_x(t)$

(b) the force of mortality, μ_{x+t} . [2]

Under the Weibull model for mortality, the survival function is of the form:

$$S_x(t) = \exp(-\alpha t^{\beta})$$

where $\alpha, \beta > 0$.

(ii) Derive an expression in terms of α and β for μ_{x+t} under this model. [2]

(iii) Determine an expression involving α for $E(T_x)$ in the case when $\beta = 1$. [1]

[Total 5]

X2.4 (i) (a) Define a Markov jump process.

(b) Explain the condition needed for such a process to be time-homogeneous. [3]

(ii) (a) Give formulae for calculating the maximum likelihood estimates of the transition rates in a time-homogeneous Markov jump process, defining any notation that you use.

(b) Outline the principal difficulties in fitting a Markov jump process model with time-inhomogeneous rates, and explain how these might be overcome. [4]

[Total 7]

X2.5 A certain variety of tomato is susceptible to blight, which is always fatal. A researcher decides to model the life cycle of the tomato using a multiple state model with the following states:

1. not suffering from blight
2. suffering from blight
3. dead.

The transition rates are dependent on the age of the plant and are as follows:

- $\mu(x)$ is the mortality rate at exact age x of a blight-free plant
- $\sigma(x)$ is the rate of contracting blight at exact age x
- $\upsilon(x)$ is the mortality rate at exact age x of a plant suffering from blight.

(i) Draw and label a transition diagram for this multiple state model. [2]

Let $p_{ij}(x,y)$ denote the probability that a plant is in State j at age y ($y \geq x$) given that it was in State i at age x.

(ii) Write down an expression involving transition rates for each of the following probabilities:

(a) $p_{11}(x,x+t)$

(b) $p_{22}(x,x+t)$ [3]

(iii) Write down an integral expression for $p_{12}(x,x+t)$ in terms of transition rates and the probabilities in (ii). [2]

[Total 7]

X2.6 (i) Explain the differences between random censoring and Type I censoring in the context of an investigation into the mortality of life insurance policyholders. Include in your explanation a statement of the circumstances in which the censoring will be random, and the circumstances in which it will be Type I, and give an example of each. [4]

(ii) Explain what is meant by non-informative censoring in the investigation in (i). Describe a situation in which censoring might be informative in this investigation. [3]

[Total 7]

X2.7 (i) State the age range over which Gompertz' Law is an appropriate model for human mortality. [1]

(ii) Show that, under Gompertz' Law, the probability of survival from exact age x to exact age $x+t$ is equal to:

$$\left[\exp\left(-\frac{B}{\ln c}\right)\right]^{c^x(c^t-1)}$$

[3]

For a certain population, estimates of survival probabilities are available as follows:

$${}_5p_{60} = 0.912$$

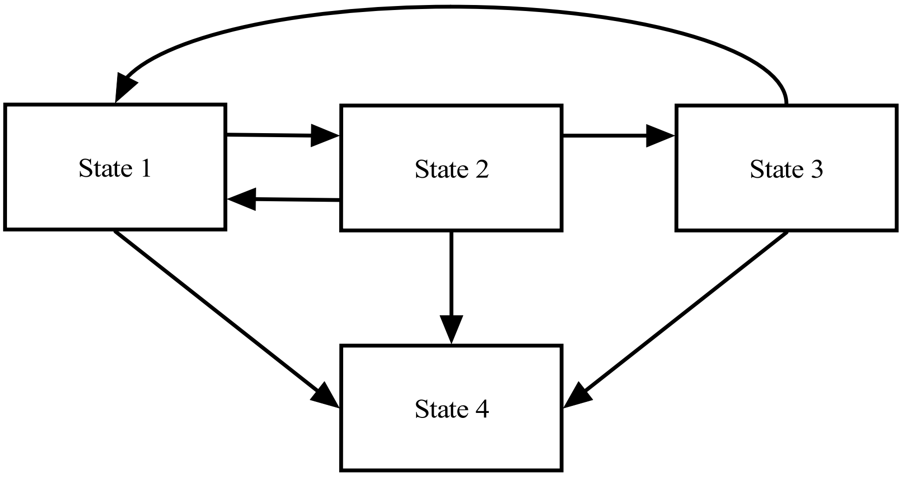
$${}_{10}p_{60} = 0.804$$

(iii) Calculate values of B and c consistent with these observations. [4]

Hint: $c^{10}-1=(c^5-1)(c^5+1)$

[Total 8]

X2.8 Consider the following multiple state model in which $S(t)$, the state occupied at time t by a life initially aged x, is assumed to follow a continuous-time Markov process.



Let μ_{x+t}^{ij} denote the force of transition at age $x+t$ ($t \geq 0$) from State i to State j, and let ${}_tp_x^{ij} = P\left(S(t)=j \mid S(0)=i\right)$.

(i) Derive from first principles the forward differential equation:

$$\frac{\partial}{\partial t}\,{}_tp_x^{21} = {}_tp_x^{22}\,\mu_{x+t}^{21} + {}_tp_x^{23}\,\mu_{x+t}^{31} - {}_tp_x^{21}\left(\mu_{x+t}^{12} + \mu_{x+t}^{14}\right)$$

stating all the assumptions that you make. [5]

(ii) Write down forward differential equations for ${}_tp_x^{23}$ and ${}_tp_x^{32}$. [3]

[Total 8]

X2.9 Mrs Pye, the baker, makes delicious custard tarts. One day last week she made 16 tarts and placed them for sale in her shop at 8am. During the rest of the day, the following tart-related events took place.

Time	Event
8.30am	A man bought a tart on his way to work.
10.00am	A woman bought two tarts.
11.00am	Mrs Pye's clumsy assistant accidentally knocked one of the tarts on to the floor, meaning that it couldn't be sold.
12.30pm	Some students from the local college bought 4 tarts.
1.00pm	Mrs Pye ate one of the tarts during her lunch break.
2.00pm	A family bought 3 tarts.
3.00pm	Two more tarts were sold.
4.00pm	The shop closed and the assistant took the remaining tarts home.

(i) Calculate the Kaplan-Meier estimate of the probability that a tart is sold before closing time. [4]

(ii) Sketch the hazard function, $h(t)$, implied by the Kaplan-Meier model of custard tart sales. [4]

[Total 8]

X2.10 As part of a clinical trial, a statistician is studying the survival rates of patients who have undergone a certain surgical procedure. Below is an extract from the statistician's data. Patients were observed from their date of operation until their date of exit.

Patient number	Date of operation	Date of exit	Reason for exit
1	1 February 2017	1 January 2018	Censored
2	1 April 2017	1 October 2017	Death
3	1 April 2017	1 January 2018	Censored
4	1 July 2016	1 July 2017	Censored
5	1 August 2017	1 January 2018	Censored
6	1 November 2016	1 January 2017	Death
7	1 January 2017	1 January 2018	Censored
8	1 March 2017	1 November 2017	Death
9	1 May 2017	1 November 2017	Death
10	1 June 2017	1 January 2018	Censored

You can assume that the censoring was non-informative with regard to the survival of any individual patient.

(i) Calculate the Nelson-Aalen estimate of the cumulative hazard function, $\Lambda(t)$, where t is the time in months since having the operation. [4]

(ii) Hence calculate an estimate of the survival function for patients who have had this operation. [2]

(iii) Construct an approximate 95% confidence interval for the probability that a patient survives for at least 10 months after having the operation. [4]

(iv) Comment on the statement that at least 90% of patients survive for 10 months or more after having the operation. [2]

[Total 12]

X2.11 Suppose that in a time-inhomogeneous Poisson process the transition rate at time t is $\lambda(t)$.

(i) Sketch the transition diagram for this process. [2]

(ii) (a) Write down the matrix forms of the Kolmogorov forward and backward differential equations for this process.

(b) Hence, or otherwise, give the Kolmogorov forward and backward differential equations for the probability $p_{ij}(s,t)$, where $0 \le i < j$. [6]

Now suppose that $\lambda(t) = 0.01(t+2)$ and that the process is in State 1 at time 5.

(iii) (a) Determine an expression for the probability that the process remains in State 1 until time r, where $r > 5$.

(b) Hence, or otherwise, calculate the probability that the process is in State 2 at time 10. [5]

[Total 13]

END OF PAPER

Subject CS2: Assignment X3

2019 Examinations

Time allowed: 2¾ hours

Instructions to the candidate

1. *Please:*

   - *attempt all of the questions, as far as possible under exam conditions*
   - ***begin your answer to each question on a new page***
   - ***leave at least 2cm margin on all borders***
   - *write in black ink using a medium-sized nib because we will be unable to mark illegible scripts*
   - *note that assignment marking is not included in the price of the course materials. Please purchase Series Marking or a Marking Voucher before submitting your script.*
   - *note that we only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2019 exams.*

2. *Please **do not**:*

   - *use headed paper*
   - *use highlighting in your script.*

At the end of the assignment

If your script is being marked by ActEd, please follow
the instructions on the reverse of this page.

In addition to this paper, you should have available actuarial tables and an electronic calculator.

Submission for marking

You should aim to submit this script for marking by the recommended submission date. The recommended and deadline dates for submission of this assignment are listed on the summary page at the back of this pack and on our website at **www.ActEd.co.uk**.

Scripts received after the deadline date will not be marked, unless you are using a Marking Voucher. *It is your responsibility to ensure that scripts reach ActEd in good time*. If you are using Marking Vouchers, then please make sure that your script reaches us by the Marking Voucher deadline date to give us enough time to mark and return the script before the exam.

When submitting your script, please:

- complete the cover sheet, including the checklist
- scan your script, cover sheet (and Marking Voucher if applicable) and save as a pdf document, then email it to: ActEdMarking@bpp.com
- **do not submit a photograph of your script**
- **do not include the question paper in the scan.**

In addition, please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* ‘CS2 Assignment X3 No. 12345’, inserting your ActEd Student Number for 12345.
- The assignment should be scanned the **right way up** (so that it can be read normally without rotation) and as a single document. We cannot accept individual files for each page.
- Please set the resolution so that the script is legible and the resulting PDF **is less than 4 MB** in size.
- Do not protect the PDF in any way (otherwise the marker cannot return the script to ActEd, which causes delays).
- Please include the ‘feedback from marker’ sheet when scanning.
- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

Subject CS2: Assignment X3

2019 Examinations

Please complete the following information:

Name:

ActEd Student Number (see Note below):

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting it will help us to process your scripts quickly. If you do not know your ActEd Student Number, please email us at ActEd@bpp.com.

Your ActEd Student Number is not the same as your IFoA Actuarial Reference Number or ARN.

Number of following pages: _______

Please put a tick in this box if you have solutions and a cross if you do not: □

Please tick here if you are allowed extra time or other special conditions in the profession's exams (if you wish to share this information): □

Time to do assignment
(see Note below): _____ hrs _____ mins

Under exam conditions
(delete as applicable): yes / nearly / no

Note: If you take more than 2¾ hours, you should indicate how much you completed within this exam time so that the marker can provide useful feedback on your progress.

Score and grade for this assignment (to be completed by marker):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	**Total**	
$\frac{\ }{2}$	$\frac{\ }{4}$	$\frac{\ }{6}$	$\frac{\ }{6}$	$\frac{\ }{8}$	$\frac{\ }{9}$	$\frac{\ }{10}$	$\frac{\ }{11}$	$\frac{\ }{10}$	$\frac{\ }{14}$	$\frac{\ }{80}$	=_____%

Grade: A B C D E **Marker's initials:** _________

Please tick the following checklist so that your script can be marked quickly. Have you:

- [ ] Checked that you are using the latest version of the assignments, *ie* 2019 for the sessions leading to the 2019 exams?
- [ ] Written your full name in the box above?
- [ ] Completed your ActEd Student Number in the box above?
- [ ] Recorded your attempt conditions?
- [ ] Numbered all pages of your script (excluding this cover sheet)?
- [ ] Written the total number of pages (excluding the cover sheet) in the space above?
- [ ] Included your Marking Voucher or ordered Series X Marking?
- [ ] Rated your X2 marker at **www.ActEd.co.uk/marking**?

Please follow the instructions on the previous page when submitting your script for marking.

Feedback from marker

Notes on marker's section

The main objective of marking is to provide specific advice on how to improve your chances of success in the exam. The most useful aspect of the marking is the comments the marker makes throughout the script, however you will also be given a percentage score and the band into which that score falls. Each assignment tests only part of the course and hence does not give a complete indication of your likely overall success in the exam. However it provides a good indicator of your understanding of the material tested and the progress you are making with your studies:

A = Excellent progress B = Good progress C = Average progress
D = Below average progress E = Well below average progress

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

X3.1 A life office is comparing the mortality of its policyholders in the age range 31 nearest birthday to 60 nearest birthday with a set of mortality rates prepared by the Continuous Mortality Investigation (CMI). The life office finds that its mortality rates are lower than those of the CMI at 18 ages and higher than those of the CMI at 12 ages.

An analyst carries out the grouping of signs test on the data using a 5% significance level and discovers that the test is only just passed – one fewer run of positive deviations would have meant that the test failed.

Determine the number of runs of positive deviations in the life office's data. [2]

X3.2 A study is being conducted, using the Cox regression model, into how smoking affects a patient's future lifetime after they have had a serious heart attack. The survival times and smoking status for 6 patients are shown in the table below. Patients have been labelled as 'censored' if they were still alive at the end of the investigation or if their death was not considered to be attributable to the heart attack.

Patient number	Time to death (weeks)	Smoker (yes/no)	Censored (yes/no)
1	3	Yes	No
2	n/a (still alive)	No	Yes
3	9	No	No
4	10	Yes	Yes
5	8	No	Yes
6	7	No	No

The force of mortality for life i at duration t is modelled as:

$$\lambda(t,Z_i)=\lambda_0(t)\exp(\beta Z_i)$$

where:

t is the duration in weeks since having a heart attack

$\lambda_0(t)$ is the baseline hazard function at time t

$$Z_i=\begin{cases}1 & \text{if life } i \text{ is a smoker}\\ 0 & \text{if life } i \text{ is a non-smoker}\end{cases}$$

β is a regression parameter

Write down the partial likelihood function of β given these data values. [4]

X3.3 In an investigation of mortality during the period 1 January 2017 to 1 July 2018, information is available about the number of lives under observation aged x next birthday on 1 January 2017, 1 January 2018 and 1 July 2018. Information is also available about the number of deaths during the period, classified by age last birthday.

(i) Derive a formula for the central exposed to risk that corresponds to the death data, stating any assumptions that you make. [5]

(ii) The force of mortality for deaths with age label x in this investigation estimates μ_{x+f}. Determine the value of f. [1]

[Total 6]

X3.4 A mortality investigation is being conducted by a life insurance company.

(i) Explain why, when investigating its mortality statistics, the company may divide the data into smaller groups. [2]

(ii) Describe the two main potential problems with subdividing mortality data. [2]

(iii) List four factors that could be used to subdivide mortality data. [2]

[Total 6]

X3.5 (i) Explain the terms ‘undergraduation’ and ‘overgraduation’. [3]

(ii) List the possible dangers to a life company of using undergraduated or overgraduated mortality rates. [5]

[Total 8]

X3.6 A medium-sized UK pension scheme has recently carried out an investigation of the mortality of its pensioners.

The data used to produce the crude rates, and the proposed graduated rates, are shown below.

Age nearest birthday	Central exposed to risk	Observed number of deaths	Crude mortality rate	Graduated mortality rate	Standardised deviation
60 - 64	1,388.9	10	0.0072	0.0061	0.5249
65 - 69	1,188.8	17	0.0143	0.0131	0.3615
70 - 74	880.5	28	0.0318	0.0262	1.0266
75 - 79	841.6	34	0.0404	0.0487	–1.0912
80 - 84	402.8	41	0.1018	0.0839	1.2394
85 - 89	123.9	19	0.1533	0.1338	0.5949
90 - 94	27.9	7	0.2509	0.1975	0.6346
95 - 99	10.0	3	0.3000	0.2706	0.1787
100+	7.5	2	0.2667	0.3455	–0.3673

(i) Test the proposed graduation for:

(a) overall goodness of fit

(b) bias over the whole age range

clearly stating any conclusions that you draw. [7]

(ii) Comment on the use of the graduated rates to value the benefits payable from the scheme. [2]

[Total 9]

X3.7 The following Lee-Carter mortality projection model is being fitted to some historical data:

$$\ln m_{x,t} = a_x + b_x\, k_t + e_{x,t}$$

where:

$m_{x,t}$ is the central mortality rate at age x in Year t

a_x and b_x are factors relating to mortality rates projected for age x

k_t is a factor relating to mortality rates projected for Year t

$e_{x,t}$ is an independent and identically distributed error term.

In this particular model there are 37 different projection years ($t = 0, 1, ..., 36$), where $t = 0$ is the base calendar year for the projection.

(i) State the constraints that are typically imposed on the estimated values of b_x and k_t when fitting the model. [1]

(ii) A model has been fitted to the data, and it is found that estimated values of k_t are related as follows:

$$\hat{k}_{t+1} = \hat{k}_t - 0.01$$

Given that these values satisfy the overall constraints specified in part (i), calculate the estimated values of k_0 and k_{10} for this model. [2]

(iii) The following ratio is used to show the projected change in mortality at a particular age x over the first ten years of the projection:

$$\frac{\hat{m}_{x,10}}{\hat{m}_{x,0}}$$

where $\hat{m}_{x,t}$ is the predicted mortality rate from the fitted model ignoring error terms.

(a) Calculate this ratio for the case where $\hat{b}_x = 1$.

(b) Three of the estimated values of b_x are:

$$\hat{b}_{50} = -0.14, \quad \hat{b}_{65} = 0.28, \quad \hat{b}_{75} = 1.30$$

Calculate the values of the above ratio for $x = 50$, 65 and 75. [3]

(iv) With reference to the values you have calculated in part (iii) or otherwise, explain how the sign and magnitude of the value of the b_x parameter influences the impact of the assumed time trend on projected mortality rates using the Lee-Carter model. [4]

[Total 10]

X3.8 An investigation has been carried out into the survival rates of patients who have just undergone a certain medical procedure at one of two major hospitals. The data recorded for each patient were sex, drug treatment received and hospital attended. A Cox proportional hazards model was fitted to the data, and the results are given below.

Covariate	*Fitted parameter value*	*Estimated standard error*
Sex:		
Male	0	
Female	–0.20	0.11
Drug treatment received:		
Treatment A	0	
Treatment B	0.12	0.05
Treatment C	–0.05	0.03
Hospital attended:		
Hospital A	0	
Hospital B	–0.06	0.04

(i) Write down a formula for the force of mortality according to this model. You should define all the terms that you use. [3]

(ii) Explain why this model is a proportional hazards model. [1]

(iii) In the context of this model state the group of lives:

(a) to which the baseline hazard refers

(b) with the lowest force of mortality. [2]

(iv) Explain whether attending Hospital B rather than Hospital A significantly improves the chances of survival. [3]

(v) Calculate the proportion, according to the fitted model, by which the force of mortality for a male patient on Treatment B who attended Hospital A exceeds that for a female patient on Treatment C who attended Hospital B. [2]

[Total 11]

X3.9 You have been given the following data relating to an insurance company mortality investigation.

Age last birthday	Policies in force on 1 July					Deaths in			
	2015	2016	2017	2018		2015	2016	2017	2018
63	4,192	4,444	4,885	4,889		104	100	117	109
64	3,998	4,200	4,664	4,334		122	114	130	124
65	3,940	4,166	4,321	4,533		118	120	129	140

(i) Calculate estimates of the force of mortality for those lives aged 63, 64 and 65 last birthday, indicating clearly the ages to which your estimates relate. State any assumptions you make. [6]

(ii) Explain the relationship between the initial mortality rate and the force of mortality under the assumptions of the Poisson model. [2]

(iii) Hence calculate estimates of the initial mortality rate for those lives aged 63, 64 and 65 last birthday, indicating clearly the ages to which your estimates relate. [2]

[Total 10]

X3.10 A large life office is investigating the recent mortality experience of its term assurance policyholders. It has been decided to graduate the data by reference to a standard table using the formula:

$$\frac{\overset{\circ}{\mu}_x}{\mu_x^s} = ax + b$$

where μ_x^s is the rate for the standard table.

(i) Outline the considerations that you would take into account in choosing an appropriate standard table. [5]

(ii) Explain how you would check whether the above formula is suitable. [3]

(iii) Describe how you would estimate a and b in the above formula using:

(a) a weighted least squares approach

(b) the method of maximum likelihood.

In each case you should state the function to be optimised. [6]

[Total 14]

END OF PAPER

Subject CS2: Assignment X4

2019 Examinations

Time allowed: 3¼ hours

Instructions to the candidate

1. *Please:*

   - *attempt all of the questions, as far as possible under exam conditions*
   - ***begin your answer to each question on a new page***
   - ***leave at least 2cm margin on all borders***
   - *write in black ink using a medium-sized nib because we will be unable to mark illegible scripts*
   - *note that assignment marking is not included in the price of the course materials. Please purchase Series Marking or a Marking Voucher before submitting your script.*
   - *note that we only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2019 exams.*

2. *Please **do not**:*

   - *use headed paper*
   - *use highlighting in your script.*

At the end of the assignment

If your script is being marked by ActEd, please follow
the instructions on the reverse of this page.

In addition to this paper, you should have available actuarial tables and an electronic calculator.

Submission for marking

You should aim to submit this script for marking by the recommended submission date. The recommended and deadline dates for submission of this assignment are listed on the summary page at the back of this pack and on our website at **www.ActEd.co.uk**.

Scripts received after the deadline date will not be marked, unless you are using a Marking Voucher. *It is your responsibility to ensure that scripts reach ActEd in good time*. If you are using Marking Vouchers, then please make sure that your script reaches us by the Marking Voucher deadline date to give us enough time to mark and return the script before the exam.

When submitting your script, please:

- complete the cover sheet, including the checklist
- scan your script, cover sheet (and Marking Voucher if applicable) and save as a pdf document, then email it to: ActEdMarking@bpp.com
- **do not submit a photograph of your script**
- **do not include the question paper in the scan.**

In addition, please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* ‘CS2 Assignment X4 No. 12345’, inserting your ActEd Student Number for 12345.
- The assignment should be scanned the **right way up** (so that it can be read normally without rotation) and as a single document. We cannot accept individual files for each page.
- Please set the resolution so that the script is legible and the resulting PDF **is less than 4 MB** in size.
- Do not protect the PDF in any way (otherwise the marker cannot return the script to ActEd, which causes delays).
- Please include the ‘feedback from marker’ sheet when scanning.
- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

Subject CS2: Assignment X4

2019 Examinations

Please complete the following information:

Name:

ActEd Student Number (see Note below):

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting it will help us to process your scripts quickly. If you do not know your ActEd Student Number, please email us at ActEd@bpp.com.

Your ActEd Student Number is not the same as your IFoA Actuarial Reference Number or ARN.

Number of following pages: _______

Please put a tick in this box if you have solutions and a cross if you do not: □

Please tick here if you are allowed extra time or other special conditions in the profession's exams (if you wish to share this information): □

Time to do assignment
(see Note below): _____ hrs _____ mins

Under exam conditions
(delete as applicable): yes / nearly / no

Note: If you take more than 3¼ hours, you should indicate how much you completed within this exam time so that the marker can provide useful feedback on your progress.

Score and grade for this assignment (to be completed by marker):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	**Total**	
$\frac{}{3}$	$\frac{}{4}$	$\frac{}{4}$	$\frac{}{4}$	$\frac{}{5}$	$\frac{}{6}$	$\frac{}{6}$	$\frac{}{6}$	$\frac{}{7}$	$\frac{}{8}$	$\frac{}{11}$	$\frac{}{13}$	$\frac{}{23}$	$\frac{}{100}$	=_____%

Grade: A B C D E **Marker's initials:** _________

Please tick the following checklist so that your script can be marked quickly. Have you:

[ ] Checked that you are using the latest version of the assignments, *ie* 2019 for the sessions leading to the 2019 exams?

[ ] Written your full name in the box above?

[ ] Completed your ActEd Student Number in the box above?

[ ] Recorded your attempt conditions?

[ ] Numbered all pages of your script (excluding this cover sheet)?

[ ] Written the total number of pages (excluding the cover sheet) in the space above?

[ ] Included your Marking Voucher or ordered Series X Marking?

[ ] Rated your X3 marker at **www.ActEd.co.uk/marking**?

Please follow the instructions on the previous page when submitting your script for marking.

Feedback from marker

Notes on marker's section

The main objective of marking is to provide specific advice on how to improve your chances of success in the exam. The most useful aspect of the marking is the comments the marker makes throughout the script, however you will also be given a percentage score and the band into which that score falls. Each assignment tests only part of the course and hence does not give a complete indication of your likely overall success in the exam. However it provides a good indicator of your understanding of the material tested and the progress you are making with your studies:

A = Excellent progress   B = Good progress   C = Average progress
D = Below average progress   E = Well below average progress

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

X4.1 Claim amounts arising from a particular group of policies follow a Pareto distribution with mean 1,000 and standard deviation 1,500.

Calculate the proportion of claims that exceed 2,000. [3]

X4.2 Claim amounts from a particular portfolio of insurance policies are believed to follow a Weibull distribution. A random sample of 2,000 claims was collected. The sample median was found to be 1,500 and 5% of claims exceeded 6,000.

Estimate the parameters of the Weibull distribution using the method of percentiles. [4]

X4.3 (i) The shape parameter, γ, in a particular generalised extreme value distribution is equal to -1. Identify the type of extreme value distribution, state its key characteristic, and give an example of a distribution of this type. [2]

(ii) State the key advantage that the generalised Pareto distribution has over the generalised extreme value distribution when modelling extreme losses. [2]

[Total 4]

X4.4 A loss amount random variable, X, follows a $Pa(\alpha,\lambda)$ distribution.

(i) Define the excess loss over the threshold u. [1]

(ii) Derive the CDF of this threshold exceedance and hence identify its distribution. [3]

[Total 4]

X4.5 A company observes its quarterly utility bills over the last 5 years $\{q_t : t = 1, 2, \ldots, 20\}$.

(i) The company decomposes the time series as follows:

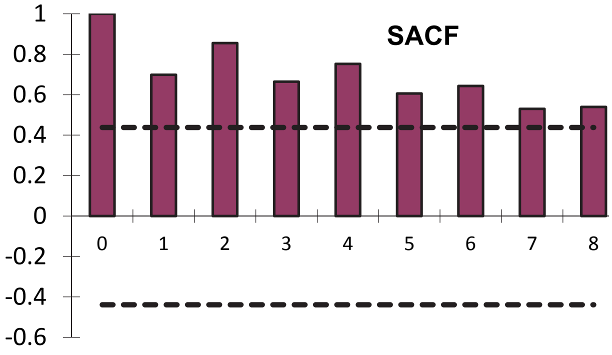
$$q_t = \mu_t + \theta_t + y_t$$

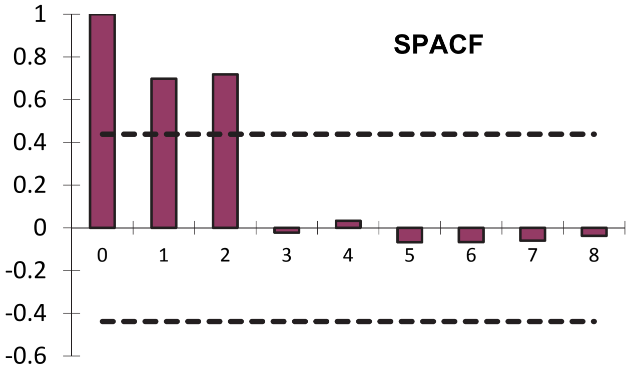
where y_t is the underlying zero-mean stationary time series.

(a) The time series exhibits quarterly seasonal variation, θ_t. Describe one method of removing this variation, giving any formula or filter used.

(b) The time series also exhibits a linear trend $\mu_t = a + bt$. Describe one method of removing this trend. [3]

The seasonal variation and linear trend are removed and the sample ACF, r_k, and sample PACF, $\hat{\phi}_k$, of the resultant zero-mean stationary time series, z_t, are obtained:





(ii) State, with reasons, an appropriate time series to model the observations z_t. [2]

[Total 5]

X4.6 Claims from a particular group of policies are thought to have a lognormal distribution with parameters μ and σ^2. Claims over the last 5 years have a sample mean of £2,000 and a sample standard deviation of £500.

(i) Obtain the method of moments estimates of μ and σ^2. [3]

(ii) Estimate the median claim amount using the fitted distribution. [3]

[Total 6]

X4.7 A moving average time series is defined by the relationship:

$$X_t = 3.1 + \varepsilon_t + 0.25\varepsilon_{t-1} + 0.5\varepsilon_{t-2} + 0.25\varepsilon_{t-3}$$

where $\varepsilon_t \sim N(0,\sigma^2)$ denotes white noise.

(i) Determine the mean and variance of X_t. [2]

(ii) Calculate the autocorrelation function ρ_k, $k = 0,1,2,\ldots$. [4]

[Total 6]

X4.8 (i) Explain why we might expect the CPI and NAEI to be cointegrated. [2]

The CPI (Consumer Price Index) is a measure of the average cost of a basket of goods and services and the NAEI (National Average Earnings Index) is a measure of the average employee pay.

(ii) A multivariate process is defined by:

$$X_n = 1.2X_{n-1} - 0.2X_{n-2} + \varepsilon_n^x$$

$$Y_n = 0.6X_{n-1} + \varepsilon_n^y$$

where ε_n^x and ε_n^y are independent white noise processes.

Show that X_n and Y_n are cointegrated with cointegrating vector $(0.6,-1)$. [4]

[Total 6]

X4.9 (i) Explain what is meant by an $I(d)$ process. [1]

(ii) Classify each of the following processes as $ARIMA(p,d,q)$ where possible:

(a) $X_t = 0.6\varepsilon_{t-1} + \varepsilon_t$

(b) $Y_t = 1.4Y_{t-2} + \varepsilon_t + 0.5\varepsilon_{t-3}$

(c) $W_t = 1.4W_{t-1} - 0.4W_{t-2} + \varepsilon_t + \varepsilon_{t-1}$

In each case ε_t denotes white noise with mean 0 and variance σ^2. [6]

[Total 7]

X4.10 (i) Show that the mean residual life of the $Gamma(2,1)$ distribution is given by:

$$e(x)=\frac{x+2}{x+1}$$ [5]

(ii) Use the mean residual life to compare the tail of the $Gamma(2,1)$ distribution with that of the $Exp(1)$ distribution. [3]

[Total 8]

X4.11 A stationary $ARMA(1,1)$ process, X_n, is defined by:

$$X_n=\alpha X_{n-1}+\varepsilon_n+\beta\varepsilon_{n-1}$$

where ε_t denotes white noise with zero mean and variance σ^2.

(i) State the range of values of β for which the process is invertible. [1]

The autocorrelation function of this process is given by:

$$\rho_k=\frac{(\alpha+\beta)(1+\alpha\beta)}{1+2\alpha\beta+\beta^2}\alpha^{k-1} \qquad k=1,2,3,\ldots$$

The sample autocorrelation coefficients at lags 1 and 2 for a time series that is believed to conform to a stationary, invertible $ARMA(1,1)$ model have been calculated to be $r_1=0.440$ and $r_2=0.264$.

(ii) (a) Obtain the method of moments estimates of α and β.

(b) Outline briefly how the method of least squares could be used to estimate the parameter values.

(c) State when the maximum likelihood estimates are equivalent to the least squares estimates. [7]

The most recently observed values of the time series are:

$$x_{79}=-0.214 \qquad x_{80}=1.087 \qquad \hat{\varepsilon}_{79}=-0.169 \qquad \hat{\varepsilon}_{80}=1.181$$

(iii) (a) Determine the 1 and 2 step ahead estimates for x_{81} and x_{82} using the fitted values of α and β obtained in part (ii)(a).

(b) The simplest form of exponential smoothing (with smoothing parameter 0.2) used at time 79 gave a forecast for x_{80} of 0.625. Determine the forecast for x_{81}. [3]

[Total 11]

X4.12 A researcher is using the Box-Jenkins approach to model an observed time series X as an $ARIMA(p,d,q)$ process.

(i) Explain what it means to say that X is an $ARIMA(p,d,q)$ process. [1]

(ii) The following table shows some information relating to the d th order differences of the observed series:

Properties of $\nabla^d X$		$d=0$	$d=1$	$d=2$	$d=3$	$d=4$
Sample autocorrelation coefficients	r_1	0.33	–0.40	–0.63	–0.74	–0.79
	r_2	0.19	–0.04	0.15	0.32	0.43
	r_3	0.12	–0.11	–0.12	–0.17	–0.23
	r_4	0.18	0.18	0.20	0.18	0.17
Sample variance		4.4	6.0	16.7	54.4	189.9

State, with reasons, which value of d you consider most appropriate if this series is to be modelled using an $ARIMA(p,d,q)$ model. [2]

(iii) Having selected an appropriate value of d and carried out some further calculations, the researcher has decided that a zero-mean $ARMA(1,1)$ model provides a good description of the series $\nabla^d X$.

The 100 residuals for this model are found to contain 74 turning points. The sample autocorrelations at lags $1,2,\dots,5$ (respectively) are calculated to be:

+0.14 –0.05 +0.10 +0.12 –0.02

Carry out each of the following tests, explaining what each test is designed to check for:

(a) the Ljung-Box (‘portmanteau’) test

(b) the turning point test

(c) inspection of the sample autocorrelation function.

You should state your conclusions clearly. [10]

[Total 13]

X4.13 A univariate $AR(2)$ process has defining equation:

$$X_n = 0.7X_{n-1} - 0.1X_{n-2} + \varepsilon_n$$

where ε_n is a white noise process with mean 0 and variance σ^2.

(i) Explain whether the process is:

(a) stationary

(b) invertible

(c) purely indeterministic

(d) Markov. [5]

(ii) (a) Calculate the values of ρ_1 and ρ_2, the autocorrelation function at lags 1 and 2, and show that the autocorrelation function for lag k $(k > 2)$ is given by:

$$\rho_k = 0.7\rho_{k-1} - 0.1\rho_{k-2}$$

(b) Show that:

$$\rho_k = \frac{A}{2^k} + \frac{B}{5^k} \qquad k > 2$$

is a solution of $\rho_k = 0.7\rho_{k-1} - 0.1\rho_{k-2}$ and hence calculate the values of the constants A and B.

(c) Calculate ϕ_1 and ϕ_2, the partial autocorrelation function for lags 1 and 2 and state what will happen for larger values of k. [12]

(iii) Explain how the univariate $AR(2)$ process $X_n = 0.7X_{n-1} - 0.1X_{n-2} + \varepsilon_n$ can be expressed as a multivariate $VAR(1)$ process $\underline{X}_n = M\,\underline{X}_{n-1} + \underline{\varepsilon}_n$. [2]

(iv) Determine whether the $VAR(1)$ process is:

(a) stationary

(b) Markov. [4]

[Total 23]

END OF PAPER

Subject CS2: Assignment X5

2019 Examinations

Time allowed: 3¼ hours

Instructions to the candidate

1. Please:

- *attempt all of the questions, as far as possible under exam conditions*
- ***begin your answer to each question on a new page***
- ***leave at least 2cm margin on all borders***
- *write in black ink using a medium-sized nib because we will be unable to mark illegible scripts*
- *note that assignment marking is not included in the price of the course materials. Please purchase Series Marking or a Marking Voucher before submitting your script.*
- *note that we only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2019 exams.*

*2. Please **do not**:*

- *use headed paper*
- *use highlighting in your script.*

At the end of the assignment

If your script is being marked by ActEd, please follow
the instructions on the reverse of this page.

In addition to this paper, you should have available actuarial tables and an electronic calculator.

Submission for marking

You should aim to submit this script for marking by the recommended submission date. The recommended and deadline dates for submission of this assignment are listed on the summary page at the back of this pack and on our website at **www.ActEd.co.uk**.

Scripts received after the deadline date will not be marked, unless you are using a Marking Voucher. *It is your responsibility to ensure that scripts reach ActEd in good time*. If you are using Marking Vouchers, then please make sure that your script reaches us by the Marking Voucher deadline date to give us enough time to mark and return the script before the exam.

When submitting your script, please:

- complete the cover sheet, including the checklist
- scan your script, cover sheet (and Marking Voucher if applicable) and save as a pdf document, then email it to: ActEdMarking@bpp.com
- **do not submit a photograph of your script**
- **do not include the question paper in the scan.**

In addition, please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* ‘CS2 Assignment X5 No. 12345’, inserting your ActEd Student Number for 12345.
- The assignment should be scanned the **right way up** (so that it can be read normally without rotation) and as a single document. We cannot accept individual files for each page.
- Please set the resolution so that the script is legible and the resulting PDF **is less than 4 MB** in size.
- Do not protect the PDF in any way (otherwise the marker cannot return the script to ActEd, which causes delays).
- Please include the ‘feedback from marker’ sheet when scanning.
- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

Subject CS2: Assignment X5

2019 Examinations

Please complete the following information:

Name:

ActEd Student Number (see Note below):

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting it will help us to process your scripts quickly. If you do not know your ActEd Student Number, please email us at ActEd@bpp.com.

Your ActEd Student Number is not the same as your IFoA Actuarial Reference Number or ARN.

Number of following pages: _______

Please put a tick in this box if you have solutions and a cross if you do not: ☐

Please tick here if you are allowed extra time or other special conditions in the profession's exams (if you wish to share this information): ☐

Time to do assignment
(see Note below): _____ hrs _____ mins

Under exam conditions
(delete as applicable): yes / nearly / no

Note: If you take more than 3¼ hours, you should indicate how much you completed within this exam time so that the marker can provide useful feedback on your progress.

Score and grade for this assignment (to be completed by marker):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	**Total**	
— 4	— 4	— 6	— 8	— 9	— 11	— 13	— 14	— 15	— 16	— 100	=_____%

Grade: A B C D E **Marker's initials:** _________

Please tick the following checklist so that your script can be marked quickly. Have you:

[ ] Checked that you are using the latest version of the assignments, *ie* 2019 for the sessions leading to the 2019 exams?

[ ] Written your full name in the box above?

[ ] Completed your ActEd Student Number in the box above?

[ ] Recorded your attempt conditions?

[ ] Numbered all pages of your script (excluding this cover sheet)?

[ ] Written the total number of pages (excluding the cover sheet) in the space above?

[ ] Included your Marking Voucher or ordered Series X Marking?

[ ] Rated your X4 marker at **www.ActEd.co.uk/marking**?

Please follow the instructions on the previous page when submitting your script for marking.

Feedback from marker

Notes on marker's section

The main objective of marking is to provide specific advice on how to improve your chances of success in the exam. The most useful aspect of the marking is the comments the marker makes throughout the script, however you will also be given a percentage score and the band into which that score falls. Each assignment tests only part of the course and hence does not give a complete indication of your likely overall success in the exam. However it provides a good indicator of your understanding of the material tested and the progress you are making with your studies:

A = Excellent progress B = Good progress C = Average progress
D = Below average progress E = Well below average progress

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

X5.1 Claims on a motor insurance policy have a gamma distribution with mean £2,000 and standard deviation £100. The insurer effects proportional reinsurance with retained proportion 85%. Determine the:

(a) mean amount paid by the insurer (after reinsurance)

(b) variance of the amount paid by the reinsurer

(c) moment generating function of the claim amount paid by the insurer (after reinsurance). [4]

X5.2 Consider a portfolio of insurance policies, on which the number of claims has a binomial distribution with parameters $n=1{,}000$ and $p=0.01$. The claim size distribution is assumed to be exponential with mean £100. Claim amounts are assumed to be independent random variables and to be independent of the number of claims.

The insurer arranges individual excess of loss reinsurance with a retention limit of 200.

Calculate the mean of S_I, where S_I is aggregate annual claims paid by the insurer net of reinsurance. [4]

X5.3 Second Life (SL) is a small life insurance company.

Responding to demand from brokers, SL is developing a new product called *Seconds Out*. This is to be a product aimed at retired couples who wish to purchase large amounts of cover to provide a tax-efficient transfer of their wealth to their dependants. The policy pays out when the second of the two policyholders dies.

SL is developing pricing assumptions for this new product, and is considering using one of the following two copulas to handle the correlated mortality expected with this product:

Clayton copula: $C[u,v]=\left(u^{-\alpha}+v^{-\alpha}-1\right)^{-1/\alpha}$

Farlie-Gumbel-Morgenstern (FGM) copula: $C[u,v]=uv\left[1+\theta(1-u)(1-v)\right]$

One of the pricing assumptions is that the probability of survival for ten years for a 70-year-old life (regardless of gender) is ${}_{10}p_{70}=0.58$.

(i) Using each of the two copulas with parameters $\alpha=0.3$ and $\theta=-0.1$ respectively, calculate the probability of paying a death benefit in the ten-year period following the issue of a *Seconds Out* policy to a couple who are both aged exactly 70. [3]

(ii) Discuss the suitability of the copulas above by comparing the results to those when independent deaths are assumed. [3]

[Total 6]

X5.4 Under a special reinsurance arrangement, a reinsurer agrees to pay an amount Z in respect of each claim X arising from a certain risk, where:

$$Z = \begin{cases} 0 & \text{if } X \le 1{,}000 \\ X - 1{,}000 & \text{if } 1{,}000 < X \le 2{,}000 \\ 1{,}000 & \text{if } X > 2{,}000 \end{cases}$$

Given that X has a lognormal distribution with parameters $\mu = 5$ and $\sigma^2 = 4$, calculate the reinsurer's expected payment amount per original claim. [8]

X5.5 The annual premium for a certain class of household insurance policies is £190. The total annual claims from a single policy has a compound Poisson distribution with Poisson parameter 0.25. Individual claim amounts have a Pareto distribution with parameters $\alpha = 4$ and $\lambda = 1{,}800$. Every time a claim is settled the insurance company incurs an expense. The amount of this expense is a random variable, uniformly distributed on the interval (£35,£85) and independent of the claim amount.

Suppose that a portfolio comprises n independent policies of this type and let S denote the total aggregate claims and expenses arising from the portfolio in a year.

(i) Show that:

$$E(S) = 165n \qquad \text{var}(S) = 288{,}952n$$ [6]

(ii) Assuming that the distribution of S is approximately normal, estimate the number of policies that the insurer must sell to be at least 99% sure of making a profit on the portfolio. [3]

[Total 9]

X5.6 Suppose that X and Y are random variables that can each take values in the range $(-\infty,\infty)$ and that have the following characteristics:

- The marginal cumulative distribution function of X is $F_X(x)=(1+e^{-x})^{-1}$.
- The marginal cumulative distribution function of Y is $F_Y(y)=(1+e^{-y})^{-1}$.
- The joint cumulative distribution function of X and Y is $F_{X,Y}(x,y)=(1+e^{-x}+e^{-y})^{-1}$.

(i) Show that the copula function for X and Y is $C[u,v]=\left(u^{-1}+v^{-1}-1\right)^{-1}$. [2]

(ii) Show that this is an Archimedean copula with generator function $\psi(t)=t^{-1}-1$. [4]

(iii) Determine the coefficients of lower and upper tail dependence for this copula.

Hint: you can use L'Hôpital's rule, $\lim\limits_{x\to a}\dfrac{f(x)}{g(x)}=\lim\limits_{x\to a}\dfrac{f'(x)}{g'(x)}$. [5]

[Total 11]

X5.7 Claims on a home insurance policy have a Pareto distribution with parameters $\alpha=4$ and $\lambda=7{,}500$. The insurer effects an individual excess of loss reinsurance treaty with a retention limit of £3,000.

(i) (a) Calculate the probability that a claim involves the reinsurer.

(b) Calculate the insurer's expected payment per claim. [5]

Next year the claim amounts on these policies are expected to increase by 10% but the reinsurance treaty will remain unchanged.

(ii) (a) Calculate the probability that a claim now involves the reinsurer.

(b) Explain whether the insurer's expected payment per claim will also increase by 10%.

(c) Calculate the reinsurer's expected claim payment next year on those claims in which it is involved. [8]

[Total 13]

X5.8 Insurance Company A has taken out an individual excess of loss reinsurance contract with a retention limit of £40,000. Individual claim amounts, gross of reinsurance, are believed to follow an exponential distribution with unknown parameter λ.

Over the last year, the following claims data are observed:

Claims below retention:	12,220	10,429	36,834	14,623
	36,932	13,205	28,506	
Claims above retention:	3 in total			

(i) (a) Estimate λ using maximum likelihood estimation.

(b) Apply the method of percentiles using the median claim to estimate λ. [7]

Insurance Company B has a policyholder excess of £50,000 on its policies. The individual claim amounts, X, are believed to have a Pareto$(\theta, 200\,000)$ distribution (before the excess is applied) with PDF:

$$f_X(x) = \frac{\theta \times 200{,}000^{\theta}}{(200{,}000 + x)^{\theta+1}}, \qquad x > 0$$

where θ is an unknown parameter.

(ii) (a) Show that the conditional distribution of the amount paid by the insurer, Y, has a Pareto$(\theta, 250\,000)$ distribution, with PDF:

$$f_Y(y) = \frac{\theta \times 250{,}000^{\theta}}{(250{,}000 + y)^{\theta+1}}, \qquad y > 0$$

The amounts paid the insurer, y_i, on the last five claims (*ie* after the £50,000 excess has been deducted) were:

£153,000 £376,000 £120,000 £20,000 £108,000

(b) Use this information and the distribution from part (a) to determine, $\hat{\theta}$, the maximum likelihood estimate of θ. [7]

[Total 14]

X5.9 A historian is classifying some old documents that have recently been discovered. These are known to be written in one of the five languages English, French, German, Spanish or Italian.

Unfortunately, the writing on some of the documents has been badly eroded and only a few letters are still legible. So the historian has asked you to help identify the correct language based on the frequency with which certain letters occur in the documents. It is known that the letter frequencies for similar documents are as shown in the table below.

LETTER	ENGLISH	FRENCH	GERMAN	SPANISH	ITALIAN
A	8%	8%	7%	12%	12%
G	2%	1%	3%	2%	2%
H	6%	1%	5%	1%	1%
I	7%	8%	7%	6%	10%
N	7%	7%	10%	7%	7%
O	7%	8%	7%	6%	10%
T	9%	7%	6%	5%	6%
U	3%	6%	4%	3%	3%
Other	51%	54%	51%	58%	49%

You intend to use a naïve Bayesian approach to identify the languages.

(i) Show that the posterior probability that a piece of text (Text i) is written in English can be calculated using the formula:

$$P(\text{Text } i \text{ is in English} \mid A_i, G_i, H_i, I_i, N_i, O_i, T_i, U_i, \Omega_i)$$

$$= \frac{P(A_i, \ldots, \Omega_i \mid \text{Text } i \text{ is in English})\ P(\text{Text } i \text{ is in English})}{\sum_k P(A_i, \ldots, \Omega_i \mid \text{Text } i \text{ is in Language } k)\ P(\text{Text } i \text{ is in Language } k)}$$

where $A_i, G_i, H_i, I_i, N_i, O_i, T_i, U_i$ and Ω_i denote the number of occurrences of the letters A, G, H, I, N, O, T, U and other letters (respectively) in Text i, and Language k $(k = 1, 2, \ldots, 5)$ denotes the five languages. [2]

(ii) (a) Calculate the posterior probability that the test message '**BONJOUR MONSIEUR DUPONT**' is written in each of the five languages using a naïve Bayes approach, assuming initially that each language is equally likely.

(b) Comment on your answer to part (ii)(a). [6]

The historian has shown you one badly damaged fragment, which has the following text:

T ????? S ?? G ??? W ??????????? L ?? H ? U ? S ? M ? O ????????????? E ??? E ???? E

The **?**'s denote letters that are illegible. You should ignore these completely in your calculations.

This fragment is believed to be written in one of the five languages with the following prior probabilities:

Language	**ENGLISH**	**FRENCH**	**GERMAN**	**SPANISH**	**ITALIAN**
Probability	40%	20%	20%	10%	10%

(iii) (a) Calculate the posterior probability that the text in this fragment is written in each of the five languages using a naïve Bayes approach, assuming the prior probabilities shown in the table above.

(b) State the conclusions you would draw from your answers to part (iii)(a). [7]

[Total 15]

X5.10 The members of an organisation are covered by group life insurance which pays a specified benefit amount if a member dies. The membership consists of two categories of member who are entitled to the following benefit amounts:

	Number of members	Benefit amount	Probability of dying during year
Active members	1,250	£50,000	0.008
Affiliated members	250	£20,000	0.012
Total	1,500		

The aggregate claims payable during the year can be assumed to conform to the individual risk model.

(i) State the individual risk model formula for modelling the aggregate claim amount for a portfolio and the assumptions underlying it. [2]

(ii) Show, from first principles, that if X denotes the claim amount payable during a given year in respect of an individual member, then:

$$E(X) = bq \qquad \text{and} \qquad \text{var}(X) = b^2 q(1-q)$$

where q is the probability that the member dies during the year and b is the benefit amount. [3]

(iii) Derive a similar formula for the skewness (*ie* the third central moment) of X. [2]

(iv) Hence calculate the mean, variance and coefficient of skewness of the total claim amount arising for all members during a given year. [4]

(v) Calculate the probability that the aggregate claims payable in a given year will exceed £1 million, using a normal approximation with no continuity correction. [3]

(vi) Comment on the likely accuracy of your approximation in part (v). [2]

[Total 16]

END OF PAPER

For the session leading to the April 2019 exams – CS2 & CM1 Subjects

Marking vouchers

Subjects	Assignments	Mocks
CS2, CM1	13 March 2019	20 March 2019

Series X and Y Assignments

Subjects	Assignment	Recommended submission date	Final deadline date
CS2, CM1	**X1**	**21 November 2018**	9 January 2019
CS2, CM1	**X2**	**5 December 2018**	23 January 2019
CS2, CM1	**X3**	**19 December 2018**	30 January 2019
CS2, CM1	**Y1**	**9 January 2019**	6 February 2019
CS2, CM1	**X4**	**23 January 2019**	20 February 2019
CS2, CM1	**X5**	**6 February 2019**	27 February 2019
CS2, CM1	**Y2**	**20 February 2019**	13 March 2019

Mock Exams

Subjects	Recommended submission date	Final deadline date
CS2 (Paper A/B), CM1 (Paper A/B)	**6 March 2019**	20 March 2019

We encourage you to work to the recommended submission dates where possible.

If you submit your mock on the final deadline date you are likely to receive your script back less than a week before your exam.

For the session leading to the September 2019 exams – CS2 & CM1 Subjects

Marking vouchers

Subjects	Assignments	Mocks
CS2	21 August 2019	28 August 2019
CM1	28 August 2019	4 September 2019

Series X and Y Assignments

Subjects	Assignment	Recommended submission date	Final deadline date
CS2	**X1**	**22 May 2019**	3 July 2019
CM1		**29 May 2019**	10 July 2019
CS2	**X2**	**5 June 2019**	10 July 2019
CM1		**12 June 2019**	17 July 2019
CS2	**X3**	**12 June 2019**	17 July 2019
CM1		**19 June 2019**	24 July 2019
CS2	**Y1**	**26 June 2019**	24 July 2019
CM1		**3 July 2019**	31 July 2019
CS2	**X4**	**10 July 2019**	31 July 2019
CM1		**17 July 2019**	7 August 2019
CS2	**X5**	**17 July 2019**	7 August 2019
CM1		**24 July 2019**	14 August 2019
CS2	**Y2**	**31 July 2019**	14 August 2019
CM1		**7 August 2019**	21 August 2019

Mock Exams

Subjects	Recommended submission date	Final deadline date
CS2 (Paper A/B)	**14 August 2019**	28 August 2019
CM1 (Paper A/B)	**21 August 2019**	4 September 2019

We encourage you to work to the recommended submission dates where possible.

If you submit your mock on the final deadline date you are likely to receive your script back less than a week before your exam.

Solution X1.1

This question is about the classification of stochastic processes according to their time set and state space. Stochastic processes are covered in Chapter 1.

A counting process has a discrete state space (as the number of events recorded up to a given time t must be a whole number). Its time set can be discrete or continuous.

A time series has a discrete time set and a continuous state space.

A compound Poisson process has a continuous time set. Its state space can be either discrete or continuous.

A simple random walk has a discrete time set and a discrete state space.

Since we must put one process into each cell of the table, the solution is as follows:

		Time set	
		Discrete	*Continuous*
State Space	*Discrete*	Simple random walk	Counting process
	Continuous	Time series	Compound Poisson process

[½ for each correct entry]

Solution X1.2

This question is about the periodicity of Markov chains. It is based on the material in Chapter 2.

(i) ***Meaning of periodic***

A state in a Markov chain is periodic with period $d > 1$ if a return to that state is possible only in a number of steps that is a multiple of d. [1]

If there is no such $d > 1$, then the state is aperiodic.

A Markov chain has period d if all the states in the chain have period d. [1]

[Total 2]

(ii) ***Periodic or aperiodic?***

Chain 1

This is aperiodic because neither state satisfies the definition of periodic. From State 1 it is not possible to return to State 1 at all, and a return to State 2 will occur after just 1 step. [1]

Chain 2

This is periodic with period 2 because a return to each state is possible only in an even number of steps. [1]

Alternatively, we could explain why State 1 is periodic with period 2, and then use the fact that the chain is irreducible to infer that State 2 must have the same periodicity as State 1. [1]

Chain 3

A return to State 1 is possible in 3 or 4 or 6 or 7 *etc* steps. These numbers are not restricted to a multiple of some number greater than 1. So State 1 is aperiodic.

Since the chain is irreducible, all the states have the same periodicity, *ie* they are all aperiodic.

So the chain is aperiodic. [1]

Alternatively, we could consider each state separately and explain why it is aperiodic – for State 2 and State 3, a return is possible in 3 or 4 or 6 or 7 etc steps; a return to State 4 is possible in 4 or 7 or 8 etc steps. Again these numbers are not restricted to a multiple of some number greater than 1. [1]

[Total 3]

Solution X1.3

The Markov property is introduced in Chapter 1, and used extensively in Chapters 2 to 5.

(i) ***Mathematical definition of the Markov property***

The Markov property says that:

$$P\left[X(t)\in A\,|\,X(s_1)=x_1, X(s_2)=x_2, \dots, X(s_n)=x_n, X(s)=x\right]=P\left[X(t)\in A\,|\,X(s)=x\right]$$

for all times $s_1 < s_2 < \cdots < s_n < s < t$ in the time set, all states $x_1, x_2, \dots, x_n$ and x in the state space, S, and all subsets A of S. [2]

(ii) ***Proof***

For all times $s_1 < s_2 < \cdots < s_n < s < t$ in the time set, all states $x_1, x_2, \dots, x_n$ and x in the state space, S, and all subsets A of S:

$$P\left[X(t)\in A\,|\,X(s_1)=x_1, X(s_2)=x_2, \dots, X(s_n)=x_n, X(s)=x\right]$$

$$=P\left[X(t)-X(s)+x\in A\,|\,X(s_1)=x_1, X(s_2)=x_2, \dots, X(s_n)=x_n, X(s)=x\right]$$ [1]

$$=P\left[X(t)-X(s)+x\in A\,|\,X(s)=x\right]$$ since the process has independent increments [1]

$$=P\left[X(t)\in A\,|\,X(s)=x\right]$$ [1]

So $X(t)$ has the Markov property.

[Total 3]

Solution X1.4

This question involves estimating transition probabilities in a Markov chain. It is based on the material in Section 7.1 of Chapter 2.

(i) ***Estimated transition probabilities***

The sequence of observations is:

SSRCSCCSCRRCSSRCCSCCS

Using the notation n_{ij} to denote the number of times that State i is followed by State j, we have:

$$\begin{array}{lll} n_{SS}=2 & n_{SC}=3 & n_{SR}=2 \\ n_{CS}=5 & n_{CC}=3 & n_{CR}=1 \\ n_{RS}=0 & n_{RC}=3 & n_{RR}=1 \end{array}$$

[1½]

The transition probability p_{ij} is estimated by $\hat{p}_{ij}=\dfrac{n_{ij}}{\sum_k n_{ik}}$. So, for example:

$$\hat{p}_{SS}=\frac{n_{SS}}{n_{SS}+n_{SC}+n_{SR}}=\frac{2}{2+3+2}=\frac{2}{7}$$

Similar calculations lead to the following matrix of estimated transition probabilities:

$$\begin{array}{c} \\ S \\ C \\ R \end{array}\begin{array}{c} \begin{array}{ccc} S & C & R \end{array} \\ \begin{pmatrix} \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \\ \frac{5}{9} & \frac{3}{9} & \frac{1}{9} \\ 0 & \frac{3}{4} & \frac{1}{4} \end{pmatrix} \end{array}$$

[1½]

[Total 3]

(ii) ***Probability of a sunny day on 23 July***

We know that 21 July was a sunny day.

21 July is the last day for which we have data (as it corresponds to the final day in the 3-week period).

The probability that it is sunny on both the 22 and 23 July is estimated to be $\frac{2}{7}\times\frac{2}{7}=\frac{4}{49}$. [½]

The probability that it is cloudy on 22 July and sunny on 23 July is estimated to be $\frac{3}{7}\times\frac{5}{9}=\frac{15}{63}$. [½]

The probability that it is rainy on 22 July and sunny on 23 July is estimated to be $\frac{2}{7}\times 0=0$. [½]

So the estimated probability of a sunny day on 23 July is $\frac{4}{49}+\frac{15}{63}+0=\frac{47}{147}=0.31973$. [½]

[Total 2]

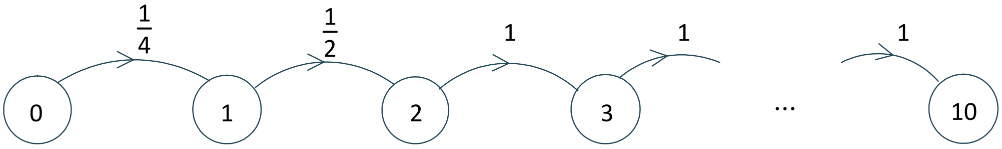
Solution X1.5

This question involves a time-homogeneous Markov jump process. It is an application of the material in Chapter 4.

(i) ***Probability***

Let $N(t)$ denote the number of breakdowns up to time t. Then the state space of $N(t)$ is the set $\{0, 1, 2, ..., 10\}$.

Drawing a transition diagram might help you see what's going on. We have:



Since we are considering a new boiler, we know that $N(0)=0$. We require $P\left(N(5)>1\right)$, which can be calculated using the equation:

$$P\left(N(5)>1\right)=1-P\left(N(5)=0\right)-P\left(N(5)=1\right)$$

First of all, we have:

$$P\left(N(5)=0\right)=e^{-5\times\frac{1}{4}}=0.28650 \qquad [1]$$

We can calculate $P\left(N(5)=1\right)$ using a differential equation or an integral equation. Integral equations are not covered until Part 2 of the course, but we include this approach here as a valid alternative.

Method 1 – differential equation

Adopting the usual notation:

$$p_{ij}(t)=P\left(N(t)=j\,|\,N(0)=i\right)$$

the forward differential equation for $P\left(N(t)=1\right)=p_{01}(t)$ is:

$$\frac{d}{dt}p_{01}(t)=\frac{1}{4}p_{00}(t)-\frac{1}{2}p_{01}(t) \qquad [1]$$

This can be solved using the integrating factor method, by first writing it in the form:

$$\frac{d}{dt}p_{01}(t)+\frac{1}{2}p_{01}(t)=\frac{1}{4}p_{00}(t)$$

For this model, $p_{ii}(t) = p_{\overline{ii}}(t)$ since, once state i has been left, a return to it is impossible. So:

$$p_{00}(t) = p_{\overline{00}}(t) = e^{-\frac{t}{4}}$$

and the differential equation becomes:

$$\frac{d}{dt}p_{01}(t) + \frac{1}{2}p_{01}(t) = \frac{1}{4}e^{-\frac{t}{4}} \qquad [½]$$

Multiplying through by the integrating factor $e^{\frac{t}{2}}$ gives:

$$e^{\frac{t}{2}}\frac{d}{dt}p_{01}(t) + \frac{1}{2}e^{\frac{t}{2}}p_{01}(t) = \frac{1}{4}e^{\frac{t}{4}} \qquad [½]$$

The LHS can also be written as:

$$\frac{d}{dt}\left(e^{\frac{t}{2}}p_{01}(t)\right)$$

This can be checked using the product rule for differentiation.

Then integrating both sides of the equation with respect to t, we get:

$$e^{\frac{t}{2}}p_{01}(t) = e^{\frac{t}{4}} + C \qquad [½]$$

where C is a constant of integration.

Setting $t = 0$:

$$0 = 1 + C \qquad \Rightarrow \qquad C = -1 \qquad [½]$$

Hence:

$$p_{01}(t) = e^{-\frac{t}{2}}\left(e^{\frac{t}{4}} - 1\right)$$

and:

$$P\left(N(5) = 1\right) = p_{01}(5) = e^{-\frac{5}{2}}\left(e^{\frac{5}{4}} - 1\right) = 0.20442 \qquad [½]$$

Method 2 – integral equation

We can write $P\left(N(5)=1\right)$ in the following integral form:

$$P\left(N(5)=1\right) = \int_0^5 \underbrace{p_{00}(t)}_{\substack{\text{stay in state}\\ \text{0 up to time } t}} \quad \underbrace{\mu_{01}}_{\substack{\text{move from}\\ \text{state 0 to}\\ \text{state 1 at}\\ \text{time } t}} \quad \underbrace{p_{11}(5-t)}_{\substack{\text{stay in state 1}\\ \text{from time } t \text{ to}\\ \text{time 5}}} \; dt \qquad \textit{[1]}$$

For this model:

$$p_{00}(t) = p_{\overline{00}}(t) = e^{-\frac{t}{4}} \qquad [½]$$

and: $$p_{11}(5-t) = p_{\overline{11}}(5-t) = e^{-\left(\frac{5-t}{2}\right)} \qquad [½]$$

So:

$$P\big(N(5)=1\big) = \int_0^5 e^{-\frac{t}{4}} \tfrac{1}{4} e^{-\left(\frac{5-t}{2}\right)} dt = \tfrac{1}{4} e^{-\frac{5}{2}} \int_0^5 e^{\frac{t}{4}}\, dt$$

$$= e^{-\frac{5}{2}} \left[e^{\frac{t}{4}} \right]_0^5 = e^{-\frac{5}{2}} \left(e^{\frac{5}{4}} - 1 \right) = 0.20442 \qquad [1½]$$

as before.

So the probability that a new boiler will break down more than once in the next 5 years is:

$$P\big(N(5)>1\big) = 1 - 0.28650 - 0.20442 = 0.50908 \qquad [½]$$

[Total 5]

A completely different approach is to use a 3-state model where:

- *State 0 = never broken down*
- *State 1 = broken down once*
- *State 2+ = broken down more than once.*

Then $P\big(N(5)>1\big) = p_{02+}(5)$. *We can calculate this using the differential equation:*

$$\frac{d}{dt} p_{02+}(t) = p_{00}(t)\mu_{02+} + p_{01}(t)\mu_{12+} + p_{02+}(t)\mu_{2+2+}$$

$$= p_{00}(t) \times 0 + p_{01}(t) \times \frac{1}{2} + p_{02+}(t) \times 0$$

$$= \frac{1}{2} p_{01}(t)$$

$$= \frac{1}{2}\big(1 - p_{00}(t) - p_{02+}(t)\big) \qquad [2]$$

This equation can be solved using the integrating factor method by first writing it in the form:

$$\frac{d}{dt} p_{02+}(t) + \frac{1}{2} p_{02+}(t) = \frac{1}{2}\big(1 - p_{00}(t)\big) = \frac{1}{2}\left(1 - e^{-\frac{t}{4}}\right) \qquad [½]$$

Multiplying through by the integrating factor $e^{\frac{t}{2}}$ *gives:*

$$e^{\frac{t}{2}}\frac{d}{dt}p_{02+}(t)+\frac{1}{2}e^{\frac{t}{2}}p_{02+}(t)=\frac{1}{2}\left(e^{\frac{t}{2}}-e^{\frac{t}{4}}\right) \qquad [½]$$

The LHS can also be written as:

$$\frac{d}{dt}\left(e^{\frac{t}{2}}p_{02+}(t)\right)$$

This can be checked using the product rule for differentiation.

Then integrating both sides of the equation with respect to t *:*

$$e^{\frac{t}{2}}p_{02+}(t)=e^{\frac{t}{2}}-2e^{\frac{t}{4}}+C \qquad [1]$$

where C *is a constant of integration.*

Setting $t=0$ *:*

$$0=1-2+C \qquad \Rightarrow \qquad C=1 \qquad [½]$$

So:

$$p_{02+}(t)=1-2e^{-\frac{t}{4}}+e^{-\frac{t}{2}}$$

and:

$$P\left(N(5)>1\right)=p_{02+}(5)=1-2e^{-\frac{5}{4}}+e^{-\frac{5}{2}}=0.50908 \qquad [½]$$

[Total 5]

(ii) ***Expected lifetime of a boiler***

For $i=0,1,...,9$, the expected holding time in State i is $\frac{1}{\lambda_i}$, where λ_i is the total force out of State i. State i must be followed by State $i+1$. So the expected lifetime of a boiler is:

$$\sum_{i=0}^{9}\frac{1}{\lambda_i}=4+2+\left(8\times 1\right)=14 \text{ years} \qquad [1]$$

Part (ii) cannot be answered using the alternative 3-state model, as this does not explicitly consider the 10th breakdown.

Solution X1.6

This question is about estimating transition rates in a time-homogeneous Markov jump process. It is based on the material in Chapter 4.

(i) ***Likelihood function***

The likelihood function is:

$$L = e^{-904(\sigma+\mu)} e^{-112(\rho+\upsilon)} \sigma^{34} \rho^{26} \mu^2 \upsilon^7$$ [2]

Full marks should be awarded if a constant factor has been included or if the equal to symbol has been replaced with the proportionality symbol, $\propto$.

(ii) ***Maximum likelihood estimate***

The log-likelihood function is:

$$\ln L = -904(\sigma+\mu) - 112(\rho+\upsilon) + 34\ln\sigma + 26\ln\rho + 2\ln\mu + 7\ln\upsilon$$ [½]

The log-likelihood function may also be given as:

$$\ln L = -112\rho + 26\ln\rho + \text{terms that don't involve } \rho$$

Differentiating with respect to ρ:

$$\frac{\partial \ln L}{\partial \rho} = -112 + \frac{26}{\rho}$$ [1]

Setting this equal to 0 gives:

$$\hat{\rho} = \frac{26}{112} = 0.23214$$ [½]

[Total 2]

(iii) ***Confidence interval***

Let $\tilde{\rho}$ denote the maximum likelihood estimator of ρ. Since $\tilde{\rho}$ is asymptotically normally distributed, an approximate 95% confidence interval for ρ is:

$$\hat{\rho} \pm 1.96\sqrt{\text{var}(\tilde{\rho})}$$ [1]

Asymptotically, the variance of the estimator is given by the CRLB:

$$\text{var}(\tilde{\rho}) = \frac{-1}{E\left(\dfrac{\partial^2 \ln L}{\partial \rho^2}\right)}$$

The formula for the CRLB is given on page 23 of the Tables.

The required second derivative is:

$$\frac{\partial^2 \ln L}{\partial \rho^2} = -\frac{26}{\rho^2}$$

So $\text{var}(\tilde{\rho})$ is estimated by:

$$\frac{-1}{\left.\dfrac{\partial^2 \ln L}{\partial \rho^2}\right|_{\rho=\hat{\rho}}} = \frac{\hat{\rho}^2}{26} = \frac{\left(\frac{26}{112}\right)^2}{26} = 0.0020727$$ [1]

Hence an approximate 95% confidence interval for ρ is:

$$0.23214 \pm 1.96\sqrt{0.0020727} = (0.14291,\ 0.32138)$$ [1]

[Total 3]

Solution X1.7

This question involves a derivation from Chapter 3.

Consider a time interval of length $t+h$, where h is a small amount. The probability that a life now aged x survives for the next $t+h$ years is ${}_{t+h}p_x$. Splitting the interval into two parts (one of length t years and the other of length h years) and using the Markov property, we have:

$$_{t+h}p_x = {}_tp_x \times {}_hp_{x+t} \qquad [1]$$

Now:

$$_hp_{x+t} = 1 - {}_hq_{x+t} = 1 - \left(h\mu_{x+t} + o(h)\right) \qquad [1]$$

So:

$$_{t+h}p_x = {}_tp_x \times {}_hp_{x+t} = {}_tp_x(1 - h\mu_{x+t}) + o(h)$$

Rearranging, we see that:

$$\frac{_{t+h}p_x - {}_tp_x}{h} = -{}_tp_x\,\mu_{x+t} + \frac{o(h)}{h} \qquad [1]$$

Then letting $h \to 0$ gives:

$$\frac{\partial}{\partial t}\,{}_tp_x = -{}_tp_x\,\mu_{x+t} \qquad [½]$$

since $\lim\limits_{h\to 0}\dfrac{o(h)}{h} = 0$. [½]

We can solve this differential equation by separating the variables:

$$\frac{\frac{\partial}{\partial t}\,{}_tp_x}{_tp_x} = \frac{\partial}{\partial t}\ln {}_tp_x = -\mu_{x+t} \qquad [1]$$

Changing notation from t to s, we have:

$$\frac{\partial}{\partial s}\ln {}_sp_x = -\mu_{x+s}$$

Now integrating with respect to s between the limits $s=0$ and $s=t$, we obtain:

$$\left[\ln {}_sp_x\right]_0^t = -\int_0^t \mu_{x+s}\,ds \qquad [1]$$

ie: $$\ln {}_tp_x - \ln {}_0p_x = -\int_0^t \mu_{x+s}\,ds$$

Since ${}_0p_x = 1$ and $\ln 1 = 0$, this simplifies to:

$$\ln {}_tp_x = -\int_0^t \mu_{x+s}\, ds$$ [½]

So, taking exponentials gives the result:

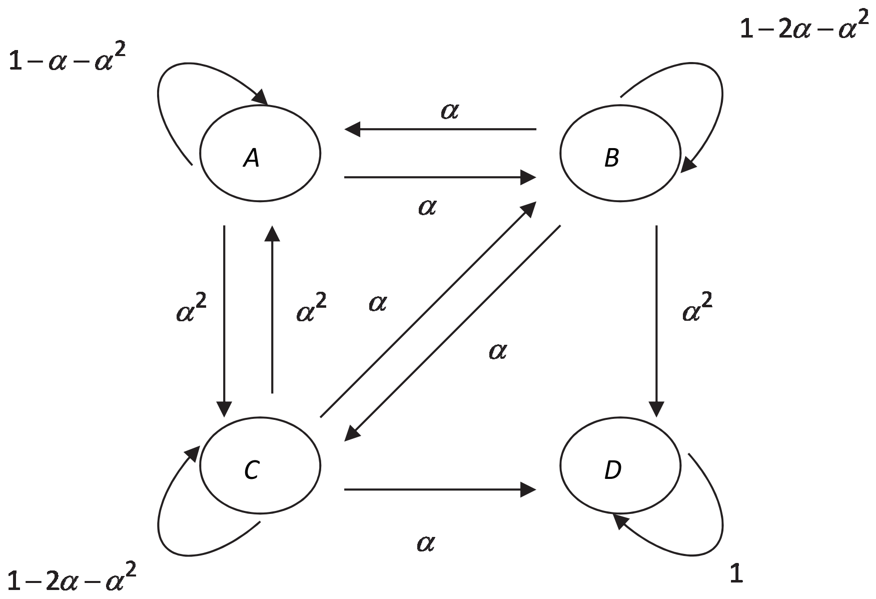
$${}_tp_x = \exp\left(-\int_0^t \mu_{x+s}\, ds \right)$$ [½]

[Total 7]

Solution X1.8

This is another question about Markov chains, based on the material in Chapter 2.

(i) ***Transition graph***



[2]

(ii) ***Range of values***

The transition matrix will be valid if the entries in each row add up to 1 (which they do) and each entry lies in the range $[0,1]$.

We can see that the α entries require $0 \le \alpha \le 1$ and that the α^2 entries require $-1 \le \alpha \le 1$. So we must have $0 \le \alpha \le 1$. [½]

The $1-\alpha-\alpha^2$ entry requires $0 \le 1-\alpha-\alpha^2 \le 1$. Since $\alpha \ge 0$, it automatically follows that $1-\alpha-\alpha^2 \le 1$. However, we need to work out the values of α for which $0 \le 1-\alpha-\alpha^2$. The roots of the quadratic equation $0 = 1-\alpha-\alpha^2$ can be obtained using the quadratic formula:

$$\alpha = \frac{1 \pm \sqrt{(-1)^2 - 4\times(-1)\times 1}}{-2} = \frac{1 \pm \sqrt{5}}{-2} = 0.618 \text{ or } -1.618 \qquad [1]$$

Since the coefficient of α^2 is negative, the graph of $1-\alpha-\alpha^2$ has an inverted U shape. This means that it takes positive values between the two roots, *ie* for the range $-1.618 \le \alpha \le 0.618$.

[½]

The $1-2\alpha-\alpha^2$ entries require $0\le 1-2\alpha-\alpha^2\le 1$. Again, since $\alpha\ge 0$, it automatically follows that $1-2\alpha-\alpha^2\le 1$. However, we need to work out the values of α for which $0\le 1-2\alpha-\alpha^2$. The roots of the quadratic equation $0=1-2\alpha-\alpha^2$ are:

$$\alpha=\frac{2\pm\sqrt{(-2)^2-4\times(-1)\times 1}}{-2}=\frac{2\pm\sqrt{8}}{-2}=0.414 \text{ or } -2.414$$ [1]

The graph of $1-2\alpha-\alpha^2$ also has an inverted U shape. So it takes positive values when $-2.414\le\alpha\le 0.414$. [½]

Putting all of this together, we see that all of the conditions:

$0\le\alpha\le 1$

$-1.618\le\alpha\le 0.618$

$-2.414\le\alpha\le 0.414$

must be satisfied. So we must have $0\le\alpha\le 0.414$. [½]

[Total 4]

(iii) ***Irreducible and/or aperiodic?***

The chain is not irreducible since State *D* is an absorbing state (*ie* it is impossible to leave State *D*). [½]

In the case when $\alpha<0.414$, every state has an arrow to itself, and so every state is aperiodic (as a return is possible in any number of steps). [1]

In the case when $\alpha=0.414$, neither State *B* nor State *C* has an arrow to itself. However, a return to State *B* is possible in 2 or 3 or 4 *etc* steps. These numbers are not restricted to a multiple of some number greater than 1. So State *B* is aperiodic. The same reasoning applies for State *C*. [1]

So the chain is aperiodic. [½]

[Total 3]

(iv)(a) ***Expected number of quarters until rating changes***

If $\alpha=0.1$, then $1-2\alpha-\alpha^2=1-0.2-0.01=0.79$. So the probability that Company XYZ is rated *B* in Quarter 2 is 0.79 and the probability that its rating changes at the end of the first quarter is 0.21. [½]

The probability that the first change happens at the end of the second quarter is 0.79×0.21. [½]

Similarly the probability that the first change happens at the end of the third quarter is $0.79^2\times 0.21$, and so on.

So, the expected number of quarters until the first rating change is:

$$(1\times 0.21)+(2\times 0.79\times 0.21)+(3\times 0.79^2\times 0.21)+(4\times 0.79^3\times 0.21)+\cdots$$ [1]

We can write this as:

$$0.21\left(1+2\times 0.79+3\times 0.79^2+4\times 0.79^3+\cdots\right) \qquad [½]$$

The expression in brackets is of the form $1+2x+3x^2+4x^3+\cdots$, where $x=0.79$. Using the binomial expansion:

$$(1-x)^{-2}=1+2x+3x^2+4x^3+\cdots \quad \text{for } -1<x<1$$

we see that:

$$1+2\times 0.79+3\times 0.79^2+4\times 0.79^3+\cdots=(1-0.79)^{-2}=\frac{1}{0.21^2} \qquad [1]$$

So the expected number of quarters until the first rating change is:

$$0.21\times\frac{1}{0.21^2}=\frac{1}{0.21}=4.76 \qquad [½]$$

Alternatively, we could say:

$$\begin{aligned}1+2\times 0.79+3\times 0.79^2+4\times 0.79^3+\cdots=1&+0.79+0.79^2+0.79^3+\cdots\\ &+0.79+0.79^2+0.79^3+\cdots\\ &\quad+0.79^2+0.79^3+\cdots\\ &\quad\quad+\cdots\end{aligned}$$

Each line on the RHS of the equation immediately above is the sum to infinity of a geometric progression. Using the formula $\dfrac{a}{1-r}$ *where* a *denotes the first term and* r *denotes the common ratio, we see that:*

$$\begin{aligned}1+2\times 0.79+3\times 0.79^2+4\times 0.79^3+\cdots&=\frac{1}{1-0.79}+\frac{0.79}{1-0.79}+\frac{0.79^2}{1-0.79}+\cdots\\ &=\frac{1}{0.21}\left(1+0.79+0.79^2+\cdots\right)\\ &=\frac{1}{0.21}\left(\frac{1}{1-0.79}\right)\\ &=\frac{1}{0.21^2}\end{aligned}$$

as before. *[1]*

Another alternative is to let:

$$S=1+2\times 0.79+3\times 0.79^2+4\times 0.79^3+\cdots \qquad (*)$$

Then:

$$0.79S = 0.79 + 2 \times 0.79^2 + 3 \times 0.79^3 + 4 \times 0.79^4 + \cdots \qquad (\dagger)$$

Subtracting (†) from (*) gives:

$$(1 - 0.79)S = 1 + 0.79 + 0.79^2 + 0.79^3 + 0.79^4 + \cdots$$

The terms on the RHS form a geometric progression with $a = 1$ *and* $r = 0.79$*, so:*

$$RHS = \frac{1}{1 - 0.79} = \frac{1}{0.21}$$

and hence:

$$S = \frac{1}{0.21^2} \qquad [1]$$

(iv)(b) ***Probability that first change is an upgrade***

The probability that the first rating change is an upgrade is the probability that Company XYZ moves to a rating of *A* when it first changes rating. This is:

$$P(\text{move to State A} \mid \text{leave State B}) = \frac{P(\text{move to State A})}{P(\text{leave State B})} = \frac{\alpha}{2\alpha + \alpha^2} = \frac{0.1}{0.21} = \frac{10}{21} = 0.47619$$

[1]

Alternatively, we can calculate this probability as follows:

$$P(\text{upgrade at end of Q1}) + P(\text{stay in B in Q2, and upgrade at end of Q2})$$

$$+ P(\text{stay in B in Q2 and Q3, and upgrade at end of Q4})$$

$$+ \cdots$$

$$= 0.1 + 0.79 \times 0.1 + 0.79^2 \times 0.1 + \cdots$$

$$= \frac{0.1}{1 - 0.79}$$

$$= 0.47619$$

[1]

[Total 5]

Solution X1.9

This question involves turning a non-Markov process into a Markov process by increasing the number of states. It is an application of the material in Chapter 2.

(i) ***Why process is not Markov***

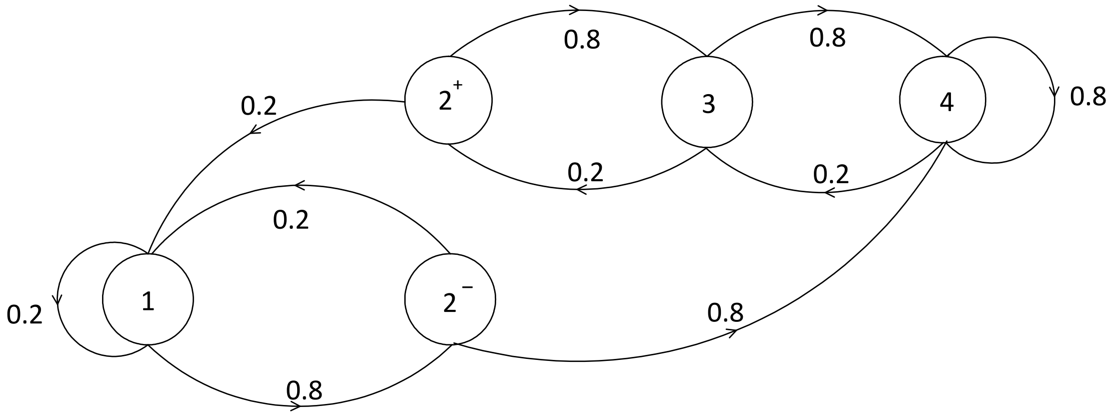
If a policyholder is on Level 2, the probability of moving to Level 4 depends on the level the policyholder was on last year. Hence the process is not Markov. [1]

(ii) ***New process***

If a policyholder on Level 2 has a claim-free year, then next year this policyholder will be on either Level 3 or Level 4, depending on whether or not the previous year was claim-free. So we need to split Level 2 into two levels. Let's call these Level 2^- (moving up from Level 1, *ie* the previous year was claim-free) and Level 2^+ (moving down from Level 3, *ie* the previous year was not claim-free). If a policyholder on Level 2^- has a claim-free year, then next year this policyholder will move to Level 4 (*ie* move up two levels). If a policyholder on Level 2^+ has a claim-free year, then next year this policyholder will move up to Level 3 (*ie* move up one level). This new process is Markov and has 5 states. [2]

(iii) ***Transition graph for Y(t)***

Using the labelling system defined in part (ii), the transition graph is as follows:



[2]

(iv) ***Transition matrix***

The one-step transition matrix is:

$$\begin{array}{c|ccccc} & 1 & 2^- & 2^+ & 3 & 4 \\ \hline 1 & 0.2 & 0.8 & 0 & 0 & 0 \\ 2^- & 0.2 & 0 & 0 & 0 & 0.8 \\ 2^+ & 0.2 & 0 & 0 & 0.8 & 0 \\ 3 & 0 & 0 & 0.2 & 0 & 0.8 \\ 4 & 0 & 0 & 0 & 0.2 & 0.8 \end{array}$$

[1]

(v) ***Sufficient conditions for unique stationary distribution***

The chain has a finite number of states, 5 …. [½]

… so it has at least one stationary distribution. [½]

The chain is also irreducible as every state can be reached from every other state … [½]

… so it has a unique stationary distribution. [½]

Because the chain is irreducible, the states will either all be aperiodic or will all have the same period. State 1 is aperiodic since it is possible to stay in state 1 in successive time periods. So all the states are aperiodic. [½]

Since the chain is aperiodic as well as having a finite state space and being irreducible, the process with settle down to its unique stationary distribution in the long run. [½]
[Total 3]

(vi) ***Probability of being on Level 2 in the long run***

The stationary distribution is the vector of probabilities $\underline{\pi}$ that satisfies the equation:

$$\underline{\pi}P = \underline{\pi}$$

where P is the transition matrix from part (iv). Writing out the system of equations in full, we have:

(1) $0.2\pi_1 + 0.2\pi_{2^-} + 0.2\pi_{2^+} = \pi_1$

(2) $0.8\pi_1 = \pi_{2^-}$

(3) $0.2\pi_3 = \pi_{2^+}$

(4) $0.8\pi_{2^+} + 0.2\pi_4 = \pi_3$

(5) $0.8\pi_{2^-} + 0.8\pi_3 + 0.8\pi_4 = \pi_4$ [1]

We will discard equation (1) and replace it with:

(6) $\pi_1 + \pi_{2^-} + \pi_{2^+} + \pi_3 + \pi_4 = 1$

We will now use equations (2) to (5) to express all the probabilities as multiples of π_{2^+}.

Rearranging (3), we see that:

$$\pi_3 = 5\pi_{2^+}$$

Using this in (4) gives:

$$\begin{aligned}\pi_4 &= 5\left(\pi_3 - 0.8\pi_{2^+}\right)\\ &= 5\left(5\pi_{2^+} - 0.8\pi_{2^+}\right)\\ &= 21\pi_{2^+}\end{aligned}$$

Then from (5):

$$\begin{aligned}0.8\pi_{2^-} &= -0.8\pi_3 + 0.2\pi_4\\ &= -0.8(5\pi_{2^+}) + 0.2(21\pi_{2^+})\\ &= 0.2\pi_{2^+}\end{aligned}$$

$$\Rightarrow \pi_{2^-} = 0.25\pi_{2^+}$$

and from (2):

$$\pi_1 = \frac{1}{0.8}\pi_{2^-} = 1.25(0.25\pi_{2^+}) = 0.3125\pi_{2^+}$$ [2]

We can now use (6) to obtain the value of π_{2^+} :

$$\left(0.3125 + 0.25 + 1 + 5 + 21\right)\pi_{2^+} = 1$$

$$\Rightarrow \pi_{2^+} = \frac{16}{441}$$ [1]

In addition:

$$\pi_{2^-} = 0.25\pi_{2^+} = \frac{4}{441}$$

So the long-run probability of being on Level 2 is:

$$\pi_{2^-} + \pi_{2^+} = \frac{20}{441}$$ [1]

There is no need to work out the complete stationary distribution as we are only interested in the long-run probability of being on 10% discount.

[Total 5]

Solution X1.10

This question involves an application of the material in Chapter 4 on time-homogeneous Markov jump processes.

(i) ***Generator matrix and transition diagram***

Let T_j denote the holding time in State j, $j = A, F, I, O$. Then each T_j is an exponential random variable. Let λ_j denote the exponential parameter for T_j, so that $E(T_j) = \frac{1}{\lambda_j}$. λ_j also represents the total force of transition out of State j.

We are told that the expected waiting time in State *A* is 1 hour. So $\lambda_A = 1$ per hour.

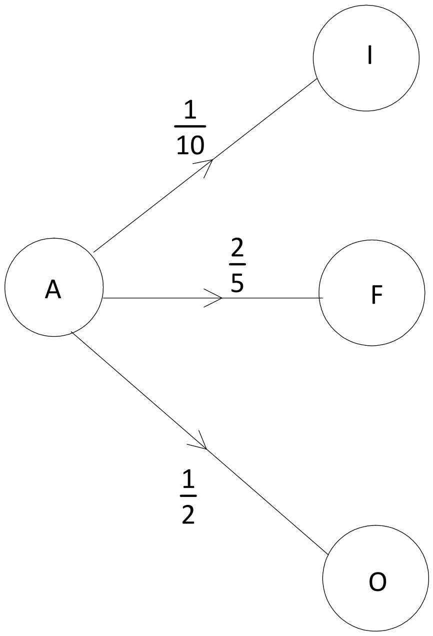
On leaving State *A*, a patient enters one of State *I*, State *O* or State *F*.

The probability of moving to State *I* is 1 in 10. So $\mu_{AI} = \frac{1}{10}\lambda_A = \frac{1}{10}$ per hour.

The probability of moving to State *O* is 5 in 10. So $\mu_{AO} = \frac{5}{10}\lambda_A = \frac{1}{2}$ per hour.

The remainder of the total force out of State *A* must be directed towards State *F*. So
$\mu_{AF} = 1 - \frac{1}{10} - \frac{1}{2} = \frac{2}{5}$ per hour. [1]

We can use this information to begin to construct the transition diagram:



Now consider transitions out of State *F*. The expected holding time in State *F* is 3 hours. So λ_F, the total force of transition out of State *F*, is $\frac{1}{3}$ per hour.

On leaving State *F*, a patient enters one of State *D*, State *I* or State *O*.

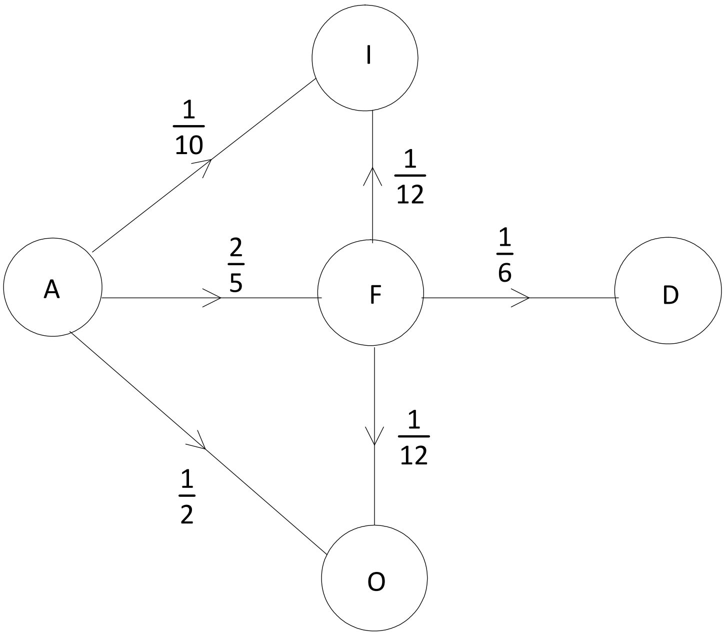
The corresponding probabilities are 50%, 25% and 25%. So:

$$\mu_{FD} = 0.5 \times \tfrac{1}{3} = \tfrac{1}{6} \text{ per hour}$$

$$\mu_{FI} = 0.25 \times \tfrac{1}{3} = \tfrac{1}{12} \text{ per hour}$$

$$\mu_{FO} = 0.25 \times \tfrac{1}{3} = \tfrac{1}{12} \text{ per hour}$$ [1]

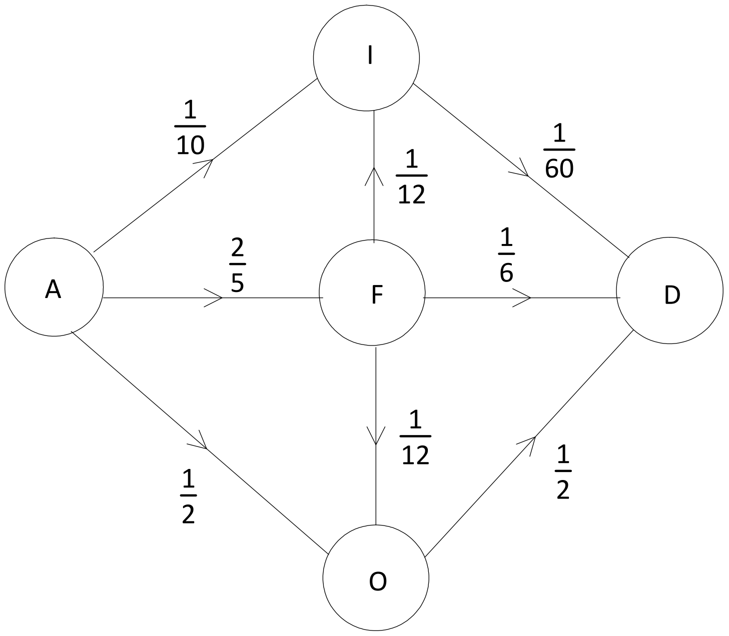
Adding these transitions to the diagram, it becomes:



The expected holding time in State *O* is 2 hours, after which the patient is discharged. So $\mu_{OD} = \tfrac{1}{2}$ per hour. [½]

The expected holding time in State *I* is 60 hours, after which the patient is discharged. So $\mu_{ID} = \tfrac{1}{60}$ per hour. [½]

So the completed transition diagram is as follows:



[1]

Markers: Please award 4 marks for the correct diagram. Students do not have to explain how the rates are calculated in order to obtain the marks.

The generator matrix is:

$$\begin{array}{c} \\ A \\ F \\ I \\ O \\ D \end{array}\begin{array}{c} \begin{array}{ccccc} A & F & I & O & D \end{array} \\ \begin{bmatrix} -1 & \frac{2}{5} & \frac{1}{10} & \frac{1}{2} & 0 \\ 0 & -\frac{1}{3} & \frac{1}{12} & \frac{1}{12} & \frac{1}{6} \\ 0 & 0 & -\frac{1}{60} & 0 & \frac{1}{60} \\ 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

[1]

[Total 5]

(ii) ***Proportion receiving in-patient treatment***

Patients will end up in State I either by going directly from State A or by going via State F. So the probability of eventually reaching state I is:

$$0.1 + 0.4 \times 0.25 = 0.2$$ [1]

(iii)(a) ***Probability that patient is yet to be classified by a junior doctor***

This is the (occupancy) probability that the patient is still in State A at time t hours:

$$p_{\overline{AA}}(t) = e^{-1\times t} = e^{-t}$$ [1]

(iii)(b) ***Probability that patient is undergoing further investigation***

Here we want an expression for $p_{AF}(t)$. We can determine this from the forward differential equation:

$$\frac{d}{dt}p_{AF}(t) = p_{AA}(t)\times\tfrac{2}{5} + p_{AF}(t)\times\left(-\tfrac{1}{3}\right) \qquad [1]$$

where $p_{AA}(t) = p_{\overline{AA}}(t) = e^{-t}$ since it is impossible to return to State A once it has been left.

We can solve this using the integrating factor method. We first need to rewrite the equation as:

$$\frac{d}{dt}p_{AF}(t) + \tfrac{1}{3}p_{AF}(t) = \tfrac{2}{5}e^{-t}$$

Then multiplying through by the integrating factor $e^{t/3}$, we obtain:

$$e^{t/3}\frac{d}{dt}p_{AF}(t) + \tfrac{1}{3}e^{t/3}p_{AF}(t) = \tfrac{2}{5}e^{-2t/3}$$

The LHS can also be written as:

$$\frac{d}{dt}\left(e^{t/3}p_{AF}(t)\right)$$

Integrating both sides of this equation with respect to t :

$$e^{t/3}p_{AF}(t) = -\tfrac{3}{5}e^{-2t/3} + C \qquad [1]$$

where C is a constant of integration.

When $t = 0$, $p_{AF}(t) = 0$. So:

$$C = \tfrac{3}{5}$$

and: $$p_{AF}(t) = \tfrac{3}{5}(e^{-t/3} - e^{-t}) \qquad [1]$$

[Total 4]

Alternatively, we could use an integral approach. Integral equations are not covered until Part 2 of the course, but we include this approach here as a valid alternative.

A patient who is in state F after t hours must have started in State A at time 0, remained in state A until some earlier time s, say, then made the transition to state F at time s and remained in state F from time s to time t. The notation that represents this sequence of events is:

$$p_{\overline{AA}}(s)\mu_{AF}\, p_{\overline{FF}}(t-s) = e^{-s}\times\tfrac{2}{5}\times e^{-\frac{1}{3}(t-s)} \qquad \textit{[1]}$$

Since s can take any value between 0 and t, we need to integrate with respect to s between $s=0$ and $s=t$. So the required probability is:

$$\int_0^t e^{-s} \times \frac{2}{5} \times e^{-\frac{1}{3}(t-s)} ds = \frac{2}{5} e^{-\frac{1}{3}t} \int_0^t e^{-\frac{2}{3}s} ds = \frac{2}{5} e^{-\frac{1}{3}t} \times \frac{3}{2}(1-e^{-\frac{2}{3}t}) = \frac{3}{5}(e^{-\frac{1}{3}t} - e^{-t})$$ *[2]*

(iv) ***Expected time until discharge***

Let m_j denote the expected time until discharge for a patient currently in State j, $j = A, F, I, O$. We have to calculate the value of m_A.

A patient is expected to spend 1 hour in State *A* before moving on to one of States *F*, *I* or *O*. The corresponding probabilities are 0.4, 0.1 and 0.5. So:

$$m_A = 1 + 0.4m_F + 0.1m_I + 0.5m_O$$ [1]

We know that $m_I = 60$ hours and $m_O = 2$ hours.

A patient in State *F* is expected to remain in that state for 3 hours before moving to one of States *I*, *O* or *D*. The corresponding probabilities are 0.25, 0.25 and 0.5. So:

$$m_F = 3 + 0.25m_I + 0.25m_O = 3 + 0.25 \times 60 + 0.25 \times 2 = 18.5 \text{ hours}$$ [1]

and hence:

$$m_A = 1 + 0.4 \times 18.5 + 0.1 \times 60 + 0.5 \times 2 = 15.4 \text{ hours}$$ [1]

[Total 3]

(v) ***Is a time-homogeneous model appropriate?***

By using a time-homogeneous model, we are assuming that the transition rates (and hence expected waiting times and proportions of patients classified in each category) are constant over time. [1]

This is unlikely to be the case in practice as the department may be particularly busy at weekends or during the winter months, resulting in longer waiting times for patients. Also, the periods of time for treatments and the proportion of patients classified in each category may change over time with medical advances or changes in medical guidelines. So a time-inhomogeneous model is likely to be more appropriate. [1]

[Total 2]

Solution X2.1

This question is about curtate future lifetime, which is introduced in Chapter 6.

The curtate expectation of life of a new-born insect is:

$$e_0 = \sum_{k=1}^{\infty} {}_k p_0 = \sum_{k=1}^{\infty} e^{-k\lambda} = e^{-\lambda} + e^{-2\lambda} + e^{-3\lambda} + \cdots \quad [1]$$

This is the sum to infinity of a geometric progression with first term $a = e^{-\lambda}$ and common ratio $r = e^{-\lambda}$. Using the formula $\frac{a}{1-r}$ for the sum to infinity, we see that:

$$e_0 = \frac{e^{-\lambda}}{1-e^{-\lambda}} \quad [1]$$

[Total 2]

Solution X2.2

This is a survival/death probability question, based on the material in Chapter 6.

The probability that the life dies between exact age 79 and exact age 82 is:

$${}_6p_{73}\ {}_3q_{79} = {}_6p_{73}\,(1 - {}_3p_{79}) \quad [½]$$

Breaking up the probability ${}_6p_{73}$ at age 75 (as the force of mortality changes at that age), we have:

$${}_6p_{73} = {}_2p_{73}\ {}_4p_{75} = e^{-2\times 0.02}\, e^{-4\times 0.04} = e^{-0.2} \quad [1]$$

Similarly, breaking up the probability ${}_3p_{79}$ at age 80, we have:

$${}_3p_{79} = p_{79}\ {}_2p_{80} = e^{-0.04}\, e^{-2\times 0.07} = e^{-0.18} \quad [1]$$

So the required probability is:

$$e^{-0.2}(1-e^{-0.18}) = 0.13487 \quad [½]$$

Markers: Please give credit for other valid approaches, eg calculating the probability as:

$${}_6p_{73}\ q_{79} + {}_7p_{73}\ {}_2q_{80}$$

[Total 3]

Solution X2.3

This question is also based on material from Chapter 6.

(i)(a) ***Survival function***

The survival function is defined as follows:

$$S_x(t) = P(T_x > t) \qquad [1]$$

(i)(b) ***Force of mortality***

In terms of probabilities involving T_x :

$$\mu_{x+t} = \lim_{h \to 0} \frac{P(T_x \le t+h \mid T_x > t)}{h} \qquad [1]$$

[Total 2]

(ii) ***Force of mortality under the Weibull model***

Using the formula:

$$\mu_{x+t} = -\frac{\partial}{\partial t} \ln {}_t p_x = -\frac{\partial}{\partial t} \ln S_x(t) \qquad [1]$$

we see that, for the Weibull model:

$$\mu_{x+t} = -\frac{\partial}{\partial t}(-\alpha t^{\beta}) = \alpha \frac{\partial}{\partial t} t^{\beta} = \alpha \beta t^{\beta-1} \qquad [1]$$

[Total 2]

The formula $\mu_{x+t} = -\frac{\partial}{\partial t} \ln S_x(t)$ *can be derived from the expression given in (i)(b) as follows:*

$$\begin{aligned}\mu_{x+t} &= \lim_{h \to 0} \frac{P(T_x \le t+h \mid T_x > t)}{h} = \lim_{h \to 0} \frac{P(T_x \le t+h \text{ and } T_x > t)}{hP(T_x > t)} \\ &= \lim_{h \to 0} \frac{P(t < T_x \le t+h)}{hP(T_x > t)} = \lim_{h \to 0} \frac{P(T_x > t) - P(T_x > t+h)}{hP(T_x > t)} \\ &= \lim_{h \to 0} \frac{S_x(t) - S_x(t+h)}{h\, S_x(t)} = -\frac{1}{S_x(t)} \lim_{h \to 0} \frac{S_x(t+h) - S_x(t)}{h} \\ &= -\frac{1}{S_x(t)} \times \frac{\partial}{\partial t} S_x(t) = -\frac{\partial}{\partial t} \ln S_x(t)\end{aligned}$$

Another way of obtaining the expression for μ_{x+t} is to compare the general formula ${}_tp_x = \exp\left(-\int_0^t \mu_{x+s}\,ds\right)$ (which appears on page 32 of the Tables) with the given survival function. Doing this, we see that:

$$\exp\left(-\alpha\, t^{\beta}\right) = \exp\left(-\int_0^t \mu_{x+s}\,ds\right)$$

$$\Rightarrow \int_0^t \mu_{x+s}\,ds = \alpha t^{\beta}$$ *[1]*

Differentiating both sides with respect to t gives the same answer as above. *[1]*

(iii) ***Expected value***

When $\beta = 1$, $\mu_{x+t} = \alpha$. Since the force of mortality is constant, the random variable T_x is exponentially distributed with parameter α. So $E(T_x) = \dfrac{1}{\alpha}$. [1]

Another way to deduce that the distribution of T_x is exponential is to compare the survival functions for the Weibull and exponential distributions:

Weibull: $S_x(t) = \exp(-\alpha t^{\beta}) = e^{-\alpha}$ *when* $\beta = 1$

Exponential: $S_x(t) = e^{-\lambda}$

These have the same form, so $T_x \sim Exp(\alpha)$ when $\beta = 1$ and hence $E(T_x) = \dfrac{1}{\alpha}$. *[1]*

The expected value can also be derived by integration:

$$E(T_x) = \int_0^{\infty} {}_tp_x\,dt = \int_0^{\infty} S_x(t)\,dt = \int_0^{\infty} e^{-\alpha t}\,dt = \left[\frac{e^{-\alpha t}}{-\alpha}\right]_0^{\infty} = \frac{1}{\alpha}$$ *[1]*

Solution X2.4

This question is based on material covered in Chapters 4 and 5.

(i)(a) ***Definition of a Markov jump process***

A Markov jump process is a stochastic process with a continuous time set and a discrete state space that satisfies the Markov property. [1]

The Markov property says that the past history of the process is irrelevant. It is only the current state that affects the transition probabilities. [1]

Alternatively, we could define the Markov property mathematically. The process $X(t)$ has the Markov property if:

$$P\left[X(t) \in A \mid X(s_1) = x_1, X(s_2) = x_2, \ldots, X(s_n) = x_n, X(s) = x\right] = P\left[X(t) \in A \mid X(s) = x\right]$$

for all times $s_1 < s_2 < \cdots < s_n < s < t$ in the time set, all states $x_1, x_2, \ldots, x_n$ and x in the state space, S, and all subsets A of S. *[1]*

(i)(b) ***Condition needed for a Markov jump process to be time-homogeneous***

A Markov jump process is time-homogeneous if its transition probabilities $P\left(X_t = j \mid X_s = i\right)$ depend only on the length of the time interval $t - s$. [1]

[Total 3]

Alternatively, we could say that a Markov jump process is time-homogeneous if its transition rates are constant, ie they do not vary over time. *[1]*

(ii)(a) ***MLEs of the transition rates in a time-homogeneous Markov jump process***

In the time-homogeneous case, the transition rates μ_{ij}, $i \neq j$, are estimated by:

$$\hat{\mu}_{ij} = \frac{n_{ij}}{t_i}$$

where n_{ij} denotes the observed number of transitions from State i to State j, and t_i denotes the total observed waiting time in state i. [1]

In addition:

$$\hat{\mu}_{ii} = -\sum_{j \neq i} \hat{\mu}_{ij}$$ [½]

(ii)(b) ***Difficulties in estimating the rates for a time-inhomogeneous process***

In the time-inhomogeneous case, it is impractical to estimate $\mu_{ij}(t)$ for all values of t since this would require a huge amount of calculation and a huge amount of data. [½]

A possible procedure is to divide the time interval into subintervals, assume that the transition rates are constant over each subinterval, and estimate the transition rates for each subinterval using the procedure described above. [1]

However, our estimates would be based on a much smaller amount of data, compared to the time-homogeneous case, and would be less reliable. [½]

Alternatively, we could select an appropriate functional form for $\mu_{ij}(t)$ and use the data to estimate the relevant parameters. This is only possible if we have an idea of what kind of formula would be appropriate. [½]

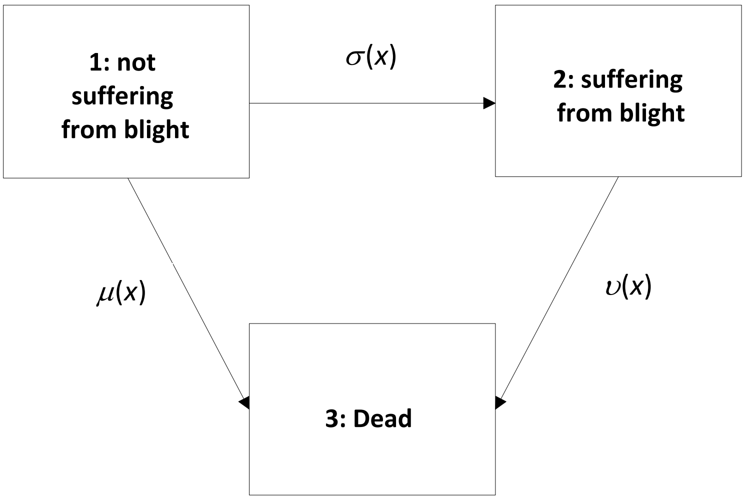
[Total 4]

Solution X2.5

This question is an application of a time-inhomogeneous Markov jump process. These processes are covered in Chapter 5.

(i) ***Transition diagram***

The transition diagram for this model is as follows:



[2]

(ii) ***Expressions involving transition rates***

(a) $p_{11}(x,x+t)$

$p_{11}(x,x+t)$ is the probability that a plant will be in State 1 at exact age $x+t$ given that it is in State 1 at exact age x. Since a return to State 1 is impossible, this is an occupancy probability, so:

$$p_{11}(x,x+t) = p_{\overline{11}}(x,x+t) = \exp\left(-\int_0^t \left[\sigma(x+s)+\mu(x+s)\right]ds\right) \qquad [2]$$

Alternatively, we could write:

$$p_{11}(x,x+t) = \exp\left(-\int_x^{x+t} \left[\sigma(s)+\mu(s)\right]ds\right) \qquad \textit{[2]}$$

(b) $p_{22}(x,x+t)$

This is another occupancy probability:

$$p_{22}(x,x+t) = p_{\overline{22}}(x,x+t) = \exp\left(-\int_0^t \upsilon(x+s)\,ds\right) \qquad [1]$$

Alternatively, we could write:

$$p_{22}(x,x+t)=\exp\left(-\int_{x}^{x+t}\upsilon(s)\,ds\right)$$ *[1]*

[Total 3]

(iii) ***Integral expression***

$p_{12}(x,x+t)$ is the probability that a plant will be in State 2 at exact age $x+t$ given that it is in State 1 at exact age x. For this to happen, the plant must remain in State 1 until exact age $x+s$, say, then transition from State 1 to State 2 at exact age $x+s$, and remain in State 2 from exact age $x+s$ to exact age $x+t$. Since s can take any value between 0 and t, we have:

$$p_{12}(x,x+t)=\int_{0}^{t}p_{11}(x,x+s)\sigma(x+s)p_{22}(x+s,x+t)\,ds$$ [2]

Alternatively, we could write:

$$p_{12}(x,x+t)=\int_{x}^{x+t}p_{11}(x,u)\sigma(u)p_{22}(u,x+t)\,du$$ *[2]*

Solution X2.6

Types of censoring are discussed at the start of Chapter 7.

(i) *Differences between random censoring and Type I censoring*

Both random censoring and Type I censoring are examples of right censoring. Right censoring occurs when a life exits the investigation for a reason other than death. [½]

With random censoring, the censoring times are not known in advance – they are not chosen by the investigator and are random variables. [1]

An example of random censoring in life insurance is the event of a policyholder choosing to surrender a policy. [1]

Type I censoring occurs when the censoring times are known in advance, *ie* the censoring times are chosen by the investigator. [1]

An example of Type I censoring is when observation ceases for all those still alive at the end of the period of investigation. [1]

[Maximum 4]

(ii) *Non-informative censoring*

Censoring is non-informative if it gives no information about the future patterns of mortality by age for the censored lives. [1]

In the context of this investigation, non-informative censoring occurs if at any given time, lives are equally likely to be censored regardless of their subsequent force of mortality. This means that we cannot tell anything about a person's mortality after the date of the censoring event from the fact that they have been censored. [1]

In this investigation withdrawals might be informative, since lives that are in better health may be more likely to surrender their policies than those in a poor state of health. Lives that are censored are therefore likely to have lighter mortality than those that remain in the investigation. [1]

[Total 3]

Markers: Please give credit for any suitable examples.

Solution X2.7

Gompertz' Law is covered at the end of Chapter 6.

(i) ***Age range***

The Gompertz model is appropriate at ages over about 30. [1]

Gompertz' Law states that $\mu_x = Bc^x$*, so it is appropriate for ages at which the force of mortality is increasing exponentially.*

(ii) ***Survival probability***

The probability of survival from exact age x to exact age $x+t$ is:

$$_tp_x = \exp\left[-\int_0^t \mu_{x+s}\,ds\right] \qquad [½]$$

Under Gompertz' Law:

$$\mu_{x+s} = Bc^{x+s}$$

So:

$$\int_0^t \mu_{x+s}\,ds = \int_0^t Bc^{x+s}\,ds = Bc^x \int_0^t c^s\,ds = Bc^x \int_0^t e^{s\ln c}\,ds \qquad [½]$$

Integrating gives:

$$\int_0^t \mu_{x+s}\,ds = Bc^x \left[\frac{e^{s\ln c}}{\ln c}\right]_0^t = \frac{Bc^x}{\ln c}\left[c^s\right]_0^t = \frac{Bc^x}{\ln c}\left[c^t - 1\right] \qquad [1]$$

Using the formula $e^{AB} = (e^A)^B$, we have:

$$_tp_x = \exp\left[-\frac{Bc^x}{\ln c}\left[c^t-1\right]\right] = \exp\left[-\frac{B}{\ln c}\,c^x\left[c^t-1\right]\right] = \left[\exp\left(-\frac{B}{\ln c}\right)\right]^{c^x\left(c^t-1\right)} \qquad [1]$$

[Total 3]

(iii) ***Values of B and c***

We have:

$$_5p_{60} = 0.912 = \exp\left[\frac{-B}{\ln c}\times c^{60}(c^5-1)\right]$$

$$_{10}p_{60} = 0.804 = \exp\left[\frac{-B}{\ln c}\times c^{60}(c^{10}-1)\right] \qquad [½]$$

Taking logs:

$$\frac{B}{\ln c}\times c^{60}(c^{5}-1)=-\ln 0.912$$

$$\frac{B}{\ln c}\times c^{60}(c^{10}-1)=-\ln 0.804$$ [½]

Dividing these equations:

$$\frac{c^{5}-1}{c^{10}-1}=\frac{\ln 0.912}{\ln 0.804}$$ [1]

Now, using the hint:

$$\frac{c^{5}-1}{(c^{5}-1)(c^{5}+1)}=\frac{1}{c^{5}+1}=\frac{\ln 0.912}{\ln 0.804}$$ [½]

$$\Rightarrow c=\left(\frac{\ln 0.804}{\ln 0.912}-1\right)^{1/5}=1.064721$$ [½]

and hence:

$$B=\frac{-\ln 0.912\times\ln 1.064721}{1.064721^{60}(1.064721^{5}-1)}=0.000364$$ [1]

[Total 4]

Solution X2.8

Differential equations for time-inhomogeneous Markov jump processes are covered in Chapter 5.

(i) ***Derivation of differential equation***

Consider the interval from age x to age $x+t+h$. Using the Markov assumption… [½]

… we can write:

(*) $${}_{t+h}p_x^{21} = {}_tp_x^{21}\ {}_hp_{x+t}^{11} + {}_tp_x^{22}\ {}_hp_{x+t}^{21} + {}_tp_x^{23}\ {}_hp_{x+t}^{31}$$ [1]

We now assume that for any two distinct states i and j, and $t \geq 0$:

$${}_hp_{x+t}^{ij} = h\mu_{x+t}^{ij} + o(h)$$

and the probability that a life makes two or more transitions in a short time interval of length h is $o(h)$. [1]

So we can write:

$${}_hp_{x+t}^{21} = h\mu_{x+t}^{21} + o(h)$$

$${}_hp_{x+t}^{31} = h\mu_{x+t}^{31} + o(h)$$

and:

$${}_hp_{x+t}^{11} = 1 - {}_hp_{x+t}^{12} - {}_hp_{x+t}^{13} - {}_hp_{x+t}^{14} = 1 - h\mu_{x+t}^{12} - h\mu_{x+t}^{14} + o(h)$$ [1]

Since the probability of more than one transition in a short time interval of length h is $o(h)$, the ${}_hp_{x+t}^{13}$ term is included in the $o(h)$ term in the equation above.

Substituting these expressions into (*) gives:

$${}_{t+h}p_x^{21} = {}_tp_x^{21}\left(1 - h\mu_{x+t}^{12} - h\mu_{x+t}^{14}\right) + {}_tp_x^{22}h\mu_{x+t}^{21} + {}_tp_x^{23}h\mu_{x+t}^{31} + o(h)$$ [½]

We can rearrange this to get:

$$\frac{{}_{t+h}p_x^{21} - {}_tp_x^{21}}{h} = {}_tp_x^{22}\mu_{x+t}^{21} + {}_tp_x^{23}\mu_{x+t}^{31} - {}_tp_x^{21}\left(\mu_{x+t}^{12} + \mu_{x+t}^{14}\right) + \frac{o(h)}{h}$$ [½]

Finally, letting $h \to 0$, we obtain the result:

$$\frac{\partial}{\partial t}\,{}_tp_x^{21} = {}_tp_x^{22}\mu_{x+t}^{21} + {}_tp_x^{23}\mu_{x+t}^{31} - {}_tp_x^{21}\left(\mu_{x+t}^{12} + \mu_{x+t}^{14}\right)$$

since $\lim\limits_{h\to 0}\dfrac{o(h)}{h} = 0$. [½]

[Total 5]

(ii) ***Other forward equations***

The corresponding differential equations for ${}_tp_x^{23}$ and ${}_tp_x^{32}$ are:

$$\frac{\partial}{\partial t}\,{}_tp_x^{23} = {}_tp_x^{22}\mu_{x+t}^{23} - {}_tp_x^{23}\left(\mu_{x+t}^{31} + \mu_{x+t}^{34}\right)$$ [1]

and:

$$\frac{\partial}{\partial t}\,{}_tp_x^{32} = {}_tp_x^{31}\mu_{x+t}^{12} - {}_tp_x^{32}\left(\mu_{x+t}^{21} + \mu_{x+t}^{23} + \mu_{x+t}^{24}\right)$$ [2]

[Total 3]

These differential equations follow the usual pattern. For example, in the first one, we are thinking about going from State 2 to State 3, and we can construct the RHS of the equation as follows:

- *Imagine that the life is in State 2 at time 0 (ie at age x).*
- *If the life is in State 2 at time t (the probability of this is ${}_tp_x^{22}$), then to get into State 3 at age $x+t$, it must instantaneously go from State 2 to State 3 at age $x+t$. So we multiply by μ_{x+t}^{23}.*
- *If the life is in State 3 at time t (probability ${}_tp_x^{23}$), then it must stay there. We need the $-\mu_{x+t}^{31}$ term to ensure that it doesn't move to State 1 at age $x+t$, and the $-\mu_{x+t}^{34}$ to ensure that it doesn't move to State 4.*
- *We don't need a term containing ${}_tp_x^{21}$ since going from State 1 to State 3 requires two transitions, and we are assuming that we can have only one transition in any given instant.*
- *We don't need a term containing ${}_tp_x^{24}$ either since it is impossible to go from State 4 to State 3.*

Solution X2.9

This question involves the Kaplan-Meier model, which is described in Chapter 7.

(i) ***Kaplan-Meier estimate***

Let $S(t)$ denote the probability that a tart has *not* been sold by time t, where time is measured in hours since 8am (the opening time of the shop). The Kaplan-Meier estimate of $S(t)$ is:

$$\hat{S}(t)=\prod_{t_j\le t}\left(1-\frac{d_j}{n_j}\right)=\prod_{t_j\le t}\left(\frac{n_j-d_j}{n_j}\right)$$

where:

- t_j is the j th sale time
- d_j is the number of tarts sold at time t_j
- n_j is the number of tarts for sale just before time t_j.

We have:

j	t_j	n_j	d_j	$\frac{n_j-d_j}{n_j}$
1	0.5	16	1	$\frac{15}{16}$
2	2	15	2	$\frac{13}{15}$
3	4.5	12	4	$\frac{8}{12}$
4	6	7	3	$\frac{4}{7}$
5	7	4	2	$\frac{2}{4}$

[2]

So the Kaplan-Meier estimate of the probability that a tart has not been sold before closing time, *ie* 4pm, is:

$$\hat{S}(8)=\tfrac{15}{16}\times\tfrac{13}{15}\times\tfrac{8}{12}\times\tfrac{4}{7}\times\tfrac{2}{4}=\tfrac{13}{84}$$ [1]

and hence the Kaplan-Meier estimate of the probability that a tart is sold before closing time is:

$$\hat{F}(8)=1-\hat{S}(8)=\tfrac{71}{84}=0.84524$$ [1]

[Total 4]

Markers: Please award full marks for the correct answer. The figures do not have to be set out in a table.

(ii) ***Sketch of hazard function***

Under the Kaplan-Meier model, the estimated hazard function is given by:

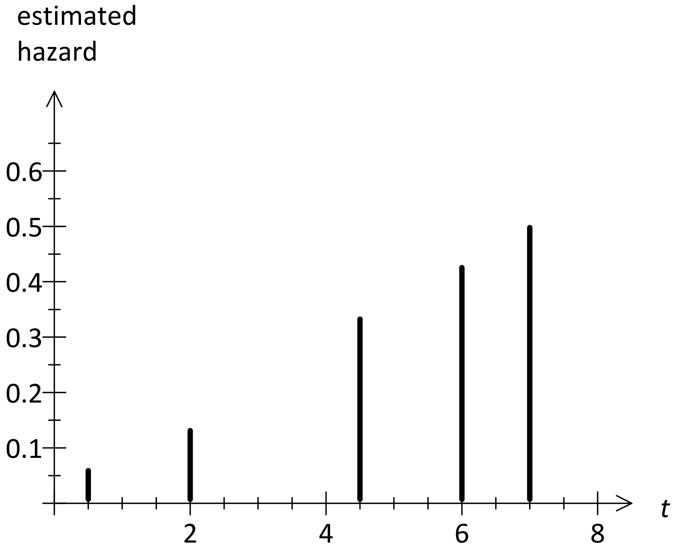
$$\hat{\lambda}_j = \frac{d_j}{n_j} \text{ at time } t_j$$

and is zero at all times at which a sale does not take place. So we have:

$$\hat{h}(t) = \begin{cases} \frac{1}{16} & \text{for } t = 0.5 \\ \frac{2}{15} & \text{for } t = 2 \\ \frac{4}{12} & \text{for } t = 4.5 \\ \frac{3}{7} & \text{for } t = 6 \\ \frac{2}{4} & \text{for } t = 7 \\ 0 & \text{otherwise} \end{cases}$$

[2]

A sketch of this function is given below:



[2]
[Total 4]

Solution X2.10

This question involves the Nelson-Aalen model, which is described in Chapter 7.

(i) ***Nelson-Aalen estimate of cumulative hazard***

We first have to work out the length of time for which each patient was observed. These figures are given in the table below.

Patient number	Length of observation period (months)	Reason for exit
1	11	Censored
2	6	Death
3	9	Censored
4	12	Censored
5	5	Censored
6	2	Death
7	12	Censored
8	8	Death
9	6	Death
10	7	Censored

[1]

So we have:

Death times: 2, 6 (two deaths) and 8

Censoring times: 5, 7, 9, 11, 12 (two lives)

No life is observed for more than 12 months.

The calculations for the cumulative hazard function are summarised in the following table:

t_j	n_j	d_j	$\hat{\lambda}_j = \dfrac{d_j}{n_j}$
2	10	1	0.1
6	8	2	0.25
8	5	1	0.2

[2]

The Nelson-Aalen estimate of the cumulative hazard function is then given by:

$$\hat{\Lambda}(t) = \sum_{t_j \le t} \hat{\lambda}_j = \begin{cases} 0 & \text{for } 0 \le t < 2 \\ 0.1 & \text{for } 2 \le t < 6 \\ 0.35 & \text{for } 6 \le t < 8 \\ 0.55 & \text{for } 8 \le t \le 12 \end{cases}$$ [1]

[Total 4]

Markers: Please award full marks for the correct answer. The figures do not have to be set out in a table.

(ii) ***Estimate of survival function***

The Nelson-Aalen estimate of the survival function is:

$$\hat{S}(t) = \exp\left(-\hat{\Lambda}(t)\right) = \begin{cases} 1 & \text{for } 0 \le t < 2 \\ 0.90484 & \text{for } 2 \le t < 6 \\ 0.70469 & \text{for } 6 \le t < 8 \\ 0.57695 & \text{for } 8 \le t \le 12 \end{cases}$$ [2]

(iii) ***Confidence interval for survival probability***

The variance of the Nelson-Aalen estimator of the integrated hazard function is given by:

$$\text{var}\left[\tilde{\Lambda}(t)\right] = \sum_{t_j \le t} \frac{d_j\left(n_j - d_j\right)}{n_j^3}$$

So:

$$\text{var}\left[\tilde{\Lambda}(10)\right] = \frac{1 \times 9}{10^3} + \frac{2 \times 6}{8^3} + \frac{1 \times 4}{5^3} = 0.064438$$ [1]

An approximate 95% confidence interval for $\Lambda(10)$ is:

$$\hat{\Lambda}(10) \pm 1.96\sqrt{\text{var}\left[\tilde{\Lambda}(10)\right]} = 0.55 \pm 1.96\sqrt{0.064438}$$

$$= 0.55 \pm 0.49754$$

$$= \left(0.05246, 1.04754\right)$$ [2]

So an approximate 95% confidence interval for $S(10)$ is:

$$\left(e^{-1.04754}, e^{-0.05246}\right) = \left(0.3508, 0.9489\right)$$ [1]

[Total 4]

(iv) ***Comment***

As the confidence interval constructed in (iii) contains the value 0.9, there is insufficient evidence (at the 2.5% significance level) to reject the hypothesis that at least 90% of patients survive for 10 months or more after the operation. [2]

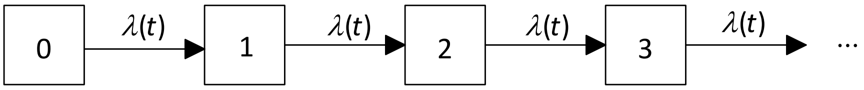
The significance level here is 2.5% because the 95% confidence interval in part (iii) has 2.5% in each tail. The hypothesis here refers to 'at least 90%', which is one-sided.

Solution X2.11

This question on a time-inhomogeneous Poisson process is an application of the material in Chapter 5.

(i) ***Transition diagram***

The transition diagram for a time-inhomogeneous Poisson process with rate $\lambda(t)$ is:



[2]

(ii)(a) ***Matrix form of differential equations***

The matrix form of the Kolmogorov forward differential equations for a time-inhomogeneous Markov jump process is:

$$\frac{\partial}{\partial t}P(s,t)=P(s,t)\,A(t)$$

where $P(s,t)$ is the matrix of transition probabilities and $A(t)$ is the generator matrix at time t .

So, for this model, the matrix form of the forward differential equations is:

$$\frac{\partial}{\partial t}\begin{pmatrix} p_{00}(s,t) & p_{01}(s,t) & p_{02}(s,t) & \cdots \\ 0 & p_{11}(s,t) & p_{12}(s,t) & \cdots \\ 0 & 0 & p_{22}(s,t) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$=\begin{pmatrix} p_{00}(s,t) & p_{01}(s,t) & p_{02}(s,t) & \cdots \\ 0 & p_{11}(s,t) & p_{12}(s,t) & \cdots \\ 0 & 0 & p_{22}(s,t) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}\begin{pmatrix} -\lambda(t) & \lambda(t) & & 0 \\ & -\lambda(t) & \lambda(t) & \\ & & -\lambda(t) & \lambda(t) \\ 0 & & & \ddots \end{pmatrix}$$

[2]

The matrix form of the Kolmogorov backward differential equations for a time-inhomogeneous Markov jump process is:

$$\frac{\partial}{\partial s}P(s,t)=-A(s)\,P(s,t)$$

For this model, we have:

$$\frac{\partial}{\partial s}\begin{pmatrix} p_{00}(s,t) & p_{01}(s,t) & p_{02}(s,t) & \cdots \\ 0 & p_{11}(s,t) & p_{12}(s,t) & \cdots \\ 0 & 0 & p_{22}(s,t) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$= -\begin{pmatrix} -\lambda(s) & \lambda(s) & & 0 \\ & -\lambda(s) & \lambda(s) & \\ & & -\lambda(s) & \lambda(s) \\ 0 & & & \ddots \end{pmatrix}\begin{pmatrix} p_{00}(s,t) & p_{01}(s,t) & p_{02}(s,t) & \cdots \\ 0 & p_{11}(s,t) & p_{12}(s,t) & \cdots \\ 0 & 0 & p_{22}(s,t) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

[2]

(ii)(b) ***Component form***

For $0 \le i < j$, the forward differential equation is:

$$\frac{\partial}{\partial t}p_{ij}(s,t) = p_{i,j-1}(s,t)\lambda(t) - p_{ij}(s,t)\lambda(t)$$ [1]

The terms on the RHS of the equation above can be obtained from the matrix expression $\frac{\partial}{\partial t}P(s,t) = P(s,t)A(t)$ by multiplying the i th row of the matrix $P(s,t)$ by the j th column of the matrix $A(t)$.

Alternatively, we could think about this as follows. Take the probability that the process goes from State i at time s to State $j-1$ at time t, and multiply this by the force of transition from State $j-1$ to State j at time t. Then add the probability that the process goes from State i at time s to State j at time t, multiplied by the force of transition that keeps the process in State j at time t, ie $-\lambda(t)$. (These are the only non-zero terms.)

The corresponding backward differential equation is:

$$\frac{\partial}{\partial s}p_{ij}(s,t) = -\left[-\lambda(s)p_{ij}(s,t) + \lambda(s)p_{i+1,j}(s,t)\right]$$

$$= \lambda(s)p_{ij}(s,t) - \lambda(s)p_{i+1,j}(s,t)$$ [1]

[Total 6]

The terms on the RHS of the equation above can be obtained from the matrix expression $\frac{\partial}{\partial s}P(s,t) = -A(s)P(s,t)$ by multiplying the i th row of the matrix $-A(s)$ by the j th column of the matrix $P(s,t)$.

Alternatively, we could think about this as follows. Start with the force of transition that keeps the process in State i at time s, ie $-\lambda(s)$, and multiply this by the probability that the process goes from State i at time s to State j at time t. Then add the force of transition from State i to State $i+1$ at time s multiplied by the probability that the process goes from State $i+1$ at time s to State j at time t. (There are no other non-zero terms.) Finally, since the process is time-inhomogeneous and this is a backward differential equation, we include an extra factor of -1 in the RHS.

(iii)(a) ***Occupancy probability***

The probability that the process stays in State 1 from time 5 to time r is:

$$p_{\overline{11}}(5,r)=\exp\left(-\int_5^r \lambda(t)\,dt\right)=\exp\left(-\int_5^r 0.01(t+2)\,dt\right) \qquad [1]$$

Now:

$$\int_5^r (t+2)\,dt=\left[\frac{1}{2}t^2+2t\right]_5^r=\left(\frac{1}{2}r^2+2r\right)-\left(\frac{1}{2}\times 5^2+2\times 5\right)=\frac{1}{2}r^2+2r-22.5$$

So:

$$p_{\overline{11}}(5,r)=\exp\left[-0.01\left(\frac{1}{2}r^2+2r-22.5\right)\right]=\exp\left(-0.005r^2-0.02r+0.225\right) \qquad [1]$$

(iii)(b) ***Probability that the process is in State 2 at time 10***

The probability that the process is in State 2 at time 10 given that it is in State 1 at time 5 can be written in integral form as follows:

$$p_{12}(5,10)=\int_5^{10} p_{\overline{11}}(5,r)\,\lambda(r)\,p_{\overline{22}}(r,10)\,dr \qquad [1]$$

The factors in the integrand represent the probability that the process stays in State 1 until time r, transitions from State 1 to State 2 at time r, and then stays in State 2 from time r to time 10. Integrating over all possible values of r (ie from 5 to 10) gives the required probability.

From part (iii)(a), we know that:

$$p_{\overline{11}}(5,r)=\exp\left(-0.005r^2-0.02r+0.225\right)$$

An expression for $p_{\overline{22}}(r,10)$ can be derived in a similar way:

$$p_{\overline{22}}(r,10)=\exp\left(-\int_r^{10}\lambda(t)\,dt\right)$$

$$=\exp\left(-\int_r^{10}0.01(t+2)\,dt\right)$$

$$=\exp\left(-\left[0.01\left(\frac{1}{2}t^2+2t\right)\right]_r^{10}\right)$$

$$=\exp\left(-0.01\left(50+20-\frac{1}{2}r^2-2r\right)\right)$$

$$=\exp\left(0.005r^2+0.02r-0.7\right) \qquad [1]$$

So:

$$p_{12}(5,10)=\int_5^{10}e^{-0.005r^2-0.02r+0.225}\,0.01(r+2)\,e^{0.005r^2+0.02r-0.7}dr$$

$$=0.01e^{-0.475}\int_5^{10}(r+2)\,dr$$

$$=0.01e^{-0.475}\left[\frac{1}{2}r^2+2r\right]_5^{10}$$

$$=0.01e^{-0.475}\left[(50+20)-(12.5+10)\right]$$

$$=0.01e^{-0.475}\times 47.5$$

$$=0.29540 \qquad [1]$$

[Total 5]

Alternatively, you might notice that:

$$p_{\overline{11}}(5,r)p_{\overline{22}}(r,10)=\exp\left(-\int_5^r\lambda(t)\,dt\right)\exp\left(-\int_r^{10}\lambda(t)\,dt\right)=\exp\left(-\int_5^{10}\lambda(t)\,dt\right) \qquad \textit{[1]}$$

Then, since:

$$\int_5^{10} \lambda(t)\,dt = \int_5^{10} 0.01(t+2)\,dt = \left[0.01\left(\frac{1}{2}t^2+2t\right)\right]_5^{10} = 0.475 \qquad [1]$$

it follows that:

$$p_{12}(5,10) = \int_5^{10} e^{-0.475}\,0.01(r+2)dr = 0.01e^{-0.475}\int_5^{10}(r+2)\,dr = 0.29540 \qquad [1]$$

as before.

Solution X3.1

This question covers graduation tests and is based on the material in Chapter 10.

Here we have:

$$n_1 = \text{ number of positive deviations } = 12$$

$$n_2 = \text{number of negative deviations} = 18$$ [½]

From page 189 of the *Tables*, the critical value for this test is 4. This means that the data will fail the test if there are 4 or fewer runs of positive deviations. [1]

The data have only just passed the test, so there must be 5 runs of positive deviations. [½]

[Total 2]

Solution X3.2

The form of the partial likelihood function for a Cox regression model is covered in Chapter 8.

From the given data we see that Patient 1 (a smoker) dies first, and at time 3. Since there were 2 smokers and 4 non-smokers in the at-risk group just before time 3, the contribution to the partial likelihood from the first death is:

$$\frac{e^{\beta}}{2e^{\beta}+4}$$ [1]

The second life to die is Patient 6 (a non-smoker). Just before this death, there were 1 smoker and 4 non-smokers at risk. So the contribution to the partial likelihood from the second death is:

$$\frac{1}{e^{\beta}+4}$$ [1]

The third life to die is Patient 3 (a non-smoker). Just prior to this death there are 1 smoker and 2 non-smokers at risk. Note that Patient 5 is censored at time 8, so is no longer part of the at-risk group. The contribution to the partial likelihood from this death is therefore:

$$\frac{1}{e^{\beta}+2}$$ [1]

So the partial likelihood function is:

$$L(\beta) = \frac{e^{\beta}}{2e^{\beta}+4} \times \frac{1}{e^{\beta}+4} \times \frac{1}{e^{\beta}+2} = \frac{e^{\beta}}{2\left(e^{\beta}+2\right)^2\left(e^{\beta}+4\right)}$$ [1]

[Total 4]

Solution X3.3

This question tests the census approximation, which is covered in Chapter 9.

(i) ***Central exposed to risk***

Let $P_x(t)$ denote the number of lives at time t aged x next birthday and suppose that time is measured in years from 1 January 2017. [½]

*We know the values of $P_x(0)$, $P_x(1)$ and $P_x(1½)$ for all x, and we are given the numbers of deaths during the investigation aged x **last** birthday.*

Since the death data and the census data don't match, define $P'_x(t)$ to be the number of lives at time t aged x last birthday. [½]

Then:

$$E_x^c = \int_0^{1½} P'_x(t)\,dt \qquad [½]$$

Assuming that $P'_x(t)$ varies linearly between time 0 and time 1, and also between time 1 and time 1½ … [1]

$$E_x^c = \frac{1}{2}\left[P'_x(0) + P'_x(1)\right] + \frac{1}{4}\left[P'_x(1) + P'_x(1½)\right] \qquad [½]$$

Now:

$$P'_x(0) = \text{number of lives at time 0 aged } x \text{ last birthday}$$

So:

$$P'_x(0) = \text{number of lives at time 0 aged } x+1 \text{ next birthday} = P_{x+1}(0) \qquad [½]$$

Similarly:

$$P'_x(1) = P_{x+1}(1) \quad \text{and} \quad P'_x(1½) = P_{x+1}(1½) \qquad [½]$$

So:

$$E_x^c = \frac{1}{2}\left[P_{x+1}(0) + P_{x+1}(1)\right] + \frac{1}{4}\left[P_{x+1}(1) + P_{x+1}(1½)\right]$$

$$= \frac{1}{2}P_{x+1}(0) + \frac{3}{4}P_{x+1}(1) + \frac{1}{4}P_{x+1}(1½) \qquad [1]$$

[Total 5]

(ii) ***Value of f***

Since deaths are classified according to age last birthday, the rate interval starts at exact age x and ends at exact age $x+1$. So the age in the middle of the rate interval is $x+½$, *ie* $f = ½$. [1]

Solution X3.4

This bookwork question tests the material at the start of Chapter 9.

(i) ***Why the company may subdivide its mortality data***

Mortality risk varies between individuals for many reasons. However, mortality models assume that we are dealing with identical lives, *ie* groups of people who have the same mortality characteristics. [1]

Companies often subdivide the data according to characteristics that are known to have a significant effect on mortality. This helps to reduce the heterogeneity within each group. [1]

[Total 2]

(ii) ***Two main problems***

One problem with subdividing data is that some of the subgroups may be very small, containing only a few individuals. [½]

Estimates of mortality rates derived from the small groups will be unreliable, as the small group size could make it difficult to ascertain the true underlying rates. [½]

The other main problem is that there may be missing data or the data may be inaccurate or may contain mistakes. [½]

This could result in unreliable mortality estimates. [½]

[Total 2]

(iii) ***Factors used to subdivide the data***

Any four of the following:

- sex [½]
- age [½]
- smoker status [½]
- occupation [½]
- nationality or ethnic group [½]
- type of policy [½]
- level of underwriting [½]
- duration in force [½]
- sales channel [½]
- policy size [½]
- known impairments [½]
- current state of health [½]
- disabilities [½]

- postcode/geographical location [½]
- residential status (*eg* homeowner, renting) [½]
- marital status [½]

[Maximum 2]

Solution X3.5

This is a bookwork question about graduation. It is based on the material at the beginning of Chapter 10.

(i) ***Undergraduation and overgraduation***

When graduating a set of crude mortality rates, there is a trade off between smoothness and close adherence to the crude rates (goodness of fit). [½]

A satisfactory graduation must achieve an appropriate balance between these two extremes. [½]

Overgraduation occurs when too much emphasis is given to smoothness. Overgraduated rates show a smooth progression from age to age, but the resulting rates do not adhere closely to the crude rates. [1]

Undergraduation occurs when too much emphasis is given to goodness of fit. Undergraduated rates adhere closely to the crude rates, but the resulting rates do not show a smooth progression from age to age. [1]

[Total 3]

(ii) ***The dangers of overgraduation***

Inadequate premium rates

The office may make losses through underestimating mortality for death benefits or overestimating mortality for survival benefits (since the graduated rates do not accurately reflect the true mortality rates at all ages). [1]

Excessive premium rates

The reverse occurs where the office may lose business through setting excessively high premium rates. [½]

Selection

The office may be exposed to selection from other offices whose premium rates more accurately reflect the true mortality rates. [½]

Reserves

Using biased rates can also lead to inappropriate levels of reserves being held. Holding insufficient reserves can endanger the company's solvency, whereas holding excessive reserves will reduce the company's profitability. [1]

The dangers of undergraduation

Inappropriate premium rates

The office may make losses or lose business if the premium rates at particular ages have been distorted by random sampling errors that were not smoothed out. [1]

Anomalies

The office may lose business or incur unnecessary alteration expenses if the rates do not show a consistent progression from age to age. (Policyholders may wait a few years because the rates become cheaper, or they may surrender and take out a new policy to take advantage of an anomaly in the rates at a particular age.) [1]

[Total 5]

Solution X3.6

This question concerns graduation tests. These tests are covered in Chapter 10.

(i)(a) ***Test for overall goodness of fit***

We are testing the null hypothesis:

H_0 : the graduated rates are the true mortality rates underlying the data

against the alternative hypothesis:

H_1 : the graduated rates are not the true mortality rates underlying the data [1]

The test statistic for the chi-squared goodness-of-fit test is:

$$\sum z_x^2 = 0.5249^2 + 0.3615^2 + \cdots + (-0.3673)^2 = 5.1104$$ [1]

We compare this against an appropriate chi-squared distribution. The method of graduation has not been stated, so it is unclear how many degrees of freedom to deduct. However, since the test covers 9 age groups, we must have fewer than 9 degrees of freedom. [½]

This is a one-tailed test with large values of the test statistic being significant.

From page 169 of the *Tables*, we see that:

- the upper 5% point of χ_8^2 is 15.51
- the upper 5% point of χ_7^2 is 14.07
- the upper 5% point of χ_6^2 is 12.59.

Most methods of graduation would involve the loss of 2 or 3 degrees of freedom, so for any reasonable number of degrees of freedom this result is not significant at the 5% level. [1]

So we conclude that the graduated rates are a good overall fit to the observed rates. [½]

(i)(b) ***Test for overall bias***

The signs test can be used to check whether or not the graduated rates are biased, ie whether they are consistently higher or lower than the observed rates.

The cumulative deviations test can also be used to check for overall bias in the graduated rates.

Only one of these tests is required here.

The null and alternative hypotheses are as stated above for the chi-squared test.

Signs test

Under the null hypothesis, the number of positive deviations, N, has a $Binomial(9, 0.5)$ distribution. [1]

Looking at the data, we see that there are 7 positive and 2 negative deviations.

This is a two-tailed test, so the p-value of the test is:

$$2P(N \geq 7) = 2\left[1 - P(N \leq 6)\right]$$

From page 187 of the *Tables*:

$$P(N \leq 6) = 0.9102$$

So the p-value is:

$$2(1 - 0.9102) = 0.1796$$ [1]

This is not significant at the 5% level. So we conclude that the graduated rates are not biased and are a good fit to the observed rates. [½]

However, inspection of the standardised deviations shows that the graduated rates are generally less than the observed rates (even though this bias is not statistically significant). The graduated rates display lighter mortality than the observed rates. [½]

[Total 7]

Cumulative deviations test

The calculations needed to calculate the observed value of the test statistic are as follows. The expected number of deaths in each age group is given by the central exposed to risk multiplied by the graduated mortality rate.

Age group	*Central exposed to risk*	*Observed number of deaths*	*Graduated mortality rate*	*Expected number of deaths*
60 – 64	*1,388.9*	*10*	*0.0061*	*8.4723*
65 – 69	*1,188.8*	*17*	*0.0131*	*15.5733*
70 – 74	*880.5*	*28*	*0.0262*	*23.0691*
75 – 79	*841.6*	*34*	*0.0487*	*40.9859*
80 – 84	*402.8*	*41*	*0.0839*	*33.7949*
85 – 89	*123.9*	*19*	*0.1338*	*16.5778*
90 – 94	*27.9*	*7*	*0.1975*	*5.5103*
95 – 99	*10.0*	*3*	*0.2706*	*2.7060*
100+	*7.5*	*2*	*0.3455*	*2.5913*
Total		*161*		*149.2809*

[1]

The observed value of the standardised test statistic is:

$$\frac{\sum \text{observed deaths} - \sum \text{expected deaths}}{\sqrt{\sum \text{expected deaths}}} = \frac{161 - 149.2809}{\sqrt{149.2809}} = 0.9592 \qquad [½]$$

We compare this against the standard normal distribution. This is a two-tailed test with both large positive and negative values of the test statistic being significant. Since the value of the test statistic lies between –1.96 and +1.96 (the lower and upper 2.5% points of the standard normal distribution), the observed value of the test statistic is not significant at the 5% level. So we conclude that the graduated rates are not biased and are a good fit to the observed rates. *[1]*

However, inspection of the standardised deviations shows that the graduated rates are generally less than the observed rates (even though this bias is not statistically significant). The graduated rates display lighter mortality than the observed rates. *[½]*

(ii) ***Comment***

Using lower mortality rates will tend to overstate the value of pension fund liabilities. So using these graduated rates would not be risky for the scheme, but it may lead to larger (usually employer) contributions than are necessary. [1]

However the observed rates will reflect the current mortality rates and not the future mortality rates that will be experienced by the scheme's pensioners. Mortality rates may improve over time. If the valuation does not anticipate this improvement then the scheme's pension liabilities may be undervalued. This problem could be mitigated by projecting the observed rates. [1]

[Total 2]

Solution X3.7

Mortality projection is covered in Chapter 12.

(i) ***Usual constraints***

The usual parameter constraints imposed when fitting the Lee-Carter model are:

$$\sum_{\text{all } x} \hat{b}_x = 1 \quad \text{and} \quad \sum_{\text{all } t} \hat{k}_t = 0$$ [1]

(ii) ***Values of k_0 and k_{10}***

If $\sum_{\text{all } t} \hat{k}_t = 0$, then the value of $\hat{k}_t$ at the median value of t is zero because $\hat{k}_t$ is a linear function of t. The median year is $t = 18$, so we have:

$$\hat{k}_{18} = 0$$ [1]

which gives:

$$\hat{k}_0 = 0 - 18 \times (-0.01) = 0.18$$

and: $$\hat{k}_{10} = 0 - 8 \times (-0.01) = 0.08$$ [1]

[Total 2]

(iii)(a) ***Value of ratio when $b_x = 1$***

We have:

$$\frac{\hat{m}_{x,10}}{\hat{m}_{x,0}} = \frac{\exp\left[\hat{a}_x + \hat{b}_x \hat{k}_{10}\right]}{\exp\left[\hat{a}_x + \hat{b}_x \hat{k}_0\right]} = \exp\left[\hat{b}_x\left(\hat{k}_{10} - \hat{k}_0\right)\right]$$ [1]

When $b_x = 1$:

$$\frac{\hat{m}_{x,10}}{\hat{m}_{x,0}} = e^{\hat{k}_{10} - \hat{k}_0} = e^{0.08 - 0.18} = e^{-0.1} = 0.905$$ [½]

(iii)(b) ***Value of ratio for ages 50, 65 and 75***

In general for this fitted model we have:

$$\frac{\hat{m}_{x,10}}{\hat{m}_{x,0}} = \exp\left[\hat{b}_x\left(\hat{k}_{10} - \hat{k}_0\right)\right] = e^{-0.1\hat{b}_x}$$

So:

$$\frac{\hat{m}_{50,10}}{\hat{m}_{50,0}} = e^{-0.1\times(-0.14)} = e^{0.014} = 1.014$$ [½]

$$\frac{\hat{m}_{65,10}}{\hat{m}_{65,0}} = e^{-0.1\times 0.28} = e^{-0.028} = 0.972$$ [½]

$$\frac{\hat{m}_{75,10}}{\hat{m}_{75,0}} = e^{-0.1\times 1.30} = e^{-0.13} = 0.878$$ [½]

[Total 3]

(iv) ***How the b_x parameter affects the projected time trend***

When $\hat{b}_x = 1$, the projected change in mortality over time directly reflects the change in the time trend function $\hat{k}_t$ over the specified time period (*eg* in this model this leads to a 9.5% reduction in mortality over the first ten years of the projection). [1]

When $\hat{b}_x$ is positive, the change in mortality over time is in the same direction as the time trend function (*eg* in this model positive $\hat{b}_x$ apply at ages 65 and 75 and so mortality is projected to reduce over the ten-year projection period at these ages). [1]

When $\hat{b}_x$ is negative, the trend in mortality assumed at that age is in the opposite direction to the time trend function in the model (*eg* in this model a negative value of $\hat{b}_x$ applies at age 50 and so mortality rates are predicted to rise over the ten- year period at this age). [1]

When $0 < \left|\hat{b}_x\right| < 1$, the change in mortality over time is smaller in absolute terms than the change in the time trend function (*eg* this applies at ages 50 and 65 in this model, where changes in mortality of $+1.4\%$ and -2.8% respectively are projected, both of which are less in absolute terms than the 9.5% change obtained when $\hat{b}_x = 1$). [1]

When $\left|\hat{b}_x\right| > 1$, the change in mortality over time is greater in absolute terms than the change in the time trend function (*eg* in this model this applies at age 75, where a reduction of 12.2% in mortality is projected for the ten-year period). [1]

[Maximum 4]

Markers: Please award ½ mark for each description and ½ mark for suitable evidence in each case.

Solution X3.8

This question is based on the Cox regression model, which is covered in Chapter 8.

(i) ***Model of force of mortality***

The model for the force of mortality is:

$$\mu(t,\mathbf{Z}) = \mu_0(t)\exp\left(-0.20Z_1 + 0.12Z_2 - 0.05Z_3 - 0.06Z_4\right)$$

where:

$t =$ time since patient underwent procedure

$\mu_0(t) =$ baseline hazard at time t

$\mathbf{Z} = \left(Z_1, Z_2, Z_3, Z_4\right)$

$$Z_1 = \begin{cases} 1 & \text{if patient is female} \\ 0 & \text{if patient is male} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if patient received Treatment B} \\ 0 & \text{if patient did not receive Treatment B} \end{cases}$$

$$Z_3 = \begin{cases} 1 & \text{if patient received Treatment C} \\ 0 & \text{if patient did not receive Treatment C} \end{cases}$$

$$Z_4 = \begin{cases} 1 & \text{if patient attended Hospital B} \\ 0 & \text{if patient attended Hospital A} \end{cases}$$

[3]

Markers: Please give credit for alternative correct solutions. Deduct ½ mark for each omission, subject to a minimum of 0.

(ii) ***Proportional hazards model***

The model is a proportional hazards model since the hazards of different lives with covariate vectors $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ are in the same proportion at all times, *ie* the ratio $\dfrac{\mu\left(t,\mathbf{Z}^{(1)}\right)}{\mu\left(t,\mathbf{Z}^{(2)}\right)}$ does not depend on t. [1]

(iii)(a) ***Baseline hazard group***

The baseline hazard refers to the lives whose Z values are all 0, *ie* to male patients on Treatment A who attended Hospital A. [1]

(iii)(b) ***Group with lowest force of mortality***

Here we must make the power in the exponential as negative as possible.

The lives with the lowest force of mortality according to this model are those for which $Z_1 = 1$, $Z_2 = 0$, $Z_3 = 1$ and $Z_4 = 1$, *ie* female patients on Treatment C who attended Hospital B. [1]

[Total 2]

(iv) ***Comparison of Hospital A with Hospital B***

Suppose that β_4 is the parameter associated with covariate Z_4. We want to test the null hypothesis:

$H_0 : \beta_4 = 0$ (*ie* hospital is not significant)

against the alternative hypothesis:

$H_1 : \beta_4 < 0$ (*ie* Hospital B is better) [1]

The estimated value of β_4 is –0.06, and the standard error of the estimator is 0.04. So the value of our test statistic is:

$$\frac{-0.06-0}{0.04} = -1.5$$ [1]

Comparing this with the lower 5% point of the standard normal distribution (–1.6449), we find that it does not fall into the rejection region. So there is insufficient evidence to conclude that attending Hospital B improves the chances of survival. [1]

[Total 3]

Alternatively, we could construct an approximate one-sided 95% confidence interval for β_4 *and check whether or not 0 lies within the interval.*

The interval is:

$$\left(-\infty, \hat{\beta}_4 + 1.645\, se\left(\tilde{\beta}_4\right)\right) = \left(-\infty, -0.06 + 1.645 \times 0.04\right) = \left(-\infty, 0.0058\right)$$

Since 0 lies in this interval, there is insufficient evidence to reject the null hypothesis (as stated above).

(v) ***Proportion***

According to the model, the force of mortality at time t since the procedure for a male patient on Treatment B who attended Hospital A is:

$$\mu(t,0,1,0,0) = \mu_0(t)\exp\left(0.12\right)$$ [½]

Also, the force of mortality at time t since the procedure for a female patient on Treatment C who attended Hospital B is:

$$\mu(t,1,0,1,1) = \mu_0(t)\exp\left(-0.20-0.05-0.06\right) = \mu_0(t)\exp\left(-0.31\right)$$ [½]

Dividing the first of these expressions by the second, we obtain:

$$\frac{\mu_0(t)\exp\left(0.12\right)}{\mu_0(t)\exp\left(-0.31\right)} = e^{0.43} = 1.5373$$

So the force of mortality for the male patient exceeds the force of mortality for the female patient by 53.73%. [1]

[Total 2]

Solution X3.9

This question is about estimating mortality rates using the census method, and the relationship between μ and q under the assumptions of the Poisson model. The census method is covered in Chapter 9 and the Poisson model is covered in Chapter 3.

(i) *Estimates of the force of mortality*

The most reasonable assumption we can make here is to assume that the average number of policies in force throughout the year can be approximated by the number in force on 1 July each year. [1]

We can then estimate the force of mortality by dividing the number of deaths (θ_x) by the central exposed to risk:

$$\hat{\mu} = \frac{\theta_x}{E_x^c}$$ [½]

Deaths are classified by age last birthday at the date of death. So at the start of the rate interval, all lives are aged exactly x. [½]

In the middle of the year of age $(x, x+1)$ the lives will be aged $x+½$. So this will give us an estimate of $\mu_{x+½}$. [1]

This leads to the following results:

Age	E_x^c	θ_x	$\hat{\mu}_{x+½}$
63	18,410	430	0.0234
64	17,196	490	0.0285
65	16,960	507	0.0299

[3]

For example, when $x = 63$, we have:

$$E_x^c = 4,192 + 4,444 + 4,885 + 4,889 = 18,410$$

$$\theta_x = 104 + 100 + 117 + 109 = 430$$

[Total 6]

(ii) *Relationship between the initial rate of mortality and the force of mortality*

The initial rate of mortality at age x last birthday is q_x.

The general formula for deriving survival probabilities from the force of mortality is:

$${}_tp_x = \exp\left(-\int_0^t \mu_{x+s} ds\right)$$ [½]

The Poisson model assumes that μ is constant over the year of age $(x, x+1)$. [½]

So:

$$q_x = 1 - p_x = 1 - \exp\left(-\int_0^1 \mu ds\right) = 1 - e^{-\mu}$$ [1]

[Total 2]

(iii) ***Estimates of the initial rates of mortality***

We can estimate the initial rates of mortality using the formula in part (ii) and the estimated values of μ from part (i):

Age	$\hat{\mu}_{x+½}$	$\hat{q}_x = 1 - \exp(-\hat{\mu}_{x+½})$
63	0.0234	0.0231
64	0.0285	0.0281
65	0.0299	0.0295

[1]

At the start of the year of age $(x, x+1)$ the lives will be aged x. So these will give us estimates of q_x. [1]

[Total 2]

Solution X3.10

This question is about the standard table method of graduation. The topic is covered in Chapter 11.

(i) ***Considerations in choosing table***

Whatever table is chosen it must satisfy several key criteria, in particular:

- it must be available for all classes of lives, *eg* males and females [1]
- it must relate to a similar class of lives, *eg* assurances and not annuities in this case [1]
- it must be a 'benchmark' table, *ie* generally acceptable to all other actuaries [½]
- it should be up-to-date, *ie* relate to fairly recent experience [½]
- it must cover the age range for which rates are required. [½]

In addition it should have the correct pattern of rates by age (not necessarily the correct level of rates though). [1]

It should not have any special features that are unlikely to be present in the experience being graduated. [1]

[Maximum 5]

(ii) ***Checking suitability of formula***

The formula implies that the ratio of the rates varies linearly with age. Is there any external evidence to indicate that this is the correct pattern? [1]

A check could be made by plotting $\dfrac{\overset{\circ}{\mu}_x}{\mu_x^s}$ against x. The plot should be roughly linear. [2]

[Total 3]

(iii)(a) ***Weighted least squares estimation***

The function to be optimised is:

$$S = \sum_x w_x \left[\hat{\mu}_x - \mu_x^s(ax+b)\right]^2$$ [1]

w_x should be inversely proportional to the variance of the estimator $\tilde{\mu}_x$. [1]

Alternatively, we could say that the weights should be proportional to the exposure at each age.

To determine the estimates of a and b, we calculate the partial derivatives $\dfrac{\partial S}{\partial a}$ and $\dfrac{\partial S}{\partial b}$, set these equal to 0, and solve the resulting simultaneous equations. [1]

(iii)(b) ***Maximum likelihood estimation***

We have:

$\theta_x =$ observed number of deaths at age x

$E_x^c =$ central exposed to risk at age x

and the graduated rates are to be calculated from the relationship:

$$\overset{\circ}{\mu}_x = \mu_x^s(ax+b)$$

The function to be optimised is:

$$L = \prod_x P(D_x = \theta_x)$$

where $D_x \sim Poisson(E_x^c \overset{\circ}{\mu}_x)$. [1]

So:

$$L = \prod_x \frac{e^{-E_x^c \mu_x^s (ax+b)} \left[E_x^c \mu_x^s (ax+b) \right]^{\theta_x}}{\theta_x !} = C \prod_x e^{-E_x^c \mu_x^s (ax+b)} (ax+b)^{\theta_x}$$

where C is a constant that does not depend on a or b. [1]

We usually take logs before attempting to optimise L, as this makes the differentiation easier.

To determine the estimates of a and b, we calculate the partial derivatives $\dfrac{\partial \log L}{\partial a}$ and $\dfrac{\partial \log L}{\partial b}$, set these equal to 0, and solve the resulting simultaneous equations. [1]

Alternatively, we could set $\dfrac{\partial L}{\partial a}$ and $\dfrac{\partial L}{\partial b}$ equal to 0 and solve the resulting equations.

[Total 6]

Solution X4.1

The Pareto distribution is introduced in Chapter 15.

Let X denote the claim amount random variable. Then $X \sim Pa(\alpha, \lambda)$ for some values of α and λ to be determined. Using the formulae for the mean and variance of a Pareto random variable from page 14 of the *Tables*, we have:

$$E(X) = \frac{\lambda}{\alpha - 1} = 1{,}000$$

$$\text{var}(X) = \frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)} = 1{,}500^2 \qquad [½]$$

Squaring the equation for the mean and substituting into the equation for the variance gives:

$$\frac{1{,}000^2\alpha}{\alpha - 2} = 1{,}500^2 \qquad [½]$$

Rearranging:

$$1{,}000^2\alpha = 1{,}500^2(\alpha - 2)$$

$$\Rightarrow (1{,}500^2 - 1{,}000^2)\alpha = 2 \times 1{,}500^2$$

$$\Rightarrow \alpha = 3.6 \qquad [½]$$

Hence:

$$\lambda = 1{,}000(\alpha - 1) = 1{,}000 \times 2.6 = 2{,}600 \qquad [½]$$

and the proportion of claims that exceed 2,000 is:

$$P(X > 2{,}000) = 1 - F_X(2{,}000) = 1 - \left[1 - \left(\frac{2{,}600}{2{,}600 + 2{,}000}\right)^{3.6}\right] = \left(\frac{2{,}600}{4{,}600}\right)^{3.6} = 0.12823 \qquad [1]$$

[Total 3]

Solution X4.2

The method of percentiles is introduced in Section 3 of Chapter 15.

Let X denote the claim amount random variable.

The median of the distribution is the value of x such that $F_X(x)=0.5$. Assuming that $X \sim W(c,\gamma)$, this is the value of x such that:

$$1-e^{-cx^{\gamma}}=0.5$$

Setting $x=1,500$ (the sample median) gives:

$$\hat{c}(1,500)^{\hat{\gamma}}=-\ln 0.5 \qquad (*) \qquad [1]$$

The 95th percentile of the distribution is the value of y such that $F_X(y)=0.95$, *ie* $P(X>y)=0.05$. Assuming that $X \sim W(c,\gamma)$, we have:

$$1-e^{-cy^{\gamma}}=0.95$$

Setting $y=6,000$ (the 95th percentile from the sample) gives:

$$\hat{c}(6,000)^{\hat{\gamma}}=-\ln 0.05 \qquad (\dagger) \qquad [1]$$

Dividing (†) by (*):

$$\frac{\hat{c}(6,000)^{\hat{\gamma}}}{\hat{c}(1,500)^{\hat{\gamma}}}=4^{\hat{\gamma}}=\frac{\ln 0.05}{\ln 0.5}$$

$$\Rightarrow \hat{\gamma}\ln 4=\ln\left(\frac{\ln 0.05}{\ln 0.5}\right)$$

$$\Rightarrow \hat{\gamma}=1.055838 \qquad [1]$$

Substituting this into (*):

$$\hat{c}=\frac{-\ln 0.5}{1,500^{1.055838}}=0.000307 \qquad [1]$$

[Total 4]

Solution X4.3

Generalised extreme value distributions and generalised Pareto distributions are studied in Chapter 16.

(i) ***Type of distribution***

Since $\gamma < 0$, this is a Weibull-type GEV distribution. [1]

The key characteristic of these distributions is that they have finite upper limits. [½]

Examples include the beta, uniform and triangular distributions. [½]
[Total 2]

(ii) ***Advantage of generalised Pareto distribution***

The key advantage of the generalised Pareto distribution is that it makes use of *all* the data in the tail. [1]

The generalised extreme value distribution, however, might exclude some extreme data values because these are not the most extreme within a particular block. [1]
[Total 2]

Solution X4.4

Threshold exceedance is defined in Chapter 16.

(i) *Excess over the threshold u*

The excess of X over the threshold u is given by:

$$X-u\,|\,X>u$$ [1]

[Total 1]

(ii) *Distribution of threshold exceedance*

Let $W=X-u\,|\,X>u$. Then the CDF of W is:

$$P(W\le w)=P(X-u\le w\,|\,X>u)$$ [½]

Using the conditional probability formula, we have:

$$P(W\le w)=\frac{P(X-u\le w \text{ and } X>u)}{P(X>u)}$$

$$=\frac{P(u<X\le u+w)}{P(X>u)}$$

$$=\frac{F_X(u+w)-F_X(u)}{1-F_X(u)}$$ [1]

Since $X\sim Pa(\alpha,\lambda)$:

$$P(W\le w)=\frac{\left[1-\left(\frac{\lambda}{\lambda+u+w}\right)^\alpha\right]-\left[1-\left(\frac{\lambda}{\lambda+u}\right)^\alpha\right]}{1-\left[1-\left(\frac{\lambda}{\lambda+u}\right)^\alpha\right]}$$

$$=\frac{\left(\frac{\lambda}{\lambda+u}\right)^\alpha-\left(\frac{\lambda}{\lambda+u+w}\right)^\alpha}{\left(\frac{\lambda}{\lambda+u}\right)^\alpha}$$

$$=1-\left(\frac{\lambda+u}{\lambda+u+w}\right)^\alpha$$ [1]

This is the CDF of the Pareto distribution with parameters α and $\lambda+u$. So:

$$X-u\,|\,X>u\sim Pa(\alpha,\lambda+u)$$ [½]

[Total 3]

Solution X4.5

The Box-Jenkins approach to modelling time series is covered in Chapter 14.

(i)(a) ***Removing seasonal variation***

Any ***one*** *of the following three methods for* ***two marks in total****.*

1. Seasonal differencing [½]

   Quarterly variation means that the period is four quarters (*ie* $\theta_t = \theta_{t+4}$). [½]

   So we subtract the value from four quarters ago:

   $$\nabla_4 (q_t) = q_t - q_{t-4}$$ [1]

2. Method of moving averages [½]

   Quarterly variation means that the period is four quarters (*ie* $\theta_t = \theta_{t+4}$). [½]

   So we calculate a symmetrical average of four terms about q_t:

   $$\tfrac{1}{4}\left(\tfrac{1}{2}q_{t-2} + q_{t-1} + q_t + q_{t+1} + \tfrac{1}{2}q_{t+2}\right)$$ [1]

3. Method of seasonal means [½]

   Quarterly variation means that the period is four quarters (*ie* $\theta_t = \theta_{t+4}$). [½]

   We first calculate estimates of the seasonal means from the data $\{q_1, \dots, q_{20}\}$:

   $$\hat{\theta}_1 = \tfrac{1}{5}(q_1 + q_5 + q_9 + q_{13} + q_{17}) - \hat{\mu}$$

   ... [½]

   $$\hat{\theta}_4 = \tfrac{1}{5}(q_4 + q_8 + q_{12} + q_{16} + q_{20}) - \hat{\mu}$$

   where $\hat{\mu} = \tfrac{1}{20}\sum q_i$ is the sample mean of the data.

   Then we subtract the appropriate seasonal mean from the data to produce the series:

   $$z_t = \begin{cases} q_t - \hat{\theta}_1 & t = 1, 5, \dots, 17 \\ \dots & \\ q_t - \hat{\theta}_4 & t = 4, 8, \dots, 20 \end{cases}$$ [½]

   which has no seasonal component.

Alternatively, we could subtract from each observation the estimated mean for that period, obtained by averaging the corresponding observations in the sample:

$$z_t = \begin{cases} q_t - \frac{1}{5}(q_1 + q_5 + q_9 + q_{13} + q_{17}) & t = 1, 5, \dots, 17 \\ \dots & \\ q_t - \frac{1}{5}(q_4 + q_8 + q_{12} + q_{16} + q_{20}) & t = 4, 8, \dots, 20 \end{cases}$$ *[1]*

(i)(b) ***Linear trend***

*Any **one** of the following two methods for **one mark in total**.*

1. Least squares trend removal

Estimate a and b using least squares regression. [½]

(*ie* determine a and b that minimise $\sum y_t^2 = \sum (x_t - a - bt)^2$).

Then subtract the regression line from the observed values, $q_t - \hat{a} - \hat{b}t$. [½]

2. Differencing

Subtract the previous observed value:

$$\nabla q_t = q_t - q_{t-1}$$ [1]

[Total 3]

(ii) ***Fitted time series***

An appropriate time series is an $AR(2)$ process ($z_t = \alpha_1 z_{t-1} + \alpha_2 z_{t-2} + \varepsilon_t$). [1]

This is because the SPACF cuts off after lag 2. [½]

Also, the SACF decays slowly to 0, which is consistent with an autoregressive process. [½]

[Total 2]

Solution X4.6

Method of moments estimation is covered in Chapter 15.

(i) ***Method of moments estimates***

Let X denote the claim amount random variable. From page 14 of the *Tables*:

$$E(X)=e^{\mu+½\sigma^2} \quad \text{and} \quad \text{var}(X)=e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$

The method of moments estimates of μ and σ^2 are obtained by equating the sample and population moments:

$$e^{\hat{\mu}+½\hat{\sigma}^2}=2{,}000 \qquad e^{2\hat{\mu}+\hat{\sigma}^2}(e^{\hat{\sigma}^2}-1)=500^2$$ [1]

Substituting the square of the first equation into the second equation gives:

$$2{,}000^2(e^{\hat{\sigma}^2}-1)=500^2 \quad\Rightarrow\quad \hat{\sigma}^2=\ln\left(1+\frac{500^2}{2{,}000^2}\right)=0.0606246$$ [1]

Substituting this back into the first equation gives:

$$\hat{\mu}=\ln 2{,}000-½\hat{\sigma}^2=7.57059$$ [1]

Alternatively we could equate the first two non-central moments:

$$e^{\hat{\mu}+½\hat{\sigma}^2}=2{,}000 \qquad e^{2\hat{\mu}+2\hat{\sigma}^2}=500^2+2{,}000^2$$ *[1]*

Substituting the square of the first equation into the second equation gives:

$$2{,}000^2e^{\hat{\sigma}^2}=500^2+2{,}000^2 \quad\Rightarrow\quad \hat{\sigma}^2=\ln 1.0625=0.0606246$$ *[1]*

Substituting this back into the first equation gives:

$$\hat{\mu}=\ln 2{,}000-½\hat{\sigma}^2=7.57059$$ *[1]*

[Total 3]

(ii) ***Median claim amount***

The median claim amount, M, satisfies the equation:

$$P(X<M)=0.5$$ [½]

Since $X\sim\log N(\mu,\sigma^2)$, it follows that $\dfrac{\ln X-\mu}{\sigma}\sim N(0,1)$. [½]

Now:

$$P\left(N(0,1)<\frac{\ln M-\mu}{\sigma}\right)=0.5$$

$$\Rightarrow \frac{\ln M-\mu}{\sigma}=0$$

$$\Rightarrow M=e^{\mu}$$ [1]

Using the fitted distribution, the median is estimated to be:

$$e^{\hat{\mu}}=e^{7.57059}=1,940.285$$ [1]

[Total 3]

Solution X4.7

Moving average processes are covered in Chapter 13.

(i) *Mean and variance*

The mean and variance are as follows:

$$E(X_t)=E(3.1+\varepsilon_t+0.25\varepsilon_{t-1}+0.5\varepsilon_{t-2}+0.25\varepsilon_{t-3})$$

$$=3.1+E(\varepsilon_t)+0.25E(\varepsilon_{t-1})+0.5E(\varepsilon_{t-2})+0.25E(\varepsilon_{t-3})=3.1 \qquad [1]$$

$$\text{var}(X_t)=\text{var}(3.1+\varepsilon_t+0.25\varepsilon_{t-1}+0.5\varepsilon_{t-2}+0.25\varepsilon_{t-3})$$

$$=\text{var}(\varepsilon_t)+0.25^2\,\text{var}(\varepsilon_{t-1})+0.5^2\,\text{var}(\varepsilon_{t-2})+0.25^2\,\text{var}(\varepsilon_{t-3})$$

$$=1.375\sigma^2 \qquad [1]$$

Alternatively, we could calculate the variance by writing it in terms of covariance:

$$\text{var}(X_t)=\text{cov}(X_t,X_t)$$

$$=\text{cov}(3.1+\varepsilon_t+0.25\varepsilon_{t-1}+0.5\varepsilon_{t-2}+0.25\varepsilon_{t-3},$$

$$3.1+\varepsilon_t+0.25\varepsilon_{t-1}+0.5\varepsilon_{t-2}+0.25\varepsilon_{t-3})$$

$$=\sigma^2+0.25^2\sigma^2+0.5^2\sigma^2+0.25^2\sigma^2$$

$$=1.375\sigma^2 \qquad \textit{[1]}$$

The constant of 3.1 can be omitted as it does not affect the variance or covariance.

[Total 2]

(ii) *Autocorrelation function*

The process is stationary as it is the sum of stationary white noise terms, so we can calculate the autocovariance function as follows (ignoring the 3.1's as they will not affect the results):

$$\gamma_0=\text{cov}(X_t,X_t)=\text{var}(X_t)=1.375\sigma^2 \text{ from (i)}$$

$$\gamma_1=\text{cov}(X_t,X_{t-1})$$

$$=\text{cov}(\varepsilon_t+0.25\varepsilon_{t-1}+0.5\varepsilon_{t-2}+0.25\varepsilon_{t-3},\varepsilon_{t-1}+0.25\varepsilon_{t-2}+0.5\varepsilon_{t-3}+0.25\varepsilon_{t-4})$$

$$=0.25\sigma^2+(0.5)(0.25)\sigma^2+(0.25)(0.5)\sigma^2=0.5\sigma^2 \qquad [½]$$

$$\gamma_2=\text{cov}(X_t,X_{t-2})$$

$$=\text{cov}(\varepsilon_t+0.25\varepsilon_{t-1}+0.5\varepsilon_{t-2}+0.25\varepsilon_{t-3},\varepsilon_{t-2}+0.25\varepsilon_{t-3}+0.5\varepsilon_{t-4}+0.25\varepsilon_{t-5})$$

$$=0.5\sigma^2+0.25^2\sigma^2=0.5625\sigma^2 \qquad [½]$$

$$\gamma_3 = \text{cov}(X_t, X_{t-3})$$

$$= \text{cov}(\varepsilon_t + 0.25\varepsilon_{t-1} + 0.5\varepsilon_{t-2} + 0.25\varepsilon_{t-3}, \varepsilon_{t-3} + 0.25\varepsilon_{t-4} + 0.5\varepsilon_{t-5} + 0.25\varepsilon_{t-6})$$

$$= 0.25\sigma^2$$ [½]

$\gamma_k = 0$ for $k > 3$ [½]

Since $\rho_k = \gamma_k / \gamma_0$, the autocorrelation function is:

$$\rho_1 = \frac{0.5\sigma^2}{1.375\sigma^2} = \frac{4}{11} \quad (= 0.364)$$ [½]

$$\rho_2 = \frac{0.5625\sigma^2}{1.375\sigma^2} = \frac{9}{22} \quad (= 0.409)$$ [½]

$$\rho_3 = \frac{0.25\sigma^2}{1.375\sigma^2} = \frac{2}{11} \quad (= 0.182)$$ [½]

Finally, $\rho_0 = 1$ and $\rho_k = 0$ for $k > 3$. [½]

[Total 4]

Solution X4.8

Cointegration is covered in Chapter 14.

(i) ***Why CPI and NAEI might be cointegrated***

CPI is price inflation, which drives wage inflation (the NAEI). So we would expect them to 'move together'. [1]

Neither process is stationary – they both have a trend (as prices and wages increase over time), so they may both be $I(1)$. [1]

[Total 2]

(ii) ***Cointegrated***

We need to show that X and Y are both $I(1)$ and that $0.6X - Y$ is stationary.

We can check the stationarity of X by calculating the roots of its characteristic equation. Rewriting the equation as:

$$X_n - 1.2X_{n-1} + 0.2X_{n-2} = \varepsilon_n^X$$

we see that its characteristic equation is:

$$1 - 1.2\lambda + 0.2\lambda^2 = 0$$ [½]

The roots of this equation are:

$$\lambda = \frac{1.2 \pm \sqrt{(-1.2)^2 - 4\times 0.2 \times 1}}{2\times 0.2} = 5,\ 1$$ [½]

Differencing once will eliminate the root of 1. The only remaining root is strictly greater than 1 in absolute value, so ∇X is stationary and hence X is $I(1)$. [½]

Alternatively, to show that ∇X is stationary, we could difference the process as follows:

$$X_n - 1.2X_{n-1} + 0.2X_{n-2} = \varepsilon_n^X$$

$$\Rightarrow (X_n - X_{n-1}) - 0.2(X_{n-1} - X_{n-2}) = \varepsilon_n^X$$

$$\Rightarrow \nabla X_n - 0.2\nabla X_{n-1} = \varepsilon_n^X$$ *[½]*

The characteristic equation of ∇X is:

$$1 - 0.2\lambda = 0$$ *[½]*

Since the root of this equation is strictly greater than 1 in absolute value, ∇X is stationary. *[½]*

It is also acceptable to deduce stationarity from the fact that the coefficient of ∇X_{n-1} in the equation $\nabla X_n - 0.2\nabla X_{n-1} = \varepsilon_n^X$ is less than 1 in magnitude. *[1½]*

The process Y is defined by the equation:

$$Y_n = 0.6X_{n-1} + \varepsilon_n^y$$

Since Y is a linear combination of the $I(1)$ process X and the stationary process ε^y, Y is $I(1)$. [1]

Alternatively, we could consider differencing the process Y :

$$(Y_n - Y_{n-1}) = (0.6X_{n-1} + \varepsilon_n^y) - (0.6X_{n-2} + \varepsilon_{n-1}^y)$$

$$\Rightarrow \nabla Y_n = 0.6\nabla X_{n-1} + \nabla\varepsilon_n^y$$

Since ∇X is stationary and the white noise terms are stationary, it follows that ∇Y is also stationary. Hence Y is $I(1)$. *[1]*

We now consider the process $0.6X - Y$. The defining equation for this process is:

$$0.6X_n - Y_n = 0.6(1.2X_{n-1} - 0.2X_{n-2} + \varepsilon_n^x) - (0.6X_{n-1} + \varepsilon_n^y) \qquad [½]$$

$$= 0.12X_{n-1} - 0.12X_{n-2} + 0.6\varepsilon_n^x - \varepsilon_n^y$$

$$= 0.12\nabla X_{n-1} + 0.6\varepsilon_n^x - \varepsilon_n^y \qquad [½]$$

Since ∇X is stationary and the white noise process is stationary, it follows that $0.6X - Y$ is also stationary. [½]

[Total 4]

Solution X4.9

ARIMA processes are covered in Chapter 13.

(i) ***I(d)***

A process, *X,* is said to be $I(d)$ ('integrated of order d') if the d th difference, $\nabla^d X$, is a stationary process. [1]

[Total 1]

Note that unless $d=0$, X is not stationary.

(ii)(a) ***Classify X***

This is an $MA(1)$ process and hence it is stationary (as it is the sum of stationary white noise terms). So it is an $ARIMA(0,0,1)$ process. [1]

(ii)(b) ***Classify Y***

This is an $ARMA(2,3)$ process. [½]

We check stationarity by calculating the roots of the characteristic equation of the autoregressive part. Rewriting the equation as:

$$Y_t - 1.4Y_{t-2} = \varepsilon_t + 0.5\varepsilon_{t-3}$$

we see that the characteristic equation is:

$$1 - 1.4\lambda^2 = 0$$ [½]

The roots of this equation are $\pm\sqrt{\dfrac{1}{1.4}}$, *ie* ± 0.8452. [½]

These roots are less than 1 in magnitude, so the process is not stationary. Differencing will not eliminate these roots, so this process is not an *ARIMA* process. [½]

(ii)(c) ***Classify W***

This is an $ARMA(2,1)$ process. [½]

We check stationarity by calculating the roots of the characteristic equation of the autoregressive part. Rewriting the equation as:

$$W_t - 1.4W_{t-1} + 0.4W_{t-2} = \varepsilon_t + \varepsilon_{t-1}$$

we see that the characteristic equation is:

$$1 - 1.4\lambda + 0.4\lambda^2 = 0$$ [½]

The roots of this equation are:

$$\lambda = \frac{1.4 \pm \sqrt{(-1.4)^2 - 4 \times 1 \times 0.4}}{2 \times 0.4} = 2.5,\ 1 \qquad [½]$$

Differencing once will eliminate the root of 1. The only remaining root is strictly greater than 1 in absolute value, so ∇W is stationary $ARMA(1,1)$. [1]

Here the value of p is 1 since there is only one remaining root. Differencing has no effect on the value of q.

Alternatively, to show that ∇W is stationary, we could difference the process as follows:

$$W_t - 1.4W_{t-1} + 0.4W_{t-2} = \varepsilon_t + \varepsilon_{t-1}$$

$$\Rightarrow (W_t - W_{t-1}) - 0.4(W_{t-1} - W_{t-2}) = \varepsilon_t + \varepsilon_{t-1}$$

$$\Rightarrow \nabla W_t - 0.4\nabla W_{t-1} = \varepsilon_t + \varepsilon_{t-1} \qquad [1]$$

The characteristic equation of ∇W is:

$$1 - 0.4\lambda = 0 \qquad [½]$$

The root of this equation is 2.5, which is strictly greater than 1 in absolute value. So ∇W is stationary. *[½]*

So W is an $ARIMA(1,1,1)$ process. [½]

[Total 6]

Solution X4.10

Mean residual life and tail weights are covered in Section 4 of Chapter 16.

(i) ***Mean residual life***

The mean residual life is given by:

$$e(x)=\frac{\int_x^{\infty}\left(1-F(y)\right)dy}{1-F(x)}\qquad [1]$$

The PDF of the $Gamma(2,1)$ distribution is:

$$f(x)=\frac{1^2}{\Gamma(2)}xe^{-x}=xe^{-x}\,,\ x>0$$

Here we are using the fact that $\Gamma(2)=1!=1$.

The CDF can be obtained by integrating the PDF:

$$F(x)=\int_0^x f(t)\,dt=\int_0^x te^{-t}dt\qquad [½]$$

Integrating by parts:

$$F(x)=\left[-te^{-t}\right]_0^x-\int_0^x\left(-e^{-t}\right)dt$$

$$=-xe^{-x}+\int_0^x e^{-t}\,dt$$

$$=-xe^{-x}+\left[-e^{-t}\right]_0^x$$

$$=-xe^{-x}-e^{-x}+1$$

So:

$$1-F(x)=xe^{-x}+e^{-x}=(x+1)e^{-x}\qquad [1½]$$

and:

$$\int_x^{\infty}\left(1-F(y)\right)dy=\int_x^{\infty}(y+1)e^{-y}\,dy$$

Integrating by parts again:

$$\int_x^\infty \left(1-F(y)\right)dy = \left[-(y+1)e^{-y}\right]_x^\infty - \int_x^\infty \left(-e^{-y}\right)dy$$

$$= \left(0+(x+1)e^{-x}\right)+\int_x^\infty e^{-y}\,dy$$

$$= (x+1)e^{-x}+\left[-e^{-y}\right]_x^\infty$$

$$= (x+1)e^{-x}+(0+e^{-x})$$

$$= (x+2)e^{-x}$$ [1½]

So the mean residual life is:

$$e(x)=\frac{(x+2)e^{-x}}{(x+1)e^{-x}}=\frac{x+2}{x+1}$$ [½]

[Total 5]

(ii) ***Comparison with exponential distribution***

By the memoryless property of the exponential distribution, the mean residual life of $Exp(\lambda)$ is:

$$E(X)=\frac{1}{\lambda}$$

So the mean residual life of $Exp(1)$ is 1. [1]

This can also be obtained as follows:

$$e(x)=\frac{\int_x^\infty \left(1-F(y)\right)dy}{1-F(x)}=\frac{\int_x^\infty e^{-y}\,dy}{e^{-x}}=\frac{\left[-e^{-y}\right]_x^\infty}{e^{-x}}=\frac{0+e^{-x}}{e^{-x}}=1$$ *[1]*

The mean residual life of $Gamma(2,1)$ is:

$$e(x)=\frac{x+2}{x+1}=1+\frac{1}{x+1}$$

This is a decreasing function of x, whereas the mean residual lifetime of $Exp(1)$ is a constant. [1]

This indicates that $Gamma(2,1)$ has a lighter tail than $Exp(1)$. [1]

[Total 3]

Solution X4.11

Invertibility is covered in Chapter 13. Estimating the parameters in a time series process and forecasting are covered in Chapter 14.

(i) ***Invertiblility***

The process is invertible provided $|\beta|<1$. [1]

[Total 1]

This can be deduced as follows, but no explanation is required to obtain the mark. The characteristic equation of the moving average part is:

$$1+\beta\lambda=0$$

The root of this equation is $\lambda=-\dfrac{1}{\beta}$. *The process is invertible if* $|\lambda|>1$, *ie if* $|\beta|<1$.

(ii)(a) ***Method of moments***

The method of moments estimates of α and β are obtained by equating the expressions for ρ_1 and ρ_2 with the corresponding sample autocorrelation coefficients:

$$\frac{(\hat{\alpha}+\hat{\beta})(1+\hat{\alpha}\hat{\beta})}{1+2\hat{\alpha}\hat{\beta}+\hat{\beta}^2}=0.440 \quad \text{and} \quad \frac{(\hat{\alpha}+\hat{\beta})(1+\hat{\alpha}\hat{\beta})}{1+2\hat{\alpha}\hat{\beta}+\hat{\beta}^2}\hat{\alpha}=0.264$$ [½]

Dividing gives:

$$\hat{\alpha}=\frac{0.264}{0.440}=0.6$$ [½]

Substituting $\alpha=0.6$ into the first equation:

$$\frac{(0.6+\hat{\beta})(1+0.6\hat{\beta})}{1+1.2\hat{\beta}+\hat{\beta}^2}=0.440$$ [½]

$$\Rightarrow (0.6+\hat{\beta})(1+0.6\hat{\beta})=0.440(1+1.2\hat{\beta}+\hat{\beta}^2)$$

$$\Rightarrow 0.16+0.832\hat{\beta}+0.16\hat{\beta}^2=0$$ [½]

The roots of this quadratic equation are:

$$\hat{\beta}=\frac{-0.832\pm\sqrt{0.832^2-4\times0.16^2}}{2\times0.16}=-0.2,-5$$ [1]

Since the process is invertible, the method of moments estimate of β is –0.2. [½] (ii)(b) ***Least squares estimation***

The steps in the process are as follows:

- Assume $\varepsilon_0 = 0$, and obtain (iteratively) expressions for ε_1, ε_2, … [½]
- Obtain $\sum \varepsilon_i^2$ in terms of α and β. [½]
- Calculate values of α and β that minimise this expression. [½]
- Use these values of α and β to work backwards to time zero from the most recent known values of the time series, to determine an updated value for ε_0. [½]
- Repeat the process to find improved estimates for α and β. [½]

(ii)(c) ***Maximum likelihood estimation***

The maximum likelihood estimates and least squared estimates are equivalent if we assume that $\varepsilon_t \sim N(0,\sigma^2)$. [1]

[Total 7]

(iii)(a) ***1 and 2 step ahead estimates***

The fitted model is:

$$x_n = 0.6x_{n-1} + \varepsilon_n - 0.2\varepsilon_{n-1}$$

Hence the 1 and 2 step-ahead forecasts are:

$$\hat{x}_{80}(1) = 0.6x_{80} + \hat{\varepsilon}_{81} - 0.2\hat{\varepsilon}_{80} = (0.6 \times 1.087) + 0 - (0.2 \times 1.181) = 0.416$$ [1]

$$\hat{x}_{80}(2) = 0.6\hat{x}_{80}(1) + \hat{\varepsilon}_{82} - 0.2\hat{\varepsilon}_{81} = (0.6 \times 0.416) + 0 - (0.2 \times 0) = 0.2496$$ [1]

(iii)(b) ***Exponential smoothing***

Exponential smoothing uses the formula:

$$\hat{x}_n(1) = (1-\alpha)\hat{x}_{n-1}(1) + \alpha x_n$$

We are told that the estimate at time 79 for x_{80} is $\hat{x}_{79}(1) = 0.625$ and the smoothing parameter is $\alpha = 0.2$. Hence, our estimate of x_{81} is:

$$\hat{x}_{80}(1) = (1-\alpha)\hat{x}_{79}(1) + \alpha x_{80}$$ [½]

$$= (0.8 \times 0.625) + (0.2 \times 1.087)$$

$$= 0.7174$$ [½]

[Total 3]

Solution X4.12

ARIMA processes are defined in Chapter 13. Selecting an appropriate value of d and testing whether the residuals conform to white noise are covered in Chapter 14.

(i) ***Definition of ARIMA process***

X is an $ARIMA(p,d,q)$ process if the d th difference, $\nabla^d X$, is a stationary $ARMA(p,q)$ process. [1]

This means that $\nabla^d X$ is a stationary process whose defining equation is of the form:

$$\nabla^d X_t - \mu = \alpha_1\left(\nabla^d X_{t-1} - \mu\right) + \alpha_2\left(\nabla^d X_{t-2} - \mu\right) + \cdots + \alpha_p\left(\nabla^d X_{t-p} - \mu\right)$$

$$+\varepsilon_t + \beta_1\varepsilon_{t-1} + \beta_2\varepsilon_{t-2} + \cdots + \beta_q\varepsilon_{t-q}$$

where ε is a white noise process.

(ii) ***An appropriate value of d***

There are two things to consider when selecting an appropriate value for d.

We want the value of d that minimises the sample variance. This would lead us to set $d = 0$. [1]

We should also examine the sample autocorrelation coefficients. If these decay slowly from 1, this is an indication that differencing is required. However, this is not the case for the (undifferenced) process X, which again leads us to choose $d = 0$. [1]

[Total 2]

(iii) ***Statistical tests***

For each of these tests, the null hypothesis is:

H_0 : the residuals form a white noise process with zero mean [½]

(a) ***Ljung-Box test***

The formula for the test statistic is given on page 42 of the *Tables*.

This test checks for correlation between the residuals. If the residuals form a white noise process, then the sample autocorrelations will be small. [1]

Here we have $m = 5$ and $n = 100$. The observed value of the test statistic is:

$$n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k} = 100\times 102\left\{\frac{0.14^2}{99} + \frac{(-0.05)^2}{98} + \frac{0.1^2}{97} + \frac{0.12^2}{96} + \frac{(-0.02)^2}{95}\right\} = 4.904 \qquad [1]$$

Since the fitted model is $ARMA(1,1)$, *ie* $p = 1$ and $q = 1$, we compare the value of the test statistic with the χ_3^2 distribution. [½]

This is a one-tailed test. From page 169 of the *Tables*, we see that the upper 5% point of the χ_3^2 distribution is 7.815. [½]

Since $4.904 < 7.815$, we have insufficient evidence at the 5% level to reject the null hypothesis. Hence we conclude that the residuals are uncorrelated. [½]

(b) ***Turning point test***

This test checks whether the residuals are patternless. [½]

Formulae for the expected value and variance of the number of turning points are given on page 42 of the *Tables*. Here we have $n = 100$ and:

$$E(T) = \tfrac{2}{3} \times 98 = 65\tfrac{1}{3} \qquad [½]$$

$$\text{var}(T) = \frac{16 \times 100 - 29}{90} = \frac{1{,}571}{90} \quad (= 17.456) \qquad [½]$$

The test is carried out using a normal approximation to the distribution of T. Since T is a discrete random variable, we should use a continuity correction when calculating the value of the test statistic. This gives a test statistic of:

$$\frac{73\frac{1}{2} - 65\frac{1}{3}}{\sqrt{\frac{1{,}571}{90}}} = 1.955 \qquad [1]$$

This is a two-tailed test (as too many or too few turning points would suggest a non-random pattern). The upper and lower 2.5% points of the standard normal distribution are ± 1.96. [½]

Since $1.955 < 1.960$, there is insufficient evidence to reject the null hypothesis at the 5% significance level. So we conclude that the residuals are patternless. [½]

Alternatively, we could calculate the approximate p-value of the test as follows:

$$2P\left(T > 74\right) = 2P\left(N(0,1) > \frac{73\frac{1}{2} - 65\frac{1}{3}}{\sqrt{\frac{1{,}571}{90}}} \right) = 2\left[1 - \Phi(1.9547)\right] = 2\left[1 - 0.9747\right] = 0.0506 \qquad [1]$$

Since this is more than 5%, there is insufficient evidence to reject the null hypothesis. So we conclude that the residuals are patternless. *[1]*

If no continuity correction is used, the value of the test statistic is 2.074 and the p-value is 3.80%. This would lead us to reject the null hypothesis and we would conclude that the residuals are not patternless.

Markers: Deduct 1 mark for omission of the continuity correction.

(c) ***Inspection of the SACF***

This test also checks whether the residuals are uncorrelated. [½]

The ACF of the residuals should be zero for all lags except 0. An approximate 95% confidence interval for ρ_k, $k \geq 1$, is $\pm 2/\sqrt{n} = \pm 0.2$ (or more accurately $\pm 1.96/\sqrt{n} = \pm 0.196$). [1]

Since all the values lie within this confidence interval, there is insufficient evidence to reject the null hypothesis. [½]

The tests suggest that the residuals form a white noise process and hence the fitted $ARMA(1,1)$ model is satisfactory. [½]

[Total 10]

Solution X4.13

Univariate processes are covered in Chapter 13. Multivariate processes are covered in Chapter 14.

(i)(a) ***Stationary?***

To check stationarity, we examine the roots of the characteristic equation. Rewriting the defining equation as:

$$X_n - 0.7X_{n-1} + 0.1X_{n-2} = \varepsilon_n$$

We see that the characteristic equation is:

$$1 - 0.7\lambda + 0.1\lambda^2 = 0$$ [½]

The roots of this equation are:

$$\lambda = \frac{0.7 \pm \sqrt{(-0.7)^2 - 4\times 0.1\times 1}}{2\times 0.1} = 5,2$$ [½]

Since both of the roots are greater than one in magnitude the process is stationary. [½]

(i)(b) ***Invertible?***

A process X is invertible if we can express the white noise as a linear combination of terms involving X. In this case, we can write:

$$\varepsilon_n = X_n - 0.7X_{n-1} + 0.1X_{n-2}$$

So X is an invertible process. [½]

AR processes are always invertible. When fitting a time series model, we calculate the residuals by inverting the process.

(i)(c) ***Purely indeterministic?***

The value of X_n is a linear combination of the past two values of the process itself plus a random white noise term, ε_n. The further into the future we go, the more random terms are added and so past values of the process become less and less useful for predicting future values. [½]

So the process is purely indeterministic. [½]

(i)(d) ***Markov?***

The Markov property states that future probabilities depend only on the current state of the process. [1]

Here X_n depends on X_{n-1} and X_{n-2}. [½]

So the process is not Markov. [½]

[Total 5]

(ii)(a) ***ACF***

We know from part (i) that the process is stationary. So we can write down the Yule-Walker equations for the autocovariance function in the usual way. For an AR process, we do not need to determine an expression for γ_0 to be able to calculate ρ_k (as the factors of γ_0 will cancel).

The autocovariance at lag 1 is:

$$\begin{aligned}\gamma_1 &= \text{cov}(X_n, X_{n-1})\\ &= \text{cov}(0.7X_{n-1} - 0.1X_{n-2} + \varepsilon_n, X_{n-1})\\ &= 0.7\gamma_0 - 0.1\gamma_1\end{aligned}$$ [1]

Rearranging:

$$1.1\gamma_1 = 0.7\gamma_0$$

$$\Rightarrow \gamma_1 = \frac{7}{11}\gamma_0$$

$$\Rightarrow \rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{7}{11}$$ [½]

The autocovariance at lag 2 is:

$$\begin{aligned}\gamma_2 &= \text{cov}(X_n, X_{n-2})\\ &= \text{cov}(0.7X_{n-1} - 0.1X_{n-2} + \varepsilon_n, X_{n-2})\\ &= 0.7\gamma_1 - 0.1\gamma_0\end{aligned}$$ [1]

So:

$$\rho_2 = \frac{\gamma_2}{\gamma_0} = 0.7\rho_1 - 0.1 = 0.7 \times \frac{7}{11} - 0.1 = \frac{19}{55}$$ [½]

In general for $k > 2$, we have:

$$\begin{aligned}\gamma_k &= \text{cov}(X_n, X_{n-k})\\ &= \text{cov}(0.7X_{n-1} - 0.1X_{n-2} + \varepsilon_n, X_{n-k})\\ &= 0.7\gamma_{k-1} - 0.1\gamma_{k-2}\end{aligned}$$ [1]

So:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{0.7\gamma_{k-1} - 0.1\gamma_{k-2}}{\gamma_0} = 0.7\rho_{k-1} - 0.1\rho_{k-2}$$ [½]

(ii)(b) ***General solution to the difference equation***

Suppose that $\rho_k = \frac{A}{2^k} + \frac{B}{5^k}$, for $k > 2$. Then:

$$0.7\rho_{k-1} - 0.1\rho_{k-2} = 0.7\left(\frac{A}{2^{k-1}} + \frac{B}{5^{k-1}}\right) - 0.1\left(\frac{A}{2^{k-2}} + \frac{B}{5^{k-2}}\right) \quad [½]$$

The right-hand side of this equation can be written as:

$$\frac{A}{2^k}(0.7\times 2 - 0.1\times 2^2) + \frac{B}{5^k}(0.7\times 5 - 0.1\times 5^2) \quad [1]$$

and this simplifies to:

$$\frac{A}{2^k} + \frac{B}{5^k} \quad [½]$$

So:

$$0.7\rho_{k-1} - 0.1\rho_{k-2} = \rho_k$$

as required.

Markers: Please award only 1 mark if the formula for the general solution of a second-order differential equation is quoted (without proof) from page 4 of the Tables.

We now have to calculate the values of A and B. We need two equations to do this. Since the difference equation holds for $k > 2$, we can consider ρ_3 and ρ_4.

We have:

$$\rho_3 = 0.7\rho_2 - 0.1\rho_1 = 0.7\times\frac{19}{55} - 0.1\times\frac{7}{11} = \frac{49}{275} \quad [½]$$

$$\rho_4 = 0.7\rho_3 - 0.1\rho_2 = 0.7\times\frac{49}{275} - 0.1\times\frac{19}{55} = \frac{124}{1,375} \quad [½]$$

So:

$$\rho_3 = \frac{A}{2^3} + \frac{B}{5^3} = \frac{49}{275} \quad [½]$$

$$\rho_4 = \frac{A}{2^4} + \frac{B}{5^4} = \frac{124}{1,375} \quad [½]$$

Multiplying ρ_4 by 2, we see that:

$$\frac{A}{2^3} + \frac{2B}{5^4} = \frac{248}{1,375}$$

Then subtracting ρ_3:

$$\frac{2B}{5^4}-\frac{B}{5^3}=\frac{248}{1,375}-\frac{49}{275}$$

$$\Rightarrow -\frac{3B}{625}=\frac{3}{1,375}$$

$$\Rightarrow B=-\frac{625}{1,375}=-\frac{5}{11}$$ [1]

Substituting this back into the equation for ρ_3 gives:

$$\frac{A}{2^3}+\frac{1}{5^3}\left(-\frac{5}{11}\right)=\frac{49}{275}$$

$$\Rightarrow A=\frac{16}{11}$$ [1]

In fact, the general autocorrelation function $\rho_k=0.7\rho_{k-1}-0.1\rho_{k-2}$ *also holds for* $k=1$ *and* $k=2$.

For example, using the results from part (ii)(a):

$$\rho_1=0.7-0.1\rho_1=0.7\rho_0-0.1\rho_{-1}$$

and:

$$\rho_2=0.7\rho_1-0.1=0.7\rho_1-0.1\rho_0$$

Similarly, we can see that the formula:

$$\frac{A}{2^k}+\frac{B}{5^k}=0.7\left(\frac{A}{2^{k-1}}+\frac{B}{5^{k-1}}\right)-0.1\left(\frac{A}{2^{k-2}}+\frac{B}{5^{k-2}}\right)$$

holds for $k=1$ *and* $k=2$.

Hence:

$$\rho_1=\frac{A}{2^1}+\frac{B}{5^1}=\frac{7}{11}$$ *[1]*

$$\rho_2=\frac{A}{2^2}+\frac{B}{5^2}=\frac{19}{55}$$ *[1]*

Solving these simultaneously gives $B=-\frac{5}{11}$ *and* $A=\frac{16}{11}$ *as before.* *[2]*

(ii)(c) ***PACF***

Using the formulae given on page 40 with $\rho_1 = \frac{7}{11}$ and $\rho_2 = \frac{19}{55}$ gives:

$$\phi_1 = \rho_1 = \frac{7}{11}$$ [½]

$$\phi_2 = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2} = -\frac{1}{10}$$ [½]

Since this is an $AR(2)$ process, $\phi_k = 0$ for $k > 2$. [½]

[Total 12]

(iii) ***Multivariate process***

We can express $X_n = 0.7X_{n-1} - 0.1X_{n-2} + \varepsilon_n$ as a $VAR(1)$ process as follows:

$$\begin{pmatrix} X_n \\ X_{n-1} \end{pmatrix} = \begin{pmatrix} 0.7 & -0.1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X_{n-1} \\ X_{n-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_n \\ 0 \end{pmatrix}$$ [1]

This is of the form $\underline{X}_n = M\,\underline{X}_{n-1} + \underline{\varepsilon}_n$ where:

$$\underline{X}_n = \begin{pmatrix} X_n \\ X_{n-1} \end{pmatrix} \qquad M = \begin{pmatrix} 0.7 & -0.1 \\ 1 & 0 \end{pmatrix} \qquad \underline{\varepsilon}_n = \begin{pmatrix} \varepsilon_n \\ 0 \end{pmatrix}$$ [1]

[Total 2]

(iv)(a) ***Stationary***

A $VAR(1)$ process is stationary if the eigenvalues of matrix M are strictly less than 1 in magnitude.

The eigenvalues are the values of λ that satisfy the equation:

$$\det \begin{pmatrix} 0.7-\lambda & -0.1 \\ 1 & -\lambda \end{pmatrix} = 0$$ [1]

ie:

$$(0.7-\lambda)(-\lambda) - (-0.1)(1) = \lambda^2 - 0.7\lambda + 0.1 = 0$$ [½]

The roots of this equation are:

$$\lambda = \frac{0.7 \pm \sqrt{(-0.7)^2 - 4\times 1 \times 0.1}}{2\times 1} = 0.2, 0.5$$ [1]

Since both of the eigenvalues are less than one in magnitude the process is stationary. [½] (iv)(b) ***Markov***

From the matrix equation, we see that $\underline{X}_n$ depends on $\underline{X}_{n-1}$ only. [½]

So the process is Markov. [½]

[Total 4]

Solution X5.1

Reinsurance is covered in Chapter 18.

(a) ***Mean amount paid by the insurer (after reinsurance)***

Let Y be the amount paid by the insurer (after reinsurance). Then $Y = 0.85X$ and:

$$E(Y) = 0.85E(X) = 0.85 \times 2{,}000 = £1{,}700 \qquad [1]$$

(b) ***Variance of the amount paid by the reinsurer***

Let Z be the amount paid by the reinsurer. Then $Z = 0.15X$ and:

$$\text{var}(Z) = 0.15^2\,\text{var}(X) = 0.15^2 \times 100^2 = £^2 225 \qquad [1]$$

(c) ***MGF of the amount paid by the insurer (after reinsurance)***

We have:

$$M_Y(t) = E(e^{tY}) = E(e^{t\times 0.85X}) = E(e^{(0.85t)X}) = M_X(0.85t) \qquad [½]$$

We know that X has a gamma distribution. We can calculate the parameters of this distribution from the mean and variance:

$$E(X) = \frac{\alpha}{\lambda} = 2{,}000$$

$$\text{var}(X) = \frac{\alpha}{\lambda^2} = 100^2$$

Dividing these we see that $\lambda = 0.2$ and hence $\alpha = 2{,}000 \times 0.2 = 400$. [1]

From page 12 of the *Tables*, the MGF of X is:

$$M_X(t) = \left(1 - \frac{t}{0.2}\right)^{-400} = (1-5t)^{-400}, \quad t < 0.2$$

Hence:

$$M_Y(t) = M_X(0.85t) = \left(1 - \frac{0.85t}{0.2}\right)^{-400} = (1-4.25t)^{-400}, \quad t < \frac{1}{4.25} \qquad [½]$$

[Total 4]

Solution X5.2

The insurer's aggregate claim amount random variable net of reinsurance is discussed in Chapter 20.

If X is the gross individual claim amount random variable, then $X \sim Exp(0.01)$ and the amount paid by the insurer on a claim is:

$$Y = \begin{cases} X & \text{if } X < 200 \\ 200 & \text{if } X \geq 200 \end{cases}$$

The aggregate amount paid by the insurer is $S_I = Y_1 + Y_2 + \cdots + Y_N$, where $N \sim Bin(1000, 0.01)$.

We have:

$$E(N) = 1{,}000 \times 0.01 = 10 \qquad [½]$$

$$E(Y) = \int_0^{200} xf(x)\,dx + \int_{200}^{\infty} 200f(x)\,dx$$

$$= \int_0^{200} x\,0.01e^{-0.01x}\,dx + 200\left[1 - F(200)\right] \qquad [1]$$

Integrating by parts gives:

$$E(Y) = \left[-xe^{-0.01x}\right]_0^{200} + \int_0^{200} e^{-0.01x}\,dx + 200e^{-2}$$

$$= \cancel{-200e^{-2}} + \left[-100e^{-0.01x}\right]_0^{200} + \cancel{200e^{-2}}$$

$$= 100(1 - e^{-2})$$

$$= 86.466 \qquad [1½]$$

Hence:

$$E(S_I) = E(N)E(Y) = 10 \times 86.466 = £864.66 \qquad [1]$$

[Total 4]

Alternatively, we could calculate the expected amount paid by the reinsurer on an individual mean claim, $E(Z)$, and calculate $E(Y)$ as $E(X) - E(Z)$.

The amount paid by the reinsurer on an individual claim is:

$$Z = \begin{cases} 0 & \text{if } X < 200 \\ X - 200 & \text{if } X \geq 200 \end{cases}$$

The reinsurer's individual mean claim amount is:

$$E(Z)=\int_{200}^{\infty}\left(x-200\right)f(x)\,dx=\int_{200}^{\infty}\left(x-200\right)0.01e^{-0.01x}\,dx$$

Using the substitution $u=x-200$ *:*

$$E(Z)=\int_{0}^{\infty}u\,0.01e^{-0.01\left(u+200\right)}\,du=e^{-2}\int_{0}^{\infty}u\,0.01e^{-0.01u}\,du$$

The integral part of the expression above is the mean of the $Exp(0.01)$ *distribution, which is 100.*

So $E(Z)=100e^{-2}$. *[2]*

Since $X\sim Exp(0.01)$, *it follows that* $E\left(X\right)=100$.

So $E(Y)=E(X)-E(Z)=100-100e^{-2}=86.466$ *as before.* *[½]*

Solution X5.3

Copulas are covered in Chapter 17.

(i) ***Probability of paying a death benefit***

The probability of death within 10 years for each life is:

$$u = v = F(10) = 1 - {}_{10}p_{70} = 1 - 0.58 = 0.42 \qquad [1]$$

Hence, the Clayton copula gives:

$$C[0.42, 0.42] = \left(0.42^{-0.3} + 0.42^{-0.3} - 1\right)^{-1/0.3} = 0.2111 \qquad [1]$$

and the FGM copula gives:

$$C[0.42, 0.42] = 0.42^2\left[1 + (-0.1)(0.58)(0.58)\right] = 0.1705 \qquad [1]$$

[Total 3]

(ii) ***Suitability of the copulas***

If the deaths are independent, then the probability of paying a benefit is:

$$0.42 \times 0.42 = 0.1764 \qquad [1]$$

However, the two lives covered are related ('retired couples'), so we would expect the probability of paying a benefit to be higher than under the assumption of independence. Hence the Clayton copula is more appropriate. [1]

If we are intending to introduce dependence (*ie* positive correlation) between the two lives, then a model that decreases the mortality probability relative to the independent calculation (such as the FGM model) is unsuitable. [1]

[Total 3]

Solution X5.4

Excess of loss reinsurance is covered in Chapter 18.

The reinsurer's expected payment amount is given by $E(Z)$, where:

$$E(Z)=\int_{1,000}^{2,000}(x-1,000)f(x)\,dx \;+\; \int_{2,000}^{\infty}1,000f(x)\,dx \qquad [1]$$

$$=\int_{1,000}^{2,000}xf(x)\,dx \;-\; \int_{1,000}^{2,000}1,000f(x)\,dx \;+\; \int_{2,000}^{\infty}1,000f(x)\,dx \qquad [1]$$

$$=\int_{1,000}^{2,000}xf(x)\,dx-1,000P(1,000<X<2,000)+1,000P(X>2,000)$$

Using the formula for the truncated moments of the lognormal distribution (given on page 18 of the *Tables*), we have:

$$E(Z)=e^{\mu+½\sigma^2}\left[\Phi\left(\frac{\ln 2,000-\mu}{\sigma}-\sigma\right)-\Phi\left(\frac{\ln 1,000-\mu}{\sigma}-\sigma\right)\right]$$

$$-1,000\left[\Phi\left(\frac{\ln 2,000-\mu}{\sigma}\right)-\Phi\left(\frac{\ln 1,000-\mu}{\sigma}\right)\right]$$

$$+1,000\left[1-\Phi\left(\frac{\ln 2,000-\mu}{\sigma}\right)\right] \qquad [2]$$

$$=e^7[\Phi(-0.69955)-\Phi(-1.04612)]-1,000[\Phi(1.30045)-\Phi(0.95388)]$$

$$+1,000[1-\Phi(1.30045)] \qquad [1]$$

$$=e^7(0.24210-0.14776)-1,000(0.90328-0.82992)+1,000\times(1-0.90328) \qquad [2]$$

$$=103.456-73.36+96.72$$

$$=126.82 \qquad [1]$$

[Total 8]

Solution X5.5

This question is based on the collective risk model, which is covered in Chapters 19 and 20.

(i) ***Mean and variance of S***

Let N denote the number of claims in a year, X denote the amount of an individual claim, and Y denote the expense associated with the claim. Then:

$$S = T_1 + \cdots + T_N$$

where $T_i = X_i + Y_i$.

S has a compound Poisson distribution with $N \sim Poisson(0.25n)$, so using the appropriate formulae from page 16 of the *Tables*:

$$E(S) = \lambda E(T) = 0.25n\,E(X+Y) = 0.25n\left[E(X)+E(Y)\right] \qquad [1]$$

and:

$$\begin{aligned} \text{var}(S) = \lambda E(T^2) &= 0.25n\,E\left[(X+Y)^2\right] \\ &= 0.25n\,E\left[X^2 + 2XY + Y^2\right] \\ &= 0.25n\left[E(X^2) + 2E(X)E(Y) + E(Y^2)\right] \qquad [1] \end{aligned}$$

Since $X \sim Pa(4,1800)$, we have:

$$E(X) = \frac{1,800}{3} = 600 \qquad [½]$$

$$E(X^2) = \frac{\Gamma(4-2)\Gamma(1+2)}{\Gamma(4)} \times 1,800^2 = \frac{1!\,2!}{3!} \times 1,800^2 = 1,080,000 \qquad [½]$$

Alternatively, $E(X^2) = \text{var}(X) + [E(X)]^2 = \dfrac{4 \times 1,800^2}{3^2 \times 2} + 600^2 = 1,080,000$. *[½]*

Formulae for the moments of a Pareto random variable are given on page 14 of the Tables.

Also, since $Y \sim U(35,85)$, we have:

$$E(Y) = \frac{35+85}{2} = 60 \qquad [½]$$

$$E(Y^2) = \frac{1}{85-35} \times \frac{1}{2+1}(85^3 - 35^3) = 3,808\tfrac{1}{3} \qquad [½]$$

Alternatively, $E(Y^2) = \text{var}(Y) + [E(Y)]^2 = \dfrac{(85-35)^2}{12} + 60^2 = 3,808\tfrac{1}{3}$. *[½]*

Formulae for the moments of a continuous uniform random variable are given on page 13 of the Tables.

So we have:

$$E(S) = 0.25n\left[600 + 60\right] = 165n \qquad [1]$$

and:

$$\text{var}(S) = 0.25n\left[1{,}080{,}000 + (2 \times 600 \times 60) + 3{,}808\tfrac{1}{3}\right] = 288{,}952\tfrac{1}{12}n \qquad [1]$$

[Total 6]

Alternatively, we can use the formula:

$$\text{var}(S) = E(N)\,\text{var}(T) + \text{var}(N)\left[E(T)\right]^2$$

where, by independence:

$$\text{var}(T) = \text{var}(X) + \text{var}(Y)$$

However, it is quicker to use the compound Poisson formula.

Another alternative is to calculate the mean and variance of the aggregate claim and expense amount per policy, and then sum over all policies.

Note that $\text{var}(S) \neq \text{var}(S_C) + \text{var}(S_E)$ *where* $S_C = X_1 + \cdots + X_N$ *and* $S_E = Y_1 + \cdots + Y_N$ *since* S_E *and* S_C *are NOT independent. (An expense only occurs if a claim occurs.)*

(ii) ***Number of policies needed***

We now assume that:

$$S \sim N(165n, 288\,952n) \text{ approximately}$$

To make a profit we must have total outgo less than total premium income, *ie* $S < 190n$. So we require:

$$P(S < 190n) \simeq P\left(N(0,1) < \frac{190n - 165n}{\sqrt{288{,}952n}}\right) = P\left(N(0,1) < 0.04651\sqrt{n}\right) \qquad [1]$$

For this probability to be at least 99%, we must have:

$$0.04651\sqrt{n} \geq 2.3263 \qquad [½]$$

ie $n \geq 2{,}501.9$ [½]

So the smallest value of n is 2,502. [1]

[Total 3]

Solution X5.6

Copulas are introduced in Chapter 17. Archimedean copulas are defined in Section 5.4 of Chapter 17.

(i) ***Copula function***

Let $u=F_X(x)$ and $v=F_Y(y)$. Then:

$$u=(1+e^{-x})^{-1}$$

$$\Rightarrow u^{-1}=1+e^{-x}$$

$$\Rightarrow e^{-x}=u^{-1}-1$$ [½]

Similarly $e^{-y}=v^{-1}-1$. [½]

Substituting these expressions for x and y into $F_{X,Y}(x,y)$, we have:

$$C[u,v]=F_{X,Y}(x,y)=\left(1+(u^{-1}-1)+(v^{-1}-1)\right)^{-1}=\left(u^{-1}+v^{-1}-1\right)^{-1}$$

as required. [1]

[Total 2]

(ii) ***Generator function***

To show that this is an Archimedean copula, we must show that it is of the form:

$$C\left[u,v\right]=\psi^{[-1]}\left(\psi\left(u\right)+\psi\left(v\right)\right)$$

where ψ is the generator function, and $\psi^{[-1]}$ is the pseudo-inverse function. [1]

Suppose that:

$$s=\psi(t)=t^{-1}-1$$

Then:

$$t^{-1}=s+1 \quad\Rightarrow\quad t=(s+1)^{-1}$$ [1]

So:

$$\psi^{-1}(s)=(s+1)^{-1}$$

Since this formula is valid for all $s\geq 0$, the pseudo-inverse function is:

$$\psi^{[-1]}(s)=(s+1)^{-1}$$ [1]

Hence:

$$\begin{aligned}\psi^{[-1]}\left(\psi(u)+\psi(v)\right)&=\left(\psi(u)+\psi(v)+1\right)^{-1}\\&=\left(u^{-1}-1+v^{-1}-1+1\right)^{-1}\\&=\left(u^{-1}+v^{-1}-1\right)^{-1}\end{aligned}$$

as required. [1]

[Total 4]

(iii) ***Coefficients of lower and upper tail dependence***

The coefficient of lower tail dependence is:

$$\lambda_L=\lim_{u\to 0^+}\frac{C[u,u]}{u}$$ [1]

So, for this copula:

$$\lambda_L=\lim_{u\to 0^+}\frac{\left(2u^{-1}-1\right)^{-1}}{u}=\lim_{u\to 0^+}\frac{1}{u\left(2u^{-1}-1\right)}=\lim_{u\to 0^+}\frac{1}{2-u}=\frac{1}{2}$$ [1]

The coefficient of upper tail dependence is:

$$\lambda_U=\lim_{u\to 1^-}\frac{1-2u+C[u,u]}{1-u}$$ [1]

So for this copula:

$$\lambda_U=\lim_{u\to 1^-}\frac{1-2u+\left(2u^{-1}-1\right)^{-1}}{1-u}$$

The limit in this fraction has the form $\frac{0}{0}$*, which is undefined. However, we can use L'Hôpital's rule given,* $\lim_{x\to a}\frac{f(x)}{g(x)}=\lim_{x\to a}\frac{f'(x)}{g'(x)}$ *, to determine the limit.*

$$\lambda_U=\lim_{u\to 1^-}\frac{-2+(-1)\left(-2u^{-2}\right)\left(2u^{-1}-1\right)^{-2}}{-1}=\frac{-2+(-1)(-2)(2-1)^{-2}}{-1}=0$$ [2]

[Total 5]

Alternatively, we could say:

$$\lambda_U = \lim_{u\to 1^-} \frac{(1-2u)\left(2u^{-1}-1\right)+1}{(1-u)\left(2u^{-1}-1\right)}$$

$$= \lim_{u\to 1^-} \frac{2u^{-1}-1-4+2u+1}{2u^{-1}-1-2+u}$$

$$= \lim_{u\to 1^-} \frac{2u^{-1}-4+2u}{2u^{-1}-3+u}$$

$$= \lim_{u\to 1^-} \frac{2-4u+2u^2}{2-3u+u^2}$$

$$= \lim_{u\to 1^-} \frac{2(u-1)^2}{(u-1)(u-2)}$$

$$= \lim_{u\to 1^-} \frac{2(u-1)}{u-2}=0$$

[2]

As another alternative, we could set $x=1-u$. *Then:*

$$\lambda_U = \lim_{x\to 0^+} \frac{1-2(1-x)+\left(2(1-x)^{-1}-1\right)^{-1}}{x} = \lim_{x\to 0^+} \frac{-1+2x+\left(\dfrac{2}{1-x}-1\right)^{-1}}{x}$$

and:

$$\left(\frac{2}{1-x}-1\right)^{-1} = \left(\frac{2-(1-x)}{1-x}\right)^{-1} = \left(\frac{1+x}{1-x}\right)^{-1} = \frac{1-x}{1+x}$$

So:

$$\lambda_U = \lim_{x\to 0^+} \frac{-1+2x+\dfrac{1-x}{1+x}}{x}$$

$$= \lim_{x\to 0^+} \left(2-\frac{1}{x}+\frac{1-x}{x(1+x)}\right)$$

$$= \lim_{x\to 0^+} \left(2+\frac{-(1+x)+1-x}{x(1+x)}\right)$$

$$= \lim_{x\to 0^+} \left(2-\frac{2x}{x(1-x)}\right)$$

$$= \lim_{x\to 0^+} \left(2-\frac{2}{1-x}\right)$$

$$=2-2=0$$

[2]

Solution X5.7

Reinsurance is covered in Chapter 18.

(i)(a) ***Probability claim involves the reinsurer***

Let X denote the individual claim amount random variable. Then $X \sim Pa(4,7500)$. Using the CDF of the Pareto distribution from page 14 of the *Tables*, we have:

$$P(X > 3{,}000) = 1 - F(3{,}000) = \left(\frac{7{,}500}{7{,}500 + 3{,}000}\right)^4 = 0.260308 \qquad [1]$$

(i)(b) ***Insurer's expected payment per claim***

Let Y denote the amount of a claim paid by the insurer. Then:

$$Y = \begin{cases} X & \text{if } X \le 3{,}000 \\ 3{,}000 & \text{if } X > 3{,}000 \end{cases}$$

So:

$$E(Y) = \int_0^{3{,}000} x f(x)\,dx + \int_{3{,}000}^{\infty} 3{,}000 f(x)\,dx$$

$$= \int_0^{3{,}000} x \frac{4 \times 7{,}500^4}{(7{,}500 + x)^5}\,dx + 3{,}000 P(X > 3{,}000) \qquad [1]$$

Either of the following two methods could be used to evaluate $E(Y)$.

Method 1

We can evaluate the integral using integration by parts and use the value of $P(X > 3{,}000)$ from part (i)(a). This gives:

$$E(Y) = \left[-x \frac{7{,}500^4}{(7{,}500 + x)^4}\right]_0^{3{,}000} + \int_0^{3{,}000} \frac{7{,}500^4}{(7{,}500 + x)^4}\,dx + (3{,}000 \times 0.260308) \qquad [1]$$

$$= -3{,}000 \frac{7{,}500^4}{10{,}500^4} + \left[-\frac{7{,}500^4}{3(7{,}500 + x)^3}\right]_0^{3{,}000} + 780.925$$

$$= \cancel{-780.925} + \left[-\frac{7{,}500^4}{3 \times 10{,}500^3} + \frac{7{,}500^4}{3 \times 7{,}500^3}\right] + \cancel{780.925} \qquad [1]$$

$$= -911.079 + 2{,}500$$

$$= £1{,}588.92 \qquad [1]$$

Method 2

We can evaluate the integral by making the substitution $u = 7{,}500 + x$. This gives:

$$E(Y) = \int_{7{,}500}^{10{,}500} (u - 7{,}500)\frac{4 \times 7{,}500^4}{u^5}\,du + (3{,}000 \times 0.260308) \qquad [1]$$

$$= \int_{7{,}500}^{10{,}500} \frac{4 \times 7{,}500^4}{u^4} - \frac{4 \times 7{,}500^5}{u^5}\,du + 780.925$$

$$= \left[-\frac{4 \times 7{,}500^4}{3u^3} + \frac{7{,}500^5}{u^4}\right]_{7{,}500}^{10{,}500} + 780.925 \qquad [1]$$

$$= \left[-1{,}692.003 - (-2{,}500)\right] + 780.925 = £1{,}588.92 \qquad [1]$$

[Total 5]

Alternatively, if we let Z denote the amount paid by the reinsurer on a claim, then we could calculate $E(Y)$ using the formula $E(X) - E(Z)$.

Since $X \sim Pa(4, 7500)$:

$$E(X) = \frac{7{,}500}{4-1} = 2{,}500 \qquad [½]$$

Also:

$$Z = \begin{cases} 0 & X \le 3{,}000 \\ X - 3{,}000 & X > 3{,}000 \end{cases}$$

So:

$$E(Z) = \int_{3{,}000}^{\infty} (x - 3{,}000)\frac{4 \times 7{,}500^4}{(7{,}500 + x)^5}\,dx \qquad [1]$$

Using the substitution $u = x - 3{,}000$ gives:

$$E(Z) = \int_0^{\infty} u \frac{4 \times 7{,}500^4}{(10{,}500 + u)^5}\,du \qquad [½]$$

We can transform the integrand into u multiplied by the PDF of the $Pa(4, 10\,500)$ distribution as follows:

$$E(Z) = \frac{7{,}500^4}{10{,}500^4}\int_0^{\infty} u \frac{4 \times (10{,}500)^4}{(10{,}500 + u)^5}\,du \qquad [1]$$

The integral part of the expression above is the mean of the $Pa(4,10\,500)$ *distribution, which is* $\dfrac{10,500}{4-1}=3,500$. *So:*

$$E(Z)=\frac{7,500^4}{10,500^4}\times 3,500=911.08 \qquad [½]$$

and hence:

$$E(Y)=2,500-911.08=£1,588.92 \qquad [½]$$

(ii)(a) ***Probability claim next year involves the reinsurer***

The claim amount random variable for next year is $1.1X$. Using the CDF of the Pareto distribution from page 14 of the *Tables*:

$$P(1.1X>3,000)=P\left(X>\tfrac{3,000}{1.1}\right)=1-F\left(\tfrac{3,000}{1.1}\right)=\left(\frac{7,500}{7,500+\frac{3,000}{1.1}}\right)^4=0.28920 \qquad [1]$$

(ii)(b) ***Effect on insurer's mean claim payment***

The average claim amount retained by the insurance company will increase by less than 10%. [½]

This is because the retention limit is unchanged, *ie* the insurer still pays a maximum amount of £3,000 in respect of each claim. [½]

The amounts that the insurer has to pay out on small claims (that were less than $£3,000/1.1$) will increase by 10%. However, the amount paid on claims that were already more than £3,000 will not change at all, and the amounts paid on claims that were between $£3,000/1.1$ and £3,000 will increase by less than 10%. [1]

(ii)(c) ***Reinsurer's expected payment on claims in which it is involved***

Let Z' denote the reinsurer's claim payment random variable next year. Then:

$$Z'=\begin{cases}0 & \text{if } 1.1X\le 3,000 \quad \left(ie \text{ if } X\le\frac{3,000}{1.1}\right)\\ 1.1X-3,000 & \text{if } 1.1X>3,000 \quad \left(ie \text{ if } X>\frac{3,000}{1.1}\right)\end{cases} \qquad [1]$$

So:

$$E(Z')=\int_{\frac{3,000}{1.1}}^{\infty}(1.1x-3,000)f(x)\,dx=\int_{\frac{3,000}{1.1}}^{\infty}(1.1x-3,000)\frac{4\times 7,500^4}{(7,500+x)^5}\,dx \qquad [1]$$

Then any one of the following three methods could be used to evaluate the integral.

Method 1

Using integration by parts, we get:

$$E(Z')=\left[-(1.1x-3{,}000)\frac{7{,}500^4}{(7{,}500+x)^4}\right]_{\frac{3{,}000}{1.1}}^{\infty}+\int_{\frac{3{,}000}{1.1}}^{\infty}1.1\frac{7{,}500^4}{(7{,}500+x)^4}dx \qquad [1]$$

The first term on the RHS is 0. So:

$$E(Z')=\left[-1.1\frac{7{,}500^4}{3(7{,}500+x)^3}\right]_{\frac{3{,}000}{1.1}}^{\infty}=0-\left(-1.1\frac{7{,}500^4}{3\left(7{,}500+\frac{3{,}000}{1.1}\right)^3}\right)=£1{,}084.52 \qquad [1]$$

Method 2

We can use the substitution $u=7{,}500+x$ to obtain:

$$E(Z')=\int_{\frac{3{,}000}{1.1}+7{,}500}^{\infty}(1.1u-11{,}250)\frac{4\times 7{,}500^4}{u^5}du$$

$$=\int_{\frac{3{,}000}{1.1}+7{,}500}^{\infty}\frac{1.1\times 4\times 7{,}500^4}{u^4}-11{,}250\frac{4\times 7{,}500^4}{u^5}du \qquad [1]$$

$$=\left[-\frac{1.1\times 4\times 7{,}500^4}{3u^3}+11{,}250\frac{7{,}500^4}{u^4}\right]_{\frac{3{,}000}{1.1}+7{,}500}^{\infty}$$

$$=\left[0-(-1{,}084.52)\right]=£1{,}084.52 \qquad [1]$$

Method 3

We can use the substitution $u=1.1x-3{,}000$ to obtain:

$$E(Z')=\int_0^{\infty}\frac{u}{1.1}\frac{4\times 7{,}500^4}{\left(7{,}500+\frac{u+3{,}000}{1.1}\right)^5}du \qquad [½]$$

Multiplying the top and bottom of the integrand by 1.1^5 gives:

$$E(Z')=\int_0^{\infty}u\frac{4\times(8{,}250)^4}{(11{,}250+u)^5}du \qquad [½]$$

We can transform the integrand into u multiplied by the PDF of the $Pa(4,11250)$ distribution as follows:

$$E(Z')=\frac{8,250^4}{11,250^4}\int_0^\infty u\frac{4\times(11,250)^4}{(11,250+u)^5}du \qquad [½]$$

The integral part of the expression above is now the mean of the $Pa(4,11250)$ distribution, which is $\frac{11,250}{4-1}=3,750$. So:

$$E(Z')=\frac{8,250^4}{11,250^4}\times 3,750=1,084.52 \qquad [½]$$

$E(Z')$ is the mean amount paid by the reinsurer on *all* claims.

We want the mean amount paid by the reinsurer on claims in which it is involved. This is the reinsurer's *conditional* mean:

$$E\left(Z'\,|\,X>\tfrac{3,000}{1.1}\right)=\frac{E(Z')}{P\left(X>\tfrac{3,000}{1.1}\right)}=\frac{1,084.52}{0.28920}=3,750 \qquad [1]$$

[Total 8]

Alternatively, we could use the following result from Chapter 18 relating to a Pareto distribution:

If $X\sim Pa(\alpha,\lambda)$, then $X'=kX\sim Pa(\alpha,k\lambda)$. *[1]*

With $\alpha=4$, $\lambda=7,500$ and $k=1.1$, we have $X'\sim Pa(4,8250)$. *[1]*

Also, if $X\sim Pa(\alpha,\lambda)$ and there is a retention level of M, then the reinsurer's conditional distribution is $Z\,|\,Z>0\sim Pa(\alpha,\lambda+M)$. *[1]*

With $X'\sim Pa(4,8250)$ and a retention level of 3,000, we have $Z'\,|\,Z'>0\sim Pa(4,11250)$. *[1]*

Hence the mean amount paid by the reinsurer on claims in which it is involved is:

$$E\left(Z'\,|\,Z'>0\right)=\frac{11,250}{3}=3,750 \qquad [1]$$

Solution X5.8

This question involves parameter estimation, which is covered in Chapter 15 and reinsurance, which is covered in Chapter 18.

(i)(a) ***Maximum likelihood estimate***

The likelihood of observing the 7 known claims $(x_1,\dots,x_7)$ and the 3 unknown claims greater than £40,000 is:

$$L(\lambda)=f(x_1)\times\cdots\times f(x_7)\times P(X>40{,}000)^3 \qquad [½]$$

$$=\lambda e^{-\lambda x_1}\times\cdots\times\lambda e^{-\lambda x_7}\times[1-F(40{,}000)]^3$$

$$=\lambda^7 e^{-\lambda\sum x_i}\left[e^{-40{,}000\lambda}\right]^3 \qquad [½]$$

$$=\lambda^7 e^{-152{,}749\lambda}e^{-120{,}000\lambda} \qquad [½]$$

$$=\lambda^7 e^{-272{,}749\lambda} \qquad [½]$$

The log-likelihood is:

$$\ln L(\lambda)=7\ln\lambda-272{,}749\lambda \qquad [½]$$

Differentiating with respect to λ:

$$\frac{d}{d\lambda}\ln L(\lambda)=\frac{7}{\lambda}-272{,}749 \qquad [½]$$

This derivative is equal to 0 when:

$$\lambda=\frac{7}{272{,}749}=0.0000257 \qquad [½]$$

Checking that we have a maximum:

$$\frac{d^2}{d\lambda^2}\ln L(\lambda)=-\frac{7}{\lambda^2}<0 \quad\Rightarrow\quad \text{max} \qquad [½]$$

So the maximum likelihood estimate of λ is 0.0000257.

(i)(b) ***Method of percentiles estimate***

The sample median of the 10 claims is the $½(10+1)=5½$ th value. By interpolation, this is:

$$½(28{,}506+36{,}834)=32{,}670 \qquad [1]$$

The median of the distribution is the value of m such that:

$$F(m)=P(X<m)=1-e^{-\lambda m}=½ \qquad [1]$$

Equating the distribution median to the sample median:

$$1-e^{-32,670\lambda}=½ \quad\Rightarrow\quad \lambda=-\frac{\ln ½}{32,670}=0.0000212 \qquad [1]$$

So the method of percentiles estimate of λ is 0.0000212.

[Total 7]

(ii)(a) ***Conditional distribution***

Let Y be the amount paid by the insurer on a claim. Since the insurer only makes a payment if a claim is greater than the excess:

$$Y=X-50,000 \mid X>50,000 \qquad [½]$$

The PDF of Y is given by:

$$f_Y(y)=\frac{f_X(x)}{P(X>50,000)}, \qquad x>50,000 \qquad [½]$$

ie:

$$f_Y(y)=\frac{f_X(y+50,000)}{P(X>50,000)}, \qquad y>0 \qquad [½]$$

Alternatively, we can obtain the PDF of Y by differentiating the CDF. For $y>0$:

$$\begin{aligned}F_Y(y)=P(Y\le y)&=P(X-50,000\le y\mid X>50,000)\\&=\frac{P(X-50,000\le y \text{ and } X>50,000)}{P(X>50,000)}\\&=\frac{P(50,000<X\le y+50,000)}{P(X>50,000)}\\&=\frac{F_X(y+50,000)-F_X(50,000)}{P(X>50,000)}\end{aligned}$$

Differentiating then gives:

$$f_Y(y)=\frac{f_X(y+50,000)}{P(X>50,000)}$$

Now:

$$P(X>50,000)=1-F(50,000)=\left(\frac{200,000}{200,000+50,000}\right)^{\theta} \qquad [½]$$

and:

$$f_X(y+50,000)=\frac{\theta\times 200,000^{\theta}}{(200,000+y+50,000)^{\theta+1}}=\frac{\theta\times 200,000^{\theta}}{(250,000+y)^{\theta+1}} \qquad [½]$$

So:

$$f_Y(y) = \frac{\theta \times 250{,}000^{\theta}}{(250{,}000+y)^{\theta+1}}, \qquad y>0$$ [½]

This is the PDF of the Pareto distribution with parameters θ and 250,000.

(ii)(b) ***Maximum likelihood estimate***

The likelihood function of observing the sample data is:

$$\begin{aligned}L(\theta) &= f_Y(y_1)\times\cdots\times f_Y(y_5)\\ &= \frac{\theta\times 250{,}000^{\theta}}{(250{,}000+y_1)^{\theta+1}}\times\cdots\times\frac{\theta\times 250{,}000^{\theta}}{(250{,}000+y_5)^{\theta+1}}\\ &= \frac{\theta^5\times 250{,}000^{5\theta}}{\prod(250{,}000+y_i)^{(\theta+1)}}\end{aligned}$$ [1]

and the log-likelihood is:

$$\begin{aligned}\ln L(\theta) &= 5\ln\theta + 5\theta\ln 250{,}000 - (\theta+1)\ln\left[\prod(250{,}000+y_i)\right]\\ &= 5\ln\theta + 5\theta\ln 250{,}000 - (\theta+1)\sum\ln(250{,}000+y_i)\end{aligned}$$ [½]

Differentiating with respect to θ:

$$\frac{d}{d\theta}\ln L(\theta) = \frac{5}{\theta} + 5\ln 250{,}000 - \sum\ln(250{,}000+y_i)$$ [½]

This is equal to 0 when:

$$\frac{5}{\theta} + 5\ln 250{,}000 = \sum\ln(250{,}000+y_i)$$

ie when:

$$\theta = \frac{5}{\sum\ln(250{,}000+y_i) - 5\ln 250{,}000}$$ [½]

From the given data:

$$\sum\ln(250{,}000+y_i) = 64.370$$ [½]

So $\frac{d}{d\theta}\ln L(\theta) = 0$ when:

$$\theta = \frac{5}{64.370 - 5\times 12.4292} = 2.25$$ [½]

Now check that this does in fact give a maximum:

$$\frac{d^2}{d\theta^2}\ln L(\theta) = -\frac{5}{\theta^2} < 0 \quad \Rightarrow \quad \text{max}$$ [½]

So $\hat{\theta}$, the maximum likelihood estimate of θ, is 2.25.

[Total 7]

Solution X5.9

The Naïve Bayes approach to classification is covered in Chapter 21.

(i) ***Formula***

Using the definition of conditional probabilities, *ie* $P(X|Y) = \dfrac{P(X,Y)}{P(Y)}$, we have:

$$P(\text{Text } i \text{ is in English}|A_i,\ldots,\Omega_i) = \frac{P(\text{Text } i \text{ is in English, } A_i,\ldots,\Omega_i)}{P(A_i,\ldots,\Omega_i)} \qquad [½]$$

Using the same definition in reverse, *ie* $P(X,Y) = P(X|Y)P(Y)$, we can write the numerator as:

$$P(\text{Text } i \text{ is in English, } A_i,\ldots,\Omega_i) = P(A_i,\ldots,\Omega_i\,|\text{Text } i \text{ is in English})\ P(\text{Text } i \text{ is in English}) \qquad [½]$$

Using the law of total probability, we can write the denominator as:

$$P(A_i,\ldots,\Omega_i) = \sum_k P(A_i,\ldots,\Omega_i\,|\text{Text } i \text{ is in Language } k)\ P(\text{Text } i \text{ is in Language } k) \qquad [½]$$

So:

$$P(\text{Text } i \text{ is in English}|A_i,G_i,H_i,I_i,N_i,O_i,T_i,U_i,\Omega_i)$$

$$= \frac{P(A_i,\ldots,\Omega_i\,|\text{Text } i \text{ is in English})\ P(\text{Text } i \text{ is in English})}{\sum_k P(A_i,\ldots,\Omega_i\,|\text{Text } i \text{ is in Language } k)\ P(\text{Text } i \text{ is in Language } k)} \qquad [½]$$

[Total 2]

(ii)(a) ***Test message***

The test message '**BONJOUR MONSIEUR DUPONT**' contains 21 letters with the counts shown in the table below.

Letter	A	G	H	I	N	O	T	U	Other
Count	$A_i = 0$	$G_i = 0$	$H_i = 0$	$I_i = 1$	$N_i = 3$	$O_i = 4$	$T_i = 1$	$U_i = 3$	$\Omega_i = 9$

[1]

Since we are assuming that the prior probabilities, *ie* $P(\text{Text } i \text{ is in English})$ *etc*, are all equal here, these will cancel out in the equation we derived in part (i), so that:

$$P(\text{Text } i \text{ is in English}|A_i,G_i,H_i,I_i,N_i,O_i,T_i,U_i,\Omega_i)$$

$$= \frac{P(A_i,\ldots,\Omega_i\,|\text{Text } i \text{ is in English})}{\sum_k P(A_i,\ldots,\Omega_i\,|\text{Text } i \text{ is in Language } k)}$$

The naïve Bayes approach assumes that the events $A_i = 0,\ G_i = 0,\ \ldots,\ \Omega_i = 9$ are independent. So we can now calculate (using slightly abbreviated notation and the letter frequencies given for each language):

$$P(A_i = 0,\ldots,\Omega_i = 9 \,|\, \text{English}) = 0.07^1 \times 0.07^3 \times 0.07^4 \times 0.09^1 \times 0.03^3 \times 0.51^9 C$$

$$= 3.27 \times 10^{-18} C \qquad [½]$$

where C is the multinomial coefficient $\dfrac{21!}{3! \times 4! \times 3! \times 9!}$ that arises from repeated letters.

We can actually ignore the constant C as it will cancel out when we calculate the posterior probabilities.

Similarly:

$$P(A_i = 0,\ldots,\Omega_i = 9 \,|\, \text{French}) = 0.08^1 \times 0.07^3 \times 0.08^4 \times 0.07^1 \times 0.06^3 \times 0.54^9 C$$

$$= 6.63 \times 10^{-17} C \qquad [½]$$

$$P(A_i = 0,\ldots,\Omega_i = 9 \,|\, \text{German}) = 0.07^1 \times 0.10^3 \times 0.07^4 \times 0.06^1 \times 0.04^3 \times 0.51^9 C$$

$$= 1.51 \times 10^{-17} C \qquad [½]$$

$$P(A_i = 0,\ldots,\Omega_i = 9 \,|\, \text{Spanish}) = 0.06^1 \times 0.07^3 \times 0.06^4 \times 0.05^1 \times 0.03^3 \times 0.58^9 C$$

$$= 2.67 \times 10^{-18} C \qquad [½]$$

$$P(A_i = 0,\ldots,\Omega_i = 9 \,|\, \text{Italian}) = 0.10^1 \times 0.07^3 \times 0.10^4 \times 0.06^1 \times 0.03^3 \times 0.49^9 C$$

$$= 9.05 \times 10^{-18} C \qquad [½]$$

Markers: If the prior probabilities are included, all these probabilities (and the total below) will be divided by 5. Please award full marks to students who include prior probabilities (correctly) in their calculations.

The sum of the probabilities above is $9.64 \times 10^{-17} C$. So the posterior probabilities are:

$$P(\text{English} \,|\, A_i = 0,\ldots,\Omega_i = 9) = \frac{3.27 \times 10^{-18}}{9.64 \times 10^{-17}} = 3\%$$

$$P(\text{French} \,|\, A_i = 0,\ldots,\Omega_i = 9) = \frac{6.63 \times 10^{-17}}{9.64 \times 10^{-17}} = 69\%$$

$$P(\text{German} \,|\, A_i = 0,\ldots,\Omega_i = 9) = \frac{1.51 \times 10^{-17}}{9.64 \times 10^{-17}} = 16\%$$

$$P(\text{Spanish} \mid A_j = 0,\ldots,\Omega_j = 9) = \frac{2.67\times 10^{-18}}{9.64\times 10^{-17}} = 3\%$$

$$P(\text{Italian} \mid A_j = 0,\ldots,\Omega_j = 9) = \frac{9.05\times 10^{-18}}{9.64\times 10^{-17}} = 9\%$$ [1]

(ii)(b) ***Comment***

The highest probability (69%) corresponds to French. So we would conclude that the test message is most likely to be French. [1]

We can see that the test message is indeed in French. So the model has identified the language correctly in this case. [½]

[Total 6]

(iii)(a) ***Probabilities for the fragment***

If we remove the **?**'s, the fragment says '**TSGWLHUSMOEEE**', which contains 13 letters with the counts shown in the table below. (We've called this fragment Text j.)

Letter	A	G	H	I	N	O	T	U	Other
Count	$A_j = 0$	$G_j = 1$	$H_j = 1$	$I_j = 0$	$N_j = 0$	$O_j = 1$	$T_j = 1$	$U_j = 1$	$\Omega_j = 8$

[1]

We can then calculate:

$$P(A_j = 0,\ldots,\Omega_j = 8 \mid \text{English}) = 0.02^1 \times 0.06^1 \times 0.07^1 \times 0.09^1 \times 0.03^1 \times 0.51^8 K$$

$$= 1.04\times 10^{-9} K$$ [½]

where K is the multinomial coefficient $\frac{13!}{8!}$. Similarly:

$$P(A_j = 0,\ldots,\Omega_j = 8 \mid \text{French}) = 0.01^1 \times 0.01^1 \times 0.08^1 \times 0.07^1 \times 0.06^1 \times 0.54^8 K$$

$$= 2.43\times 10^{-10} K$$ [½]

$$P(A_j = 0,\ldots,\Omega_j = 8 \mid \text{German}) = 0.03^1 \times 0.05^1 \times 0.07^1 \times 0.06^1 \times 0.04^1 \times 0.51^8 K$$

$$= 1.15\times 10^{-9} K$$ [½]

$$P(A_j = 0,\ldots,\Omega_j = 8 \mid \text{Spanish}) = 0.02^1 \times 0.01^1 \times 0.06^1 \times 0.05^1 \times 0.03^1 \times 0.58^8 K$$

$$= 2.31\times 10^{-10} K$$ [½]

$$P(A_j=0,\ldots,\Omega_j=8\,|\,\text{Italian})=0.02^1\times 0.01^1\times 0.10^1\times 0.06^1\times 0.03^1\times 0.49^8 K$$

$$=1.20\times 10^{-10} K$$ [½]

Applying the prior probabilities gives:

$$P(A_j=0,\ldots,\Omega_j=8\,|\,\text{English})\ P(\text{English})=1.04\times 10^{-9} K\times 0.40=4.15\times 10^{-10} K$$

$$P(A_j=0,\ldots,\Omega_j=8\,|\,\text{French})\ P(\text{French})=2.43\times 10^{-10} K\times 0.20=4.86\times 10^{-11} K$$

$$P(A_j=0,\ldots,\Omega_j=8\,|\,\text{German})\ P(\text{German})=1.15\times 10^{-9} K\times 0.20=2.31\times 10^{-10} K$$

$$P(A_j=0,\ldots,\Omega_j=8\,|\,\text{Spanish})\ P(\text{Spanish})=2.31\times 10^{-10} K\times 0.10=2.31\times 10^{-11} K$$

$$P(A_j=0,\ldots,\Omega_j=8\,|\,\text{Italian})\ P(\text{Italian})=1.20\times 10^{-10} K\times 0.10=1.20\times 10^{-11} K$$ [1]

The sum of these is $7.29\times 10^{-10} K$. So the posterior probabilities are:

$$P(\text{English}\,|\,A_j=0,\ldots,\Omega_j=8)=\frac{4.15\times 10^{-10}}{7.29\times 10^{-10}}=57\%$$

$$P(\text{French}\,|\,A_j=0,\ldots,\Omega_j=8)=\frac{4.86\times 10^{-11}}{7.29\times 10^{-10}}=7\%$$

$$P(\text{German}\,|\,A_j=0,\ldots,\Omega_j=8)=\frac{2.31\times 10^{-10}}{7.29\times 10^{-10}}=32\%$$

$$P(\text{Spanish}\,|\,A_j=0,\ldots,\Omega_j=8)=\frac{2.31\times 10^{-11}}{7.29\times 10^{-10}}=3\%$$

$$P(\text{Italian}\,|\,A_j=0,\ldots,\Omega_j=8)=\frac{1.20\times 10^{-11}}{7.29\times 10^{-10}}=2\%$$ [1]

(iii)(b) *Conclusions*

The model predicts that the fragment is most likely to be written in English (with probability 57%). [1]

German is also a possibility (with probability 32%), but the other three languages are unlikely. [½]
[Total 7]

In fact the original message in the fragment was:

THISMESSAGEISWRITTENINENGLISHBUTSOMEOFTHELETTERSAREILLEGIBLE

To disguise the message, we chose 70% of the letters at random and replaced them with ?'s.

Solution X5.10

The individual risk model is discussed in Chapter 20.

(i) ***Formula and assumptions***

The formula for the total claims from the portfolio is:

$$S = X_1 + X_2 + \cdots + X_n$$

where X_i is the claim amount from the i th member (which may be zero). [½]

The assumptions underlying this model are:

- there are a fixed number of risks (*ie* members), n [½]
- claims occur independently for each member [½]
- the number of claims for each member is either 0 or 1. [½]

[Total 2]

(ii) ***Mean and variance***

Let $X = bI$, where I is an indicator random variable denoting whether or not a claim is paid, *ie* $P(I=0)=1-q$, $P(I=1)=q$ and b is the fixed benefit amount.

Then $I \sim Bin(1,q)$, so:

$$E(I) = q \quad \text{and} \quad \text{var}(I) = q(1-q)$$ [1]

Since b is a constant:

$$E(X) = E(bI) = bE(I) = bq$$ [1]

and:

$$\text{var}(X) = \text{var}(bI) = b^2\text{var}(I) = b^2 q(1-q)$$ [1]

[Total 3]

Alternatively, we could use the conditional expectation formula from page 16 of the Tables. Since:

$$E(X \mid I) = \begin{cases} 0 & \text{if } I=0 \\ b & \text{if } I=1 \end{cases}$$ *[½]*

it follows that:

$$\begin{aligned} E(X) &= E[E(X \mid I)] \\ &= E(X \mid I=0)P(I=0) + E(X \mid I=1)P(I=1) \\ &= 0 \times (1-q) + bq = bq \end{aligned}$$ *[½]*

We can derive the formula for $\text{var}(X)$ *using the conditional variance formula:*

$$\text{var}(X) = E[\text{var}(X \mid I)] + \text{var}[E(X \mid I)]$$

This is also give on page 16 of the Tables.

Since the claim amount is fixed:

$$\text{var}(X \mid I) = 0 \qquad [½]$$

So:

$$E[\text{var}(X \mid I)] = 0$$

$$\begin{aligned}\text{var}[E(X \mid I)] &= E\left[(E(X \mid I))^2\right] - \left[E[E(X \mid I)]\right]^2 \\ &= 0^2 \times P(I=0) + b^2 \times P(I=1) - (bq)^2 \\ &= b^2q - b^2q^2 \\ &= b^2q(1-q)\end{aligned} \qquad [1]$$

and hence:

$$\text{var}(X) = 0 + b^2q(1-q) \qquad [½]$$

Another alternative is to use the fact that X *is a compound binomial random variable. The number of claims,* N*, has a* $Binomial(1,q)$ *distribution and the individual claim amount is the constant,* b*.* [1]

Then using the formulae for the mean and variance of a compound random variable from page 16 of the Tables:

$$E(X) = E(N)E(b) = bq \qquad [1]$$

$$\begin{aligned}\text{var}(X) &= E(N)\,\text{var}(b) + \text{var}(N)[E(b)]^2 \\ &= q \times 0 + q(1-q) \times b^2 \\ &= b^2q(1-q)\end{aligned} \qquad [1]$$

(iii) ***Skewness***

We have:

$$\text{skew}(I) = E[(I - E(I))^3]$$

$$= E\left[I^3 - 3I^2E(I) + 2E^3(I)\right]$$

$$= E(I^3) - 3E(I)E(I^2) + 2E^3(I)$$

$$= q - 3q^2 + 2q^3$$

$$= q(1-q)(1-2q)$$ [1]

So:

$$\text{skew}(X) = \text{skew}(bI) = b^3\text{skew}(I) = b^3q(1-q)(1-2q)$$ [1]

[Total 2]

Alternatively, we could use the fact that X is a compound binomial random variable and consider its CGF. This is a rather long-winded approach, however. Using the notation Y to denote the individual claim amount random variable, the MGF of X is:

$$M_X(t) = M_N(\ln M_Y(t))$$

This formula is given on page 16 of the Tables.

Now, since $N \sim Binomial(1,q)$:

$$M_N(t) = (1-q) + qe^t$$

In addition:

$$M_Y(t) = E(e^{tY}) = e^{tb}$$

So the MGF of X is:

$$M_X(t) = (1-q) + qe^{\ln M_Y(t)} = (1-q) + qM_Y(t) = (1-q) + qe^{tb}$$ *[½]*

and its CGF is:

$$C_X(t) = \ln M_X(t) = \ln[(1-q) + qe^{bt}]$$

The skewness is equal to the third derivative of the CGF evaluated at the point $t = 0$:

$$C'_X(t) = \frac{qbe^{bt}}{[(1-q)+qe^{bt}]}$$

$$C''_X(t) = \frac{[(1-q)+qe^{bt}]qb^2e^{bt} - qbe^{bt}qbe^{bt}}{[(1-q)+qe^{bt}]^2} = \frac{q(1-q)b^2e^{bt}}{[(1-q)+qe^{bt}]^2}$$

$$C'''_X(t) = \frac{[(1-q)+qe^{bt}]^2q(1-q)b^3e^{bt} - q(1-q)b^2e^{bt}2[(1-q)+qe^{bt}]qbe^{bt}}{[(1-q)+qe^{bt}]^4}$$

$$= \frac{q(1-q)(1-2q)b^3e^{bt}}{[(1-q)+qe^{bt}]^4}$$ *[1]*

So:

$$skew(X) = C'''_X(0) = q(1-q)(1-2q)b^3$$ *[½]*

(iv) ***Mean, variance and skewness of the total claim amount***

If S is the total claim amount, then:

$$E(S) = 1{,}250 \times 50{,}000 \times 0.008 + 250 \times 20{,}000 \times 0.012 = 560{,}000$$ [1]

$$\text{var}(S) = 1{,}250 \times 50{,}000^2 \times 0.008 \times 0.992 + 250 \times 20{,}000^2 \times 0.012 \times 0.988$$

$$= 2.59856 \times 10^{10}$$ [1]

$$skew(S) = 1{,}250 \times 50{,}000^3 \times 0.008 \times 0.992 \times 0.984 + 250 \times 20{,}000^3 \times 0.012 \times 0.988 \times 0.976$$

$$= 1.2433029 \times 10^{15}$$

[1]

Hence, the coefficient of skewness is:

$$\frac{skew(S)}{(\text{var}(S))^{3/2}} = \frac{1.2433029 \times 10^{15}}{\left(2.59856 \times 10^{10}\right)^{3/2}} = 0.297$$ [1]

Markers: Please award follow-through marks for the coefficient of skewness if an incorrect formula for the skewness is derived in part (iii).

[Total 4]

(v) ***Probability***

We now assume that:

$$S \sim N(560\,000,\ 2.59856 \times 10^{10}) \text{ approximately}$$ [½]

So the required probability is:

$$P(S>1{,}000{,}000)\approx P\left(N(0,1)>\frac{1{,}000{,}000-560{,}000}{\sqrt{2.59856\times10^{10}}}\right)$$ [1]

$$=1-P(N(0,1)<2.72952)$$ [½]

$$=0.00317$$ [1]

[Total 3]

Markers: Please award follow-through marks if an incorrect mean and/or variance is calculated in part (iv).

(vi) ***Comment***

A normal distribution gives the most accurate answers if the distribution is symmetrical. The coefficient of skewness calculated in part (iv) shows that the distribution is positively skewed, but not by very much. So the value is probably not that inaccurate. [1]

On the other hand, we are looking at a probability relating to the distribution of values in the upper tail, where a normal distribution is likely to approximate less well than at the centre of the distribution. [1]

[Total 2]