Mount Kenya University

P.O. Box 342-01000
Thika

Email: Info@mku.ac.ke

Web: www.mku.ac.ke

DEPARTMENT OF FINANCE AND ACCOUNTING

COURSE CODE: BBM 312

COURSE TITLE: BUSINESS STATISTICS

Instructional Material for BBM- Distance Learning

COURSE OUTLINE

PRE –REQUISITE: - BBM 112, BBM 212 and BBM 223.

PURPOSE: - To equip students with data collections skills, correlations, analysis hypothesis formulation and testing methods

OBJECTIVES: - By the end of this module, the learner should be able to: -

(a) Compute the correlation coefficient, equation of the regression line and the coefficient of determination.

(b) Test the hypothesis $H_o : P = 0$

(c) Draw the scatter plots and correlation.

(d) Test the difference between two variances or standard deviations.

(e) Test the difference between sample means, and between two proportions using the z – test.

(f) State the null and alternative hypothesis

(g) State the five steps used in hypothesis testing.

(h) Test means when σ is known using the z- test and chi-square.

(i) Test hypothesis, using the sign test, wilcoxon rank sum test and the Kruskal – Wallis test.

WEEK	CONTENT
1	Measure of variation correlation, Pearson product moment coefficient r.
2	Regression analysis. Coefficient of determination. The rank correlation coefficient
3	Sampling theory, types of sampling mean and standard deviation in simple sampling attributes.
4	Large sample testing the significance for a single proportion.
5	Definition used in hypothesis testing, steps in hypothesis testing – traditional method. Z-test for a mean.
6	Z – Test for a proportion chi-square testing. CAT 1 sit in.
7	Test of proportions. Test the difference between sample means using z-test; test the difference between two means for independent samples, using the t-test.

8	Testing hypothesis about proportion hypothesis testing of the difference between proportions. Testing the hypothesis for equity of two variances.
9	Advantages and disadvantages of non parametric methods. Test of hypothesis using the Wilcoxon rank sum test.
10	Test hypothesis using the Kruskal Wallis test
11	Revision: using revision questions
12	Tutorials and revision consultations
13	Examination and consultation
14	Examination and consultation

Course evaluation:

Cat 1 seat in 1 hour examination from week one to six taking 15% of overall marks.

Cat II distance cat to b handed in week 12 taking 15% of overall marks.

Final examination at the end of the semester 70% overall examination, any one of the examinations not done will incomplete results

TABLE OF CONTENTS

CHAPTER 1: CORRELATION AND REGRESSION

Purpose:- to equip the learner with methods of correlation and regression used to determine the relationship between two or more numerical or quantitative variables.

Objectives: After completing this chapter you should be able to compute the following

- ***Pearson's moment correlation coefficient***
- ***The coefficient of determinations***
- ***Equation of the regression line***
- ***Draw scatter graph***
- ***Draw the best of fit line***
- ***Compute the Spearman's rank correlation coefficient***

A Statistician collects information for variables, which describe the situation. A variable is a characteristics or attribute that can assume different values. That is Data are values, measurements or observations that the variables can assume. Variables whose values are determined by chance are called random variable.

A collection of data values forms a data set. Each value in the data set is called a data value or a datum.

1.1 WHAT IS CORRELATION?

In statistics, correlation is a measure of the strength of a linear relationship between two sets of numbers, i.e. if a change in one number is accompanied by a change in the other. The range of the correlation coefficient is from -1 to +1. If there is a strong positive linear relationship between the variables, the value of r will be close +1.

1.2 PEARSON'S PRODUCT MOMENT COEFFICIENT OF CORRELATION

r (PPMCC)

This provides a measure of the strength of association between two variables; one the dependent variable, the other the independent variable, r, can range from +1 i.e. perfect positive correlation where the variables change value in the same direction as each other, to -1 i.e. perfect negative correlation where y decreases linearly as x increases.
There are several possible formulae but practical ones are: -

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}}$$

Where x = independent variable, $\bar{x}$= mean of the independent.

y = dependent variable, $\bar{y}$= mean of the dependent variable.

Or

$$s_{xx} = \sum (x-\bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum (y-\bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$S_{xy} = \sum (x-\bar{x})(y-\bar{y}) = \sum x_i y_i - \frac{\sum x \sum y}{n}$$

The correlation coefficient computed from the sample data measures the strength and direction of a linear relationship between two variables. The symbol for the sample correlation coefficient is r; the symbol for population correlation coefficient is ρ. The correlation coefficient explained is called Pearson product moment correlation coefficient (PPMCC), Named after Statistician Karl Pearson, pioneered the research in this area.

Gives $r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$

Example 1.1: A football team has kept records of hours per month spent training and the goals scored per month and wish to know if there is any correlation between training hours and goals scored.

Hours per month x	12	21	27	16	19	22	33	10
Goals per month y	8	11	13	11	9	15	17	12

Table 1.1

Solution using table 1.1

Sum up the values and calculate the mean as column 1 and 2 table 2

$$\bar{x} = \frac{160}{8} = 20 \qquad \bar{y} = \frac{96}{8} = 12$$

Subtract the respective means from the independent and dependent variable to form column three and four.

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
12	8	-8	-4	32	64	16
21	11	1	-1	-1	1	1
27	13	7	2	14	49	4
16	11	-4	-2	8	16	4
19	9	-1	-3	3	1	9
22	15	2	3	6	4	9
33	17	13	5	65	169	25
10	12	-10	0	0	100	0
160	**96**			**127**	**404**	**68**

Table 1.2

Column five is the product three and four, while column six and seven are squares of the column three and four respectively.

$S_{xy} = 127 \qquad S_{xx} = 404 \qquad S_{yy} = 68$

$$r = \frac{127}{\sqrt{404 \times 68}} = \frac{127}{165 \cdot 75} = +0 \cdot 77$$

The correlation coefficient is +0.77 which indicates a reasonably strong positive linear association between training hours and goals scored.

"Alternative"

Instead of finding the mean first you find the product of both independent and dependent variable, square the variables respectively as table 3

x	y	xy	x^2	y^2
12	8	96	144	64
21	11	231	441	121
27	13	351	729	169
16	11	176	256	121
19	9	171	361	81
22	15	330	484	225
33	17	561	1089	289
10	12	120	100	144
160	**96**	**2036**	**3604**	**1214**

Table1. 3

$$S_{xx} = 3604 - \frac{160 \times 160}{8} = 404$$

$$S_{yy} = 1214 - \frac{96 \times 96}{8} = 62$$

$$S_{xy} = 2036 - \frac{160 \times 96}{8} = 116$$

$$r = \frac{116}{\sqrt{404 \times 62}} = 0 \cdot 7329$$

Example1.2: The following marks have been obtained by a class of students in statistics (out of 100), table 4

x statistics I	45	55	56	58	60	65	68	70	75	80	85
y statistics II	56	50	48	60	62	64	65	70	74	82	90

Table1.4

Solution

Compute the coefficient of correlation for the above data.

Find the mean of both dependent and independent variable. Then subtract each variable respectively.

$$\bar{x} = \frac{717}{11} = 65.\dot{1}\dot{8} = 65\frac{2}{11} \qquad \bar{y} = 65.5\dot{4} = 65\frac{6}{11}$$

Table 1.5, Shows the working of the following working in respective columns

$\sum (x - \bar{x})(y - \bar{y}) = 1401.89$, $\sum (x - \bar{x})^2 = 1413.63$ and $\sum (y - \bar{y})^2 = 1646.47$

$$r = \frac{1401.89}{\sqrt{1413.64 \times 1646.47}} = 0 \cdot 91889 \approx 0 \cdot 92$$

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
45	56	$-20\frac{2}{11}$	$-9\frac{6}{11}$	192.64	407.31	91.12
55	50	$-10\frac{2}{11}$	$-15\frac{6}{11}$	158.28	103.67	241.86
56	48	$-9\frac{2}{11}$	$-17\frac{6}{11}$	161.10	84.31	307.84
58	60	$-7\frac{2}{11}$	$-5\frac{6}{11}$	39.83	51.58	30.75
60	62	$-5\frac{2}{11}$	$-3\frac{6}{11}$	18.37	26.85	12.57
65	64	$-\frac{2}{11}$	$-1\frac{6}{11}$	0.28	0.03	2.39
68	65	$2\frac{9}{11}$	$-\frac{6}{11}$	-1.54	7.94	0.05
70	70	$4\frac{9}{11}$	$4\frac{5}{11}$	21.46	23021	19.84
75	74	$9\frac{9}{11}$	$8\frac{5}{11}$	83.01	96.40	71.48
80	82	$14\frac{9}{11}$	$16\frac{5}{11}$	243.82	219.58	270.75
85	90	$19\frac{9}{11}$	$24\frac{5}{11}$	484.64	392.76	598.02
717	**721**			**1401.89**	**1413.64**	**1646.47**

Table 1.5

Or “alternative “

Instead of finding the mean first you find the product of both independent and dependent variable, square the variables respectively as table 6

x	y	xy	x^2	y^2
45	56	2520	2025	3136
55	50	2750	3025	2500
56	48	2688	3136	2304
58	60	3480	3364	3600
60	62	3720	3600	3844
65	64	4160	4225	4096
68	65	4420	4624	4225
70	70	4900	4900	4900
75	74	5550	5625	5476
80	82	6560	6400	6724
85	90	7650	7225	8100
717	**721**	**48398**	**48149**	**48905**

Table 1.6

$$s_{xx} = 48149 - \frac{717^2}{11} = 1413 \cdot \dot{6}\dot{3}$$

$$s_{yy} = 48905 - \frac{721^2}{11} = 1646 \cdot \dot{7}\dot{2}$$

$$s_{xy} = 48398 = \frac{717 \times 721}{11} = 1401 \cdot \dot{9}\dot{0}\dot{9}$$

$$r = \frac{1401.89}{\sqrt{1413.64 \times 1646.47}} = 0 \cdot 91889 \approx 0 \cdot 92$$

The performance of both statistics papers has a strong positive linear relationship which means the papers are related.

The significance of the correlation coefficient

As stated before the range of the correlation is between -1 and +1. When the value of r is near +1 or -1, there is a strong linear relationship. When the value is near zero, the linear relationship is weak and nonexistent. Since r is computed from the samples, there are two possibilities when r is not: either the value of is high enough to conclude that there is a significant linear relationship between the variables, or the value of r is due to chance.

To make this decision, you use hypothesis testing procedure. The traditional method is:-

- State the hypothesis
- Find the critical value
- Compute the test value
- Make the decision
- Summarize the result

1.3 REGRESSION ANALYSIS

Regression analysis is a statistical technique which can be used for short to medium term forecasting which seeks to establish the line of "best fit" to be observed data. The data could be shown as a scatter diagram and the line of best fit. The purpose of the scatter plot, as indicated above is to determine the nature of the relationship. The possibilities include a positive linear relationship, a negative linear relationship, a curvilinear relationship, or discernible relationship. If the value of the correlation coefficient is significant, the next step is to determine the equation of the regression line which is the data's line of best fit. The purpose of the regression line is to enable the researcher to see the trend and make predictions on the basis of the data.

Example 1.3: the following data was collect for the sales of wedges for a period of 7 years

Year	1	2	3	4	5	6	7
Sales of wedges in thousands	14	17	15	23	18	22	27

Table1.7

As there seven years of readings in table 7, the data are set out as follows

Year x	Sales of wedges y	xy	x^2	y^2
1	14	14	1	196
2	17	34	4	289
3	15	45	9	225
4	23	92	16	529
5	18	90	25	324
6	22	132	36	484
7	27	189	49	729
28	**136**	**596**	**140**	**2776**

Table 1.8

Draw a scatter diagram?

The general form the equation of any straight line on a graph is y = α + βx, where α and β are constant and where α represents the fixed element and β represents the slope of the line.

To find the values of α and β, it is necessary to solve two simultaneous equations known as the normal equations which are

$$\sum y = \alpha n + \beta \sum x$$

$$\sum xy = \alpha \sum x + \beta \sum x^2$$

Where n is the number of pairs of figures?

$$136 = 7\alpha + 28\beta \qquad 596 = 28\alpha + 140\beta$$

$$596 = 28\alpha + 140\beta \qquad 544 = 28\alpha + 112\beta$$

$$28\beta = 52 \qquad \therefore \qquad \beta = \frac{52}{28} = 1 \cdot 86$$

$$\alpha = \frac{136 - 28 \times 1.86}{7} = 12$$

Regression line y = 12 + 1.86x

Sales (in 000's of units) = 12 + 1.86 (no of years).

To use this expression for forecasting, we merely need to insert the number of the year required. For example 8^{th} year sales = 12 + 1.86(8) = 26.88

The slope of the line, β, is called the **regression coefficient.**

To transpose the normal equation so as to be able to find α and β

$$\alpha = \frac{\sum y}{n} - \beta \frac{\sum x}{n} \qquad \beta = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$\alpha = \bar{y} - \beta \bar{x} \qquad \beta = \frac{s_{xy}}{s_{xx}}$$

$$\beta = \frac{7 \times 596 - 28 \times 136}{7 \times 140 - (28)^2} = 1 \cdot 86$$

$$\alpha = \frac{140}{7} - \frac{1 \cdot 86 \times 28}{7} = 12 \cdot 6$$

$$y = 12 \cdot 6 + 1 \cdot 86x$$

Or

There are several methods for finding the equation of the regression line. Two formulae are given here. These formulae use the same values that are used in computing the value of the correlation coefficient. The development of these formulae is beyond the scope of this module.

Formulas for the regression line y = α + βx

$$\alpha = (\frac{\sum y \sum x^2 - (\sum x \sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$\beta = \frac{n(\sum xy) - (\sum x \sum y)}{n(\sum x^2) - (\sum x)^2}$$

Where α is the y intercept and β is the slope of the line.

Example 1.4: The values of the regression line for the data are given by

N =6, $\sum x = 153.8$ $\sum y = 18.7$, $\sum xy = 682.77$ and $\sum x^2 = 5859.26$.

Find the equation of the regression line.

Solution

Substituting in the formulae above you get

$$\alpha = \frac{(18.7)(5859.26) - (153.8)(682.77)}{(6)(5859.26) - (153.8)^2} = 0.396$$

$$\beta = \frac{(6)(682.77 - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$$

hence the equation of the regression line is y = 0.396 + 0.106x

1.4 COEFFICIENT OF DETERMINATION

This measure denoted by r^2 (because it is the square of the correlation coefficient r) calculates what proportion of the variation in the actual values of *y* May be predicted by change in value of *x*.

$$r^2 \text{ is the ratio} = \frac{\text{explained variation}}{\text{total variation}} = \frac{\sum (Y_E - \bar{Y})^2}{\sum (y - \bar{y})^2}$$

Where Y_E = estimate of *y* given by the regression equation for each value of *x*.

Table 8, example 3 is modified using regression *formula obtained* $y_E = 12 + 1.86x$ to table 9

$$\bar{Y} = \frac{136}{7} = 19.43$$

Year *x*	Sales of wedges *y*	Y_E	$Y_E - \bar{Y}$	$(Y_E - \bar{Y})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$
1	14	13.86	-5.57	31.02	-5.43	29.48
2	17	15.72	-3.71	13.76	-2.43	5.90
3	15	17.58	-1.85	3.42	-4.43	19.62

4	23	19.44	0.01	0	3.57	12.74
5	18	21.30	1.87	3.49	-1.43	2.04
6	22	23.16	3.73	13.91	2.57	6.60
7	27	25.00	5.59	31.24	7.57	57.30
	136			**96.84**		**133.68**

Table 1.9

$$r^2 = \frac{96 \cdot 84}{133 \cdot 68} = 0 \cdot 7244 \cong 72\%$$

Alternative formula for r^2 is given

$$r^2 = \frac{(n\sum xY - \sum x \sum Y)^2}{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}$$

$$= \frac{(7 \times 596 - 28 \times 136)^2}{(7(140) - 28^2)(7(2776) - 136^2} = 0 \cdot 7222$$

Pearson Product Moment correlation coefficient = $\sqrt{\text{coefficient of determinant}}$

= 0.8498

Example 1.5

Fit a straight line to the following points

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Table1. 10

Solution

The regression line is obtained by using the table 11 below

						$\sum$
x	0	1	2	3	4	10
y	1	1.8	3.3	4.5	6.3	16.9
xy	0	1.8	6.6	13.5	25.2	47.1
x^2	0	1	4	9	16	30
y^2	1	3.24	10.89	20.25	39.69	

Table 1.11

The gradient (slope) is (the values are obtain from table 14

$$\beta = \frac{5 \times 47.1 - 10 \times 16.9}{5 \times 30 - (10)^2} = \frac{66.5}{50} \cong 1.33$$

The y –intercept (fixed value) is

$$\alpha = \frac{30}{5} - \frac{1.33 \times 10}{5} = 3.34$$

$$y = 3 \cdot 34 + 1 \cdot 33x$$

1.5 THE RANK CORRELATION COEFFICIENT (R)

This provides a measure of the association between two sets of ranked or ordered data R can also vary from +1, perfect positive rank, Correlation to -1, perfect negative rank correlation. This coefficient is also known as the **spearman rank correlation coefficient.**

The formula is as follows

$$\mathrm{R} = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Where d = difference between the pairs of ranked values.

n = Number of pairs of rankings

Example1. 6: A group of 8 mathematics students are tested in statistics and calculus. Their rankings in the tests were: -

Student	A	B	C	D	E	F	G	H
Statistics	2	7	6	1	4	3	5	8
Calculus	3	6	4	2	5	1	8	7

Table 1.12

Find the Spearman's correlation between Statistics and calculus**.**

Solution

The table 13 shows the solution of the above problem by finding the difference and squares in column four and five respectively

student	statistics	calculus	d	d^2
A	2	3	-1	1
B	7	6	1	1
C	6	4	2	4
D	1	2	-1	1
E	4	5	-1	1
F	3	1	2	4
G	5	8	-3	9
H	8	7	1	1
				22

Table1. 13

$$R = 1 - \frac{6 \times 22}{8(64-1)} = 0 \cdot 738$$

As the rank correlation coefficient is +0.74, we are able to say that there is a reasonable agreement between the student's performances in the two types of tests.

1.6 TIED RANKINGS

The table above has the students when the student E and F tied same 3rd position so $\frac{3+4}{2} = 3\frac{1}{2}$

Example1.7: A group of 8 mathematics students are tested in statistics and calculus. Their rankings in the tests were: -

student	calculus	statistics	d^2
A	3	2	1
B	6	7	1
C	4	6	4
D	2	1	2
E	5	$3\frac{1}{2}$	$2\frac{1}{4}$
F	1	$3\frac{1}{2}$	$6\frac{1}{4}$
G	8	5	9
H	7	8	1
			$\mathbf{25\frac{1}{2}}$

Table1.14

A slight adjustment to the formula is necessary if some students obtain the same ranking.

The adjustment is $\frac{t^3 - t}{12}$, where t is number of tied rankings.

And adjusted formula is

$$R = \frac{1 - 6(\sum d^2 + \frac{t^3 - t}{12})}{n(n^2 - 1)}$$

For example assume that students E and F achieved equal marks in statistics and were given joint third place.

$$R = \frac{1 - 6(25\frac{1}{2} + \frac{2^3 - 2}{12})}{8(64 - 1)} = +0 \cdot 69$$

As will be seen, the R-value has moved also from +0.74 to 0.69

1.7 TIME SERIES ANALYSIS

Time series analysis uses some form of mathematical or statistical analysis on past data arranged in a time series, e.g. sales by month for the last ten years. Time series analyses have the advantage of relative simplicity but certain factors need to be considered.

(a) Are the past data representative, for example, do they contain the results of a recession/boom, a major shift of taste etc.
(b) Time series methods are more appropriate where short term forecasts are required. Over the longer period external pressures, internal policy changes make historical data less appropriate.
(c) Time series methods are best suited to relatively stable situations. Where substantial fluctuations are common and or conditions are expected to change, then the time series methods may give relatively poor results.

1.7.1 Time series analysis – moving average

If the forecast for the next month's sales say December was actual sales for November, then the forecasts obtained would fluctuate up and down with every random fluctuation. If he forecast for

next month's sales was the average of sales for several preceding months then, hopefully, the random fluctuations would cancel each other out, i.e. would be smoothed away.

Example1. 8

Month	Actual sales	3 monthly moving average	6 monthly moving average	12 monthly moving averages.
January	450			
February	440			
March	460	450		
April	410	437		
May	380	417		
June	400	397		
July	370	383	423	
August	360	377	410	
September	410	380	397	
October	450	407	388	
November	470	443	395	
December	490	470	410	
January	460		425	424

Table 19

Any month's forecast is the average of the preceding, n, month's actual sales:

For example, 3 monthly moving average forecasts were prepared as follows:

april's forecast = jan sales + feb sales + march sales

3

$$= \frac{450 + 440 + 460}{3} = 450$$

May forecast $= \frac{440 + 460 + 410}{3} = 437$

Similar logic applies for 6 and 12 monthly moving averages.

NB

A moving average can be used as a forecast as shown above but when graphing moving averages it is important to realize that being averages , they must be plotted at the midpoint of the period to which they relate.

1.7.2 CHARACTERISTICS OF MOVING AVERAGES

(a) The different moving averages produce different forecasts.

(b) The greater the number of periods in the moving average, the greater the smoothing effect.

(c) If the underlying trend of the past data is thought to be fairly constant with substantial randomness, then the greater number of periods should be chosen.

(d) Alternatively, if there is thought to be some change in the underlying state of the data, more responsiveness is needed, therefore fewer periods should be included in the moving average.

1.7.3 LIMITATIONS OF MOVING AVERAGES

(a) Equal weighing is given to each of the values used in the moving average calculations, whereas it is reasonable to suppose that the most recent data is more relevant to current conditions.

(b) A n period moving average requires the storage $n-1$ values to which is added the latest observation.

(c) The moving average calculation takes no account of data outside the period of the average, so full use is not made of all the data available.

(d) The use of the unadjusted moving average as a forecast can cause misleading results when there is an underlying seasonal variation.

Exercise 1

i. In a competition to win a new car, entrants were asked to rank 8 features of the car in order of importance. The features were labeled A, B, C, D, E, F, G and H. The entries for a randomly selected insurance salesman and a randomly selected mother with young children are given below.

Rank	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
Insurance salesman	***A***	***E***	***B***	***H***	***D***	***G***	***F***	***C***
Mother	***D***	***H***	***A***	***B***	***E***	***F***	***G***	***C***

Table 1.15

(a) Calculate spearman's rank correlation coefficient for these data.

(b) Stating your hypothesis clearly tests, at the 5 % level of significance, whether or not there is evidence of a correlation.

(c) Explain what could be said about the criteria these two entrants used when making their choices.

The values of spearman's rank correlation coefficient between judges' orders were 0.6190 and 0.8571 respectively.

(d) State, giving a reason, which of these two types of persons you believe the competition was aimed at.

ii. Given the data, table 16

x	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
y	*9*	*8*	*10*	*12*	*11*	*13*	*14*	*16*	*15*

Table 1.16

(a) Calculate the coefficient of correlation?

(b) Obtain the line of regression.

(c) Estimate the value of y which should correspond to x = 6.2

iii. Calculate the coefficient of correlation between the values of x and y and obtain the lines of regression for the following data.

x	*78*	*89*	*97*	*69*	*59*	*79*	*61*	*61*
y	*125*	*137*	*156*	*112*	*107*	*136*	*123*	*108*

Table 1.17

iv. Two judges in a beauty contest rank the ten competitors in the following order.

X	*6*	*4*	*3*	*1*	*2*	*7*	*9*	*8*	*10*	*5*
Y	*4*	*1*	*6*	*7*	*5*	*8*	*10*	*9*	*3*	*2*

Table 1.18

Do the two judges appear to agree in their standard?

Suggested References

i. DAVID S. Moore and George P. McCabe (1993), Introduction to the Practice of Statistics (second Edition) W.H. FREEMAN AND COMPANY NEW YORK, p94 -117 and p161 -183

ii. Ajit C. Tamhane and Dorothy D. Dunlop (2000) Statistics and Data Analysis (from Elementary to Intermediate) PRENTICE HALL, Upper Saddle River , p347 - 384

iii. Allan G. Bluman, (2009), Elementary Statistics A Step by Step Approach, (seventh Edition) MAGRAW –HILL INTERNATIONAL EDITION, P533 – 573

iv. Murray R. Spiegel and Larry J. Stephens (2009) Statistics (fourth Edition) SCHAUM'S SERIES P316 -352,

CHAPTER TWO: SAMPLING THEORY

Purpose: to equip the learner with sampling techniques and identifying the confidence interval for the mean, proportion and standard deviation.

Objective: By the end of his chapter, the learner should be able to:-

a) Find the confidence interval for the mean when σ is known.

b) Determine the maximum sample size for finding a confidence interval for the mean.

c) Find the confidence interval for the mean when σ is known.

d) Find the confidence interval for a proportion.

e) Determine the minimum sample size for finding a confidence interval for a proportion.

f) Find a confidence interval for variance and a standard deviation

2.1 SAMPLING:

Population or universe or census in statistics I used to refer to any collection of individuals or their attributes or of results of operations which can be numerically specified. A population containing a finite number of individuals or members is called a finite population. A population with infinite number of members is known as infinite population.

A part or small section selected from the population is called a sample and the process of such selection is called **sampling**. The aim of the theory of sampling is to get as much information as possible, ideally the whole of the information about the population from which the sample has been drawn. Given the form of the parent population we would like to estimate the parameters of the population or specify the limits within which the population parameters are expected to lie with a specified degree of confidence.

The fundamental assumption underlying most of the theory of sampling is random sampling which consists in selecting the individuals from the population in such a way that each individual of the population has the same chance of being selected.

2.2 TYPES OF SAMPLING

(i) Purposive sampling in which sampling attributes may be regarded as the drawing of samples from a population whose members possess the attribute A or not –A for example in sampling from a population of men, persons who are smokers and non-smokers. The choosing or drawing of an individual in sampling may be called an 'event' or 'trial' and the possession of the specified attribute A by the individual selected a "success".

(ii) Random sampling -

(iii) Simple sampling we mean random sampling in which each event has the same probability ρ of success and the probability of an event is independent of the success or failure of events in the preceding trials. For example counting the number of success in the throwing of a dice or tossing of a coin is a case of simple sampling.

(iv) Stratified sampling.

2.3 Mean and standard deviation in simple sampling of attributes

A simple sample of , n, members

This is clearly identical with that a series of, n, independent trials and with constant probability ρ of success. The probabilities of 0, 1, 2, ------ n success are the terms in the binomial expansion of

$$(q+p)^n \text{ where } (q=1-p)$$

The distribution so obtained is called sampling distribution of the number of successes in the sample. The expected value or the mean value, of the number of successes is therefore np and the standard deviation (also called the standard error) of the number of successes is $\sqrt{npq}$

By the proportion of successes in a sample, we mean the number of successes divided by the number of members in the sample. The mean and the standard error of the proportion of successes can be obtained by dividing the corresponding results for the number of success by n .

Mean of the proportion of successes = p

Standard error of the proportion of success = $\sqrt{\frac{pq}{n}}$

2.4 Large samples: Testing the significance for a single proportion

Suppose large number n of independent x trials. The probability of success in each success in each trial is p

$$X \sim B(n, p)$$

Assuming the hypothesis to be correct

The mean of the sampling distribution is np

The variance of the sampling distribution is npq , $q = 1 - p$

For large, n, then, $X \sim N(np, npq)$

$Z = \dfrac{x - np}{\sqrt{npq}}$ is distributed as a standard normal variation.

NB:

(i) We use the proportion of successes p

(ii) For large samples, $n \to \infty$, and small p, the binomial distribution may be approximated by normal distribution.

(iii) By the testing of a statistical hypothesis is meant to a procedure for deciding whether to accept or reject the hypothesis. The procedure usually consists in assuming or accepting the hypothesis as correct and then calculating the probability of getting the observed or more extreme sample.

(iv) The probability level below which we reject the hypothesis is called the level of significance usually, two levels are used i.e. 5% and 1%

Example 2.1: In a locality of 18000 families a sample of 840 families was selected, of these 840 families, 206 families were found to have a monthly income of Ksh7000 or less. It is desired to estimate how many out of the 18000 families have a monthly income of Ksh3000 or less. Within what limits would you place your estimate.

p = proportion of families having monthly income of Ksh3000

$$p = \frac{206}{840} = \frac{103}{420} = 0 \cdot 245 \qquad\qquad q = 0 \cdot 755$$

Assume that the condition of this problem will give a simple sample.

$$\text{Mean} = \quad np \quad = 18000 \times \frac{103}{420} = 4414.29$$

Standard Error (S.E) of the proportion of families having monthly income of Ksh3000

$$\sqrt{\frac{0.245\times 0755}{840}} \cong 0.015 = 1.5\%$$

Hence taking 0.245 to be the estimate of the families having a monthly income of Ksh3000 or less the limits are $24.5\% \pm (3\times 1.5)\%$ = 20% and 29% between 3600 and 5220 families.

Exercise 2

1. ***A sample of 10 measurements of the diameter of a sphere gave a mean*** $\bar{x} = 438$ ***centimeters (cm) and a standard deviation S = 0.06 cm. find the (a) 95% and (b) 99% confidence limits for the actual diameter.***

2. ***The intelligence quotients (IQs) of 16 students from one area of a city showed a mean of 107 and a standard deviation of 10, while IQ of 14 students from another area of the city showed a mean of 112 and a standard deviation of 8. Is there a significant difference between IQs of the two groups at significance level of (a) 0.01 and (b) 0.05?***

3. ***The standard deviation of the life times of a sample of 200electric light bulbs in 100 hours. Find the (a) 95% and (b) 99% confidence limits for the standard deviation of all such electric light bulbs.***

4. ***A survey of 1721 people found that 15.9% of individuals purchase religious book at a Christian bookstore. Find the 95% confidence of the true proportion of people who purchase their religious books at a Christian book store.***

Suggested References

1. DAVID S. Moore and George P. McCabe (1993), Introduction to the Practice of Statistics (second Edition) W.H. FREEMAN AND COMPANY NEW YORK p432 -440

2. Ajit C. Tamhane and Dorothy D. Dunlop (2000) Statistics and Data Analysis (from Elementary to Intermediate) PRENTICE HALL, Upper Saddle River ,p197 -208

3. Allan G. Bluman, (2009), Elementary Statistics A Step by Step Approach, (seventh Edition) MAGRAW –HILL INTERNATIONAL EDITION, p355

4. Murray R. Spiegel and Larry J. Stephens (2009) Statistics (fourth Edition) SCHAUM'S SERIES P203 -207,

CHAPTER THREE: TEST OF HYPOTHESIS

Purpose: - The learner should be able to state and test hypothesis to make decision based on the sample data.

Objective:- After completing this chapter, the learner should be able to:-

a) Define the terms used in hypothesis testing

b) State the null and alternative hypotheses

c) Find critical values for z-test

d) State five steps used in hypothesis testing

e) Test means when σ is known, using the z –test

f) Test proportions, using the z –test

g) Test variances or standard deviations, using the chi- square test

h) Test hypothesis, using confidence intervals

i) Explain the relationship between type I and type II errors and the powers of a test.

3.1 Introduction:-

In hypotheses testing, the researcher must define the population under study, state the particular hypotheses that will be investigated, give the significance level, select a sample from the population, collect the data, perform the calculations required for the statistical test, and reach a conclusion.

Hypotheses concerning parameter such as means and proportions can be investigated. The three methods used to test hypotheses are:-

- Traditional method
- The p –value method
- The confidence interval method

A hypothesis is an assumption, belief, or opinion which may or may not be true. E.g. it may be believed that a given drug cures 90% of the patients taking it or the average height of soldiers in the army is 168cms. The testing of a statistical hypothesis is the process by which this belief or opinion is tested by statistical means. We accept the hypothesis as being true, when it is supported by the sample data. We reject the hypothesis when the sample data fail to support it.

3.2 Traditional method

Every hypothesis testing situation begins with a statement of the hypothesis. A statistical hypothesis is a conjecture about a population parameter. This conjecture may or may not be true. There are two types of statistical hypothesis and alternative hypothesis.

3.3 Null and alternative hypothesis

A null hypothesis, generally denoted by the symbol H_o, is a statistical hypothesis that states tht there is no difference between a parameter and specific value, or that these arte no difference two parameters. I.e. Hypothesis to be tested for possible rejection or nullification under the assumption that it is true, the hypothesis is assigned a numerical value H_o: μ = 150cms.

An alternative hypothesis is any other hypothesis which we are willing to accept when the null hypothesis H_o is rejected, denoted by H_1 or H_A

$H_o : \mu = 150cms$, then $H_1 : \mu \neq 150cms$ or $H_1 : \mu > 150cms$ or $H_1 : \mu < 150cms$

3.4 Simple and composite hypothesis

A statistical hypothesis is said to be simple hypothesis if it completely specifies the value(s) of the parameter(s). for example, if we hypothesize that $\mu = 150cms$ or $p = 0.4$, then we have stated simple hypothesis. If a hypothesis, e.g. if we hypothesize that μ > 150**cms**, the hypothesis is composite as it does not assign a specified value to the parameter μ

3.5 Test – Statistic

A statistics which provides a basis for testing a null hypothesis is called a **test-statistic**. Every test-statistic has a probability (sampling) distribution which gives the probability of obtaining a specified value of the test – statistic when the null hypothesis is true. The most commonly used test – statistic are $z, t, x^2 or F$

3.6 Acceptance and Rejection Regions

All possible values which a test – statistic may assume can be divided into two mutually exclusive groups: one group consisting of values which appear to be consistent with null hypothesis, and the other having values which lead to the rejection of the null hypothesis. The first group is called acceptance region and the second rejection region is also called critical

region. The value(s) that separates the critical region from the acceptance region is called the critical value(s).

3.7 Type I and Type II errors

When testing hypothesis, two types of errors are likely to be made. On the basis of sample information, we may reject a null hypothesis H_o, when it is, in fact true or we may accept a null hypothesis H_o, when it is actually false.

Decision	Accept H_o	Reject H_o (or accept H_1)
H_o is true	Correct decision (no error)	Wrong decision (type I error)
H_o is false	Wrong decision (type II error)	Correct decision (no error)

Table 3.1

A legal analogy

In court trial, the supposition of law is that the accused (the defendant) is innocent. (H_o).

After having heard the evidence presented during the trial, the judge arrives at a decision. Suppose that the accused is, in fact, innocent (H_o is true but the finding of the judge is guilty. The judge has rejected a true null hypothesis and in so doing has made a type I error. If, on the hand, the accused is, in fact, guilty (i.e. H_o is false) and finding of the judge is innocent, the judge has accepted a false null hypothesis and by accepting a false hypothesis, he has committed a type II error.

These situations may be summarized as under: -

(i) The hypothesis is true but our test rejects it.

(ii) The hypothesis is false but our test accepts it

(iii)The hypothesis is true and our tests accepts it

(iv)The hypothesis is false and our test rejects it.

The probability of making a type I error is conventionally denoted by α and that of committing α type II error is indicated by β.

α= P (type I error) = p (reject H_o/H_o is true)

β= p (type II error) = p (accept t H_o/H_o is false)

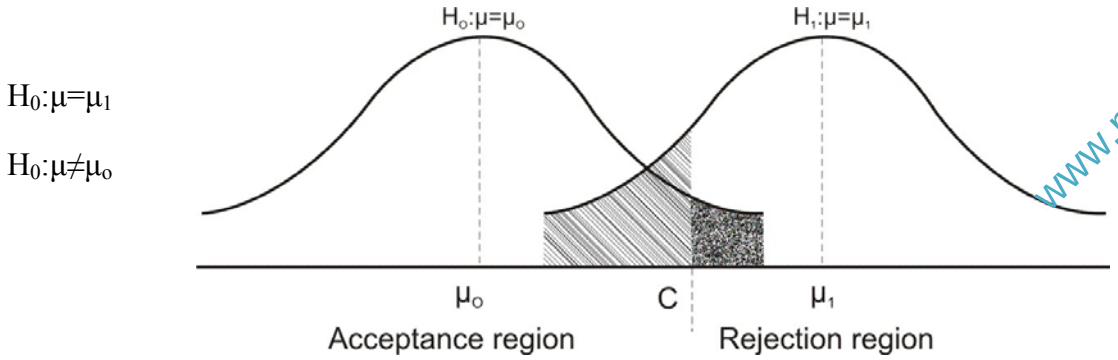

Figure 3.1

The probabilities of α and β are the shaded and dotted areas respectively of the distributions under the null hypothesis and under the alternative hypothesis.

THE SIGNIFICANCE LEVEL

The significance level of a test is the probability used as a standard for rejecting a null hypothesis H_o is assumed to be true. The value α is known as the size of the critical region, which is most frequently

$\alpha = 0.05$ & $\alpha = 0.01$

A test of significance is a rule or procedure by which sample results are used to decide whether to accept or reject a null hypothesis. A value of the statistic is said to be statistically significant when the probability of its occurrence under H_o is equal to or less than the significance level α, that is the value falls in the rejection region H_o in this case is rejected.

3.8 One-tailed and two –tailed tests

Three kinds of problems in test of hypothesis, they include

(i) Two tailed test,
(ii) Right-tailed test and
(iii)Left-tailed test.

Two tailed test is that where the hypothesis about the population mean is rejected for value of x falling into either tail of the sampling distribution. When the hypothesis about population mean is rejected only for value of x falling into one of the tails of the sampling distribution, then it is known as one tailed test. If it is the right tail then it is called right-tailed test or one –sided alternative to the right and if it is on the left tail then it is one sided alternative to the left and is called tailed test.

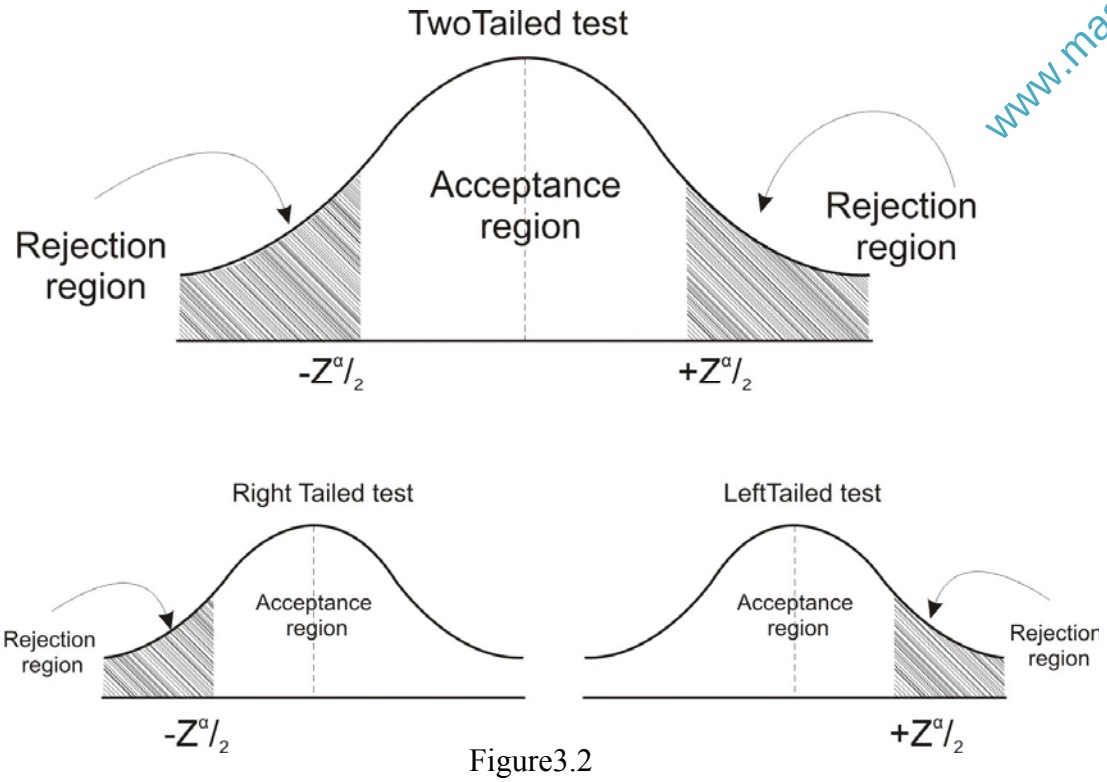

Figure3.2

The following table gives critical values of z for both one- tailed and two- tailed tests at various levels of significance

Level of significance	0.10	0.05	0.01	0.005	0.0002
Critical value of z for one tailed tests	-1.28 1.28	-1.650 1.650	-2.33 2.33	-2.58 2.58	-2.88 2.88
Critical value of z for two tailed tests	-1.645 And 1.645	-1.96 And 1.96	-2.58 And 2.58	2.81 And 2.81	-3.08 And 3.08

Table 3.2

3.9 Tests of hypothesis concerning large samples

Though it is difficult to draw a clear-cut line of demarcation between large and small samples, it is generally agreed that if the size of sample exceeds 30 it should be regarded as a sample. The tests of significance used for large samples are different from the ones used for small samples for the reason that the assumptions we make in case of large samples do not hold for small samples.

3.10 Tests of hypothesis involving large samples are based on the following assumptions.

(i) The sampling distribution of statistic is approximately normal
(ii) Values given by the samples are sufficiently close to the population value and can be used in its place for the standard error of the estimate.

3.11 Testing hypothesis about population mean

For testing hypothesis about population mean, the following procedure is adopted: -

(a) State the null hypothesis that there is no contradiction between the sample mean and population mean.
(b) Find the standard error of the mean by using the formula: -

$$\sigma\bar{x} = \frac{\sigma x}{\sqrt{n}}$$

(c) Compute limits within which the sample mean will fall at 95% or 99% confidence levels if the population mean is true.
(d) Find out whether the sample mean does lie within those limits or not. If, the sample mean lies within those limits then null hypothesis is accepted, otherwise, it is rejected.

Example 3.1: The mean lifetime of a sample of 100 light tubes produced by a company is found to be 1570 hours with standard deviation of 80 hours. Test the hypothesis that the mean lifetime of the tubes produced by the company is 1600 hrs.

Solution:

$H_0: \mu = 1600$

$H_0: \mu \neq 1600$

At 95% level of confidence

$$\text{Population mean} = \text{sample mean} \pm 1.96\sigma\bar{x}$$

$$\sigma\bar{x} = \frac{\sigma x}{\sqrt{n}} = \frac{80}{\sqrt{100}} = 8$$

Population must lie between the following ranges:

$$1570 \pm 1.96(8)$$
$$1570 \pm 15.7$$

1554.3 to 1585.7

As the population mean 1600 lies outside these limits so the null hypothesis is rejected.

Alterative method

The null hypothesis is that there is no difference between the sample mean and hypothetical population mean

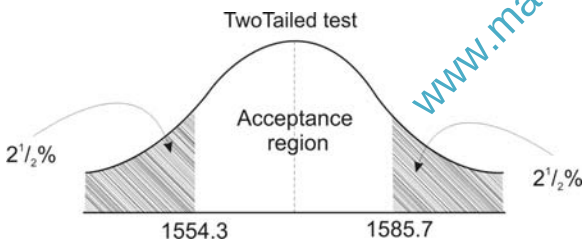

Figure 3.3

$$H_0: \mu = \mu_o$$

$$H_0: \mu \neq \mu_o$$

$z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}}$, where $\sqrt{\bar{x}} = \dfrac{s}{\sqrt{n}}$ [since σ is unknown for large sample]

$$z = \frac{1570 - 1600}{80/\sqrt{100}} = -3.75$$

The critical value $z = \pm 1.96$ for a two tailed test at 5% level of significance

Since the computed value of $z = -3.75$ falls in the rejection region

Example 3.2: A child welfare officer asserts that the mean sleep of young babies is 14 hours a day. A random sample of 64 babies shows that the mean sleep was only 13 hours 30 minutes with standard deviation of 3 hours. At 5% level of significance, test the assertion that mean sleep of babies is less than 14 hours a day.

Solution:

$$H_0: \mu = 14 \text{ hours}$$

$$H_0: \mu < 14 \text{ hours}$$

This is one tailed test as we are interested in the left hand tail of the distribution.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{13\,\frac{1}{3} - 14}{3/\sqrt{64}} = \frac{-2/3}{3/8} = -\frac{2}{3} \times \frac{8}{3} = -\frac{16}{9}$$

z Should be within 1.645 at 5% level of significance H_1 is accepted

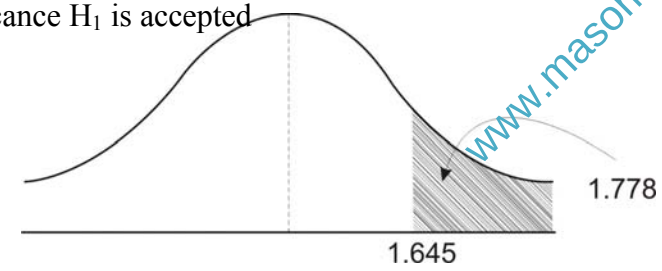

Figure 3.4

I.e. population mean is less than 14 hours a day.

Exercise 3

1. ***(a) Write short notes on the following: -***

   (i) Null and alternative hypothesis

   (ii) One tailed and two tailed tests

   (iii) Type I and type II errors

   (iv) Acceptance and rejection regions.

   (b) A sample of 100 motorcar tyres has a mean of 20,000 km and a standard deviation of 800 km. A second sample of 150 tyres has a mean life of 22,000 km and a standard deviation of 900 km. is it true to say that the two samples were drawn from the same population?

2. ***A researcher reports that the average salary of assistant professors is more than £42,000. A sample of 30 assistant professors has a mean salary of £43,260 at $\alpha = 0.05$, test the claim that assistant professors earn more than £42,000 per year. The standard deviation of the population is £ 5,230.***

3. ***(a) What is meant by a p- value?***

   (b) State whether the null hypothesis should be rejected on the basis of the given P-value

   (i) P-value = 0.258, $\alpha = 0.05$, one tailed test

   (ii) P-value = 0.0684, $\alpha = 0.10$, two tailed test

   (iii) P-value = 0.0153, $\alpha = 0.01$, one tailed test

   (iv) P-value = 0.0232, $\alpha = 0.05$, two tailed test

   (v) P-value = 0.002, $\alpha = 0.01$, one tailed test

Suggested References

1. DAVID S. Moore and George P. McCabe (1993), Introduction to the Practice of Statistics (second Edition) W.H. FREEMAN AND COMPANY NEW YORK p 455 -458

2. Ajit C. Tamhane and Dorothy D. Dunlop (2000) Statistics and Data Analysis (from Elementary to Intermediate) PRENTICE HALL, Upper Saddle River , p237 – 256

3. Allan G. Bluman, (2009), Elementary Statistics A Step by Step Approach, (seventh Edition) MAGRAW –HILL INTERNATIONAL EDITION, p399 -452

4. Murray R. Spiegel and Larry J. Stephens (2009) Statistics (fourth Edition) SCHAUM'S SERIES P245- 277,

CAT 1

1. A group of 25 students took examinations in both pure mathematics and statistics. Their marks out of 150 in mathematics, x, and in statistics, y, were recorded and are summarised below.
   $\sum x = 1978$, $\sum x^2 = 175840$ $\sum y = 2123$ $\sum y^2 = 202257$ $\sum xy = 181572$

   i. Calculate S_{xx}, S_{yy} and S_{xy}
   [6marks]
   ii. Find the product moment correlation coefficient between the marks in pure Mathematics and Statistics,
   [2marks]
   iii. Starting your hypotheses clearly tests, at the 5% level of significance, whether or not there is evidence of a correlation.
   [4marks]
   iv. State an assumption needed for the test in part (iii) to be valid.
   [1mark]

2. Explain the following terms
   i. The difference between one tailed and two tailed test
   [2marks]
   ii. Null and Alternative Hypothesis
   [2marks]

3. The breaking strengths of cables produced by manufacturer have a mean of 1800 Kg and a standard deviation of 100 Kg. by a new technique in the manufacturing process, it is claimed that breaking strength can be increased. To test this claim, a sample of 50 cables is tested and it is found that the mean breaking strength is 1850Kg. can we support the claim at the 0.01 significance level?
   [6 marks]
4. Given that $H_{0:}\ \mu \leq 15$
   $H_1:\ \mu > 15$

   A sample of 40 provides a sample mean of 16.5 and a sample standard deviation of 7

   i. At $\alpha = 0.02$, what is the critical value for z and what is the rejection rule?
   ii. Compute the value of the test statistics z
   [7marks]

CHAPTER 4: TEST OF PROPORTIONS

Purpose: - The learner should be able to test differences between two means, proportions and variances using z-Test

OBJECTIVES:

After the end of this chapter, you should be able to:-

- ***Test the difference between sample means using the z –test***
- ***Test the difference between two means for independent samples, using the t –test***
- ***Test the difference between two means for dependent samples***
- ***Test the differences between two proportions.***
- ***Test the difference between two variances or standard deviations***

4.1 Testing hypothesis about population proportion

For testing hypothesis about population proportion, the following procedure is adopted: -

(a) State the null hypothesis that there is no contradiction proportion. (The required measure as found from the sample.)
(b) Select the confidence level on the basis of one tail or two tests.
(c) Compute the standard error of proportion of using the following formula : -

Standard error of proportion $\sigma_p = \sqrt{\frac{pq}{n}}$

Where p = sample proportion, $q = 1 - p$, n = sample size.

(d) Compute the values of : $p \pm$ appropriate number σ_p
(e) Check where the sample proportion lies. If the sample means lies between the above limits then null hypothesis is accepted and vice versa.

Example 4.1 A sample of 400 electors selected at random gives 51% majority to political party XY. Could such a sample have been drawn from population with a 50 – 50 division of political opinion

Solution:

$p = 0.5$ $\quad$ $\hat{p} = 0.51$ $\quad$ $n = 400$
$q = 0.5$ $\quad$ $\hat{q} = 0.49$

H_0: Po=0.5

H_0:$P_o \neq 0.5$

P_o is the population proportion.

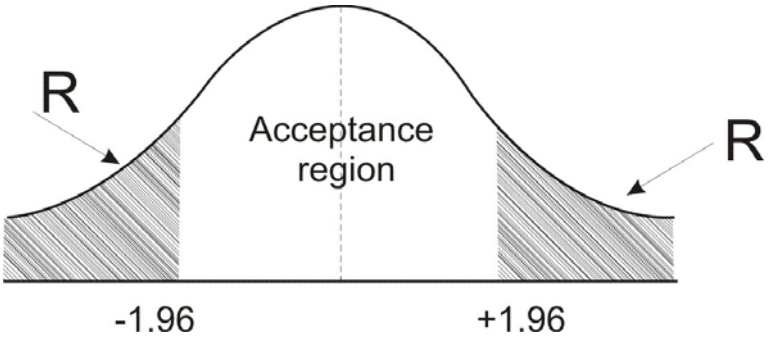

Figure 4.1

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.50)(0.49)}{400}} = 0.025$$

At 5 % significance sample proportion should fall within the following limits.

$$p \pm 1.96\sigma_p = 0.50 \pm 1.96(0.025) = 0.50 \pm 0.049$$

Or 0.451 to 0.0549

A sample proportion lies between these limits, the null hypothesis is accepted.

Example 4.2 A sales clerk in a departmental store claims that 60% of the shoppers entering the store leave without make a purchase. A random sample of 50 shoppers showed that 35 of them left without buying anything. Are the sample results consistent with the claim of the sales clerk? Use a level of significance of 0.05

Solution

H_0: Po=0.6

H_0:$P_o \neq 0.6$

Test: two – tails

$$p = \frac{35}{50} = 0.7$$

Using the z statistic we have

$$z = \frac{\bar{p} - \bar{p_o}}{\sqrt{\frac{p_o q_o}{n}}} = \frac{0.7 - 0.60}{\sqrt{\frac{(0.6)(0.4)}{50}}} = 1.44$$

This is one tailed test so the critical value of z is 1.645 at 5% level of confidence. Since the computed value of z =1.44 which is less than the critical value of $z = 1.645$, therefore the null hypothesis cannot be rejected. Hence based on this sample data, we cannot reject the claim of the sales clerk.

Example 4.3: A firm purchases a very large quantity of metal off cuts and wishes to know the average weight an off cut. A random sample of 625 off cuts is weighed and it is found that the mean sample weight is 150 grams with a sample standard deviation of 30 grams. What is the estimate of the population mean and what is the standard error of the mean. What would be the standard error if the sample size was 1225?

Solution:

The sample mean is 150g so that the estimate of the population mean is 150g

$$\bar{x} = 150 = \hat{\mu} = 150g \quad \text{Since n >30}$$

Where $\hat{\mu}$ means best "estimate of" where n=625

$$\text{Standard error of the mean} = \frac{s}{\sqrt{n}} = \frac{30}{\sqrt{625}} = 1.2gms$$

When $n = 1225$

$$s_{\bar{x}} = \frac{30}{\sqrt{1225}} = 0.857gms.$$

It will be seen that increasing the sample size reduces the standard error. This accord with common sense, we would expect a larger sample to be better than a smaller one.

$\bar{x} = 150g$ And $s_{\bar{x}} = 1.2g$

$\therefore$ At 95% confidence level

$$\mu = \bar{x} \pm 1.96 s_{\bar{x}}$$

$$= 150 \pm 1.96(1.2)$$
$$= 150 \pm 2.35$$

This means that we are 95% confident that the populations mean lies within the confidence zone; somewhere between 147.65gm and 152.35g

At 99% confidence level the limits are $\mu = 150 \pm 2.58(1.2gms)$

Range from 146.9 to 153.1gms.

Example 4.4 A random sample of 400 rail passengers is taken and 55% are in favour of proposed new timetables. With 95% confidence what proportion of all rail passengers are In favour of the timetables?

Solution:

$$n = 400 \qquad p = 0.55 \qquad q = 1 - 0.55 = 0.45$$

As $np = 220$ i.e. well over s, the normal approximation can be used

$$s_{ps} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.55 \times 0.45}{400}} = 0.025$$

$\therefore$ We are 95% confident that the population proportion is $ps \pm 1.96 s_{ps}$

$$= 0.55 \pm 1.96(0.025)$$
$$= 0.55 \pm 0.049$$
$$= 0.501 \text{ to } 0.599$$

Example 4.5 It is required to test hypothesis that 50% of households have freezers. A random sample of 400 households found that 54% of the sample had freezers. The significance level is 5%

Solution:

H_0: $\prod$=50% of all house holds.

H_0: $\prod \neq$ 50% of all house holds.

$$s_{ps} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.5 \times 0.5}{400}} = 0.025$$

$$z = \frac{0.54 - 0.50}{0.025} = 1.6$$

At the 5% level of significance for a two- tailed the appropriate value 1.96

$\therefore$ As the calculated z score is 1.6 we can say that difference is not significant and that H_o should not be rejected.

4.2 Hypothesis Testing of the Difference between Two Means

Where two random samples are taken, frequently it is required to know if there is a significant difference between the two means. This hypothesis test follows the general pattern except that, the standard error calculations differs.

The distribution of sample mean differences is normally distributed and remains normally distributed whatever the distribution of the populations from which the samples are drawn. Where n >30 i.e. large samples, the normal area tables are used, when n < 30, the t distribution applies.

Standard error of the difference of means = $s_{\overline{X_A}-\overline{X_B}} = \sqrt{\frac{s_A^{\,2}}{n_A} + \frac{s_B^{\,2}}{n_B}}$

s_A = standard deviation of sample A, size n_A

s_B = standard deviation of sample B, size n_B

$$z = \frac{\overline{X_A} - \overline{X_B}}{S_{(\overline{X_A}-\overline{X_B})}}$$

Example 4.6 Machine A and machine B produce identical components and it is required to test if the mean diameter of the components is the same. A random sample of 144 from machine A had a mean of 36.40mm and a standard deviation of 3.6mm, whilst a random sample of 225 from machine B had a mean of 36.90mm and a standard deviation of 2.9mm. Are the means significantly different at the 5% level?

Solutions:

H_0: mean of A= mean of B

H_0: mean of A≠ mean of B

I.e. as we are not concerned with the direction of the variation this is a two tail test.

$$s_{\overline{X_A}-\overline{X_B}} = \sqrt{\frac{3.6^2}{144} + \frac{2.9^2}{225}} = 0.357$$

$$z = \frac{36.40 - 36.90}{0.3569} = -1.401 \qquad (H_o \text{ would be accepted})$$

The score for a two failed test at 5% level is ± 1.96 $\therefore$ as the calculated z score of 1.4 is within this value there is nothing to suggest that there is any difference between.

4.3 Hypothesis Testing of The Difference Between Proportions

In a similar manner it may be required to test the differences between the proportions of a given attribute found in two random samples.

	Sample 1	Sample 2
Sample size	n_1	n_2
Sample proportion of success	p_1	p_2
Population proportion of success	π_1	π_2

Table 4.1

The null hypothesis will be $\prod_1 = \prod_2$ i.e. that the two samples are from the same population. This being so the best estimate of the standard error of the difference of p_1 and p_2 is given by polling the samples and finding the pooled sample proportion thus.

$$P = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$$

And the Standard error is

$$s_{P_1 - P_2} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$

And $$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{s_{P - P_2}}$$

But where the null hypothesis is $\prod_1 = \prod_2$ the second part of he numerator disappears.

Example 4.7 A market research agency take a sample of 1000 people and finds that 200 know of brand X after an advertising campaign a further sample of 1091 people is taken and it is found that 240 know of brand X.

Solution:

It is required to know if there has been an increase in the number of people having an awareness of brand X at the 5% level.

$$H_o : \pi_2 = \pi_1$$
$$H_1 : \pi_2 > \pi_1 \qquad \text{(One tail test)}$$

$$P_1 = \frac{200}{1000} = 0.2 \qquad P_2 = \frac{240}{1091} = 0.22$$

Pooled sample proportion $P = \frac{200 + 240}{1000 + 1091} = 0.21$ and $q = 1 - p = 1 - 0.21 = 0.79$

$$\therefore s_{(P_1 - P_2)} = \sqrt{\frac{0.21 \times 0.79}{1000} + \frac{0.21 \times 0.79}{1091}} = 0.0178$$

$$z = \frac{0.20 - 0.22}{0.0178} = -1.12$$

The critical value for a one-tailed test at the 5% level is -11.64 so that as the calculated value is within this value we conclude there is insufficient evidence to reject the null hypothesis.

Example 4.8 in random sample of 100 persons taken from village A, 60 are found to be consuming tea. In another sample of 200 persons taken from village B, 100 persons are found to be consuming tea. Do the data reveal significant difference between the two villages so far as habit of taking tea is concerned?

Solution:

Let us take the hypothesis that there is no significant difference between the two as far as the habit of taking tea is concerned i.e. $\prod_1 = \prod_2$

We are given

$$P_1 = \frac{x_1}{n_1} = \frac{60}{100} = 0.6, \qquad n_1 = 100$$

$$P_1 = \frac{x_2}{n_2} = \frac{100}{200} = 0.5, \qquad n_2 = 200$$

The appropriate statistics to be used here is given by

$$P = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{(0.6)(100) + (0.5)(200)}{100 + 200} = \frac{60 + 100}{300} = 0.53$$

$q = 1 - 0.53 = 0.47$

$$\sigma_{P_1-P_2} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} = \sqrt{\frac{(0.53)(0.47)}{100} + \frac{(0.53)(0.47)}{200}} = 0.0608$$

$$z = \frac{(p_1 - p_2)}{\sigma_{P-P_2}} = \frac{0.6 - 0.5}{0.0608} = \frac{0.1}{0.0608} = 1.64$$

Accept the null hypothesis since computed value **z** is less the 5%

Example 4.9 Suppose that you are as a purchase manager for a company. The following information has been supplied to you by two manufactures of electric bulbs: -

	Company A	Company B
Mean life (in hours)	1300	1248
Standard deviation (in hours)	82	93
Sample size	100	100

Table 4.3

Which brand of bulbs are you going to purchase if you desire to take a risk of 5%?

Solution: let us take the hypothesis that there is no significant difference in the quality of the brands of bulbs.

i.e.

$$H_o : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}}} = \frac{1300 - 1248}{\sqrt{\frac{82^2}{100} + \frac{93^2}{100}}} = \frac{52}{\sqrt{67.24 + 86.49}} = \frac{52}{12.399} = 4.19$$

Since our computed value z = 4.19 is greater than critical value of z= 1.96

(5% level) we reject the null hypothesis, hence the quality of two brands of bulbs differ significantly.

Example 4.10: A sample of 100 motor car tyres has a mean of 20,000 miles and a standard deviation of 800 miles. A second sample of 150 miles has a mean life of 22,000 miles and standard deviation of 900 miles.

Is it true to say that the two samples were drawn from the same population?

	Sample A	Sample B
$\bar{x}$	20,000	22,000
S^2	800^2	900^2
n	100	150

Table 4.4

Solution:

The samples were drawn from the same sample

$H_o : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

Samples are not drawn from the sample population.

Standard error of difference between means:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}} = \sqrt{\frac{800^2}{100} + \frac{900^2}{150}} = \sqrt{11{,}800}$$

$$= 108.6$$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{20000 - 22000}{108.6}$$

$$= -18.4$$

$$z - values = \pm 1.96$$

Under H_o, and at 0.05 level of significance with a two-tailed test, critical

Conclusion: H_o is rejected i.e. samples could not have been drawn. From the

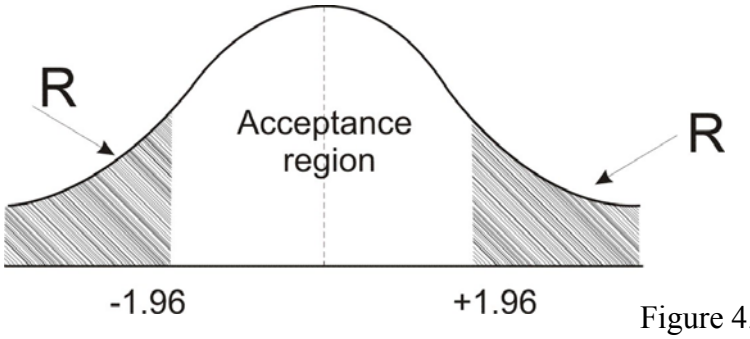

Figure 4.2

Same population, since the calculated z value falls in the rejection region

4.3 Chi- Square Test (χ^2)

The chi-square is denoted by the Greek letter χ^2. It is frequently used as a test statistic in testing a hypothesis concerning the difference concerning the difference of a sample and a corresponding set of expected or theoretical frequencies called the number of degrees of freedom (d.f), where the term degree of freedom represents the number of independent random variables that express the chi-square.

PROPERTIES

(1) For every increase in the number of degrees of freedom, there is a new χ^2 – distribution.

(2) This possesses additive property so that when χ^2 and χ_2^2 are independent and have a chi-square distribution with n_1 and n_2 degree of freedom, $\chi^2_1 + \chi^2_2$ will also be distributed as a chi-square distribution n_1+n_2 degrees of freedom.

(3) Where the d.f is 30 and less, the distribution of χ^2 is skewed. But, for degrees of freedom greater than 30 in a distribution, the values of $\chi 2$ are normally distributed.

(4) The χ^2 function has only one parameter, the number of degrees of freedom. The χ^2 distribution is positively skewed on the right, especially when the number of degrees of freedom is small.

(5) χ^2 distribution is a continuous probability distribution which has the value zero at its lower limit and extends to infinity in the positive direction. Negative value of $\chi 2$ is not possible because the differences between the observed and expected frequencies are always squared.

4.4 Chi-square test

The χ^2 test is one of the simplest and most widely used non-parametric tests in statistical work. It makes no assumptions about the population being sampled. The quantity $\chi 2$ describes the magnitude of discrepancy between theory and observation, i.e. with the help of χ^2 test we can know whether a given discrepancy between theory and observation can be attributed to chance or whether it results from the inadequacy of the theory to fit the observed facts. If χ^2 is zero, it means that the observed and expected frequencies completely coincide. The greater the value of χ^2 the greater would be the discrepancy between observed and expected frequencies.

The formula for computing square is: -

$$\chi^2 = \frac{\sum (O-E)^2}{E}$$, where O = Observed frequency

E = Expected or theoretical frequency

The calculated value of χ^2 is compared with the table value of χ^2 for given degrees of freedom at specified level of significance. If the calculated value of $\chi 2$ is greater than the table value, the difference between theory and observation is considered to be significant.

On the other hand, if the calculated value of χ^2 is less than the table value, the difference between theory and observation is not considered significant, i.e. it could have arisen due to fluctuations of sampling.

Conditions for application of $\chi 2$ test

Basic conditions must be met in order for chi-square analysis to be applied.

(1) The experimental data (sample observation) must be independent of each other.

(2) The sample data must be drawn at random from the target population

(3) The data should be expressed in original units for convenience of comparison, and not in percentage or ratio form.

(4) The sample should contain at least 50 observations.

(5) There should not be less than five observations in any one cell.

Example 4.12: A random sample of 400 householders is classified by two characteristics: whether they own a colour television and by what type of householder (i.e. owner –occupier, private tenant, council tenant). The results of this investigation are given below:

	OWNER OCCUPIER	COUNCIL TENANT	PRIVATE TENANT	TOTAL
Colour TV	150	60	20	230
No colour TV	45	68	57	170
	195	128	77	400

Table 4.6

It is required to test at the 5% level, the following hypothesis.

H_o: The two classifications are independent. (I.e. no relation between classes of householder and colour

TV ownership)

H_1: the classifications are NOT independent

The expected frequency in each cell in the table is found by apportioning the total of the type of house holder in the ratio of colour TV: no occurs TV

$$230: 170 = 23: 17$$

The 195 owner occupiers are split in the 23: 17 proportion

$$\frac{195\times 23}{40} : \frac{195\times 17}{40} = 112 : 83$$

$$\frac{128\times 23}{40} : \frac{128\times 17}{40} = 74 : 54$$

$$\frac{77\times 23}{40} : \frac{77\times 17}{40} = 44 : 33$$

EXPECTED FREQUENCIES

	OWNER OCCUPIER	COUNCIL TENANT	PRIVATE TENANT	TOTAL
Colour TV	112	74	44	230
No colour TV	83	54	33	170
	195	128	77	400

Table 4.7

The $\chi 2$ calculations can now be made

Observed frequencies (O)	Expected frequencies (E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
150	112	38	1444	12.89
45	83	-38	1444	17.40
60	74	-14	196	2.65
68	54	14	196	3.63
20	44	-24	576	13.09
57	33	24	576	17.45
				$\chi 2$ =67.11

Table 4.8

To find the appropriate $\chi 2$ value from tables, this is done by establishing v, the degree of freedom. This is found by multiplying the number of columns less one.

V = (rows-1) (columns-1)

= (2-1) (3-1) = 2 degrees of degrees of freedom.

The value of the cut-off point of $\chi 2$ for 2 degrees of freedom is 5.999. As calculated value 67.11 is greater than the table value we reject the null hypothesis and accept there is a connection between the type of house holder and colour TV ownership.

Example 4.13

Observed frequencies	364	376	218	89	33	13	21
Expected frequencies (based on Poisson distribution)	339	397	234	92	27	6	10

Table 4.10

The last three classes should be combined together after grouping the position would be as follows

Observed frequencies (O)	Expected frequencies (E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
364	339	25	625	1.844
376	397	-21	441	1.111
218	234	-16	256	1.094
89	92	-3	9	0.0978
33	27	6	36	1.091
16	7	9	81	5.0625
				10.3002

Table 4.11

V = (2-1) (6-2) = 4

X^2 value is 11.070 calculated values 10.300, we accept the null hypothesis

Example 23: Test the hypothesis that the number of parts demanded does not depend on the day of t he week. In a sample study the following information was obtained:

Day	Mon	Tue	Wed	Thur	Fri	Sat	Total
No of parts demanded	1124	1125	1110	1120	1126	1115	6270

Table 4.12

Test the hypothesis that the number of parts demanded does not depend on the day of the week. (The table value of $\chi 2$ for S d.f at 5% level of significance is 11.071

Solution:

H_o: The number of parts demanded does not depend on the day of the week

H_1: The number depends on the day of the week.

The numbers of spare parts demanded in a week are 6720 and if all days are same we should expect $\frac{6720}{6} = 1120$ i.e. 1120 spare parts on each day of the week.

Days	Observed frequencies (O)	Expected frequencies (E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Monday	1124	1120	16	0.014
Tuesday	1125	1120	25	0.022
Wednesday	1110	1120	100	0.089
Thursday	1120	1120	0	0
Friday	1126	1120	36	0.032
Saturday	1115	1120	25	0.022

Table 4.13

The table value of $\chi 2$ for 5 d.f at 5% level of significance is 11.07. The computed value of $\chi 2$ is much less than the table value. The hypothesis is accepted and we conclude that the demand for spare parts is independent of the day of the week.

Example4.14: A survey of 320 families with S children each revealed the following distribution.

No of boys	5	4	3	2	1	0
No of girls	0	1	2	3	4	5
No of families	14	56	110	88	40	12

Table 4.14

Is this result consistent with the hypothesis that make and female births are equally probable?

Solution:

$$p = \frac{1}{2}$$

On the assumption that male and female births are equally probable, the probability of a male is

The expected number of families can be calculated by the use of binomial distribution.

$$\chi^2 \sim B(5, 1/2)$$

$$f(x) = 5_{C_x}\, p^x q^{5-x} \qquad x = 0,1,2,3,4,5$$

$$= 5_{C_x} (\frac{1}{2})^5 \qquad \text{, for } [\because p = q = 1/2]$$

To get the expected frequencies, multiply f(x) by the total number N=320

x	f(x)	Expected frequency N f(x)	No of families	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	$5_{C_0}(1/2)^5 = 1/32$	10	14	16	1.6
1	$5_{C_1}(1/2)^5 = 5/32$	50	56	36	0.72
2	$5_{C_2}(1/2)^5 = 10/32$	100	110	100	1.0
3	$5_{C_3}(1/2)^5 = 10/32$	100	88	144	1.44
4	$5_{C_4}(1/2)^5 = 5/32$	50	40	100	2.0
5	$5_{C_5}(1/2)^5 = 1/32$	10	12	4	0.4
					7.16

Table 4.15

$$\chi^2 = \frac{\sum (O-E)^2}{E} = 7.16$$

The table value of χ^2 for v = 6-1 = 5 at 5% level of significance is 11.07. The computed value of

$\chi 2$=7.16 is less than the table value. Therefore, the hypothesis is accepted. Thus, it can be conclude that male and female births are equally probable.

4.5 Testing The Hypothesis For Equality Of Two Variances

The test for equality of two population variances is based on the variances in two independent by selected random samples drawn from two normal populations.

Under the null hypothesis $\sigma_1^2 = \sigma_2^2$

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}, \quad \text{reduces to } F = \frac{S_1^2}{S_2^2}$$

Which follows F-distribution with v_1 and v_2 degrees of freedom, It is convenient to place larger

sample variance in the numerator for computational purpose.

If the computed value of F exceeds the table value of F, we reject the null hypothesis. I.e. The alternate hypothesis is accepted.

Example 4.15: Two sources of raw materials are under consideration by a company. Both sources seem to have similar characteristics, but the company is not sure about their respective uniformity. A sample of 10 lots from source A yields a variance of 225 and a sample of 11 lots from sources B yields a variance of 200. Is it likely that the variance of source A is significantly greater than the variance of sources B at $\alpha = 0.01$?

Solution:

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ the variance of source A and that of source B are same

$H_1 : \sigma_1^2 \neq \sigma_2^2$

$F = \frac{S_1^2}{S_2^2}$, where is $S_1^2 = 225$, $S_2^2 = 200$

$$F = \frac{225}{200} = 1.125$$

The table value of F for v_1=9 and v_2=10 at 1% level of significance is 4.94. Since the computed value of F is smaller than the table value of F. the null hypothesis is accepted. Hence the population variances of the two populations are the same.

Example4.16: In a manufacturing process a sample of 10 items is found to have a standard deviation of 5 mm. the same items are produced by a different process and a sample of 12 items is found to have a standard deviation of 4.2 mm.

Is the new process more consistent than the old at the 5% level?

Solution: The sample details are

Sample 1	Sample 2
$S_1 = 5$	$S_2 = 4.2$
$S_1^{\,2} = 25$	$S_2^{\,2} = 17.64$
$n_1 = 10$	$n_2 = 12$

Table 4.16

Degrees of freedom = 10-1 = 9 and 12-1 = 11

Null hypothesis $H_0 : \sigma_1^{\,2} = \sigma_2^{\,2}$, there is no difference between the population variability's.

Alternative hypothesis $H_1 : \sigma_2^{\,2} < \sigma_1^{\,2}$, that the new process is more consistent than the old.

$$Fscore = \frac{S_1^{\,2}}{S_2^{\,2}} = \frac{25}{17.64} = 1.42$$

The F score where v_1=9 and v_2=11, at the 5% level is 2.90.

Therefore, the null hypothesis cannot be rejected. Based on the sample results there is no significant difference between the two processes.

Exercise 4

1. ***(a) A survey found that the average hotel room rate in Mombasa is KSh7, 506 and the average room rate in Nakuru is KSh6850. Assume that the data were obtained from two samples of 50 hotels each and that the standard deviations of the populations are KSh475 and KSh410 respectively. At α = 0.05, can it be concluded that there is a significant difference in the rates?***

   (b) Find the 95% confidence interval for the difference between the means for the data in example above.

2. ***Explain the difference between testing a single mean and testing the difference between two means.***

A random sample of size 36 was taken from a population distributed N (μ, 3.9^2). The value of Sample mean $\bar{x}$ was 15.6
the

(a) Find a 90% confidence interval for μ. it is believed that the value of μ is 17.0

(b) Use your confidence interval to comment on this belief.

3. ***A social historian believes that changes in attitudes to left-handedness around 1980 have led to a change in the number of left handed people less than 40 years. A random sample of 200 people was asked their age and whether they were left or right handed. The results are summarized in the table, 36 below.***

	Left handed	*Right handed*
Age ≤ 40	*17*	*93*
Age > 40	*5*	*85*

Table 4.17

Stating your hypothesis clearly test the social historian's belief. Use a 5% level of significance.

Suggested References

1. DAVID S. Moore and George P. McCabe (1993), Introduction to the Practice of Statistics (second Edition) W.H. FREEMAN AND COMPANY NEW YORK, p530

2. Ajit C. Tamhane and Dorothy D. Dunlop (2000) Statistics and Data Analysis (from Elementary to Intermediate) PRENTICE HALL, Upper Saddle River , p 591 – 593

3. Allan G. Bluman, (2009), Elementary Statistics A Step by Step Approach, (seventh Edition) MAGRAW –HILL INTERNATIONAL EDITION, p399 -452, 589

4. Murray R. Spiegel and Larry J. Stephens (2009) Statistics (fourth Edition) SCHAUM'S SERIES

CHAPTER FIVE: NON PARAMETRIC STATISTICS

Purpose: - to Equip the student with information how to use nonparametric tests

OBJECTIVES: By the end of the chapter you should be able to:-

a) ***State the advantage and advantages of nonparametric methods***

b) ***Test hypothesis, using sign test***

c) ***Test hypothesis, using the Wilcoxon rank sum test.***

d) ***Test hypothesis using the Kruskal Wallis test***

e) ***Compute the Spearman rank correlation Coefficient μ, σ and p.***

Introduction

Statistical tests such as z, t and F tests are called parametric tests. Parametric tests are statistical tests for population parameters such as means, variances and proportions that involve assumptions about populations from which the sample were selected. However, Statisticians have developed a branch of statistics known as nonparametric statistics or distribution statistics free to use when the population from which the samples are drawn is not normally distributed. Nonparametric statics can also be used to test hypothesis that do not involve specific population parameters, such as

5.1. Advantages and Disadvantages

There are five advantages that nonparametric methods have over parametric methods:

1. They can be used to test population parameters when the variable is not normally distributed.

2. They can be used when the data nominal or ordinal.

3. They can be used to test hypotheses that do not involve population parameters.

4. In some cases, the computations are easier than those for the parametric counterparts.

5. They are easy to understand.

5.2 There are three disadvantages of nonparametric methods:

1. They are less sensitive than their parametric counterparts when the assumptions of the parametric methods met. Therefore, larger differences are needed before the null hypotheses can be rejected.

2. They tend to use less information than parametric tests

3. They are less efficient than their parametric counterparts when the assumptions of the parametric methods are met.

5.3 The Sign Test

The simplest nonparametric test, the sign test for single samples, is used to test the value of a median for a specific sample. When using the sign test, the researcher hypothesizes the specific value for the median of a population, and then he or she selects a sample of data and compares each value with the conjectured median. If it is below the conjecture median, it is assigned minus sign. If the data value is above the conjectured median it is assigned a plus sign. And if it is exactly the same as the conjectured median, it is assigned zero. Then the number of plus and minus signs are compared. If the null hypothesis is true, the number of plus signs should be approximately equal to the number of minus signs. If the null hypothesis is not true, there will be a disproportionate number of plus or minus signs.

The test value is the smaller number of plus or minus signs

Example5.1. A convenience store owner hypothesis that the median number of scones she sells per day is 40. A random sample of 20days yields the following data for the number of scones sold each day.

18	43	40	16	22
30	29	32	37	36
39	34	39	45	28
36	40	34	39	52

At α =0.05, test the owner's hypothesis.

Solution

Step 1: State the hypotheses and identify the claim

H_0: median = 40(claim) and H_1: median ≠ 40

Step 2: Find the critical value. Compare each value of the data with the median. If the value is greater than the median, replace the value with a plus sign and vice versa. The completed table follows.

-	+	0	-	-
-	-	-	-	-
-	-	-	+	-
-	0	-	-	+

Table 5.2

Using n =18 (omit the zeros) and α =0.05 for two -tailed test, the critical value is 4.

Step 3: Compute the test value. Count the number of plus and minus signs obtained in step, and uses the smaller value as he test value. Since there are 3 plus signs and 15 minus signs, 3 is the test value.

Step 4: Make the decision. Compare the test value 3 with critical value 4. The null hypotheses is rejected since 3 < 4

Step 5: Summarize the results. There is enough evidence to reject the claim that the median number of scones sold per day is 40

5.4 The Wilcoxon Rank Sum Test

- Wilcoxon signed Rank test the population of differences assumed to have a symmetric distribution. In the wilcoxon tests, the values of the data for both samples are combined and then ranked. If the null hypothesis is true – meaning that there is no difference in the population distributions – then the values in each sample should be ranked approximately the same. Therefore, when the ranks are summed for each sample, the sums should be approximately equal, and the null hypothesis will not be rejected. If there is a large difference in the sums of the ranks, then the distributions are not identical and the null hypothesis will be rejected.

- The first test to be considered is the wilcoxon rank sum tests for independent samples. For this test, both sample sizes must be greater than or equal to 10. The formulas needed for the test are given next.

- Formula for the Wilcoxon Rank Sum Test when samples are independent.

$$z = \frac{R - \mu_R}{\sigma_R}$$

Where

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

R = Sum of ranks for smaller sample size (n_1)

n_1 = smaller of sample sizes

n_2 = larger of sample sizes

$n_1 \geq 10$ and $n_2 \geq 10$

Note that if both samples are the same size, either size can be used as n_1

Example 5.2: Two independent samples of army and navy recruits are selected and the time in minutes it takes each recruit to complete an obstacle course is recorded as shown in the table. At α = 0.05, is there any difference in the times it takes the recruits to complete the course?

Army	15	18	16	17	13	22	24	17	19	21	26	28
Navy	14	9	16	19	10	12	11	8	15	18	15	

Table 5.3

Solution:

Step 1: State the hypothesis and identify the claim

H_o: there is no difference in the times it takes the recruits to complete the obstacle course.

H_1: there is a difference in the times it takes the recruits to complete the obstacle course.

Step 2: Find the critical value. Since α = 0.05

And this test is a two-tailed test; use the z-values of +1.96 and -1.96 from normal tables.

Step 3: Compute the test value.

(a) Combine the data from the two samples, arrange the combined data in order and rank each value. Be sure to indicate the group.

Time	8	9	10	11	12	13	14	15	15	16	16	17
Group	N	N	N	N	N	A	N	A	N	A	N	A
Rank	1	2	3	4	5	6	7	8.5	8.5	10.5	10.5	12.5

Time	17	18	18	19	19	21	22	24	25	26	28
Group	A	N	A	A	N	A	A	A	N	A	A
Rank	12.5	14.5	14.5	16.5	16.5	18	19	20	21	22	23

Table 5.4

(b) Sum the ranks of the group with the smaller sample size. The sample size for navy is smaller

R = 1+2+3+4+5+7+8.5+10.5+14.5+16.5+21=93

(c) Substitute in the formulas to find the test value

$$\mu_R = \frac{n_1(n_1+n_2+1)}{2} = \frac{11(11+12+1)}{2} = 132$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}} = \sqrt{\frac{(11)(12)(11+12+1)}{12}} = \sqrt{264} = 16.2$$

$$z = \frac{R-\mu_R}{\sigma_R} = \frac{93-132}{16.2} = -2.41$$

Step 4: Make the decision. The decision is to reject the null hypothesis, since -2.41 < -1.96

Step 5: Summarize the results. There is enough evidence to support the claim that there is a difference in times it takes the recruits to complete the course.

5.5 THE WILCOXON SIGNED-RANK TEST

When the samples are dependent, as they would be in a before-and-after test using the same subjects, the wilcoxon signed-rank test can be used in place of the t-test for dependent sample.

Example 29: in a large supermarket, the owner wishes to see whether the number of shop lifting incidents per day will change if the number of uniformed security officers is doubled. A sample of 7 days before security is increased and 7 days after the increase shows the number of shoplifting incidents.

	Number of shoplifting incidents	
Day	Before	After
Monday	7	5
Tuesday	2	3
Wednesday	3	4
Thursday	6	3
Friday	5	1
Saturday	8	6
Sunday	12	4

Table 5.5

Is there enough evidence to support the claim, at $\alpha = 0.05$ that there is a difference in the number of shoplifting incidents before and after the increase in security?

Solution:

5.6 THE KRUSKAL – WALLIS TEST

The U test is non parametric test for deciding whether or not two samples come from the same population. A generalization of this for k samples is to be provided by the Kruskal-Wallis H test. Suppose that we have k samples of sizes N_1, N_2… N_k with the total size of all samples taken together being by $N = N_1+N_2+\ldots\ldots\ldots +N_k$. suppose further that the data taken together are

ranked and that the sums of the ranks for the samples are R_1, R_2… R_k respectively. If we define statistic

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^{\,2}}{N_j} - 3(N+1)$$

$$= \frac{12}{N(N+1)} \left(\frac{R_1^{\,2}}{n_1} + \frac{R_2^{\,2}}{n_2} + \ldots\ldots\ldots\ldots + \frac{R_k^{\,2}}{n_k}\right) - 3(N+1)$$

Where

R_1 = sum of ranks in sample 1.

n_1 = size of sample 1 .etc.

Then it can be shown that the sampling distribution of H is very nearly a chi-square distribution with k-1 degrees of freedom, provided N_1, N_2, ….N_k are all at least 5.

The H test provides a non parametric method in the analysis of variance for one-way classification, or one- factor experiments, and generalizations can be made.

Example5.3: a company wishes to purchase one of five different machines: A, B, C, D or E. in an experiment designed to determine whether these is a performance difference between the machines, five experienced operators each work on the machines for equal times. Table below shows the no of units produced by each machine. Test the hypothesis that there is no difference between the machines at the (a) 0.05 and (b) 0.01 significance levels.

A	68	72	77	42	53
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

Table 5.6

Solution:

Since there are five samples (A, B, C, D, and E)

k=5

And since each sample consists of five values, we have $N_1 = N_2 = N_3 = N_4 = N_5$ and $N = 5 \times 5 = 25$.

By arranging all the values in increasing order of magnitude and assigning appropriate ranks to the ties, we replace above table with the table below.

						Sum of ranks
A	17.5	21	24	1	6.5	70
B	21	6.5	12	6.5	2.5	48.5
C	10	25	14	23	21	93
D	2.5	11	9	14	4	40.5
E	14	16	19	17.5	6.5	73

Table 5.7

$$R_1 = 70,\ R_2 = 48.5,\ R_3 = 93,\ R_4 = 40.5,\ R_5 = 75$$

Thus

$$H = \frac{12}{25(25+1)}\left[\frac{70^2}{5} + \frac{48.5^2}{5} + \frac{93^2}{5} + \frac{40.5^2}{5} + \frac{73^2}{5}\right] - 3(25+1)$$

$$= 6.44$$

For k-1 = 4 degrees of freedom at the 0.05 significance level.

We have $\chi^2_{0.95} = 9.49$ since $6.44 < 9.49$

We cannot reject the hypothesis of no difference between machines at the 0.05 level and certainly cannot reject it at the 0.01 level. Therefore, we can either accept the null hypothesis or reserve the judgment, that there is no difference between the machines at both levels.

Exercise 5

1. ***Referring to the table below, test the hypothesis H_o that there is no difference between machines I and II against the alternative hypothesis H_1 that there is a difference at the 0.05 significance level.***

Day	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***

Machine I	*47*	*56*	*54*	*49*	*36*	*48*	*51*	*38*	*61*	*49*	*56*	*52*
Machine II	*71*	*63*	*45*	*64*	*50*	*55*	*42*	*46*	*53*	*57*	*75*	*60*

Table 5.8

2. ***Work out problem 1. Using a normal approximation to the binomial distribution.***

3. ***A sample of 40 grades from a statewide examination is shown in table below. Test the hypothesis at the 0.05 significance level that the median grade for all participants is (a) 66 and (b) 75. Work the problem first by hand, supplying all the details for the sign test. Follow this with the MINITAB solution to the problem.***

71	*67*	*55*	*64*	*82*	*66*	*74*	*58*	*79*	*61*
78	*46*	*84*	*93*	*72*	*54*	*78*	*86*	*48*	*52*
67	*95*	*70*	*43*	*70*	*73*	*57*	*64*	*60*	*83*
73	*40*	*78*	*70*	*64*	*86*	*76*	*62*	*95*	*66*

Table 5.9

4. ***Referring to the table below, determine whether there is a difference the 0.05 significance level between cables made of alloy I and alloy II. Work the problem first by hand, supplying all the details for the Mann-Witney U test. Follow this with the MINITAB solution to the problem.***

Alloy I				*Alloy II*				
18.3	*16.4*	*22.7*	*17.8*	*12.6*	*14.1*	*20.5*	*10.7*	*15.9*
18.9	*25.3*	*16.1*	*24.2*	*19.6*	*12.9*	*15.2*	*11.8*	*14.7*

5.A company wishes to purchase one of five different machines A, B, C, D, OR E. in an experiment designed to determine whether there is a performance difference between the machines, five experienced operators each work on the machines for equal times. Table below shows the number of units produced by each machine. Test the hypothesis that there is no difference between the machines at the (a) 0.05 and (b) 0.01 significance levels. Work the problem first by hand, supplying all the details for the Kruskal-Wallis H test. Follow this with the MINITAB solution to the problem.

A	***68***	***72***	***77***	***42***	***53***
B	***72***	***53***	***63***	***53***	***48***
C	***60***	***82***	***64***	***75***	***72***
D	***48***	***61***	***57***	***64***	***50***
E	***64***	***65***	***70***	***68***	***53***

Table 5.11

Suggested References

1. DAVID S. Moore and George P. McCabe (1993), Introduction to the Practice of Statistics (second Edition) W.H. FREEMAN AND COMPANY NEW YORK. p562-570
2. Ajit C. Tamhane and Dorothy D. Dunlop (2000) Statistics and Data Analysis (from Elementary to Intermediate) PRENTICE HALL, Upper Saddle River p 580 -594
3. Allan G. Bluman, (2009), Elementary Statistics A Step by Step Approach, (seventh Edition) MAGRAW –HILL INTERNATIONAL EDITION, p669 -692
4. Murray R. Spiegel and Larry J. Stephens (2009) Statistics (fourth Edition) SCHAUM'S SERIES

SAMPLE PAPERS

Mount Kenya University

BBM 312: STATISTICS II ASSIGNMENT

Answer all questions

1. ***A random sample $X_1, X_{2,...} X_{10}$ is taken from a normal population with mean 100 and standard deviation 14.***
   a) ***Write down the distribution of $\overline{X}$, the mean of this sample.*** ***(2marks)***
   b) ***Find P(\| $\overline{X}$ -100\|) > 5*** ***(3marks)***

2. ***A random sample of the invoices, for books purchased by customers of a large bookshop, was classified by book cover (hardback, paperback) and type of book (novel, textbook, general interest). As part of the analysis of these invoices, an approximate χ^2 statistic was calculated and found to be 11.09.***
   Assuming that there was no need to amalgamate any of the classification, carry out an appropriate test to determine whether or not there was any association between book cover and type of book. State your hypotheses clearly and use a 5% level of significance. ***(6marks)***

3. ***For one of the activities at a gymnastics completion, 8 gymnasts were awarded marks out of 10 for each of artistic performance and technical ability. The results were as follows:-***

Gymnast	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Technical ability	*8.5*	*8.6*	*9.5*	*7.5*	*6.8*	*9.1*	*9.4*	*9.2*
Artistic performance	*6.2*	*7.5*	*8.2*	*6.7*	*6.0*	*7.2*	*8.0*	*9.1*

The value of the product moment correlation coefficient for these data is 0.774.
   a) ***Stating your hypothesis clearly and using a 1% level of significance, interpret this value.***
   b) ***Calculate the value of the rank correlation coefficient for these data*** ***(6marks)***

c) *Stating your hypotheses clearly and using a 1% level of significance, interpret this coefficient.* *(3marks)*

d) *Explain why the rank correlation coefficient might be better one to use with these data.*

4. *The results of an experiment to investigate the amount of chemical compound, y grams, that dissolved in 100grams of water as x^0C are recorded below*

x	*5*	*10*	*15*	*20*	*25*	*30*	*35*	*40*	*45*	*50*	*55*	*60*	
y	*9*	*11*	*13*	*17*	*21*	*24*	*27*	*29*	*31*	*35*	*38*	*42*	

$\sum x^2 = 16250$, $\sum y^2 = 8641$, $\sum xy = 11295$)

a) *Calculate S_{xx} , S_{yy} and S_{xy}* *(3marks)*

b) *Find the equation of the regression line of y on x in the form y =a + bx* *(3marks)*

c) *Test, at the 5% level of significance, whether or not there is evidence that the gradient of the regression line is positive.* *(8marks)*

d) *Explain how residuals can be used to decide on the suitability of a regression model.* *(3marks)*

5. *Past records from a large supermarket show that 20% of people who buy chocolate bars buy the family size bars. On one particular day a random sample of 30 people was taken from those that had bought chocolate bars and 2 of them were found to have bought a family size bar.*

a) *Test at the 5% significance level, whether or not the proportion p, of people who bought a family size bar of chocolate that day had decreased. State your hypotheses clearly.*

(6 marks)

The manager of the supermarket thinks that the probability of a person buying a gigantic chocolate bar is only 0.02. To test whether this hypothesis is true the manager decides to take a random sample of 200 people who bought bars.

b) *Find the critical region that would enable the manager to test whether or not there is evidence that the probability is different from 0.02. The probability of each tail should be as close to 2.5% as possible.* *(6 marks)*

c) *Write down the significance level of this test.* *(1mark)*

Mount Kenya University

UNIVERSITY EXAMINATIONS 2010

SCHOOL OF APPLIED SOCIAL SCIENCES

DEPARTMENT OF INFORMATION TECHNOLOGY

BACHELOR OF BUSINESS INFORMATION TECHNOLOGY

UNIVERSITY EXAMINATION AUGUST 2010

BBM 312: STATISTICS II

TIME 2HOURS
AUGUST SERIES

INSTRUCTIONS: ANSWER QUESTIONS ONE (COMPULSORY) AND ANY OTHER TWO QUESTIONS

QUESTION ONE (COMPULSORY) 30 MARKS

a) Define a statistic
(2 marks)

A random sample $x_1, x_2 \ldots\ldots x_n$ is taken from a population with unknown mean μ

b) For each of the following state with reason whether or not it is a statistic

i. $\frac{x_1 + x_4}{2}$ (2 marks)

ii. $\frac{\sum x^2}{N} - \mu^2$ (2marks)

c) Two judges in a beauty contest rank the ten competitors in the following order.

x	6	4	3	9	2	7	1	5	10	8
y	4	1	6	7	5	8	2	3	9	10

Do the two judges appear to agree in their standards? (6marks)

d) A random sample of size 36 was taken from a population distributed N (μ, 3.9^2). The value of the sample $\bar{x}$ was 15.6.

i. Find a 90% confidence interval for μ (5 marks)

It is believed that value of μ is 17.0

ii. Use your confidence interval to comment on this belief. (2marks)

e) A random sample $x_1, x_2 \ldots x_{10}$ is taken from a normal population with 100 and standard deviation 14.

f) A tennis coach believes that taller players are generally capable of hitting faster serves. To investigate this hypothesis he collects data on the 20 adult male players he coaches. The height, h, in metres and the speed of each player's fastest serve, v, in Km per hour were recorded and summarized as follows:-
$\sum h = 36.22$, $\sum v = 2275$, $\sum h^2 = 65.7396$, $\sum v^2 = 259853$ and $\sum hv = 4128.03$

i. Calculate the Pearson Moment Correlation Coefficient for these data. (4marks)

ii. Comment on the coach's hypothesis (2marks)

Question 2 (20marks)

a) Penshop have stores selling stationary in each of 6 towns. The population, P, in tens of thousands and monthly turnover, T, in thousands of Pounds for each of the shops are as recorded below

Town	Agnet	Bonns	Claster	Doggis	Edgeton	Figland
P(0000's)	3.2	7.6	5.2	9.0	8.1	4.8
T(000's)	11.1	12.4	13.3	19.3	17.9	11.8

i. Represent these data on a scatter diagram with T on the vertical axis (4 marks)

ii. Which town's shop might appear to be under achieving given the populations of the towns?
(2marks)

iii. Suggest two other factors that might affect each shop's turnover. (2 marks)

You may assume that $\sum P^2 = 264.69$, $\sum T^2 = 1286$ and $\sum PT = 574.25$.

b) Find the equation of the regression line of T on P
(8marks)

c) Estimate the monthly turnover that might be expected if a shop were opened in Gratton, a town with a population of 68000 (2marks)

d) Why might the management of Penshop be reluctant to use the regression line to estimate the monthly turnover they could expect if a shop were opened in Haggin, a town with a population of 172,000?

Question 3 (20marks)

a) Write shot notes on the following :-

i. Null and alternative hypothesis (2marks)

ii. One tailed and two tailed tests (2marks)

iii. Type I and Type II errors (2marks)

iv. Acceptance and rejection regions (2marks)

b) A sample of 100 motorcar Tyres has a mean of 20000Km and a standard deviation of 800Km. a second sample of 150 Tyres, has a mean life of 22000Km and a standard deviation of 900 Km.

i. Is it true to say that the two samples were drawn from the same population? (10marks)

ii. Which is your rejection or acceptance as you conclude about the two populations (2marks)

Question 4 (20 marks)

a) Explain briefly what you understand by the terms

i. Population

ii. Statistics (2marks)

b) A rugby player scores an average of 0.4 tries per march in which he plays

i. Find the probability that he scores 2 or more tries in a march (5marks)

The team's coach moves the player to a different position in the team believing he will then score more frequently. In the next five marches he scores 6 tries.

ii. Stating your hypothesis clearly, test at the 5% level of significance whether or not there is evidence of an increase in the number of tries the player scores per match as a result of playing in a different position (5marks)

c) A child welfare officer asserts that the mean sleep of young babies is 14 hours a day. A random sample of 64 babies shows that their mean sleep was only 13hours 30 minutes, with a standard deviation of 3 hours. At 5% level of significance, test the assertion that mean sleep of babies is less than 14 hours a day. (8marks)

Question 5 (20 Marks)

a) What is time series analysis? (3marks)

b) What is the moving averages system of forecasting (2marks)

c) The table below shows the number of bags of sugar per week

Week	1	2	3	4	5	6	7	8
No of Bags	340	330	323	343	350	345	365	342

Calculate thee weeks moving averages and six weeks moving averages (5marks)

d) A new diet program claims that participants wi8ll lose on average at least eight pounds during the week of the program. A random sample of 40 people participating in the program showed a sample mean weight loss of seven pounds. The sample standard deviation was 3.2 pounds.

i. What is the rejection rule with $\alpha = 0.05$ [3]

ii. What is your conclusion about the claim made by the diet program? [4]

iii. What is the p-value? [3]