CHAPTER EIGHT

ESTIMATION AND TEST OF HYPOTHESIS

Specific Objectives

At the end of this topic the trainee should be able to:

- Define estimation
- Differentiate between the two types of estimation
- Dertemine the sampling distribution of a statistics
- Determine the confidence interval for a parameter;
- Design a simple hypothesis;
- Define errors in hypothesis testing;
- Test various hypothesis.

Introduction

Statistical inference

It is the process of drawing conclusions about attributes of a population based upon information contained in a sample (taken from the population). It is divided into estimation of parameters and testing of hypothesis. Symbols for statistic of population parameters are as follows.

	Sample Statistic	Population Parameter
Arithmetic mean	$\bar{x}$	μ
Standard deviation	s	σ
Number of items	n	N

Statistical estimation

Def: It is the procedure of using statistic to estimate a population parameter. It is divided into point estimation (where an estimate of a population parameter is given by a single number) and interval estimation (where an estimate of a population is given by a range in which the parameter may be considered to lie) e.g. a bus meant to take a class of 100 students (population N) for trip has a limit to the maximum weight of 600kg of which it can carry, the teacher realizes he has to find out the weight of the class but without enough time to weigh everyone he picks 25 students selected at random (sample n = 25). These students are weighed and their average weight recorded as 64kg ($\overline{X}$ - mean of a sample) with a standard deviation (s), now using this the teacher intends to estimate the average weight of the whole class (μ – population mean) by using the statistical parameters standard deviation (s), and mean of the sample ($\bar{x}$).

Characteristic of a good estimator

(i) Unbiased: where the expected value of the statistic is equal to the population parameter e.g. if the expected mean of a sample is equal to the population mean

(ii) Consistency: where an estimator yields values more closely approaching the population parameter as the sample increases

(iii) Efficiency: where the estimator has smaller variance on repeated sampling.

(iv) Sufficiency: where an estimator uses all the information available in the data concerning a parameter

Confidence Interval

The interval estimate or a 'confidence interval' consists of a range (upper confidences limit and lower confidence limit) within which we are confident that a population parameter lays and we assign a probability that this interval contains the true population value

The confidence limits are the outer limits to a confidence interval. Confidence interval is the interval between the confidence limits. The higher the confidence level the greater the confidence interval. For example

A normal distribution has the following characteristic

i. Sample mean ± 1.960 σ includes 95% of the population
ii. Sample mean ± 2.575 σ includes 99% of the population

Sampling distribution

LARGE SAMPLES

These are samples that contain a sample size greater than 30(i.e. n>30)

(a) Estimation of population mean

Here we assume that if we take a large sample from a population then the mean of the population is very close to the mean of the sample

Steps to follow to estimate the population mean includes

i. Take a random sample of n items where (n>30)

ii. Compute sample mean ($\overline{X}$) and standard deviation (S)

iii. Compute the standard error of the mean by using the following formula

$$S_{\overline{x}} = \frac{s}{\sqrt{n}}$$

Where $S_{\overline{x}}$ = Standard error of mean

S = standard deviation of the sample

n = sample size

iv. Choose a confidence level e.g. 95% or 99%

v. Estimate the population mean as under

Population mean $\mu = \overline{\chi} \pm$ (appropriate number) $\times S_{\overline{x}}$

'Appropriate number' means confidence level e.g. at 95% confidence level is 1.96 this number is usually denoted by Z and is obtained from the normal tables.

Example

The quality department of a wire manufacturing company periodically selects a sample of wire specimens in order to test for breaking strength. Past experience has shown that the breaking strengths of a certain type of wire are normally distributed with standard deviation of 200 kg. A random sample of 64 specimens gave a mean of 6200 kgs. Find out the population mean at 95% level of confidence

Solution

Population mean = $\overline{\chi} \pm 1.96\, S_{\overline{x}}$

Note that sample size is already n > 30 whereas s and $\overline{x}$ are given thus step i), ii) and iv) are provided.

Here: $\overline{X}$ = 6200 kgs

$$S_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{200}{\sqrt{64}} = 25$$

Population mean = 6200 ± 1.96(25)

= 6200 ± 49

= 6151 to 6249

At 95% level of confidence, population mean will be in between 6151 and 6249

Finite Population Correction Factor (FPCF)

If a given population is relatively of small size and sample size is more than 5% of the population then the standard error should be adjusted by multiplying it by the finite population correction factor

FPCF is given by $= \sqrt{\frac{N-n}{n-1}}$

Where N = population size

n = sample size

Example

A manager wants an estimate of sales of salesmen in his company. A random sample 100 out of 500 salesmen is selected and average sales are found to be Shs. 75,000. If a sample standard deviation is Shs. 15000 then find out the population mean at 99% level of confidence

Solution

Here N = 500, n = 100, $\overline{X}$ = 75000 and S = 15000

Now

Standard error of mean

$$= S_{\overline{x}} = \frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{n-1}}$$

$$= \frac{15000}{\sqrt{100}} \times \sqrt{\frac{(500-100)}{(500-1)}}$$

$$= \frac{15000}{10} \times \sqrt{\frac{400}{499}}$$

$$= \frac{15000}{10}(0.895)$$

$S_{\overline{x}}$ = 1342.50 at 99% level of confidence

Population mean $= \overline{X} \pm 2.58\, S_{\overline{x}}$

=shs 75000 ± 2.58(1342.50)

=shs 75000 ± 3464

= Shs 71536 to 78464

(b) Estimation of difference between two means

We know that the standard error of a sample is given by the value of the standard deviation (σ) divided by the square root of the number of items in the sample ($\sqrt{n}$).

But, when given two samples, the standard errors is given by

$$S_{(\bar{x}A-\bar{x}B)} = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}$$

Also note that we do estimate the interval not from the mean but from the difference between the two sample means i.e. $\left(\overline{X}_A - \overline{X}_B\right)$.

The appropriate number of confidence level does not change

Thus the confidence interval is given by;

$$\left(\overline{X}_A - \overline{X}_B\right) \pm \text{Confidence level } S_{(\overline{x}A - \overline{x}B)}$$

$$= \left(\overline{X}_A - \overline{X}_B\right) \pm Z\ S_{(\overline{x}A - \overline{x}B)}$$

Example

Given two samples A and B of 100 and 400 items respectively, they have the means $\overline{X}_1$ = 7 ad $\overline{X}_2$ = 10 and standard deviations of 2 and 3 respectively. Construct confidence interval at 70% confidence level?

Solution

Sample	A	B
	$\overline{X}_1 = 7$	$\overline{X}_2 = 10$
	$n_1 = 100$	$n_2 = 400$
	$S_1 = 2$	$S_2 = 3$

The standard error of the samples A and B is given by

$$S_{(\overline{x}A - \overline{x}B)} = \sqrt{\frac{4}{100} + \frac{9}{400}}$$

$$= \sqrt{\frac{25}{400}} = \frac{5}{20}$$

$$= ¼ = 0.25$$

At 70% confidence level, then appropriate number is equal to 1.04 (as read from the normal tables)

$$\overline{X}_1 - \overline{X}_2 = 7 - 10 = -3 = 3$$

We take the absolute value of the difference between the means e.g. the value of $|X|$ = absolute value of X i.e. a positive value of X.

Confidence interval is therefore given by

= 3± 1.04 (0.25) From the normal tables a z value of 1.04 gives a value of 0.7.

= 3± 0.26

= 3.26 and 2.974

Thus 2.974 ≤ X ≤ 3.26

Example 2
A comparison of the wearing out quality of two types of tyres was obtained by road testing. Samples of 100 tyres were collected. The miles traveled until wear out were recorded and the results given were as follows

Tyres	T1	T2
Mean	$\overline{X}_1$ = 26400 miles	$\overline{X}_2$ = 25000 miles
Variance	S^2_1= 1440000 miles	S^2_2= 1960000 miles

Find a confidence interval at the confidence level of 70%

Solution
$\overline{X}_1$ = 26400
$\overline{X}_2$ = 25000
Difference between the two means
$\left(\overline{X}_1 - \overline{X}_2\right)$ = (26400 – 25000)
= 1,400

Again we take the absolute value of the difference between the two means
We calculate the standard error as follows

$$S_{(\overline{x}A - \overline{x}B)} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$= \sqrt{\frac{1,440,000}{100} + \frac{1,960,000}{100}}$$

= 184.4
Confidence level at 70% is read from the normal tables as 1.04 (Z = 1.04).
Thus the confidence interval is calculated as follows
= 1400 ± (1.04) (184.4)

= 1400 ± 191.77

or (1400 – 191.77) to (1400 + 191.77)

1,208.23 ≤ X ≤ 1591.77

c) Estimation of population proportions
This type of estimation applies at the times when information cannot be given as a mean or as a measure but only as a fraction or percentage

The sampling theory stipulates that if repeated large random samples are taken from a population, the sample proportion "p' will be normally distributed with mean equal to the population proportion and standard error equal to

$S_p = \sqrt{\frac{Pq}{n}}$ = Standard error for sampling of population proportions

Where n is the sample size and q = 1 – p.

The procedure for estimating a proportion is similar to that for estimating a mean, we only have a different formula for calculating standard.

Example 1

In a sample of 800 candidates, 560 were male. Estimate the population proportion at 95% confidence level.

Solution

Here

Sample proportion (P) = $\frac{560}{800}$ = 0.70

q = 1 – p = 1 – 0.70 = 0.30

n = 800

$\sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.70)(0.30)}{800}}$

S_p = 0.016

Population proportion

= P ± 1.96 S_p where 1.96 = Z.

= 0.70 ± 1.96 (0.016)

= 0.70 ± 0.03

= 0.67 to 0.73

= between 67% to 73%

Example 2

A sample of 600 accounts was taken to test the accuracy of posting and balancing of accounts where in 45 mistakes were found. Find out the population proportion. Use 99% level of confidence

Solution

Here

n = 600; p = $\frac{45}{600}$ = 0.075

q = 1 – 0.075 = 0.925

$S_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.075)(0.925)}{600}}$

= 0.011

Population proportion

= P ± 2.58 (S_p)

= 0.075 ± 2.58 (0.011)

= 0.075 ± 0.028

= 0.047 to 0.10

= between 4.7% to 10%

d) <u>Estimation of difference between population proportions</u>

Let the two proportions be given by P_1 and P_2, respectively

Then the difference (absolute) between the two proportions is given by (P_1 – P_2)

The standard error is given by

$S_{(P1-P2)} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$ where $p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$ and q = 1 - p

Then given the confidence level, the confidence interval between the two population proportions is given by

(P_1 – P_2) ± Confidence level $S_{(P1-P2)}$

= (P_1 - P_2) ± Z $\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$

Where P = $\frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$ always remember to convert P_1 & P_2 to P.

2. SMALL SAMPLES

(a) <u>Estimation of population mean</u>

If the sample size is small (n<30) the arithmetic mean of small samples are not normally distributed. In such circumstances, students'
t-distribution must be used to estimate the population mean.
In this case

Population mean μ = $\overline{X} \pm ts_{\bar{x}}$

$\overline{X}$ = Sample mean

$S_{\bar{x}} = \frac{s}{\sqrt{n}}$

S = standard deviation of samples = $\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}}$ for small samples.

n = sample size
v = n – 1 degrees of freedom.

The value of t is obtained from students t-distribution tables for the required confidence level

Example
A random sample of 12 items is taken and is found to have a mean weight of 50 grams and a standard deviation of 9 grams
What is the mean weight of population?

a) with 95% confidence
b) with 99% confidence

Solution

$\overline{X} = 50;$ S = 9; v = n - 1 = 12 - 1 = 11; $S_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{9}{\sqrt{12}}$

μ = x' ± $ts_{\bar{x}}$

At 95% confidence level

μ = 50 ± 2.262$\left(\frac{9}{\sqrt{12}}\right)$

= 50 ± 5.72 grams

Therefore we can state with 95% confidence that the population mean is between 44.28 and 55.72 grams
At 99% confidence level

μ = 50 ± 3.25$\left(\frac{-9}{\sqrt{12}}\right)$

= 50 ± 8.07 grams

Therefore we can state with 99% confidence that the population mean is between 41.93 and 58.07 grams

Note: To use the t distribution tables it is important to find the degrees of freedom ($v = n - 1$). In the example above $v = 12 - 1 = 11$
From the tables we find that at 95% confidence level against 11 and under 0.05, the value of $t = 2.201$

HYPOTHESIS TESTING

Definition
- A hypothesis is a claim or an opinion about an item or issue. Therefore it has to be tested statistically in order to establish whether it is correct or not correct
- Whenever testing a hypothesis, one must fully understand the 2 basic hypotheses to be tested namely
  i. The null hypothesis (H_0)
  ii. The alternative hypothesis(H_1)

The null hypothesis
This is the hypothesis being tested, the belief of a certain characteristic e.g. Kenya Bureau of Standards (KBS) may walk to a sugar making company with an intention of confirming that the 2kgs bags of sugar produced are actually 2kgs and not less, they conduct hypothesis testing with the null hypothesis being: H_0 = each bag weighs 2kgs. The testing will set out to confirm this or to refute it.

The alternative hypothesis
While formulating a null hypothesis we also consider the fact that the belief might be found to be untrue hence we will reject it. We therefore formulate an alternative hypothesis which is a contradiction to the null hypothesis, thus when we reject the null hypothesis we accept the alternative hypothesis.
In our example the alternative hypothesis would be
H_1 = each bag does not weigh 2kg

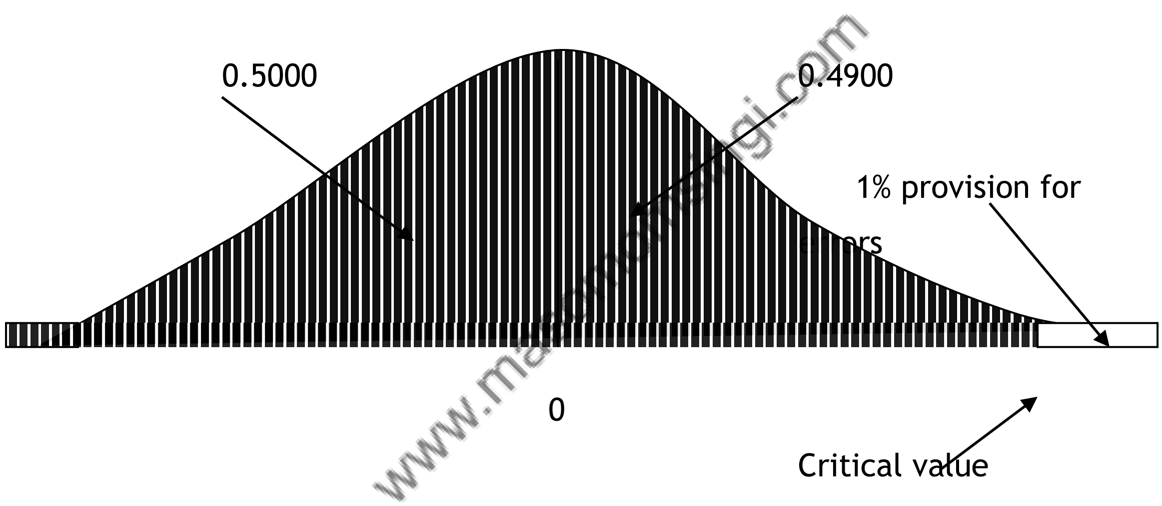
Acceptance and rejection regions
All possible values which a test statistic may either assume consistency with the null hypothesis (acceptance region) or lead to the rejection of the null hypothesis (rejection region or **critical region)**

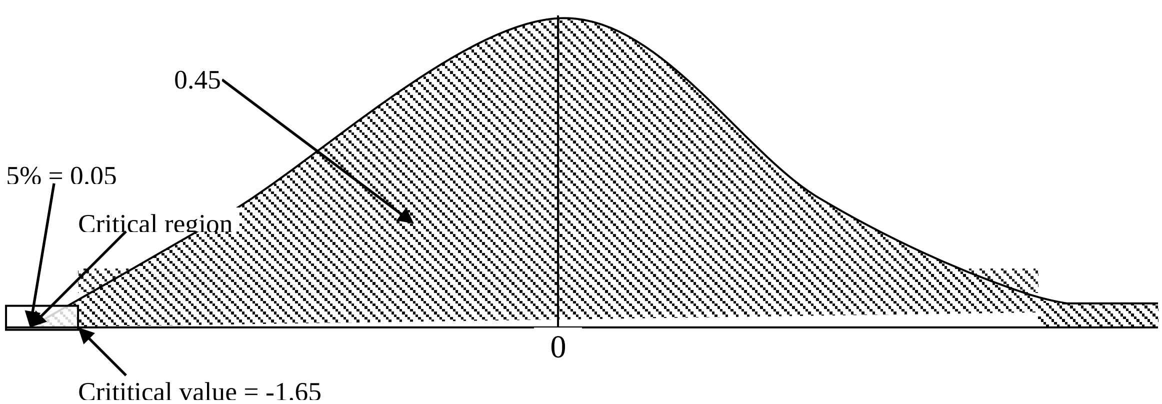
The values which separate the rejection region from the acceptance region are called **critical values**

Type I and type II errors
While testing hypothesis (H_0) and deciding to either accept or reject a null hypothesis, there are four possible occurrences.

a) Acceptance of a true hypothesis (correct decision) – accepting the null hypothesis and it happens to be the correct decision. Note that statistics does not give absolute information, thus its conclusion could be wrong only that the probability of it being right are high.
b) Rejection of a false hypothesis (correct decision).
c) Rejection of a true hypothesis – (incorrect decision) – this is called type I error, with probability = α.
d) Acceptance of a false hypothesis – (incorrect decision) – this is called type II error, with probability = β.

<u>Levels of significance</u>
A level of significance is a probability value which is used when conducting tests of hypothesis. A level of significance is basically the probability of one making an <u>incorrect decision</u> after the statistical testing has been done. Usually such probability used are very small e.g. 1% or 5%

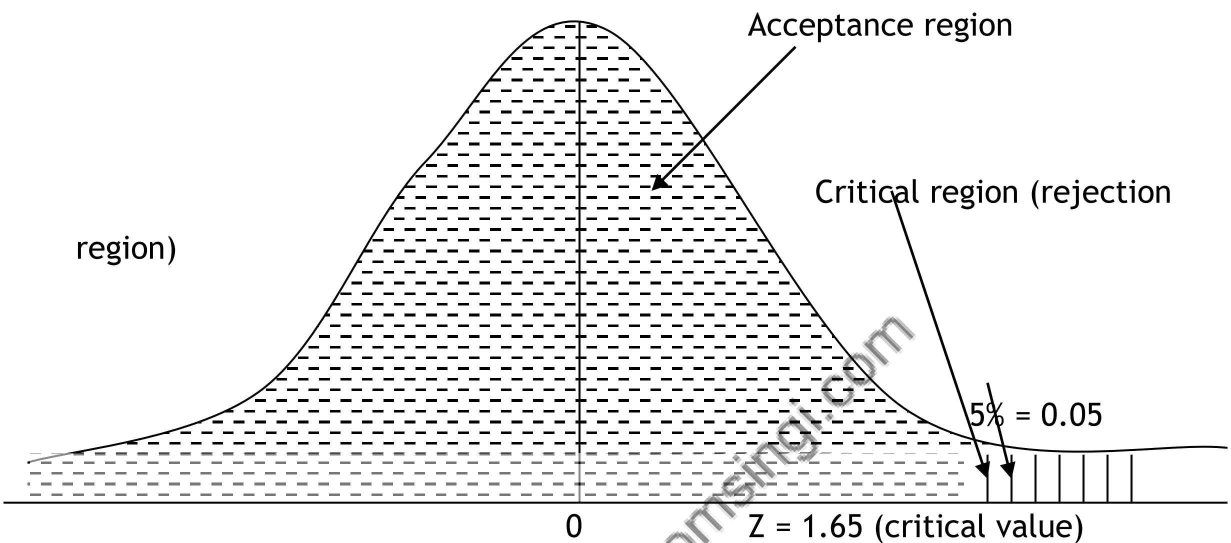




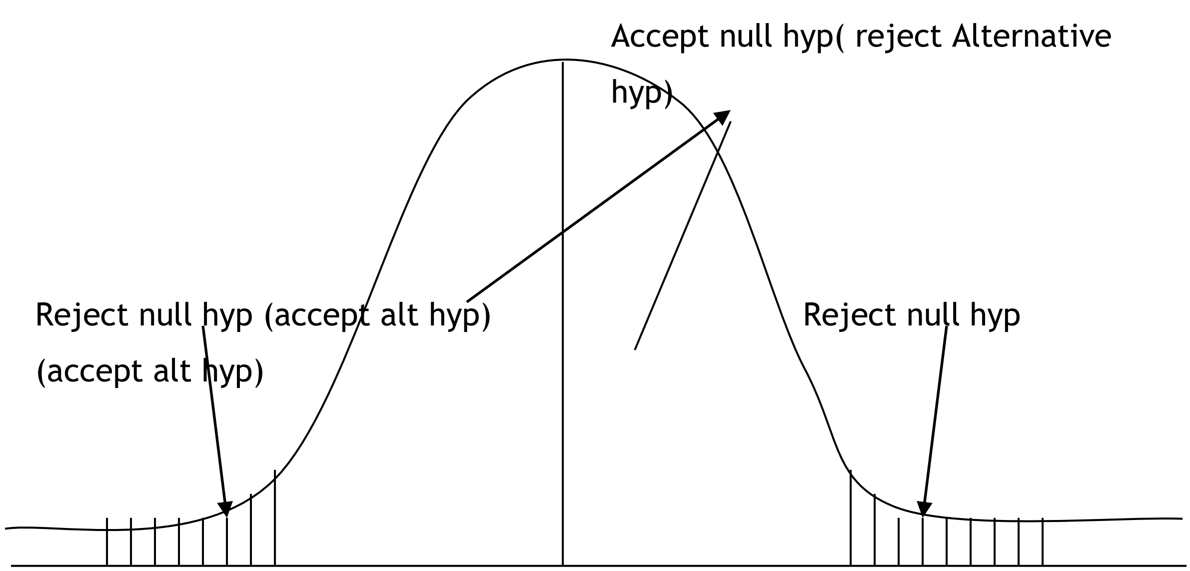
NB: If the standardized value of the mean is less than –1.65 we reject the null hypothesis (H_0) and accept the alternative Hypothesis (H_1) but if the standardized value of the mean is more than –1.65 we accept the null hypothesis and reject the alternative hypothesis

The above sketch graph and level of significance are applicable when the sample mean is < (i.e. less than the population mean)

The following is used when sample mean > population mean



NB: If the sample means standardized value < 1.65, we accept the null hypothesis but reject the alternative. If the sample means value > 1.65 we reject the null hypothesis and accept the alternative hypothesis
The above sketch is normally used when the sample mean given is greater than the population mean



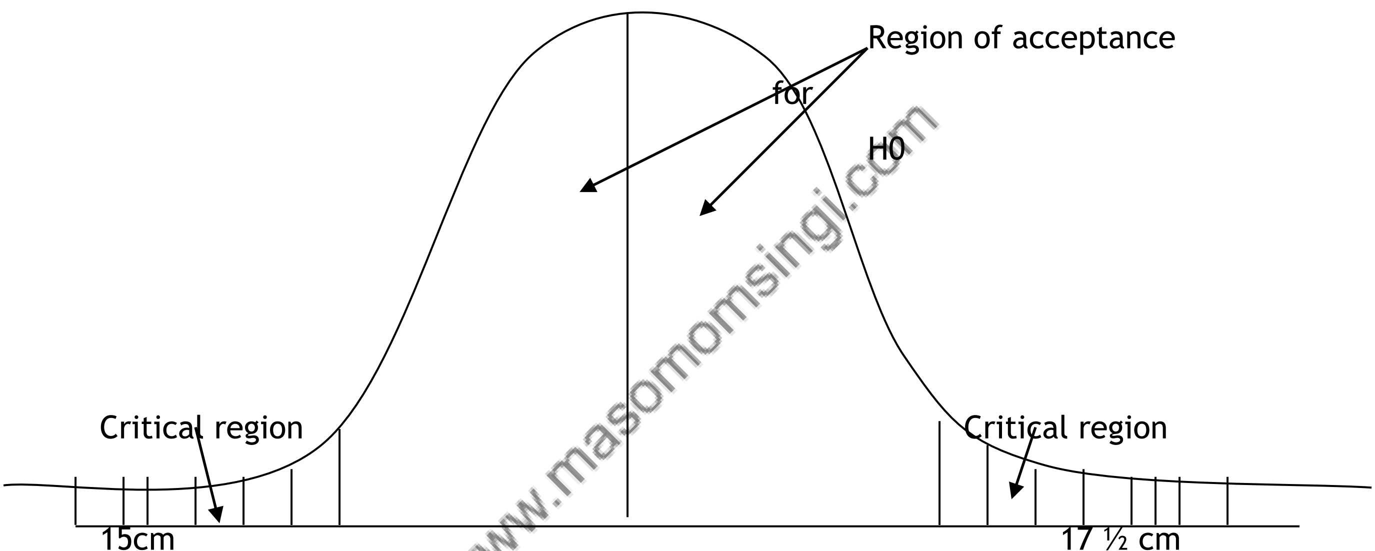
0.05% = 0.05 0.495 0.495 0.5% = 0.05

-2.58 +2.58

NB: if the standardized value of the sample mean is between –2.58 and +2.58 accept the null hypothesis but otherwise reject it and therefore accept the alternative hypothesis

TWO TAILED TESTS

A two tailed test is normally used in statistical work(tests of significance) e.g. if a complaint lodged by the client is about a product not meeting certain specifications i.e. the item will generate a complaint if its measurements are below the **lower tolerance limit** or above the **upper tolerance limit**

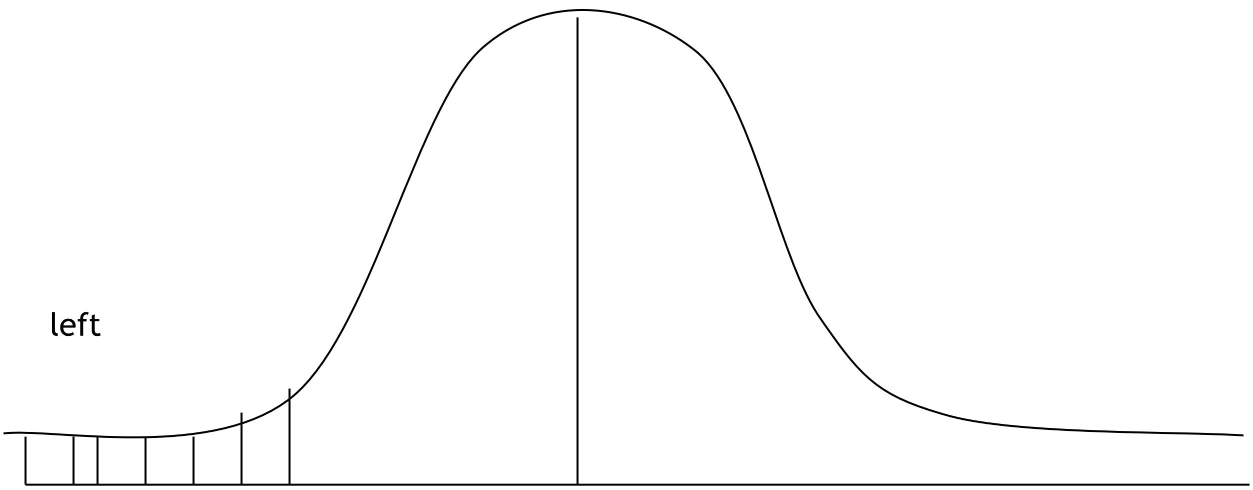


NB: Alternative hypothesis is usually rejected if the standardized value of the sample mean lies beyond the tolerance limits (15cm and 17 ½ cm).

ONE TAILED TEST

This is a test where the alternative hypothesis (H_1:) is only concerned with one of the tails of the distribution e.g. to test a business complaint if the complaint is above the measurements of item being shorter than is required.

E.g. a manufacturer of a given brand of bread may state that the average weight of the bread is 500 gms but if a consumer takes a sample and weighs each of the pieces of bread and happens to have a mean of 450 gms he will definitely complain about the bread which is underweight. The statistical analysis to be done will concentrate on the left tail of the normal distribution in which one will have to establish whether 450 gms being less than 500g is statistically significant. Such a test therefore is referred to as one tailed test.



On the other hand the test may compuliate on the right hand tail of the normal distribution when this happens the major complaint is likely to do with oversize items bought. Therefore the test is known as one tailed as the focus is on one end of the normal distribution.

	Number of standard errors	
	Two tailed test	One tailed test
5% level of significance	1.96	1.65
1% level of significance	2.58	2.33

HYPOTHESIS TESTING PROCEDURE

Whenever a business complaint comes up there is a recommended procedure for conducting a statistical test. The purpose of such a test is to establish whether the null hypothesis or alternative hypothesis is to be accepted.

The following are steps normally adopted

1. Statement of the null and alternative hypothesis
2. Statement of the level of significance to be used.
3. Statement about the test statistic i.e. what is to be tested e.g. the sample mean, sample proportion, difference between sample means or sample proportions
4. Type of test whether two tailed or one tailed.
5. Statement on critical values using the appropriate level of significance
6. Standardizing the test statistic
7. Conclusion showing whether to accept or reject the null hypothesis

STANDARD HYPOTHESIS TESTS

In principal, we can test the significance of any statistic related to any probability distribution. However we will be interested in a few standard cases. The sample statistics mean, proportion and variance, are related to the normal, t, F, and chi squared distributions

Thus

1. **Normal test**

   Test a sample mean ($\overline{X}$) against a population mean (μ) (where samples size n > 30 and population variance σ^2 is known) and sample proportion, P(where sample size np >5 and nq >5 since in this case the normal distribution can be used to approximate the binomial distribution

2. t test

Tests a sample mean ($\overline{X}$) against a population mean and especially where the population variance is unknown and n < 30.

3. Variance ratio test or f test

It is used to compare population variances and it is used with samples of any size drawn from normal populations.

4. Chi squared test

It can be used to test the association between attributes or the goodness of fit of an observed frequency distribution to a standard distribution

Example 1

A certain NGO carried out a survey in a certain community in order to establish the average at which the girls are married. The results of the survey indicated that the marriage age for the girls is 19 years

In order to establish the validity of the mean marital age, a sample of 50 women was interviewed and the average age indicated that they got married at the age of 16 years. However the different ages at which they were married differed with the standard deviation of 2.1years

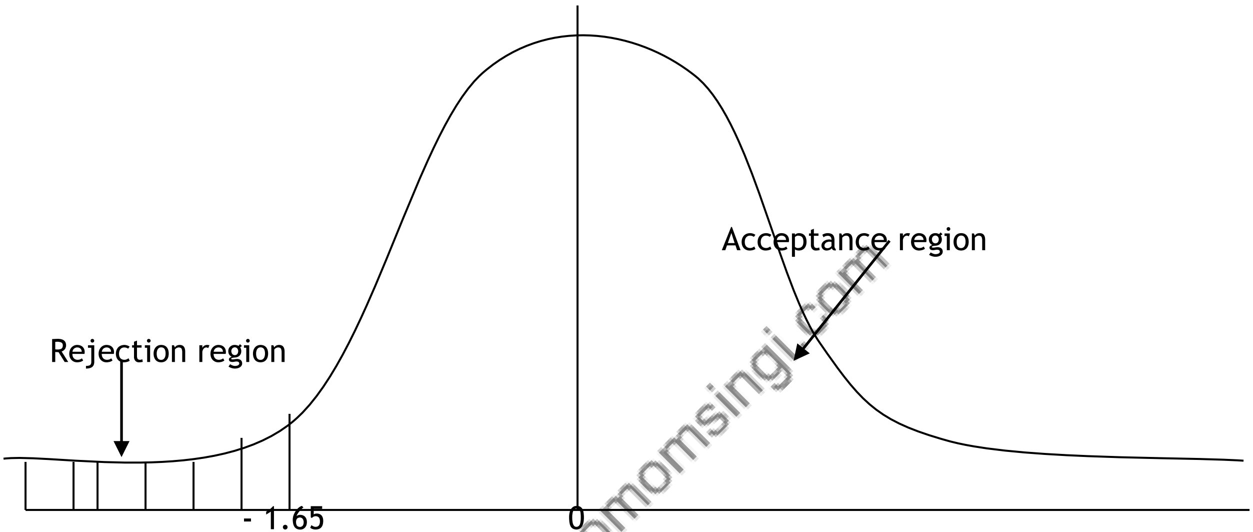
The sample data indicates that the marital age is less 19 years. Is this conclusion true or not?

Required

Conduct a statistical test to either support the above conclusion drawn from the sample statistics i.e. the marriage age is less than 19 years, use a level of significance of 5%

Solution

1. Null hypothesis
   H_0: μ (mean marital age) = 19 years
   Alternative hypothesis H_1: μ (mean marital age) < 19 years
2. The level of significance is 5%
3. The test statistics is the sample mean age, $\overline{X}$ = 16 years
4. The critical value of the one tailed test (one tailed because the alternative hypothesis is an inequality) at 5% level of significance is –1.65



5. The standardizes value of the sample mean is

   $$Z = \frac{\overline{X} - \mu}{S_{\overline{X}}} \quad \text{where } S_{\overline{x}} = \frac{S}{\sqrt{n}}$$

   Where, $\overline{X}$ = Sample mean
   μ = Population mean
   S = sample standard deviation
   n = sample size
   z = standard value (as per computation)

   The standard value Z must fall within the acceptance region for us to accept the null hypothesis. Thus it must be > - 1.65 otherwise we accept the alternative hypothesis.

   $$Z = \frac{16-19}{\frac{2.1}{\sqrt{50}}} = -10.1$$

6. Since –10.1 < -1.65, we reject the null hypothesis but accept the alternative hypothesis at 5% level of significance i.e. the marriage age in this community is significantly lower than 19 years

Example 2

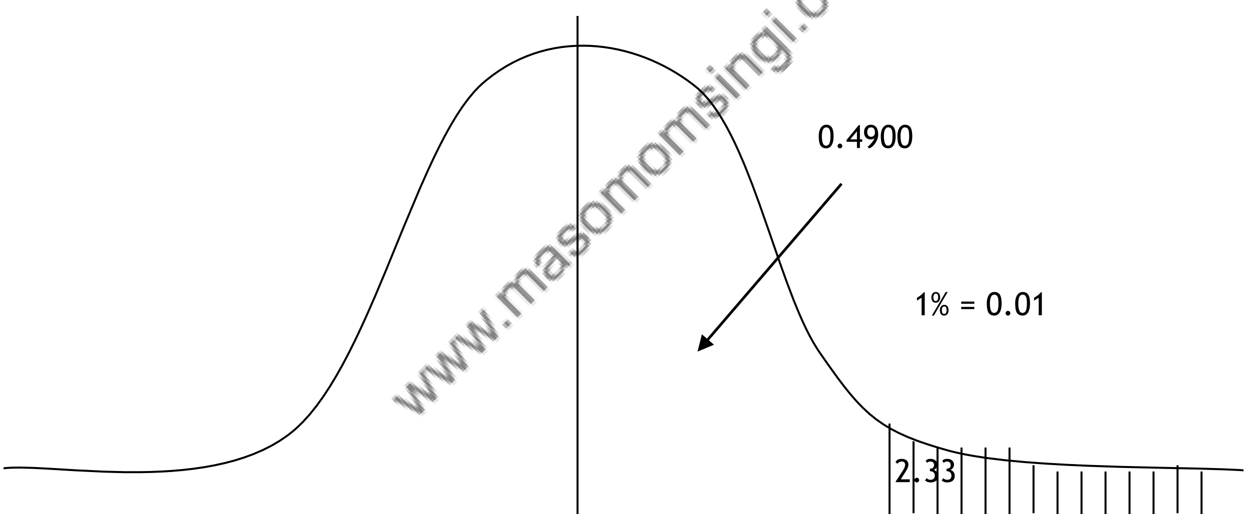
A foreign company which manufactures electric bulbs has assured its customers that the lifespan of the bulbs is 28 month with a standard deviation of 4months

Recently the company embarked on a quality improvement research for their product. After the research using new technology, a sample of 70 bulbs was tested and they gave a mean lifespan of 30.2 months

Does this justify the research undertaken? Use 1% level of significance to conduct a statistical test in order to establish the truth about the above question.

Testing procedure

1. Null hypothesis H_0: $\mu = 28$
   Alternative hypothesis H_1: $\mu > 28$
2. The level of significance is 1% (one tailed test)
3. The test statistics is the sample mean age, x' = 30.2
4. The critical value of the one tailed test at 5% level of significance is + 2.33



5. The standardized value of the sample mean is

$$Z = \frac{\overline{X} - \mu}{S_{\bar{x}}} = \frac{30.2 - 28}{\frac{4}{\sqrt{70}}} = 4.6$$

6. Since 4.6 > 2.33, we reject the null hypothesis but accept the alternative hypothesis at 1% level of significance i.e. the new sample mean life span is statistically significant higher than the population mean
   Therefore the research undertaken was worth while or justified

Example 3

A construction firm has placed an order that they require a consignment of wires which have a mean length of 10.5 meters with a standard deviation of 1.7 m

The company which produces the wires delivered 90 wires, which had a mean length of 9.2 m., The construction company rejected the consignment on the grounds that they were different from the order placed.

Required

Conduct a statistical test to indicate whether you support or not support the action taken by the construction company at 5% level of significance.

Solution

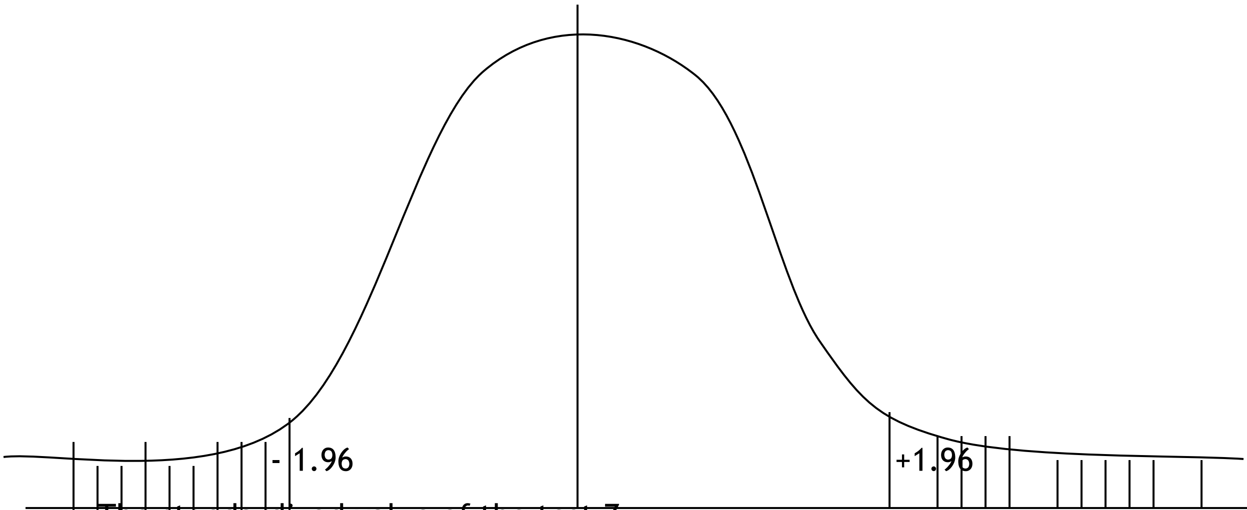
Null hypothesis $\mu = 10.5$ m

Alternative hypothesis $\mu \neq 10.5$ m

Level of significance be 5%

The test statistics is the sample mean $\overline{X}$ = 9.2m

The critical value of the two tailed test at 5% level of significance is $\pm$ 1.96 (two tailed test).



The standardized value of the test Z =

$$Z = \left|\frac{\bar{X} - \mu}{S_{\bar{X}}}\right| = \frac{9.2 - 10.5}{\frac{1.7}{\sqrt{90}}} = -7.25$$

Since 7.25 < 1.96, reject the null hypothesis but accept the alternative hypothesis at 5% level of significance i.e. the sample mean is statistically different from the consignment ordered by the construction company. Therefore support the action taken by the construction company

TESTING THE DIFFERENCE BETWEEN TWO SAMPLE MEANS (LARGE SAMPLES)

A large sample is defined as one which contains 30 or more items ($n \geq 30$) Where n is the sample size

In a business those involved are constantly observant about the standards or specifications of the item which they sell e.g. a trader may receive a batch of items at one time and another batch at a later time at the end he may have concluded that the two samples are different in certain specifications e.g. mean weight mean lifespan, mean length e.t.c. further it may become necessary to establish whether the observed differences are statistically significant or not. If the differences are statistically significant then it means that such differences must be explained i.e. there are known causes but if they are not statistically significant then it means that the difference observed have no known causes and are mainly due to chance

If the differences are established to be statistically significant then it implies that the complaints, which necessitated that kind of test, are justified

Let X_1 and X_2 be any two samples whose sizes are n_1 and n_2 and mean $\bar{X}_1$ and $\bar{X}_2$. Standard deviation S_1 and S_2 respectively. In order to test the difference between the two sample means, we apply the following formulas

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{S\left(\overline{X}_1 - \overline{X}_2\right)} \quad \text{where } S\left(\overline{X}_1 - \overline{X}_2\right) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Example 1

An agronomist was interested in the particular fertilizer yield output. He planted maize on 50 equal pieces of land and the mean harvest obtained later was 60 bags per plot with a standard deviation of 1.5 bags. The crops grew under natural circumstances and conditions without the soil being treated with any fertilizer. The same agronomist carried out an alternative experiment where he picked 60 plots in the same area and planted the same plant of maize but a fertilizer was applied on these plots. After the harvest it was established that the mean harvest was 63 bags per plot with a standard deviation of 1.3 bags

Required

Conduct a statistical test in order to establish whether there was a significant difference between the mean harvest under the two types of field conditions. Use 5% level of significance.

Solution

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

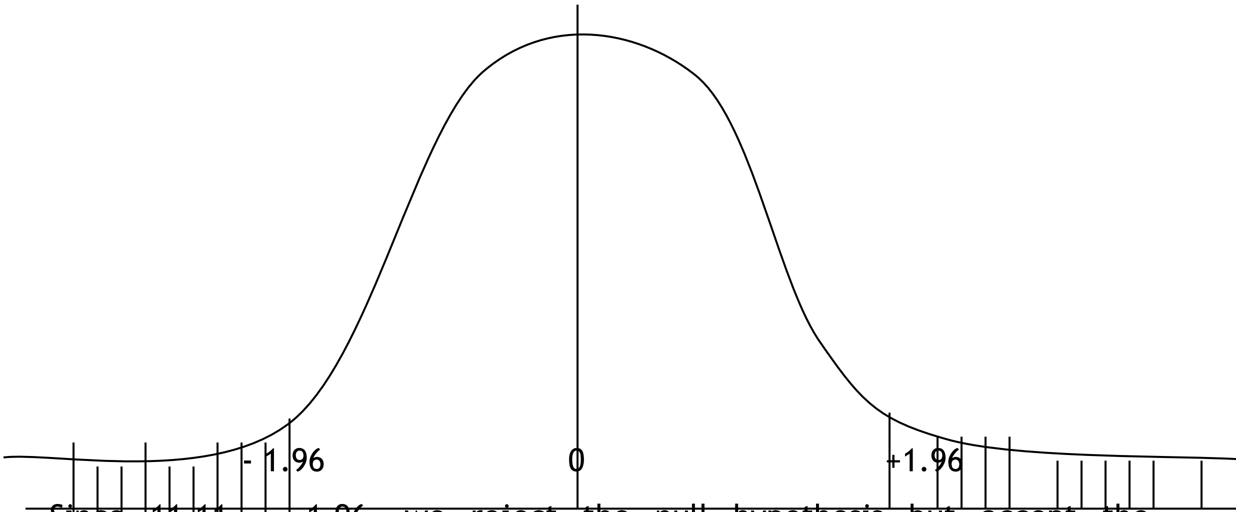
Critical values of the two tailed test at 5% level of significance are 1.96

The standardized value of the difference between sample means is given by Z where

$$Z = \left|\frac{\overline{X}_1 - \overline{X}_2}{S\left(\overline{X}_1 - \overline{X}_2\right)}\right| \quad \text{where } S\left(\overline{X}_1 - \overline{X}_2\right) = \sqrt{\frac{1.5^2}{50} + \frac{1.3^2}{60}}$$

$$Z = \left|\frac{(60-63)}{\sqrt{0.045+0.028}}\right|$$

$$= 11.11$$



Since 11.11 < -1.96, we reject the null hypothesis but accept the alternative hypothesis at 5% level of significance i.e. the difference between the sample mean harvest is statistically significant. This implies that the fertilizer had a positive effect on the harvest of maize

Note: You don't have to illustrate your solution with a diagram.

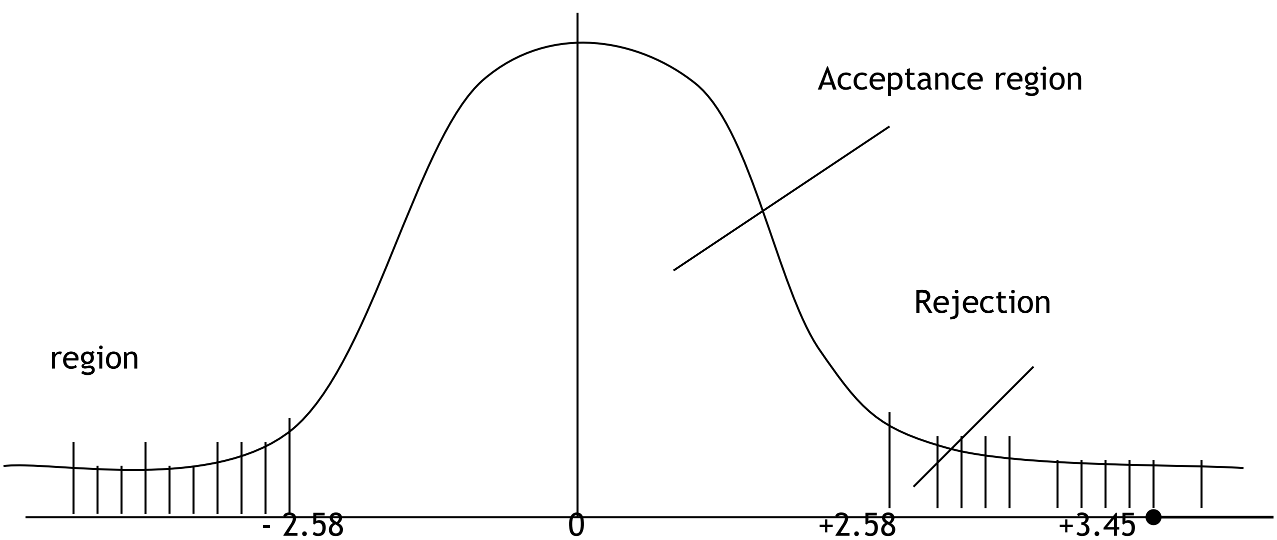
Example 2
An observation was made about reading abilities of males and females. The observation lead to a conclusion that females are faster readers than males. The observation was based on the times taken by both females and males when reading out a list of names during graduation ceremonies.
In order to investigate into the observation and the consequent conclusion a sample of 200 men were given lists to read. On average each man took 63 seconds with a standard deviation of 4 seconds
A sample of 250 women were also taken and asked to read the same list of names. It was found that they on average took 62 seconds with a standard deviation of 1 second.

Required
By conducting a statistical hypothesis testing at 1% level of significance establish whether the sample data obtained does support earlier observation or not

Solution
H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$
Critical values of the two tailed test is at 1% level of significance is 2.58.

$$Z = \left|\frac{\overline{X}_1 - \overline{X}_2}{S\left(\overline{X}_1 - \overline{X}_2\right)}\right|$$

$$Z = \left|\frac{63-62}{\sqrt{\frac{4^2}{200}+\frac{1^2}{250}}}\right| = 3.45$$



Since 3.45 > 2.33 reject the null hypothesis but accept the alternative hypothesis at 1% level of significance i.e. there is a significant difference between the reading speed of Males and females, thus females are actually faster readers.

TEST OF HYPOTHESIS ON PROPORTIONS

This follows a similar method to the one for means exept that the standard error used in this case:

$$Sp = \sqrt{\frac{Pq}{n}}$$

Z score is calculated as, $Z = \frac{P - \Pi}{Sp}$ Where P = Proportion found in the sample.

Π – the hypothetical proportion.

Example

A member of parliament (MP) claims that in his constituency only 50% of the total youth population lacks university education. A local media company wanted to acertain that claim thus they conducted a survey taking a sample of 400 youths, of these 54% lacked university education.

Required:

At 5% level of significance confirm if the MP's claim is wrong.

Solution.

Note: This is a two tailed tests since we wish to test the hypothesis that the hypothesis is different (≠) and not against a specific alternative hypothesis e.g. < less than or > more than.

H_0 : π = 50% of all youth in the constituency lack university education.

H_1 : π ≠ 50% of all youth in the constituency lack university education.

$$Sp = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.5x0.5}{400}} = 0.025$$

$$Z = \left|\frac{0.54 - 0.50}{0.025}\right| = 1.6$$

at 5% level of significance for a two-tailored test the critical value is 1.96 since calculated Z value < tabulated value (1.96).

i.e. 1.6 < 1.96 we accept the null hypothesis.

Thus the MP's claim is accurate.

HYPOTHESIS TESTING OF THE DIFFERENCE BETWEEN PROPORTIONS

Example

Ken industrial manufacturers have produced a perfume known as "fianchetto." In order to test its popularity in the market, the manufacturer carried a random survey in Back rank city where 10,000 consumers were interviewed after which 7,200 showed preference. The manufacturer also moved to area Rook town where he interviewed 12,000 consumers out of which 1,0000 showed preference for the product.

Required

Design a statistical test and hence use it to advise the manufacturer regarding the differences in the proportion, at 5% level of significance.

Solution

$H_0 : \pi_1 = \pi_2$
$H_1 : \pi_1 \neq \pi_2$

The critical value for this two tailed test at 5% level of significance = 1.96.

Now Z = $\left|\frac{(P_1 - P_2) - (\Pi_1 - \Pi_2)}{S(P_1 - P_2)}\right|$

But since the null hypothesis is $\pi_1 = \pi_2$, the second part of the numerator disappear i.e.
$\pi_1 - \pi_2 = 0$ which will always be the case at this level.

Then Z = $\left|\frac{(P_1 - P_2)}{S(P_1 - P_2)}\right|$

Where;

	Sample 1	Sample 2
Sample size	n_1 = 10,000	n_2 = 12,000
Sample proportion of success	P_1 =0.72	P_2 = 0.83
Population proportion of success.	Π_1	Π_2

Now $S(p_1 - p_2) = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$

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

And q = 1 – p

$\therefore$ in our case

$$P = \frac{10,000(0.72)+12,000(0.83)}{10,000+12,000}$$

$$= \frac{84,000}{22,000}$$

= 0.78

$\therefore$ q = 0.22

$$S(P_1 - P_2) = \sqrt{\frac{0.78(0.22)}{10,000} + \frac{0.78(0.22)}{12,000}}$$

= 0.00894

$$Z = \left|\frac{0.72 - 0.83}{0.00894}\right| = 12.3$$

Since 12.3 > 1.96, we reject the null hypothesis but accept the alternative. The differences between the proportions are statistically significant. This implies that the perfume is much more popular in Rook town than in back rank city.

HYPOTHESIS TESTING ABOUT THE DIFFERENCE BETWEEN TWO PROPORTIONS

Is used to test the difference between the proportions of a given attribute found in two random samples.

The null hypothesis is that there is no difference between the population proportions. It means two samples are from the same population.

Hence

$H_0 : \pi_1 = \pi_2$

The best estimate of the standard error of the difference of P1 and P2 is given by pooling the samples and finding the pooled sample proportions (P) thus

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

Standard error of difference between proportions

$$S(p_1 - p_2) = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$

$$\text{And } Z = \left|\frac{P_1 - P_2}{S(p_1 - p_2)}\right|$$

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

Solution
Let us take the hypothesis that there is no significant difference between the two villages as far as the habit of taking tea is concerned i.e. $\pi_1 = \pi_2$
We are given

$P_1 = 0.6;$ $n_1 = 100$
$P_2 = 0.5;$ $n_2 = 200$

Appropriate statistic to be used here is given by

$$P = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$$
$$= \frac{(0.6)(100)+(0.5)(200)}{100+200} = \frac{60+100}{300}$$
$$= 0.53$$
$$q = 1 - 0.53$$
$$= 0.47$$
$$S(P_1 - P_2) = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$
$$= \sqrt{\frac{(0.53)(0.47)}{100} + \frac{(0.53)(0.47)}{200}}$$
$$= 0.0608$$
$$Z = \left|\frac{0.6-0.5}{0.0608}\right|$$
$$= 1.64$$

Since the computed value of Z is less than the critical value of Z = 1.96 at 5% level of significance therefore we accept the hypothesis and conclude that there is no significant difference in the habit of taking tea in the two villages A and B

t distribution (student's t distribution) tests of hypothesis (test for small samples n < 30)
For small samples n < 30, the method used in hypothesis testing is exactly similar to the one for large samples except that t values are used from t distribution at a given degree of freedom v, instead of z score, the standard error Se statistic used is also different.
Note that v = n – 1 for a single sample and $n_1 + n_2 - 2$ where two sample are involved.

a) **Test of hypothesis about the population mean**

When the population standard deviation (S) is known then the t statistic is defined as

$$t = \left|\frac{\overline{X}-\mu}{S_{\overline{X}}}\right| \qquad \text{where } S_{\overline{X}} = \frac{S}{\sqrt{n}}$$

Follows the students t distribution with (n-1) d.f. where

$\overline{X}$ = Sample mean
μ = Hypothesis population mean
n = sample size

and S is the standard deviation of the sample calculated by the formula

$$S = \sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{n-1}} \qquad \text{for } n < 30$$

If the calculated value of t exceeds the table value of t at a specified level of significance, the null hypothesis is rejected.

Example

Ten oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 15.8 kg and the standard deviation is 0.5kg. Does the sample mean differ significantly from the intended weight of 16kgs? Use 5% level of significance.

Solution

Given that n = 10; $\bar{x}$ = 15.8; S = 0.50; μ = 16; v = 9

$H_0 : \mu = 16$
$H_1 : \mu \neq 16$

$$= S_{\overline{X}} = \frac{0.5}{\sqrt{10}}$$

$$t = \left|\frac{15.8-16}{\frac{0.5}{\sqrt{10}}}\right|$$

$$= \left|\frac{0.2}{0.16}\right|$$

$$= -1.25$$

The table value for t for 9 d.f. at 5% level of significance is 2.26. the computed value of t is smaller than the table value of t. therefore, difference is insignificant and the null hypothesis is accepted.

b) **Test of hypothesis about the difference between two means**

The t test can be used under two assumptions when testing hypothesis concerning the difference between the two means; that the two are normally distributed (or near normally distributed) populations and that the standard deviation of the two is the same or at any rate not significantly different.

Appropriate test statistic to be used is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{(\overline{X}_1 - \overline{X}_2)}} \quad \text{at } (n_1 + n_2 - 2) \text{ d.f.}$$

The standard deviation is obtained by pooling the two sample standard deviation as shown below.

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Where S_1 and S_2 are standard deviation for sample 1 & 2 respectively.

Now $S_{\overline{X}_1} = \frac{Sp}{\sqrt{n_1}}$ and $S_{\overline{X}_2} = \frac{Sp}{\sqrt{n_2}}$

$$S_{(\overline{X}_1 - \overline{X}_2)} = \sqrt{S_{\overline{X}_1}^2 + S_{\overline{X}_2}^2}$$

Alternatively $S_{(\overline{X}_1 - \overline{X}_2)} = Sp\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$

Example

Two different types of drugs A and B were tried on certain patients for increasing weights, 5 persons were given drug A and 7 persons were given drug B. the increase in weight (in pounds) is given below

Drug A	8	12	16	9	3		
Drug B	10	8	12	15	6	8	11

Do the two drugs differ significantly with regard to their effect in increasing weight? (Given that v= 10; $t_{0.05} = 2.23$)

Solution

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$$t = \left| \frac{\overline{X}_1 - \overline{X}_2}{S_{(\overline{X}_1 - \overline{X}_2)}} \right|$$

Calculate for $\overline{X}_1$, $\overline{X}_2$ and S

X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)$	$(X_2 - \bar{X})^2$
8	-1	1	10	0	0
12	+3	9	8	-2	4
13	+4	16	12	+2	4
9	0	0	15	+5	25
3	-6	36	6	-4	16
			8	-2	4
			11	+1	1
$\Sigma X_1 = 45$	$\Sigma(X_1 - \bar{X}_1) = 0$	$\Sigma (X_1 - \bar{X}_1)^2 = 62$	$\Sigma X_2 = 70$	$\Sigma (X_2 - \bar{X}_2) = 0$	$\Sigma (X_2 - \bar{X}_2)^2 = 54$

$X_1 = \frac{\sum X_1}{n_1} = \frac{45}{5} = 9$ $\qquad$ $X_2 = \frac{\sum X_2}{n_2} = \frac{70}{7} = 10$

$S_1 = \sqrt{\frac{62}{4}} = 3.94$ $\qquad$ $S_2 = \sqrt{\frac{54}{6}} = 3$

$$S_p = \sqrt{\frac{(4)15.4 + (6)9}{10}}$$

$= 3.406$

$$S_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{11.6}{5} + \frac{11.6}{7}} \text{ Or } 3.406\sqrt{\frac{7+5}{5(7)}}$$

$= 1.99$

$$t = \left|\frac{\bar{X}_1 - \bar{X}_2}{S_{(\bar{X}_1 - \bar{X}_2)}}\right| = \left|\frac{9-10}{1.99}\right|$$

$= 0.50$

Now $t_{0.05}$ (at $v = 10$) $= 2.23 > 0.5$

Thus we accept the null hypothesis.

Hence there is no significant difference in the efficacy of the two drugs in the matter of increasing weight

Example

Two salesmen A and B are working in a certain district. From a survey conducted by the head office, the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen at 5% level of significance.

	A	B
No. of sales	20	18
Average sales in shs	170	205
Standard deviation in shs	20	25

Solution

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Where

$$S_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$S_{(\bar{X}_1 - \bar{X}_2)} = S_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

Where: $\bar{X}_1 = 170$, $\bar{X}_2 = 205$, $n_1 = 20$, $n_2 = 18$, $S_1 = 20$, $S_2 = 25$, V = 36

$$S_p = \sqrt{\frac{(19)(20^2) + (17)(25^2)}{20 + 18 - 2}}$$

$$= 22.5$$

$$S_{(\bar{X}_1 - \bar{X}_2)} = 22.5\sqrt{\frac{38}{360}}$$

$$= 7.31$$

$$t = \left|\frac{170 - 205}{7.31}\right|$$

$$= 4.79$$

$t_{0.05}(36) = 1.9$ (Since d.f > 30 we use the normal tables)

The table value of t at 5% level of significance for 36 d.f. when d.f. >30, that t distribution is the same as normal distribution is 1.9. Since the value computed value of t is more than the table value, we reject the null hypothesis. Thus, we conclude that there is significant difference in the average sales between the two salesmen

Testing the hypothesis equality of two variances

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

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

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$ Now under the H_0 : $\sigma_1^2 = \sigma_2^2$ it follows that

$$F = \frac{S_1^2}{S_2^2}$$ which is the test statistic.

Which follows F – distribution with V_1 and V_2 degrees of freedom. The larger sample variance is placed in the numerator and the smaller one in the denominator

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

In one sample of observations the sum of the squares of the deviations of the sample values from sample mean was 120 and in the other sample of 12 observations it was 314. test whether the difference is significant at 5% level of significance

Solution

Given that $n_1 = 10$, $n_2 = 12$, $\Sigma(x_1 - \overline{X}_1)^2 = 120$

$\Sigma(x_2 - \overline{X}_2)^2 = 314$

Let us take the null hypothesis that the two samples are drawn from the same normal population of equal variance

H_0 : $\sigma_1^2 = \sigma_2^2$

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

Applying F test i.e.

$$F = \frac{S_1^2}{S_2^2}$$

$$= \frac{\frac{\sum(X_1 - \overline{X}_1)^2}{n_1 - 1}}{\frac{\sum(X_2 - \overline{X}_2)^2}{(n_2 - 1)}}$$

$$= \frac{\frac{120}{9}}{\frac{314}{11}}$$

$$= \frac{13.33}{28.55}$$

since the numerator should be greater than denominator

$$F = \frac{28.55}{13.33} = 2.1$$

The table value of F at 5% level of significance for $V_1 = 9$ and $V_2 = 11$. Since the calculated value of F is less than the table value, we accept the hypothesis. The samples may have been drawn from the two populations having the same variances.

Chi square hypothesis tests (Non-parametric test)(X^2)

They include amongst others

i. Test for goodness of fit
ii. Test for independence of attributes
iii. Test of homogeneity
iv. Test for population variance

The Chi square test (x^2) is used when comparing an actual (observed) distribution with a hypothesized or explained distribution.

It is given by; $x^2 = \sum \frac{(O-E)^2}{E}$ Where O = Observed frequency

E = Expected frequency

The computed value of x^2 is compared with that of tabulated x^2 for a given significance level and degrees of freedom.

i. <u>Test for goodness of fit</u>

These tests are used when we want to determine whether an actual sample distribution matches a known theoretical distribution

The null hypothesis usually states that the sample is drawn from the theoretical population distribution and the alternate hypothesis usually states that it is not.

Example

Mr. Nguku carried out a survey of 320 families in Ateka district, each family had 5 children and they 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

Is the result consistent with the hypothesis that male and female births are equally probable at 5% level of significance?

Solution

If the distribution of gender is equally probable then the distribution conforms to a binomial distribution with probability P(X) = ½.

Therefore

H_0 = the observed number of boys conforms to a binomial distribution with P = ½

H_1 = the observations do not conform to a binomial distribution.

On the assumption that male and female births are equally probable the probability of a male birth is P = ½. The expected number of families can be calculated by the use of binomial distribution. The probability of male births in a family of 5 is given by

$P(x) = {}^5c_x P^x q^{5-x}$ (for x = 0, 1, 2, 3, 4, 5,)

$= {}^5c_x (½)^5$ (Since P = q = ½)

To get the expected frequencies, multiply P(x) by the total number N = 320. The calculations are shown below in the tables

x	P(x)	Expected frequency = NP(x)
0	${}^5c_0 (½)^5 = 1/32$	$320 \times 1/32 = 10$
1	${}^5c_1 (½)^5 = 5/32$	$320 \times 5/32 = 50$
2	${}^5c_2 (½)^5 = 10/32$	$320 \times 10/32 = 100$
3	${}^5c_3 (½)^5 = 10/32$	$320 \times 10/32 = 100$
4	${}^5c_4 (½)^5 = 5/32$	$320 \times 5/32 = 50$
5	${}^5c_5 (½)^5 = 1/32$	$320 \times 1/32 = 10$

Arranging observed and expected frequencies in the following table and calculating x^2

O	E	$(O - E)^2$	$(O - E)^2 / E$
14	10	16	1.60
56	50	16	0.72
110	100	100	1.00
88	100	144	1.44
40	50	100	2.00
12	10	4	0.40
			$\Sigma(0 - E)^2 / E = 7.16$

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

= 7.16

The table of x^2 for V = 6 – 1 = 5 at 5% level of significance is 11.07. The computed value of x^2 = 7.16 is less than the table value. Therefore the hypothesis is accepted. Thus it can be concluded that male and female births are equally probable.

ii) Test of independence of attributes

This test disclosed whether there is any association or relationship between two or more attributes or not. The following steps are required to perform the test of hypothesis.

1. The null and alternative hypothesis are set as follows
   H_0: No association exists between the attributes
   H_1: an association exists between the attributes
2. Under H_0 an expected frequency E corresponding to each cell in the contingency table is found by using the formula
   $$E = \frac{R \times C}{n}$$
   Where R = a row total, C = a column total and n = sample size
3. Based upon the observed values and corresponding expected frequencies the x^2 statistic is obtained using the formula
   $$x^2 = \sum \frac{(O-E)^2}{E}$$
4. The characteristic of this distribution are defined by the number of degrees of freedom (d.f.) which is given by
   d.f. = (r-1) (c-1),
   Where r is the number of rows and c is number of columns corresponding to a chosen level of significance, the critical value is found from the chi squared table
5. The calculated value of x^2 is compared with the tabulated value x^2 for (r-1) (c-1) degrees of freedom at a certain level of significance. If the computed value of x^2 is greater than the tabulated value, the null hypothesis of independence is rejected. Otherwise we accept it.

Example

In a sample of 200 people where a particular devise was selected, 100 were given a drug and the others were not given any drug. The results are as follows

	Drug	No drug	Total
Cured	65	55	120
Not cured	35	45	80
Total	100	100	200

Test whether the drug will be effective or not, at 5% level of significance.

Solution

Let us take the null hypothesis that the drug is not effective in curing the disease.
Applying the x^2 test
The expected cell frequencies are computed as follows

$E_{11} = \frac{R_1C_1}{n} = \frac{120\times100}{200} = 60$

$E_{12} = \frac{R_1C_2}{n} = \frac{120\times100}{200} = 60$

$E_{21} = \frac{R_2C_1}{n} = \frac{80\times100}{200} = 40$

$E_{22} = \frac{R_2C_2}{n} = \frac{80\times100}{200} = 40$

The table of expected frequencies is as follows

60	60	120
40	40	80
100	100	200

O	E	$(O - E)^2$	$(O - E)^2 /E$
65	60	25	0.417
55	60	25	0.625
35	40	25	0.417
45	40	25	0.625
			$\Sigma(O - E)^2 /E = 2.084$

Arranging the observed frequencies with their corresponding frequencies in the following table we get

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

= 2.084

V= (r –1) (c-1) = (2 – 1) (2 –1) = 1; $\chi^2_{tabulated(0.05)}$ = 3.841

The calculated value of x^2 is less than the table value. The hypothesis is accepted. Hence the drug is not effective in curing the disease.

Test of homogeneity

It is concerned with the proposition that several populations are homogenous with respect to some characteristic of interest e.g. one may be interested in knowing if raw material available from several retailers are homogenous. A random sample is drawn from each of the population and the number in each of sample falling into each category is determined. The sample data is displayed in a contingency table

The analytical procedure is the same as that discussed for the test of independence

Example

A random sample of 400 persons was selected from each of three age groups and each person was asked to specify which types of TV programs be preferred. The results are shown in the following table

Type of program

Age group	A	B	C	Total
Under 30	120	30	50	200
30 – 44	10	75	15	100
45 and above	10	30	60	100
Total	140	135	125	400

Test the hypothesis that the **populations** are homogenous with respect to the types of television program they prefer, at 5% level of significance.

Solution

Let us take hypothesis that the populations are homogenous with respect to different types of television programs they prefer

Applying x^2 test

O	E	$(O – E)^2$	$(O – E)^2 /E$
120	70.00	2500.00	35.7143
10	35.00	625.00	17.8571
10	35.00	625.00	17.8571
30	67.50	1406.25	20.8333
75	33.75	1701.56	50.4166
30	33.75	14.06	0.4166

50	62.50	156.25	2.500
15	31.25	264.06	8.4499
60	31.25	826.56	26.449
			$\Sigma(O - E)^2 /E$ = 180.4948

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

The table value of x^2 for 4d.f. at 5% level of significance is 9.488
The calculated value of x^2 is greater than the table value. We reject the hypothesis and conclude that the populations are not homogenous with respect to the type of TV programs preferred, thus the different age groups vary in choice of TV programs.

SUMMARY OF FORMULAE IN HYPOTHESIS

Testing

(a) Hypothesis testing of mean

For n>30

Z = $\left|\frac{\overline{X}-\mu}{S_{\overline{X}}}\right|$ Where $S_{\overline{X}} = \frac{S}{\sqrt{n}}$ at $\propto$ level of significance.

For n < 30

t = $\left|\frac{X-\mu}{S_{\overline{X}}}\right|$ where $S_{\overline{X}} = \frac{S}{\sqrt{n}}$

at n – 1 d.f

$\propto$ level of significance

(b) Difference between means (Independent samples)

For n > 30

Z = $\left|\frac{\overline{X}_1-\overline{X}_2}{S_{(\overline{X}_1-\overline{X}_2)}}\right|$

Where $S_{(\overline{X}_1-\overline{X}_2)} = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$

At $\propto$ = level of significance

For n < 30

t = $\frac{\overline{X}_1-\overline{X}_2}{S_{(\overline{X}_1-\overline{X}_2)}}$ at $n_1 + n_2 - 2$ d.f

where $S_{(\overline{X}_1-\overline{X}_2)} = S_p\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

and $S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$

(c) Hypothesis testing of proportions

Z = $\frac{p - \pi}{S_p}$

Where: S_p = $\sqrt{\frac{pq}{n}}$

p = Proportion found in sample

q = 1 - p

π = hypothetical proportion

(d) Difference between proportions

Z = $\frac{P_1 - P_2}{S_{(P_1 - P_2)}}$

Where:

$$S_{(P_1 - P_2)} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$

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

q = 1 - P

(e) Chi-square test

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

Where O = observed frequency

E = $\frac{\text{Column total} \times \text{Row total}}{\text{Sample Size}}$ = expected frequency

(f) F - test (variance test)

F = $\frac{S_1^2}{S_2^2}$

Here the bigger value between the standard deviations makes the numerator.

LESSON 6 REINFORCING QUESTIONS

QUESTION ONE

A firm purchases a very large quantity of metal off-cuts and wishes to know the average weight of 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?

QUESTION TWO

A sample of 80 is drawn at random from a population of 800. The sample standard deviation was found to be 6 grams.

- What is the finite population correction factor?
- What is the approximation of the correction factor?
- What is the standard error of the mean?

QUESTION THREE

State the Central Limit Theorem

QUESTION FOUR

a) What is statistical inference?
b) What is the purpose of estimation?
c) What are the properties of good estimators?
d) What is the standard error of the mean?
e) What are confident limits?
f) When is the Finite Population Correction Factor used? What is the formula?
g) How are population proportions estimated?
h) What are the characteristics of the t distribution?

QUESTION FIVE

A market research agency takes a sample of 1000 people and finds that 200 of them know of Brand X. After an advertising campaign a further sample of 1091 people is taken and it's found that 240 know of Brand X.
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.

QUESTION SIX

The monthly bonuses of two groups of salesmen are being investigated to see if there is a difference in the average bonus received. Random samples of 12 and 9 are taken from the two groups and it can be assumed that the bonuses in both groups are approximately normally distributed and that the standard deviations are about the same. The same level of significance is to be used.

n_1=12	n_2=9
x_1=£1060	x_2=£970
s_1=£63	s_2=£76

The sample results were

QUESTION SEVEN

Torch bulbs are packed in boxes of 5 and 100 boxes are selected randomly to test for the number of defectives

Number of Defectives	Number of boxes	Total defectives
0	40	0
1	37	37
2	17	34
3	5	15
4	1	4
5	0	0
	100	90

The number of any individual bulb being a reject is

$$\frac{90}{100} \div 5 = 0.18$$

and it is required to test at the 5% level whether the frequency of rejects conforms to a binomial distribution.

QUESTION EIGHT

a) Define type I and type II errors.
b) What is a two-tail test?
c) What is the best estimate of the population standard deviation when the two samples are taken

QUESTION NINE