11

LARGE SAMPLE TESTS

1. Introduction :

The main objective of taking a sample from a population is to get reliable information about the population. From the information obtained from the sample, conclusions are drawn about the population. This is called statistical inference. It mainly consists of two parts

(1) Estimation of parameters

(2) Tests of statistical hypothesis

We shall first of all understand certain terms associated with the study of statistical inference.

Parameters and statistics

A constant obtained from all the observations of a population is called a parameter. Population mean is a parameter. Similarly population median population standard deviation, population proportion of some attribute etc are parameters. Parameters are generally denoted by greek letters or capital letters e.g. population mean is denoted by μ, population proportion is denoted by P.

In order to estimate the parameter of a population a sample is drawn from the population. The constant obtained from a sample is called a statistic. Thus, sample mean $\bar{x}$, sample standard deviation S, sample proportion of some attribute p are statistics.

The following are some of the parameters and their statistics.

	Parameter	Statistic
Mean	μ	$\bar{x}$
S.D	σ	S
Proporation of some attribute	P	p

mpling distribution of a statistic

From a population of size N, number of samples of size n can be drawn. ese samples will give different values of a statistic. e.g. if different mples of size n are drawn from a population different values of sample an $\bar{x}$ are obtained. The various values of a statistic thus obtained can be ranged in form of a frequency distribution; known as sampling distribution. hus, we can have sampling distribution of sample mean $\bar{x}$, sampling stribution of proportion p etc.

tandard error of a statistic

The standard deviation of the sample statistic is called standard error of at statistic. e.g. if different samples of the same size n are drawn from a opulation, we get different values of sample mean $\bar{x}$. The S. D. of $\bar{x}$ is alled standard error of $\bar{x}$. It is obvious that the standard error of $\bar{x}$ will epend upon the size of the sample and the variability of the population. It an be derived that S.E. of $\bar{x} = \frac{\sigma}{\sqrt{n}}$. The following are the standard errors f some of the well-known statistics

Statistic	S.E.
Mean $\bar{x}$	$\frac{\sigma}{\sqrt{n}}$
Difference between two means $\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Sample proportion p	$\sqrt{\frac{PQ}{n}}$
Difference between two proportions $p'_1 - p'_2$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n^2}}$

Standard error plays a very important part in large sample tests. The following are its main uses.

Uses of S.E.

(1) To test whether a given value of a statistic differs significantly from the intended population parameter. i.e. whether the difference between value of the sample statistic and population parameter is significant and the difference may be attributed to chance or sampling fluctuations.

(2) To test the randomness of a sample i.e. to test whether the given sample be regarded as a random sample from the population.

(3) To obtain confidence interval for the parameter of the population.

(4) To determine the precision of the sample estimate, because precision

$$\text{of a statistic} = \frac{1}{\text{S.E. of the statistic}}$$

Statistical hypothesis :

A statistical hypothesis is an assumption about the parameter of the population or the nature of the population. e.g.

(i) The population mean $\mu = 25$

(ii) The average weights of students of college A and college B are same

(iii) 20% of the students of a college are non vegetarains.

(iv) The given population is a Binominal population

are some of the hypothesis.

Null hypothesis :

A statistical hypothesis which is taken for the possible acceptance is called a null hypothesis and it is denoted by H_0. The neutral attitude of the decision maker, before the sample observations are taken is the keynote of the null hypothesis. e.g.

(1) Mean of the population is 60

$H_0 : \mu = 60.$

(2) Means of both populations are equal

$H_0 : \mu_1 = \mu_2$

(3) The coin is unbiased

$$H_0 : P = \frac{1}{2}$$

(4) The proportions of drinkers in both the cities are equal

$H_0 : P_1 = P_2$

Alternative hypothesis :

A hypothesis complementary to the null hypothesis is called alternative hypothesis and it is denoted by H_1. e.g.

(i) $H_1 : \mu \neq 60$

(ii) $H_1 : \mu_1 \neq \mu_2$

(iii) $H_1 : \mu_1 > \mu_2$

(iv) $H_1 : P_1 \neq P_2$

are alternative hypothesis

Test of a statistical hypothesis :

The test of a statistical hypothesis is a procedure to decide whether to accept the null hypothesis or to reject it. In large samples i.e. in samples having number of observations $n > 30$, normal distribution is used. The value of standard normal variate Z is calculated from the given data and it is compared with the table value. The decision regarding acceptance or rejection of the null hypothesis is then taken.

Type I and Type II errors :

In testing of a statistical hypothesis the following situations may arise :

(i) The hypothesis may be true but it is rejected by the test

(ii) The hypothesis may be false but it is accepted by the test.

(iii) The hypothesis may be true and is accepted by the test.

(iv) The hypothesis may be false and is rejected by the test.

(iii) and (iv) are correct decisions while (i) and (ii) are errors.

The error committed in rejecting a hypothesis which is true is called Type I error and its probability is denoted by α.

The error committed in accepting a hypothesis which is false is called Type-II error and its probability is denoted by β.

The above four possibilities can be represented in a table as follows :

	Accept	Reject
H_0 is true	Correct decision	Type-I error
H_0 is false	Type-II error	Correct decision

Level of significance :

In any test procedure both the types of errors should be kept minimum. But as they are inter-related it is not possible to minimize both the errors simultaneously. Hence in practice, the probability of type-I error is fixed and the probability of type-II error is minimized. The fixed value of type-I error

is called level of significance and it is denoted by α. Thus, level of significance is the probability of rejecting a hypothesis which ought to be accepted. The most commonly used level of significance are 5% and 1%. When a decision is taken at 5% level of significances, it is meant that in 5 cases out of 100, it is likely to reject a hypothesis which ought to be accepted. In other words, our decision to reject H_0 is 95% correct.

Critical region :

In large sample tests the standard normal variate Z is used as a test statistic. The total area under the standard normal curve is 1. The area or the region of standard normal curve is divided into two regions by predetermined level of significance. The area of the normal curve corresponding to type-I error, i.e. the probability of rejecting a hypothesis which is true is known as area of rejection or critical region. If the computed value of Z falls in the critical region, the null hypothesis may be rejected. The region of standard normal curve other than critical region is called acceptance region when the computed value of Z falls in this region the null hypothesis may be accepted.

Two-tailed and One-tailed test :

In a standard normal curve the critical region may be given in two ways :

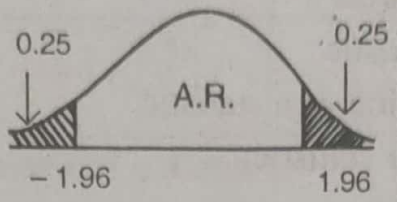
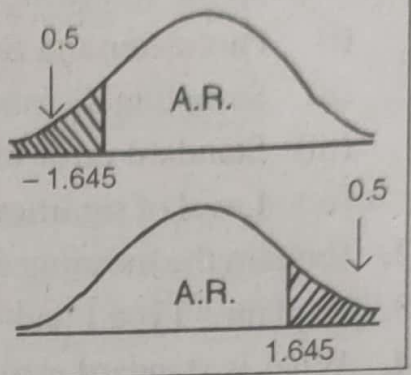
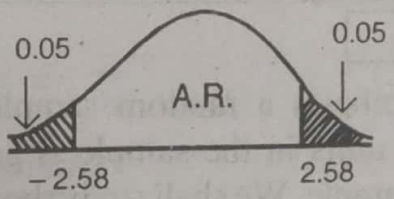
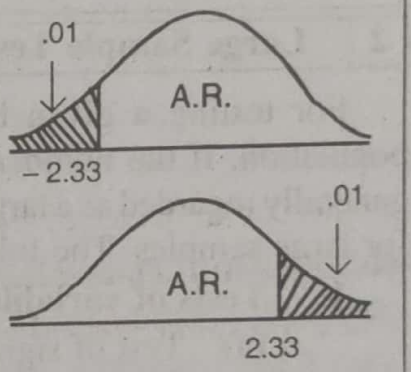
(i) on both the tails of the curve

(ii) on one tail of the curve i.e. either on the right tail of the curve or on the left tail of the curve.

If the given null hypothesis is tested using the critical region represented by both the tails of the normal curve, it is called two tailed test and if the null hypothesis is tested using the critical region represented by only one tail of the normal curve, it is called one-tailed test.

Whenever it is required to test whether the sample statistic is significantly different from the population parameter. i.e. whether the difference may be regarded as due to chance or due to sampling fluctuations, the two tail test is used. If we want to test that the mean of the population is greater than a specified value or the mean of one population is greater than the mean of another population or the proportion of defectives in the population is less than a specified proportion, one tail test is used.

The following figures show the critical regions corresponding to 5% and 1% levels of significance for two-tailed and one tailed test.

Level of significance	Two tailed C.R.	One tailed C.R.
5 %		
1 %		

Decision making :

From the given data the value of test statistic Z is computed. If the computed value of Z falls in the acceptance region i.e. for two tailed test the value of $|Z|$ is less than 1.96 (for 5% level significance), the null hypothesis may be accepted. If the computed value of Z falls in the critical region i.e. for two tailed lest $|Z|$ is greater than 1.96 (for 5% level of significance) the null hypothesis may be rejected. For one tail test the computed value of Z should be compared will the critical value of one tail, i.e. with 1.645 (for 5% level of significance.)

Generally the decisions are taken at 5% level of singnifcane or at 1% level of significance. If required the decision can be taken at any specified level of significance.

EXERCISE 11.1

1. Explain the following terms :
   (i) Parameter and Statistic
   (ii) Sampling distribution of a statistic
   (iii) Standard error of a statistic
   (iv) Level of significance
2. Explain the meaning of Null hypothesis and Alternative hypothesis
3. Explain : Type I and Type II errors.
4. What is standard error. State the uses of S.E.
5. Explain clearly the difference between one-tailed test and two-tailed test.

2 Large Sample Tests :

For testing a given hypothesis a random sample is drawn from a population. If the number of units in the sample is greater than 30, it is generally regarded as a large sample. We shall study the tests of significance for large samples. The following tests will be discussed in this chapter

1. **Tests of variables :**
   (a) Test of significance of a mean
   (b) Test of significance of difference between two means.
   (c) Test of significance of difference between two standard deviations
2. **Tests of attributes :**
   (d) Test of significance of proportion of successes
   (e) Test of significance of difference between two proportions

In all these tests we shalll follow the steps given below for testing a given hypothesis.

Step 1 : Stating clearly null and alternative hypothesis.
Step 2 : Finding out the difference between observed value and the value taken in null hypothesis.
Step 3 : Calculating S.E. of statistic
Step 4 : Computing the value of Z

$$Z = \frac{\text{Difference}}{\text{S.E.}}$$

The computed value of Z is then compared with the table value of Z at a required level of significance and the decision regarding acceptance or rejection of the null hypothesis is taken.

For ready reference the critical values at important levels of significance are given below :

	1 %	5 %	10 %
Two-tailed test	2.58	1.96	1.645
One-tailed test	2.33	1.645	1.282

(a) Test of significance of a mean :

Suppose a random sample of size n is drawn from a large population. Suppose the mean of the sample is $\bar{x}$. If we want to test the hypothesis that population mean is μ_0 i.e. $H_0 : \mu = \mu_0$. We can use the following steps :

(i) $H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

(ii) Difference $= |\bar{x} - \mu_0|$

(iii) S.E. of $\bar{x} = \dfrac{\sigma}{\sqrt{n}}$

(iv) $Z = \dfrac{\text{Difference}}{\text{S.E.}}$

The calculated value of $|Z|$ is then compared with the critical value of Z. If the calculated value of Z falls in the critical region (i.e. $|Z| > 1.96$ at 5% level of significance) the null hypothesis may be rejected and it may be concluded that the difference between the sample mean and the population mean may be regarded significant or the given sample may not be regarded as a random sample from the population. If the computed value of $|Z|$ is less than the critical value of Z (i.e. $|Z| < 1.96$) the null hypothesis is accepted and it may be concluded that the difference between sample mean and population mean is insignificant or the sample may be regarded as a random sample from the population.

[**Note :** If S.D. σ of the population is not given, the S.D.S of the sample can be taken as its estimate]

Illustration 1 : A sample of 400 students have a mean height of 171.38 cms. Can it be reasonably regarded as a random sample from a large population with mean height 171.17 and standard deviation 3.3. cms ?

Ans. : $H_0 : \mu = 171.17$

$H_1 : \mu \neq 171.17$

Difference $= |\bar{x} - \mu|$

$= |171.38 - 171.17|$

$= 0.21$

$$\text{S.E. of } \bar{x} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{3.3}{\sqrt{400}}$$

$$= 0.165$$

$$Z = \frac{\text{Diff.}}{\text{S.E.}} = \frac{0.21}{0.165} = 1.27 < 1.96$$

$\therefore$ H_0 may be accepted at 5% level of significance

$\therefore$ The sample may be regarded as a random sample from a population with mean 171.17

Illustration 2 : A stenographer claims that he can write at an average speed of 120 words per minute. In 100 trials he obtained an average speed of 116 words per minute with a standard deviation of 15 words. Is the claim justified ? Use 5% level of significance.

Ans. : $H_0 : \mu = 120$

$H_1 : \mu < 120$

$$\text{Difference} = |\bar{x} - \mu|$$

$$= |116 - 120|$$

$$= 4$$

$$\text{S. E. of } \bar{x} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{15}{\sqrt{100}}$$

$$= 1.5$$

$$Z = \frac{\text{Diff.}}{\text{S.E.}} = \frac{4}{1.5} = 2.67 > 1.645$$

The value of Z > critical value (one tail)

$\therefore$ H_0 may be rejected at 5% level of significance.

$\therefore$ The claim that the average speed is 120 words per minute may be rejected. It may be less than 120 words per minute.

Illustration 3 : A random sample of 400 items gave mean 4.45 and variance 4. can the sample be regarded as drawn from a normal population with mean 4 ?

Ans. $H_0 : \mu = 4$

$H_1 : \mu \neq 4$

Difference $= |\bar{x} - \mu|$
$= |4.45 - 4|$
$= 0.45$

S. E. of $\bar{x} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{400}} = 0.1$

$Z = \frac{\text{Diff.}}{\text{S.E.}} = \frac{0.45}{0.1} = 4.5 > 1.96$

$\therefore$ At 5% level of significance H_0 may be rejected

$\therefore$ The sample may not be regarded as drawn from a population with mean 4.

EXERCISE 11.2

1. How will you test the significance of the difference between mean of a sample and the mean of the population in case of a large sample ?
2. A random sample of 900 members is found to have a mean of 4.45 cms. Can it be reasonably regarded as a sample from a large population whose mean is 5 cms and variance is 4 square cms. **[Ans. : Z = 8.25]**
3. The mean life time of 100 fluorescent light tubes produced by a company is computed to be 1570 hours with a standard deviation of 120 hours. The company claims that the average life of the tubes produced by the company is 1600 hours. Is the claim justified ? Use 5% level of significance. *(I.C.W.A., Final)*
[Ans. : Z = 2.5]
4. The mean of a random sample of 400 units is 82 and S.D. is 18. Test the hypothesis that population mean is 80. **[Ans. : Z = 2.22]**
5. A sample of size 400 was drawn and sample mean was found to be 99. Test whether the sample could have come from normal population with mean 100 and variance 64 ? *(Delhi Uni., M.A.)*
[Ans. : Z = 2.5]
6. A random sample of size 36 was taken from a universe of size 1000 and it was found that the sample average was 20 with variance 16. How will you examine that the average in the universe is 22 ?
[Ans. : Hint : Here the population is finite hence the following formula of S.E. of $\bar{x}$ should be taken S.E. of $\bar{x} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = 0.654$; $Z = 3.05$]

7. A manufacturer of ball pens claims that a certain pen he manufactures has a mean writing of 400 pages with a S.D. of 20 pages. A purchasing agent selects a sample of 100 pens and puts them for test. The mean writing life of the sample was 390 pages. Should the purchasing agent reject the manufacturer's claim at 5% level of significance ?

(C.A., Inter)

[Ans. : Z = 5]

8. A sample of 900 members is found to have a mean of 3.4 cm. Can it be reasonably regarded a simple sample from a population with mean 3.25 and S.D. 2.61 cm ? **[Ans. Z = 1.72]**

9. The mean I.Q. of a sample of 1600 children was 99. Is it likely that this was a random sample from a population with mean I. Q. 100 and S. D. 15 ? **[Ans. : Z = 2.67]**

10. 100 observations of a sample gave the following results. $\Sigma x_i = 3000$, $\Sigma x_i^2 = 1,80,000$. Test the hypothesis that the mean of the population is 31.2. **[Ans. : Z = 0.4]**

(b) Test of significance of difference between two means :

Suppose two independent random samples are drawn from two different populations and their means are respectively $\bar{x}_1$ and $\bar{x}_2$. If we want to test the hypothesis that the population means are equal i.e. $H_0 : \mu_1 = \mu_2$, we can follow the steps given below.

(i) $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

(ii) Difference $= |\bar{x}_1 - \bar{x}_2|$

(iii) S. E. of $\bar{x}_1 - \bar{x}_2 = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(iv) $Z = \frac{\text{Difference}}{\text{S.E.}}$

Here n_1 and n_2 are the sizes of the two samples. σ_1^2 and σ_2^2 are population variances. If they are not known the sample variances S_1^2 and S_2^2 can be used as their estimates. If σ^2 is given as the variance of both the populations then $\sigma_1^2 = \sigma_2^2 = \sigma^2$ should be taken. From the given data the value of Z is computed and it is compared with the critical value of Z at a required level of significance and the decision regarding the acceptance or the rejection of the hypothesis can be taken.

Illustration 4 : The average life of 150 electric bulbs of a company A is 1400 hours with a S.D. of 120 hours while the average life of 200 electric bulbs of company B is 1200 hours with a S.D. of 80 hours. Is the difference between the average lives of the bulbs significant ?

Ans. : $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Difference $= |\bar{x}_1 - \bar{x}_2|$

$= |1400 - 1200|$

$= 200$

$$\text{S.E. of } \bar{x}_1 - \bar{x}_2 = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \sqrt{\frac{(120)^2}{150} + \frac{(80)^2}{200}}$$

$= 11.314$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{200}{11.314} = 17.68 > 1.96$$

$\therefore$ H_0 may be rejected at 5% level of significance

$\therefore$ The difference between the averages of two types of bulbs is significant.

Illustration 5 : The mean of a random sample of 1000 units is 17.6 and the mean of another random sample of 800 units is 18. Can it be concluded that both the samples come from the same population with S.D. = 2.6.

Ans. :

H_0 : Both the samples come from the same population

$\mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Difference $= |\bar{x}_1 - \bar{x}_2|$

$= |17.6 - 18|$

$= 0.4$

$$\text{S.E. of } \bar{x}_1 - \bar{x}_2 = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \sqrt{\frac{(2.6)^2}{1000} + \frac{(2.6)^2}{800}}$$

$$= 0.1233$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{0.4}{0.1233} = 3.24 > 1.96$$

$\therefore$ H_0 may be rejected at 5% level of significance. Both the samples may not be regarded as drawn from the same population.

Illustration 6 : The average daily wage of 1000 labourers of a factory A is Rs.47 with S.D. of Rs. 28. The average daily wage of 1500 labourers of a factory B is Rs 49 with S. D. of Rs. 40 can it be said that the average daily wage of factory B is more than the average daily wage of factory A ?

Ans. : $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_2 > \mu_1$

$$\text{Difference} = |\bar{x}_1 - \bar{x}_2|$$

$$= |47 - 49|$$

$$= 2$$

$$\text{S. E. of } \bar{x}_1 - \bar{x}_2 = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \sqrt{\frac{(28)^2}{1000} + \frac{(40)^2}{1500}}$$

$$= 1.36$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{2}{1.36} = 1.47 < 1.645$$

$\therefore$ H_0 is accepted at 5% level of significance.

$\therefore$ There is no significant difference in the daily wages of two factories. It cannot be said that average daily wage of factory B is more than that of A.

(C) Test of significance of difference between two standard deviations :

Here we want to test that the standard deviations of the two populations do not differ significantly. For this we shall follow the steps given below :

(1) $H_0 : \sigma_1 = \sigma_2$

$H_1 : \sigma_1 \neq \sigma_2$

Difference = $S_1 - S_2$ = difference between sample standard deviations.

$$\text{S.E. of } S_1 - S_2 = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

$$= \sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}$$

$$Z = \frac{\text{Difference}}{\text{S.E.}}$$

The value of Z is computed and it is compared with the critical value of Z, at a required level of significance and the decision regarding acceptance or rejection of the hypothesis is taken.

Illustration 7 : The information regarding marks of boys and girls of a college is given below.

Sample	Mean	S.D.	Sample size
Boys	83	10	121
Girls	81	12	81

Test whether the difference in standard deviations is significant

Ans. : $H_0 : \sigma_1 = \sigma_2$

$H_1 : \sigma_1 \neq \sigma_2$

Difference $= |S_1 - S_2|$

$= |10 - 12|$

$= 2$

$$\text{S.E. of } S_1 - S_2 = \sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}$$

$$= \sqrt{\frac{(10)^2}{2(121)} + \frac{(12)^2}{2(81)}}$$

$$= 1.1411$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{2}{1.1411} = 1.75 < 1.96$$

H_0 may be accepted at 5% level of significance i.e. the difference in standard deviations is insignificant.

13. A large corporation employs both men and women to do the same kind of work. It is hypothesised that the average hourly output of men is less than that of women. The following information was gathered by the "ABC" Team of the corporation.

	Men	Women
Variance	70	74
Sample size	36	36
Sample mean in units	150	153

Is the average hourly output of men significantly lower than that of women at 5% level at significance ? *(Sau. Uni., SYBBA, April '97)*

[Ans. : S.E. = 2, Z = 1.5]

14. A random sample of 150 villages was taken from 'Kheda' district of State Gujarat and the average population per village was found to be 745 with a standard deviation of 65. Another random sample of 175 villages from the same district gave an average population of 690 with a standard deviation of 72. Is the difference between average of two samples statistically significant at $\alpha = 1\ \%$?

(Guj. Uni., S.Y.B.B.A., 2007)

[Ans. : S.E. 7.6, Z = 7.29]

15. Two independent samples are drawn from two different populations. The information is given below :

Sample size	Mean	S.D.
100	1200	240
200	900	220

Test whether the variabilities of life of bulbs of the factories significantly differ.

[Ans. : Z = 0.99]

16. An examination was conducted for two classes consisting of 40 and 50 students respectively. In the first class; mean marks obtained by student was 72 marks with standard deviation of 6 marks, while for other class mean marks was 77 marks with standard deviation of 8 marks. At 5 % level of significant, can we conclude that performance of both the class is consistent ?

(Guj. Uni., S.Y.B.B.A., 2007)

[Ans. : S.E. = 1.4765, Z = 3.3864]

(d) Test of significance of a sample proportion :

Suppose a random sample of n units is drawn from a population and x units of them possess a particular attribute. i.e. the sample proportion of that attribute is $p = \frac{x}{n}$. In order to test the null hypothesis that the population proportion of the attribute is P the following steps are used :

H_0 : Population proportion = P

H_1 : Population proportion ≠ P

$$\text{Difference} = |p - P| = \left|\frac{x}{n} - P\right|$$

$$\text{S.E. of } p = \sqrt{\frac{PQ}{n}} \quad ; \quad Q = 1 - P$$

$$Z = \frac{\text{Difference}}{\text{S. E.}}$$

The value of Z is computed from the given data and it is compared with the critical value of Z at a required level of significance and the decision regarding the acceptance or the rejection of the hypothesis is taken.

Illustration 8 : In a hospital out of 500 new born babies, 280 are boys. Does this information support the hypothesis that the births of boys and girls are in equal proportion ? (Take 1% level of significance)

Ans. : H_0 : Proportion of boys $P = \frac{1}{2}$

$$H_1 : P \neq \frac{1}{2}$$

$$\text{Difference} = |p - P|$$

$$= \left|\frac{280}{500} - \frac{1}{2}\right|$$

$$= .06$$

$$\text{S.E. of } p = \sqrt{\frac{PQ}{n}}$$

$$= \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{500}}$$

$$= 0.02236$$

$$Z = \frac{\text{Difference}}{\text{S. E.}} = \frac{0.06}{0.02236} = 2.68 > 258$$

$\therefore$ H_0 may be rejected at 1% level of significance

i.e. the proportion of births of boys and girls may not be regarded equal.

Illustration 9 : In a large consignment of apples, 64 fruits out of a sample of 400 fruits are found to be bad. Test the hypothesis that the population proportion of bad apples in the consignment is 20%. (Use 1% level of significance.)

Ans. : $H_0 : P = 0.20$

$H_1 : P \neq 0.20$

Difference $= |p - P|$

$$= \left| \frac{64}{400} - 0.20 \right|$$

$$= 0.04$$

S.E. of $p = \sqrt{\frac{PQ}{n}}$

$$= \sqrt{\frac{0.2 \times 0.8}{400}}$$

$$= 0.02$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{0.04}{0.02} = 2 < 2.58$$

$\therefore H_0$ may be accepted at 1% level of significance. It may be concluded that 20% of apples are bad in the consignment.

Illustration 10 : In a big city 480 men out of a sample of 800 men are smokers. Does this information support the hypothesis that the majority of men in the city are smokers ?

Ans. : $H_0 : P = \frac{1}{2}$

$H_1 : P > \frac{1}{2}$

Difference $= |p - P|$

$$= \left| \frac{480}{800} - \frac{1}{2} \right|$$

$$= 0.1$$

S.E. of $p = \sqrt{\frac{PQ}{n}}$

$$= \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{800}}$$

$$= 0.017678$$

$$Z = \frac{Differenc}{S.E.} = \frac{0.1}{0.017678}$$

$$= 5.66 > 1.645$$

∴ At 5% level of significance, H_0 may be rejected.

i.e. the given information supports the hypothesis that the majority of men in the city are smokers.

EXERCISE 11.4

1. Explain how will you test the significance of the difference between sample proportion and population proportion in case of a large sample.
2. A cubical die was thrown 9000 times and a 5 or 6 was obtained 3240 times. Do the data indicate that the die is unbiased.

[Ans. : Z = 5.37]

3. In a certain city 380 men out of 800 men were found to be smokers. Discuss whether this information support the view that the majority of men in the city are smokers. *(C.A. Foundation, 1993)*

[Ans. : S.E. = 0.017, Z = 1.41]

4. In a sample of 400 parts manufactured by a company the number of defective parts was found to be 30. The company however claimed that only 5% of the products is defective. Test at 5% level of significance whether the company's claim is tenable *(C.A. Foundation, Nov., '96)*

[Ans. : S.E. = 0.109, Z = 2.29]

5. A shoe company produces 3% defective shoe pairs. In a random sample of 500 pairs, 20 pairs were found to be defective. Does this information suggest that the level of production has deteriorated ?

[Ans. : S.E. = 0.0076, Z = 1.31]

same.

(c) Test of significance of difference between two sample proportions :

Suppose a random sample of size n_1 is taken from one population and x_1 units of them possess some attribute. Hence $p_1 = \frac{x_1}{n_1}$ is the proportion of units possesing that attribute in the sample. Suppose another independent random sample of size n_2 is taken from another population and x_2 units possess the same attribute, i.e. sample proportion of that attribute in the second sample is $p_2 = \frac{x_2}{n_2}$. Now if we want to test the hypothesis that the population proportions of that attribute are equal i.e. $H_0 : P_1 = P_2$. We can apply the test in the following way

$H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2$

Difference $= p_1 - p_2$

$$\text{S.E. of } p_1 - p_2 = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n^2}}$$

$$Z = \frac{\text{Difference}}{\text{S.E.}}$$

If the population proportions P_1 and P_2 are not known their estimate is obtained by combining two sample proportions. The pooled estimate P can be obtained as

$$P = \frac{x_1 + x_2}{n_1 + n_2}$$

$$= \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$$

$$\text{Hence, S.E. of } p = \sqrt{\frac{PQ}{n_1} + \frac{PQ}{n_2}}$$

$$= \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The value of Z is computed from the given data and it is compared with the critical value of Z at a required level of significance. The decision regarding acceptance or rejection of the hypothesis is then taken.

Illustration 11 : The proportions of literates in two towns A and B are 30% and 25%. If samples of 1200 and 900 are taken from these populations, will the difference between the proportion remain hidden ?

H_0 : The difference will remain hidden ($P_1 = P_2$)

H_1 : Difference will not remain hidden.

Difference $= |P_1 - P_2|$

$= |0.30 - 0.25|$

$= 0.05$

$$\text{S.E. of } p_1 - p_2 = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n^2}}$$

$$= \sqrt{\frac{0.3 \times 0.7}{1200} + \frac{0.25 \times 0.75}{900}}$$

$= 0.0196$

$$Z = \frac{\text{Difference}}{\text{S.E.}}$$

$$= \frac{0.05}{0.0196}$$

$= 2.55 > 1.96$

$\therefore$ At 5% significance level, H_0 may be rejected i.e. the difference will not remain hidden.

Illustration 12 : In a sample of 1000 men from one city 750 were found to be smokers. In another sample of 1200 men from another city 1000 men were found to be smokers. Do the data indicate that the two cities are significantly different with respect to the prevalence of smoking habit among men ?

Ans. : $H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2$

Difference $= |p_1 - p_2|$

$$= \left| \frac{x_1}{n_1} - \frac{x_2}{n_2} \right|$$

$$= \left| \frac{750}{1000} - \frac{1000}{1200} \right|$$

$= 0.0833$

$$\text{S.E. of } p_1 - p_2 = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{where } P = \frac{x_1 + x_2}{n_1 + n_2}$$

$$= \frac{750 + 1000}{1000 + 1200}$$

$$= \frac{1750}{2200} = \frac{35}{44}, \quad Q = \frac{9}{44}$$

$\therefore$ S.E. of $p_1 - p_2$

$$= \sqrt{\frac{35}{44} \times \frac{9}{44}\left(\frac{1}{1000} + \frac{1}{1200}\right)}$$

$$= 0.0173$$

$$Z = \frac{\text{Difference}}{\text{S.E.}}$$

$$= \frac{0.0833}{0.0173}$$

$$= 4.82 > 1.96$$

$\therefore$ H_0 may be rejected at 5% level of significance

i.e. the two cities differ significantly with respect to the prevalence of smoking habit among men.

Illustration 13 : A machine produced 16 defective articles in a batch of 500 articles. After overhauling it produced 3 defective articles in a sample of 100 articles. Has the machine improved ?

$H_0 : P_1 = P_2$

$H_1 : P_2 < P_1$

Difference $= |p_1 - p_2|$

$$= \left| \frac{16}{500} - \frac{3}{100} \right|$$

$$= 0.002$$

$$\text{S.E. of } p_1 - p_2 = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{Where } p = \frac{x_1 + x_2}{n_1 + n_2}$$

$$= \frac{16+3}{500+100}$$

$$= \frac{19}{600} \; ; Q = \frac{581}{600}$$

$$\text{S.E. of } p_1 - p_2 = \sqrt{\frac{19}{600} \times \frac{581}{600}\left(\frac{1}{500} + \frac{1}{100}\right)}$$

$$= 0.01918$$

$$Z = \frac{\text{Difference}}{\text{S. E.}}$$

$$= \frac{0.002}{0.01918}$$

$$= 0.104 < 1.64$$

$\therefore$ H_0 may be accepted at 5% level of significance.

The machine may not be regarded as improved because of overhauling.

Illustration 14 : In a sample of 500 families in a city A, 30 families used a specific brand of detergent powder. In city B, 55 families used the same brand in a sample of 1000 families. Do the data prove that the use of this detergent is equal in the two cities.

Ans. : $H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2$

Difference $= |p_1 - p_2|$

$$= \left|\frac{x_1}{n_1} - \frac{x_2}{n_2}\right|$$

$$= \left|\frac{30}{500} - \frac{55}{1000}\right|$$

$$= 0.005$$

$$\text{S.E. of } p_1 - p_2 = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{Where } P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{30+55}{500+1000} = \frac{17}{300}$$

$$Q = 1 - P = 1 - \frac{17}{300} = \frac{283}{300}$$

$$\text{S. E. of } p_1 - p_2 = \sqrt{\frac{17}{300} \times \frac{283}{300}\left(\frac{1}{500} + \frac{1}{1000}\right)}$$

$$= 0.01266$$

$$Z = \frac{\text{Difference}}{\text{S.E.}}$$

$$= \frac{0.005}{0.01266}$$

$$= 0.39 < 1.96$$

$\therefore$ At 5% level of significance H_0 may be accepted

i.e. the use of the detergent is equal in the two cities.

EXERCISE 11.5

1. Explain how will you test the significance of the difference between two proportions of two large samples.
2. A machine produced 20 defective articles in a batch of 400 articles. After overhauling it produced 10 defective articles in a batch of 300 articles. Has the machine improved ? *(Sau. Uni., SYBBA., '98)*

   [Ans. : S.E. = .0155, Z = 1.08]
3. A sample survey results show that out of 800 literate people 480 are employed whereas out of 700 illiterate people only 350 are employed. Can the difference between proportion of employed persons be ascribed due to sampling fluctuations ? *(C.A., 1984)*

   [Ans. : Z = 3.89]
4. In a certain district A, 450 persons are regular consumers of tea out of a sample of 1000 persons. In another district B, 400 persons are regular consumers of tea out of a sample of 800 persons. Do these facts reveal a significant difference between two districts as far as tea drinking habit is concerned ?

   (Use 5% level)

   [Ans. : Z = 2.11]
5. In a large city A, 20% of a random sample of 900 school boys had defective eye-sight. In another large city B, 15.5% of a random sample of 1600 school boys had the same defect. Is the difference between two proportions significant.

   [Ans. : Z = 2.87]

[Ans. : S.E. = 0.6106, Z = 0.46]

3. Confidence Interval :

We know that a value of a statistic obtained from a random sample drawn from a population estimates the parameter of the population. e.g. A sample mean $\bar{x}$ estimates population mean μ i.e. $\bar{x}$ is an estimate of μ. A single value of the statistic obtained to estimate the parameter is called a **point estimate.** $\bar{x}$ is a point estimate of μ. Similarly sample proportion p is a point estimate or estimator of population proportion P.

To give one value for estimating the parameter of the population does not appear satisfactory. In practice therefore an interval is obtained which may include the parameter of the population with a certain degree of confidence. The interval developed by using standard error of the statistic is called confidence interval or fiducial interval.

Confidence interval for population mean μ

95% C.I. for μ is $\bar{x} \pm 1.96$ (S.E. of $\bar{x}$) $= \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$

99% C. I. for μ is $\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$

Most certain (99.73%) C.I. for μ is $\bar{x} \pm 3\left(\frac{\sigma}{\sqrt{n}}\right)$

Confidence interval for population proportion P

95% C.I. for P is $p \pm 1.96$ (S.E. of p)

$$= p \pm 1.96 \sqrt{\frac{pq}{n}}$$

99% C.I. for P is $p \pm 2.58 \sqrt{\frac{pq}{n}}$

99.73% C.I. for P is $p \pm 3 \sqrt{\frac{pq}{n}}$

Thus, confidence interval for any parameter and for any degree of confidence can be developed, using the standard error of the statistic.

Note : If the population is finite with N units, the S.E. should be multiplied by the factor $\sqrt{\frac{N-n}{N-1}}$

i.e. for a finite population of size N, 95% C.I. for μ is

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Illustration 15 : The mean of a sample of size 400 is 82 and S.D. is 18. Find 95% confidence limits for population mean.

Ans. :

95% confidence limits for population mean is given by

$\bar{x} \pm 1.96$ (S.E. of $\bar{x}$)

$$= \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

$$= 82 \pm 1.96 \frac{18}{\sqrt{400}}$$

$= 82 \pm 1.76$

$82 - 1.76 = 80.24$ and $82 + 1.76 = 83.76$

i.e. 95% confidence limits for μ is 80.24 to 83.76

Illustration 16 : The S.D. of a population is 20. The mean of a random sample of size 450 is 30. Obtain 99% confidence limits of the population mean.

Ans. : 99% C.I. for population mean is given by

$\bar{x} \pm 2.58$ (S.E. of $\bar{x}$)

$$= \bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$$

$$= 30 \pm 2.58 \frac{20}{\sqrt{450}}$$

$= 30 \pm 2.43$

$\therefore$ 99% confidence limits of μ are 27.57 to 32.43

Illustration 17 : A random sample of size 100 is drawn from a population of 3000 units. The mean and S.D. of the sample are 16 and 1.2. Find 95% confidence interval for the population mean.

Ans. : Here the population is finite, hence S.E. should be multiplied by

$$\text{factor } \sqrt{\frac{N-n}{N-1}}$$

95% C.I. for population mean μ is

$\bar{x} \pm 1.96$ (S.E. of $\bar{x}$)

$$= \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$= 16 \pm 1.96 \frac{1.2}{\sqrt{100}} \times \sqrt{\frac{3000-100}{3000-1}}$$

$= 16 \pm 1.96(0.12)\ (0.9834)$

$= 16 \pm 0.23$

$\therefore$ 95% C.I. for μ is 15.77 to 16.23

Illustration 18 : A sample of 400 articles from a big lot gave 40 defective articles. Find 99.73% confidence limits of the percentage of defective articles in the entire lot.

Ans. : 99.73% C.I. for the population proportion of defective articles is given by

3. Degrees of Freedom :

For testing the significance of the difference between sample mean and the population mean, t distribution can be used. The probability tables of t distribution are given for various levels of significance and for different degrees of freedom. Degrees of freedom is the number of independent observations of the variable. The number of independent observations is different for different statistics. Suppose we are asked to select any five observations. There is no restriction on the selection of these observations. We are free to select any five observations. Hence the degree of freedom is 5. Now suppose we want to select five observations whose sum is 100. Here four observations can be selected freely but the fifth observation is automatically selected by virtue of the restriction of total 100. Hence we are not free to select all the five observations but our freedom is restricted to the selection of only 4 observations. In fact our freedom is for selection of $5-1=4$ observations. Thus the degree of freedom for selecting n observations when one such restriction is given is $n-1$. If two such restrictions are given the degrees of freedom will be $n-2$.

The degrees of freedom associated with some of the important statistics are given below :

(i) D.f of $\bar{x} = \dfrac{\Sigma x_i}{n}$ is n.

(ii) D.f of $S^2 = \dfrac{\Sigma(x_i - \bar{x})^2}{n}$ is $n - 1$

(iii) D.f of correlation coefficient

$$r = \frac{\Sigma(\bar{x}_i - \bar{x})(y_i - \bar{y})}{n.S_x S_y}$$ is $n-2$

(iv) D. f of a $r \times c$ contingency table is $(r-1)(c-1)$

4. Student's t-Distribution :

This very important distribution was given by W.S. Gosset in 1908. He published his work under the pen-name of student. Hence the distribution is known as student's t distribution.

It $x_1, x_2, x_3 \ldots x_n$ is a random sample of n observations drawn from a normal population with mean μ and S. D. σ then the distribution of $t = \frac{\bar{x} - \mu}{S/\sqrt{n-1}}$ is defined as t distribution on $n-1$ degrees of freedom.

Here $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

The following is the probability density function of t distribution

$$f(t) = \frac{1}{\sqrt{n}\cdot\beta\left(\frac{1}{2}, \frac{n}{2}\right)\left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}}, \quad -\infty < t < \infty$$

Assumptions of t distribution :

(i) The population from which the sample is drawn is normal

(ii) The sample is random.

(iii) The population S.D. σ is not known.

Properties of t-distribution :

(i) The probability curve of t distribution is symmetrical

(ii) The tails of the curve are asymptotic to x-axis.

(iii) When $n \to \infty$, t distribution tends to normal distribution.

(iv) The form of the t-distribution varies with the degrees of freedom.

Uses of t-distribution :

(i) For testing the significance of the difference between sample mean and population mean.

(ii) For testing the difference between means of two samples.

(iii) For testing significance of the observed correlation coefficient.

(iv) For testing the significance of observed regression coefficient.

We shall now study the above mentioned uses of t distribution.

5. Test of Significance of a Mean of a Small Sample :

Suppose a random sample $x_1, x_2, x_3, \ldots x_n$ is drawn from a normal population and the mean and variance of the sample are $\bar{x}$ and S^2 respectively. If we want to test the hypothesis that there is no significant difference between sample mean $\bar{x}$ and the assumed mean μ of the population. We can apply t test in the following way.

H_0 : Population mean $= \mu$

H_1 : Population mean $\neq \mu$

$$t = \frac{|\bar{x} - \mu|}{S/\sqrt{n-1}} = \frac{|\bar{x} - \mu|\sqrt{n-1}}{S}$$

Where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

The d.f. associated with statistic t is $n - 1$.

The value of t is computed from the given data and it is compared with the table value of t on appropriate degrees of freedom and at a required level of significance. If the calculated value of $t <$ table value of t, the null hypothesis H_0 may be accepted and if the calculated value of t is greater than the table value of t, the null hypothesis H_0 may be rejected.

Note : $S^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$

$$= \frac{1}{n}\left\{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right\}$$

$$= \frac{1}{n}\left\{\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right\} \quad \text{where } d_i = x_i - A$$

Illustration 1 : A sample of 4 observations from a normal population gave the following results :

$$\Sigma x_i = 7,\ \Sigma x_i^2 = 15$$

Test the hypothesis that the mean of the population is 2.

Ans. :

$H_0 : \mu = 2$

$H_1 : \mu \neq 2$

Here, $\bar{x} = \frac{\Sigma x_i}{n} = \frac{7}{4} = 1.75$

$$S^2 = \frac{1}{n}\left\{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right\}$$

$$= \frac{1}{4}\left\{15 - \frac{(7)^2}{4}\right\} = 0.6875$$

$\therefore S = \sqrt{0.6875} = 0.8292$

Now, $t = \frac{|\bar{x} - \mu|\sqrt{n-1}}{S}$

$$= \frac{|1.75 - 2|\sqrt{4-1}}{0.8292}$$

$= 0.52$

D.f $= n - 1 = 3$.

At 5% level of significance the value of $t = 3.183$

$\therefore t_{cal} < t_{tab}$

$\therefore H_0$ may be accepted at 5% level of significance

$\therefore$ Population mean may be taken as 2.

Illustration 2 : Ten individuals are chosen at random from a population and their heights are found to be in inches as

63, 63, 66, 67, 68, 69, 70, 70, 71, 71

In the light of these data, test the hypothesis that the mean height of the population is 66.''

Ans. :

$H_0 : \mu = 66$

$H_1 : \mu \neq 66$

x_i	$d_i = x_i - 67$	d_i^2
63	– 4	16
63	– 4	16
66	– 1	1
67	0	0
68	1	1
69	2	4
70	3	9
70	3	9
71	4	16
71	4	16
678	8	88

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{678}{10} = 67.8$$

$$S^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

$$= \frac{1}{n}\left\{\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right\}$$

$$= \frac{1}{10}\left\{88 - \frac{(8)^2}{10}\right\}$$

$$= 8.16$$

$$S = \sqrt{8.16} = 2.8566$$

$$\text{Now } t = \frac{|\bar{x} - \mu|\sqrt{n-1}}{S}$$

$$= \frac{|67.8 - 66|\sqrt{10-1}}{2.8566}$$

$$= 1.89$$

The table value of t on $n - 1 = 9$ d.f and at 5% level of significance $= 2.26$

$\therefore t_{cal} < t_{tab}$

$\therefore H_0$ may be accepted

i.e. the mean height of the population may be regarded as 66.

Illustration 3 : The following is a random sample obtained from a normal population. Find 95% confidence limits for the mean of the population.

65, 72, 68, 71, 77, 61, 63, 69, 73, 71

Ans :

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
65	− 4	16
72	3	9
68	− 1	1
71	2	4
77	8	64
61	− 8	64
63	− 6	36
69	0	0
73	4	16
71	2	4
690	0	214

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{690}{10} = 69$$

$$S^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

$$= \frac{1}{10}(214) = 21.4$$

$$\therefore S = \sqrt{21.4} = 4.626$$

95% C.I. for the mean of the population

$$= \bar{x} \pm t_{0.05} \text{ (S.E. of } t)$$

$$= \bar{x} \pm 2.26 \frac{S}{\sqrt{n-1}}$$

$$= 69 \pm 2.26 \frac{4.626}{\sqrt{10-1}}$$

$$= 69 \pm 3.48$$

$\therefore$ 95% confidence limits for the mean of the population are from 65.52 to 72.48

Illustration 4 : A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025 cms. A random

sample of 10 washers was found to have an average thickness of 0.024 cms. with a standard deviation of 0.02 cms. Test the significance of the deviation. *(C.A., Nov. '80)*

Ans. :

$H_0 : \mu = 0.025$

$H_1 : \mu \neq 0.025$

$\bar{x} = 0.024,\ n = 10,\ S = 0.02$

$$\text{Now, } t = \frac{|\bar{x} - \mu|\sqrt{n-1}}{S}$$

$$= \frac{|0.024 - 0.025|\sqrt{10-1}}{0.02}$$

$$= 0.15$$

Table value of t on $10 - 1 = 9$ d.f and at 5% level of significance = 2.26

$\therefore\ t_{cal} < t_{tab}$

$\therefore$ H_0 may be accepted. Hence it may be said that there is no significant difference between the sample mean and the population mean.

Illustration 5 : For a sample of 9 observations the following information is obtained

$\bar{x} = 49,\ \Sigma(x_i - \bar{x})^2 = 52.$

Find 99% C.I. for population mean.

Ans. :

On $n-1 = 8$ d.f and at 1% level of significance the table value of $t = 3.355$

$$S^2 = \frac{1}{n}\ \Sigma(x_i - \bar{x})^2$$

$$= \frac{1}{9}\ (52) = 5.778$$

$$\therefore\ S = \sqrt{5.778}$$

$$= 2.4.$$

$\therefore$ 99% C.I. for population mean is

$$\bar{x} \pm t_{0.01}\ \frac{S}{\sqrt{n-1}}$$

$$= 49 \pm 3.355\ \frac{2.4}{\sqrt{9-1}}$$

$= 49 \pm 2.85$

$\therefore$ 99% C.I. for population mean is 46.15 to 51.85

EXERCISE 12.1

1. Define t statistic and give its probabilily density function.
2. Give properties and uses of t distribution.
3. Give the difference between large sample tests and small sample tests.
4. Explain how will test the significance of the difference between sample mean and the mean of the population.
5. An automatic machine is set to fill 170 tablets in a bottle. A sample of 10 bottles was examined and the number of tablets in them were 168, 164, 166, 167, 168, 169, 170, 170, 171, 170. Test whether the machine is set properly or not.

[Ans. : $\bar{x} = 168.3$, $t = 2.49$**]**

6. Ten students are selected at random from a college and their heights are found to be 100, 104, 108, 110, 118, 120, 122, 124, 126 and 128 cms. In the light of these data, discuss the suggestion that the mean height of the students of the college is 110 cms. [Use 5% level of significance] [The table value of "t" at 5% level for 8 d.f. is 2.306, for 9 d.f. is 2.262 and for 10 d.f. is 2.228 for a two tailed test]

(Sau. Uni., S.Y.B.B.A., April, 2000)

[Ans. : $t = 1.94$**]**

7. A company claims that the weight of their product is 10 kgs. A sample of 10 items from a lot supplied by the company has shown the following weights 10.2, 9.7, 10.3, 10.0, 9.8, 9.7, 9.6, 9.6, 9.6, 9.7, 9.4. Is there any statistical evidence to support the claim of the company about the weight of the items ?

(C. A. Foundation, May, '95)

[Ans. : $t = 2.24$**]**

8. The following information is obtained from

6. Test of Significance of Difference between Means of Two Small Samples :

Suppose two independent small samples of sizes n_1 and n_2 are drawn from two normal populations and the means of the samples are $\bar{x}_1$ and $\bar{x}_2$ respectively. If we want to test the hypothesis that population means are equal we can apply t test in the following way.

Under the assumption that both the population have the same variance.

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{|\bar{x}_1 - \bar{x}_2|}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$\text{Where } S^2 = \frac{1}{n_1 + n_2 - 2}\left\{\Sigma(x_i - \bar{x})^2 + \Sigma(x_2 - \bar{x}_2)^2\right\}$$

$$= \frac{1}{n_1 + n_2 - 2}\left\{n_1 S_1^2 + n_2 S_2^2\right\}$$

$$\text{Where } S_1^2 = \frac{1}{n_1}\Sigma(x_1 - \bar{x}_1)^2; \qquad S_2^2 = \frac{1}{n_2}\Sigma(x_2 - \bar{x}_2)^2$$

t is based on $n_1 + n_2 - 2$ degrees of freedom. For testing the null hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ the value of t is computed from the given data and it is compared with the table value of t on appropriate degrees of freedom and at a required level of significance. The decision regarding acceptance or rejection of the hypothesis is then taken.

Illustration 6 : Two horses A and B were tested for running a particular track. The time (in seconds) taken by them are given below :

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Can it be concluded that horse A is faster than horse B.

Ans. : $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

Here $n_1 = 7$, $n_2 = 6$

x_1	x_2	$d_1 = x_1 - 31$	d_1^2	$d_2 = x_2 - 28$	d_2^2
28	29	− 3	9	1	1
30	30	− 1	1	2	4
32	30	1	1	2	4
33	24	2	4	− 4	16
33	27	2	4	− 1	1
29	29	− 2	4	1	1
34		3	9		
219	169	2	32	1	27

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{219}{7} = 31.29$$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{169}{6} = 28.17$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left\{\Sigma(x_i - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2\right\}$$

$$= \frac{1}{n_1 + n_2 - 2}\left\{\Sigma d_1^2 - \frac{(\Sigma d_1)^2}{n_1} + \Sigma d_2^2 - \frac{(\Sigma d_2)^2}{n_2}\right\}$$

$$= \frac{1}{7 + 6 - 2}\left\{32 - \frac{(2)^2}{7} + 27 - \frac{(1)^2}{6}\right\}$$

$$= \frac{1}{11}[58.262]$$

$$= 5.2965$$

$$\therefore S = \sqrt{5.2965} = 2.3$$

$$\text{Now, } t = \frac{|\bar{x}_1 - \bar{x}_2|}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{|31.29 - 28.17|}{2.3} \times \sqrt{\frac{7 \times 6}{7 + 6}}$$

$$= 2.44$$

D.f. $= n_1 + n_2 - 2 = 11$

One tail table value of t on 11 d.f. and at 5% level of significance = 1.796

$\therefore t_{cal} > t_{tab}$

$\therefore H_0$ may be rejected

i.e. it may be concluded that horse A is faster than horse B.

Illustration 7 : Below are given the gain in weights (in lbs) of cows fed on two diets X and Y.

Diet X : 25 32 30 32 24 14 32

Diet Y : 24 34 22 30 42 31 40 30 32 35

Test at 5% level whether the two diets differ as regards their effects on mean increase in weight. *(C. A., May. '88)*

Ans. : $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

x_1	x_2	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
25	24	– 2	4	– 8	64
32	34	5	25	2	4
30	22	3	9	– 10	100
32	30	5	25	– 2	4
24	42	– 3	9	10	100
14	31	– 13	169	– 1	1
32	40	5	25	8	64
	30			– 2	4
	32			0	0
	35			3	9
189	320	0	266	0	350

$$\bar{x}_1 = \frac{\Sigma x_i}{n} = \frac{189}{7} = 27, \quad \bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{320}{10} = 32$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left\{\Sigma(x_1 - \bar{x})^2 + \Sigma(x_2 - \bar{x}_2)^2\right\}$$

$$= \frac{1}{7+10-2}\{266 + 350\}$$

$$= 41.067$$

$$S = \sqrt{41.067} = 6.41$$

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{|27 - 32|}{6.41} \times \sqrt{\frac{7 \times 10}{7+10}}$$

$$= 1.58$$

D.f $= n_1 + n_2 - 2 = 15$

Table value of t on 15 d.f and at 5% level of significance = 2.131

$\therefore t_{cal} < t_{tab}$

$\therefore$ H_0 may be accepted at 5% level of significance.

$\therefore$ Diets do not differ significantly.

Illustration 8 : Two random samples of sizes 9 and 7 respectively are drawn from two different populations. The means of the samples are 196.4 and 198.8 respectively. The sum of the squares of deviations from their respective means are 26.94 and 18.73. Test the hypothesis that population means are equal.

Ans. : $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$\bar{x}_1 = 196.4$, $\bar{x}_2 = 198.8$, $n_1 = 9$, $n_2 = 7$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left\{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2\right\}$$

$$= \frac{1}{9+7-2}\{26.94 + 18.73\}$$

$$= \frac{1}{14}(45.67) = 3.262$$

$\therefore S = \sqrt{3.262} = 1.806$

$$\text{Now, } t = \frac{|\bar{x}_1 - \bar{x}_2|}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{|196.4 - 198.8|}{1.806} \times \sqrt{\frac{9 \times 7}{9 + 7}}$$

$$= 2.637$$

D.f. $= n_1 + n_2 - 2 = 14$

Table value of t on 14 d.f and at 5% level of significance = 2.145

$t_{cal} > t_{tab}$

$\therefore$ H_0 may be rejected.

$\therefore$ Population means may not be regarded equal.

Illustration 9 : For two independent samples the following information is available.

Sample	*Size*	*Mean*	*S.D.*
I	*10*	*15*	*3.5*
II	*15*	*16.5*	*4.5*

Test the hypothesis that population means are equal

Ans. : Here, $n_1 = 10$, $n_2 = 15$, $\bar{x}_1 = 15$, $\bar{x}_2 = 16.5$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\{n_1 S_1^2 + n_2 S_2^2\}$$

$$= \frac{1}{10 + 15 - 2}\{10(3.5)^2 + 15(4.5)^2\}$$

$$= 18.5326$$

$$\therefore S = \sqrt{18.5326} = 4.3$$

$$\text{Now } t = \frac{|\bar{x}_1 - \bar{x}_2|}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{|15 - 16.5|}{4.3} \times \sqrt{\frac{10 \times 15}{10 + 15}}$$

$$= 0.85$$

D.f $= n_1 + n_2 - 2 = 23$

Table value of t on 23 d.f and at 5% level of significance = 2.069

$t_{cal} < t_{tab}$

∴ H_0 may be accepted at 5% level of significance

∴ Population means may be regarded equal.

Illustration 10 : Samples of two types of electric bulbs were tested for ngth of life and the following data were obtained

	Type I	*Type II*
Number of units	*8*	*7*
Mean (in hours)	*1134*	*1024*
S.D. (in hours)	*35*	*40*

Test at 5% level whether the difference in the sample means is gnificant

(Tables of values of *t* for 13 d.f. = 2.16 for 14 d.f. = 2.15, for 15 d.f. 2.13 at 5% level for two-tail areas and 1.77, 1.76 and 1.75 respectively r one tail area. *(C.A. Inter, Nov., '91)*

Ans. : Here $n_1 = 8, \bar{x}_1 = 1134, S_1 = 35$

$n_2 = 7, \bar{x}_2 = 1024, S_2 = 40$

$$S^2 = \frac{1}{n_1 + n_2 - 2} \{n_1 S_1^2 + n_2 S_2^2\}$$

$$= \frac{1}{8+7-2} \{8(35)^2 + 7(40)^2\}$$

$$= 1615.38$$

$$\therefore S = \sqrt{1615.38} = 40.192$$

$$\text{Now } t = \frac{|\bar{x}_1 - \bar{x}_2|}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{|1134 - 1024|}{40.192} \times \sqrt{\frac{8 \times 7}{8+7}}$$

$$= 5.288$$

D.f. $= n_1 + n_2 - 2 = 13$

Table value of t on 13 d.f. and at 5% level of significance = 2.16

$\therefore t_{cal} > t_{tab}$

∴ H_0 may be rejected

∴ The two types of bulbs differ significantly so far as their mean lives concerned.

7. Paired *t* Test for Difference of Means :

We can use *t* test for testing the significance of the difference between two sample means. If we want to test whether advertisement is effective or not, we may advertise in *n* shops selected at random. And we may not advertise in another randomly selected *n* shops. From the difference between the average sales by the two methods the null hypothesis that the sales by both the methods do not differ significantly can be tested. Sometimes it may so happen that the shops selected in the sample in which the advertisement is done may be quite different from the shops in which the advertisement is not done. Hence the result obtained may not be reliable. To overcome this difficulty *n* shops are selected at random and the advertisement is done in these shops. The differences in the sales after and before advertisement are found out. The difference between the two observations of a unit of a sample is devoted by d. Under the hypothesis that there is no difference in the population, *t* test can be applied in the following way.

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

$$t = \frac{\bar{d}\sqrt{n-1}}{S}$$

where $\bar{d} = \frac{\Sigma d_i}{n}$ and $S^2 = \frac{1}{n}\Sigma(d_i - \bar{d})^2$

$$= \frac{1}{n}\left\{\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right\}$$

The value of *t* is computed from the given data and it is compared with the table value of *t* on *n*–1 d.f and at a required level of significance. The decision regarding acceptance or rejection of the null hypothesis can be taken.

Illustration 11 : The sales data of an item in six shops before and after a special promotion campaign are as under :

Shops	A	B	C	D	E	F
Before campaign	53	28	32	48	50	42
After campaign	58	32	30	50	56	45

Can the campaign be judged as success. Test at 5% level of significance.

Ans. :

H_0 : Compaign is not effective $\mu = 0$

$H_1 = \mu \neq 0$

Sale before campaign	Sale after campaign	d	d^2
53	58	5	25
28	32	4	16
32	30	– 2	4
48	50	2	4
50	56	6	36
42	45	3	9
		18	94

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{18}{6} = 3$$

$$S^2 = \frac{1}{n}\left\{\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right\}$$

$$= \frac{1}{6}\left\{94 - \frac{(18)^2}{6}\right\}$$

$$= 6.667$$

$$\therefore S = \sqrt{6.6667} = 2.58$$

$$\text{Now } t = \frac{\bar{d}\sqrt{n-1}}{S}$$

$$= \frac{3\sqrt{6-1}}{2.58}$$

$$= 2.6$$

Table value of t on $n - 1 = 5$ d.f. and at 5% level of significance = 2.571

$\therefore t_{cal} > t_{tab}$

$\therefore H_0$ may be rejected at 5% level of significance

$\therefore$ Campaign may be judged as success.

Illustration 12 : A drug is given to 10 patients and the increments in their blood pressure were recorded as

8, 3, 6, 10, 2, – 2, 3, 0, – 6, 1

Is it reasonable to believe that the drug has no effect on change of blood pressure ?

Ans. :

H_0 : The drug has no effect on change of blood pressure $\mu = 0$

$H_1 : \mu \neq 0$

d	8	3	6	10	2	– 2	3	0	– 6	1
d^2	64	9	36	100	4	4	9	0	36	1

$\Sigma d_i = 25, \Sigma d_i^2 = 263$

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{25}{10} = 2.5$$

$$S^2 = \frac{1}{n}\left\{\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right\}$$

$$= \frac{1}{10}\left\{263 - \frac{(25)^2}{10}\right\}$$

$$= 20.05$$

$$S = \sqrt{20.05} = 4.48$$

$$\text{Now } t = \frac{\bar{d}\sqrt{n-1}}{S}$$

$$= \frac{2.5\sqrt{10-1}}{4.48}$$

$$= 1.67$$

Table value of t on $n - 1 = 9$ d.f. and at 5% level of significance = 2.26

$t_{cal} < t_{tab}$

$\therefore$ H_0 may be accepted at 5% level of significance

$\therefore$ It may be concluded that the drug has no effect on change of blood pressure.

EXERCISE 12.2

1. Explain how will you test the significance of difference between two sample means in small samples.
2. What is pair t-test ? Explain how it can applied.
3. Samples of sales in similar shops yielded the following information.
   For town A : $\bar{x}_1 = 3.45, \Sigma x_1 = 38, \Sigma x_1^2 = 228. n_1 = 11$

F TEST AND ANALYSIS OF VARIANCE

1. Introduction :

To test the hypothesis of equality of means in two small samples we use t-test. In applying t test it is assumed that the population from which the samples are drawn have equal variances. If this assumption is not correct the result obtained may not be reliable. Hence before applying t test, it is necessary to test that the population variances are equal i.e. $\sigma_1^2 = \sigma_2^2$. Snedecore's F test can be used for testing the hypothesis that the variances of the populations are equal. The statistic F is defined as

$F = \frac{\hat{s}_1^2}{\hat{s}_2^2}$ where $\hat{s}_1^2$ and $\hat{s}_2^2$ are unbiased estimates of the population variances.

2. Variance Ratio Test (F test) :

Suppose a random sample of size n_1 is drawn from a normal population having variance σ_1^2 and another independent random sample of size n_2 is drawn from another normal population having variance σ_2^2. If we are interested in testing the hypothesis that the population variances are equal, we can apply F test in the following way

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

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

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2}$$

$$= \frac{\frac{n_1}{n_1 - 1} S_1^2}{\frac{n_2}{n_2 - 1} S_2^2}$$

Where $S_1^2 = \frac{1}{n_1} \Sigma (x_1 - \bar{x}_1)^2$ and $S_2^2 = \frac{1}{n_2} \Sigma (x_2 - \bar{x}_2)^2$

F is thus the ratio of two independent unbiased estimates of population variances. F is based on $n_1 - 1, n_2 - 1$ degrees of freedom. It should be noted that F is defined as a ratio of two independent estimates of population variances where **the numerator is greater than the denominator.**

$$\therefore F = \frac{\frac{n_1}{n_1 - 1} S_1^2}{\frac{n_2}{n_2 - 1} S_2^2} \text{ on } n_1 - 1, n_2 - 1 \text{ d.f.}$$

OR

$$F = \frac{\frac{n_2}{n_2 - 1} S_2^2}{\frac{n_1}{n_1 - 1} S_1^2} \text{ on } n_2 - 1, n_1 - 1 \text{ d.f.}$$

From the given data, the value of F is computed and it is compared with the table value of F on appropriate degrees of freedom and at a required level of significance. The decision regarding acceptance or rejection of the hypothesis is then taken.

Illustration 1 : The following samples are drawn from two normal populations. Test the hypothesis that the population variances are equal.

Sample I	8	10	14	10	13		
Sample II	12	15	11	16	14	14	16

Ans. :

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

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

x_1	x_2	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
8	12	– 3	9	– 2	4
10	15	– 1	1	1	1
14	11	3	9	– 3	9
10	16	– 1	1	2	4
13	14	2	4	0	0
	14			0	0
	16			2	4
55	98	0	24		22

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{55}{5} = 11, \qquad \bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{98}{7} = 14$$

$$S_1^2 = \frac{1}{n_1}\Sigma(x_1 - \bar{x}_1)^2 = \frac{1}{5}(24) = 4.8$$

$$\frac{n_1}{n_1 - 1} S_1^2 = \frac{5}{4}(4.8) = 6$$

$$S_2^2 = \frac{1}{n_1}\Sigma(x_2 - \bar{x}_2)^2 = \frac{1}{7}(22) = 3.1428$$

$$\frac{n_2}{n_2 - 1} S_2^2 = \frac{7}{6}(3.1428) = 3.67$$

$$\text{Now, } F = \frac{\frac{n_1}{n_1 - 1} S_1^2}{\frac{n_2}{n_2 - 1} S_2^2} = \frac{6}{3.67} = 1.63$$

D.f. = $n_1 - 1, n_2 - 1 = 4, 6$

Table value of F on 4, 6 degrees of freedom and at 5% level of significance = 4.53

$\therefore F_{cal} < F_{tab}$

$\therefore H_0$ may be accepted

i.e. population variances may be regarded equal.

Illustration 2 : The following information is obtained for two samples drawn from two normal populations

Sample	Size	Mean	S.D.
I	10	12	3.162
II	12	15	5.115

Test the hypothesis that the population variances are equal

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

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

$$\frac{n_1}{n_1 - 1} S_1^2 = \frac{10}{9}(3.162)^2 = 11.11$$

$$\frac{n_2}{n_2 - 1} S_2^2 = \frac{15}{14}(5.115)^2 = 28.03$$

$$\text{Now } F = \frac{\frac{n_2}{n_2 - 1} S_2^2}{\frac{n_1}{n_1 - 1} S_1^2}$$

$$= \frac{28.03}{11.11}$$
$$= 2.52$$

D.f. = $n_2 - 1, n_1 - 1 = 14, 9$

Table value of F on 14, 9 d.f. and at 5% level of significance = 3.03

$F_{cal} < F_{tab}$

$\therefore$ H_0 may be accepted

$\therefore$ Population variances may be regarded equal.

Illustration 3 : Two independent samples provided the following results.

Sample	Size	Mean	Sum of squares of deviations from their respective means
I	10	12	120
II	12	13	144

Can the two samples be regarded as drawn from the same normal population.

Ans. :

For testing that both samples come from the same normal population we shall apply the following two tests.

(i) Population variances are equal

(ii) Population means are equal

(i) Population variances are equal

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

$$S_1^2 = \frac{1}{n_1} \Sigma(x_1 - \bar{x}_1)^2 = \frac{1}{10}(120) = 12$$

$$\frac{n_1}{n_1 - 1} S_1^2 = \frac{10}{9} (12) = 13.33$$

$$S_2^2 = \frac{1}{n_2} \Sigma(x_2 - \bar{x}_2)^2 = \frac{1}{12}(144) = 12$$

$$\frac{n_2}{n_2 - 1} S_2^2 = \frac{12}{11} (12) = 13.09$$

$$F = \frac{\frac{n_1}{n_1 - 1} S_1^2}{\frac{n_2}{n_2 - 1} S_2^2} = \frac{13.33}{13.09} = 1.02$$

D.f. = $n_1 - 1, n_2 - 1 = 9, 11$

For $F_{9, 11}$ and at 5 % level of significance table value = 2.9

$F_{cal} < F_{tab}$

$\therefore$ H_0 may be accepted

i.e. The samples may be regarded as drawn from populations with equal variances :

(ii) Population means are equal

$H_0 : \mu_1 = \mu_2$

$\bar{x}_1 = 12,\ \bar{x}_2 = 13$

$$S^2 = \frac{1}{n_1 + n_2 - 2} \{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2\}$$

$$= \frac{1}{10 + 12 - 2} \{120 + 144\}$$

$$= 13.2$$

$$\therefore \quad S = \sqrt{13.2} = 3.63$$

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{|12 - 13|}{3.63} \times \sqrt{\frac{10 \times 12}{10 + 12}}$$

$$= 0.64$$

Table value of t on $n_1 + n_2 - 2 = 20$ d.f. and at 5 % level of significance $= 2.089$

$t_{cal} < t_{tab}$

$\therefore$ H_0 may be accepted i.e. population means may be regarded equal.

As the variances of the two populations are equal and also their means are equal, the two samples may be regarded as drawn from the same normal population.

EXERCISE 13.1

1. What is variance ratio ? Explain the test based on it.
2. The following figures give the weights of products of items produced by two machines. Test the hypothesis that there is no significant variation in the products of two machines.

Machine A :	3	7	5	6	5	4	4	5	3	3
Machine B :	8	5	7	8	3	2	7	6	5	7

[Ans. : F = 2.28]

3. The following are samples drawn from two normal populations. Test the hypothesis that population variances are equal.

Sample I :	22	15	18	20	25	24	16	20		
Sample II :	27	33	40	35	32	35	37	29	41	31

[Ans. : F = 1.59]

13

CHI–SQUARE TEST

1. Definition of χ^2
2. Uses of χ^2
3. Goodness of fit test
4. Limitations of χ^2 test
5. Test of independence of two attributes
6. Yate's correction
7. To test the specified value of the variance of the population

1. Definition of χ^2 :

If $x_1, x_2, x_3 \ldots\ldots x_n$ is a random sample of size n, from a normal population with mean 0 and S.D. 1, then the distribution of Σx_i^2 is called χ^2 distribution on n degrees of freedom. Similarly if $x_1, x_2, x_3 \ldots\ldots x_n$ is a random sample of size n from a normal population with mean μ and S.D. σ, then the distribution of $\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2$ is also χ^2 distribution with n degrees of freedom.

The probability density function of χ^2 distribution is

$$f(\chi^2) = \frac{1}{2^{\frac{n}{2}} \left|\frac{n}{2}\right.} \cdot e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}, \; 0 < \chi^2 < \infty$$

χ^2 is a continuous distribution and the form of the distribution depends upon degrees of freedom n. The mean of the distribution is n and its variance is $2n$.

2. Uses of χ^2 :

χ^2 distribution has a large number of applications in statistics. We shall discuss the following three main uses of χ^2

1. To test goodness of fit
2. To test independence of attributes
3. To test a specified value of the variance of the population.

3. Goodness of Fit Test :

Suppose we have obtained an observed frequency distribution and we are interested in knowing whether the observed frequency distribution support a particular hypothesis. For this a very powerful test for testing the significance of the discrepancy between observed frequency distribution and expected frequency distribution was given by **Karl Pearson** in 1900. The test is known as χ^2 test of goodness of fit.

Under the null hypothesis that there is no significant difference between observed and expected frequencies, the value of χ^2 is calculated by the formula :

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

If all the observed frequencies and expected frequencies are equal, the value of χ^2 will be zero. This will signify a perfect agreement of observations with expectations. More the value of χ^2, more is the divergence between the observed and expected frequencies.

The value of χ^2 is calculated from the given data and it is compared with the table value of χ^2 on $n-1$ degrees of freedom and at a required level of significance. If calculated value of χ^2 is less than table value of χ^2 the null hypothesis may be accepted and it may be concluded that the given frequency distribution fits the hypothesis. And if the calculated value of χ^2 is greater than the table value of χ^2, the hypothesis may be rejected and it may be concluded that the observed frequency distribution does not fit the hypothesis.

Note : The degrees of freedom in applying goodness of fit test is $n-k-1$, where k is the number of parameters estimated.

4. Limitations of χ^2 Test :

(1) The observations of the sample should be independent.

(2) Absolute frequencies should always be used

(3) If there are any constraints on class frequencies, then they must be linear

(4) The frequency of any class should not be less than 5. If any class frequency is less than five it should be combined with the frequency of the adjoing class or classes, so that the total frequency of combined classes is more than 5.

(5) The class frequencies should be combined is such a way that degrees of freedom is more than 0.

Illustration 1 : A die is thrown for 300 times and the following distribution is obtained. Can the die be regarded unbiased.

Number on the die	1	2	3	4	5	6
Frequency	41	44	49	53	57	56

Ans. H_0 : Die is unbiased i.e. the probability of obtaining any number

$$= \frac{1}{6}$$

Number on the die	Observed frequency o_i	Expected frequency e_i	$\frac{(o_i - e_i)^2}{e_i}$
1	41	50	$\frac{81}{50} = 1.62$
2	44	50	0.72
3	49	50	0.02
4	53	50	0.18
5	57	50	0.98
6	56	50	0.72
	300	300	4.24

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

$$= 4.24$$

$$\text{D.f} = n - 1 = 6 - 1 = 5$$

Table value of χ^2 on 5 d.f. and at 5% level of significance = 11.07

$\chi^2_{cal} < \chi^2_{tab}$

$\therefore$ H_0 may be accepted.

$\therefore$ Die may be regarded unbiased.

Illustration 2 : The number of road accidents on a high way during a week is given below. Can it be concluded that the proportion of accidents are equal for all days.

Day	Mon.	Tue.	Wed.	Thurs.	Fri.	Sat.	Sun.
Number of accidents	14	16	8	12	11	9	14

Ans. : H_0 : The proportion of accidents is same for all days i.e. probability of an accident on any day is $\frac{1}{7}$.

Day	Mon.	Tue.	Wed.	Thurs.	Fri.	Sat.	Sun.	Total
Observed frequency	14	16	8	12	11	9	14	84
Expected frequency	12	12	12	12	12	12	12	84

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(14-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(12-12)^2}{12}$$

$$+ \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(14-12)^2}{12}$$

$$= \frac{4+16+16+0+1+9+4}{12}$$

$$= \frac{50}{12}$$

$$= 4.17$$

D.f. $= n - 1 = 7 - 1 = 6$

Table value of χ^2 on 6 d.f. and at 5% level of significance = 12.59

$\chi^2_{cal} < \chi^2_{tab}$

$\therefore$ H_o may be accepted.

$\therefore$ Proportion of accidents is same for all days.

Illustration 3 : The units produced by a plant are classified into four grades. The past performance of the plant shows that the respective proportions are 8 : 4 : 2 : 1. To check the run of the plant 600 parts were examined and classified as follows. Is there any evidence of a change in production standards.

Grade	First	Second	Third	Fourth	Total
Units	340	130	100	30	600

Ans : H_o : There is no change in production standards.

Grade	Units Observed o_i	Units Expected e_i	$\frac{(o_i - e_i)^2}{e_i}$
First	340	$600 \times \frac{8}{15} = 320$	1.25
Second	130	$600 \times \frac{4}{15} = 160$	5.625
Third	100	$600 \times \frac{2}{15} = 80$	5.00
Fourth	30	$600 \times \frac{1}{15} = 40$	2.50
	600		14.375

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

$= 14.375$

D.f. $= n-1 = 4-1 = 3$

Table value of χ^2 on 3 d.f and at 5 % level of significance = 7.815

$\chi^2_{cal} > \chi^2_{tab}$

$\therefore$ H_o may be rejected.

$\therefore$ There is evidence in change of production standards.

Illustrations 4 : Five coins are tossed for 320 times and the following distribution of number of heads is obtained

Number of heads :	0	1	2	3	4	5
Frequency :	8	42	116	90	52	12

Test the hypothesis that the coins are unbiased.

Ans. : H_o : Coins are unbiased i.e. $p = \frac{1}{2}$

Number of heads	Observed frequency o_i	Probability $P(xi)$	Expected frequency $e_i = N \times P(x_i)$	$\frac{(o_i - e_i)^2}{e_i}$
0	8	${}^5c_0\, p^0q^5 = \frac{1}{32}$	10	0.40
1	42	${}^5c_1\, p^1q^4 = \frac{5}{32}$	50	1.28
2	116	${}^5c_2\, p^2q^3 = \frac{10}{32}$	100	2.56
3	90	${}^5c_3\, p^3q^2 = \frac{10}{32}$	100	1.00
4	52	${}^5c_4\, p^4q = \frac{5}{32}$	50	0.08
5	12	${}^5c_5\, p^5q^0 = \frac{1}{32}$	10	0.40
	320	1	320	5.72

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

$= 5.72$

D.f. $= n - 1 = 6 - 1 = 5$

Table value of χ^2 on 5 d.f. and at 5% level of significance = 11.07

$\chi^2_{cal} < \chi^2_{tab}$

$\therefore$ H_o may be accepted.

$\therefore$ Coins may be regarded unbiased

Illustration 5 : The following is a distribution of mistakes committed by a typist in 100 pages.

Number of Mistakes	0	1	2	3	4	5	6
Number of Pages	11	31	26	17	10	4	1

Fit a poisson distribution and test the goodness of fit ($e^{-2} = 0.135$)

Ans. : H_0 : Poisson distribution fits the data

Number of mistakes x	Number of pages f		fx	$P(xi)$	Expected frequency $ei = N \times P(xi)$		$\frac{(o_i - e_i)^2}{e_i}$
0	11		0	0.135	13.5		0.4630
1	31		31	0.270	27.0		0.5926
2	26		52	0.270	27.0		0.0370
3	17		51	0.180	18.0		0.0556
4	10	15	40	0.090	9.0	13.8	
5	4		20	0.036	3.6		0.1043
6	1		6	0.012	1.2		
	100		200				1.2525

$$\text{Mean } m = \frac{\Sigma fx}{N} = \frac{200}{100} = 2$$

For poisson distribution $P(x) = \frac{e^{-m} m^x}{x!}$

$\therefore \quad P(0) = e^{-m} = e^{-2} = 0.135$

$$P(1) = \frac{m}{1} P(0) = \frac{2}{1} (0.135) = 0.270$$

$$P(2) = \frac{m}{2} P(1) = \frac{2}{2} (0.270) = 0.270$$

$$P(3) = \frac{m}{3} P(2) = \frac{2}{3} (0.270) = 0.180$$

$$P(4) = \frac{m}{4} P(3) = \frac{2}{4} (0.180) = 0.090$$

$$P(5) = \frac{m}{5} P(4) = \frac{2}{5} (0.090) = 0.036$$

$$P(6) = \frac{m}{6} P(5) = \frac{2}{6} (.036) = 0.012$$

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i} = 1.2525$$

D.f. $= n - 1 - 1 = 5 - 2 = 3$

Table value of χ^2 on 3 d.f. and at 5 % level of significance = 7.82

$\chi^2_{cal} < \chi^2_{tab}$

$\therefore$ Ho may be accepted

$\therefore$ Poisson distribution fits the data.

[Note : In fitting a poisson distribution mean m is estimated, hence one d.f. is lost.]

EXERCISE 13.1

1. Define χ^2 and give its uses
2. Explain goodness of fit test
3. Give the limitations of χ^2 lest
4. The demands of books from a library from Monday to Friday are given below :

Day	Mon.	Tue.	Wed.	Thurs.	Fri.
Demand	136	100	120	130	114

Test the hypothesis that the demands of number of books do not depend upon the day of the week.

[Ans. : $\chi^2 = 6.6$ **]**

5. A die is thrown 150 times and the following results are obtained :

Number turned up :	1	2	3	4	5	6
Frequency :	19	23	28	17	32	31

Test the hypothesis that the die is unbiased at 5% level of significant

[Ans. : $\chi^2 = 7.92$**]**

6. In an experiment of pea–breeding, the following frequencies of seeds were obtained.

Round and yellow	315
Wrinkled and yellow	101
Round and green	108
Wrinkled and green	36
Total	560

Theory predicts that the frequencies should be in the proportion 9 : 3 : 3 : 1 respectively. Are the data consistent with the theory at 5% level of significance ?

[Ans. : $\chi^2 = 0.267$**]**

7. Genetic theory states that children having one parent of blood group M and other of blood group N, will always be one of the three types M, MN, N and that the proportion of these types will be on an average 1 : 2 : 1.

5. Test of Independence of Two Attributes :

When the data are classified according to two attributes, χ^2 can also be used to test the hypothesis that the two attributes are independent.

Suppose the data are classified into r classes $A_1, A_2, A_3, \ldots A_r$ according to attribute A and into c classes $B_1, B_2 B_3 \ldots B_c$ according to attribute B. The representation of the data in a cross-classified table known as a contingency table is given below. In the $r \times c$ contingency table the observed frequencies of different cells are shown.

	B_1	B_2	B_3 ..... B_j ..... B_c	
A_1	O_{11}	O_{12}	O_{13} ..... O_{1j} ..... O_{1c}	(A_1)
A_2	O_{21}	O_{22}	O_{23} ..... O_{2j} ..... O_{2c}	(A_2)
A_3	O_{31}	O_{32}	O_{33} ..... O_{3j} ..... O_{3c}	(A_3)
....	.....			.....
....	.....			.....
....	.....			.....
A_i	O_{i1}	O_{i2}	O_{i3} ..... O_{ij} ..... O_{ic}	(A_i)
....	.....			.....
....	.....			.....
....	.....			.....
A_r	O_{r1}	O_{r2}	O_{r3} ..... O_{rj} ..... O_{rc}	(A_r)
	(B_1)	(B_2)	(B_3) ..... (B_j) ..... (B_c)	N

The total of i^{th} row is denoted by (A_i) and the total of j^{th} column is denoted by (B_j). O_{ij} denotes the frequency of the cell common to i^{th} row and j^{th} column. The total frequency is N. i.e. $\Sigma(A_i) = \Sigma(B_j) = N$

Under the null hypothesis that the two attributes A and B are independent, we shall find the expected frequency of $(i, j)^{th}$ cell.

The probability that any observation will fall in the i^{th} row $= \frac{(A_i)}{N}$

Similarly the probability that any observation will fall in the j^{th} column $= \frac{(B_j)}{N}$

Under the hypothesis of independence the probability that any observation will fall in the i^{th} row and j^{th} column $= \frac{(A_i)}{N} \times \frac{(B_j)}{N}$

$\therefore$ Expected frequency of $(i, j)^{th}$ cell

$$e_{ij} = N \times \frac{(A_i)}{N} \times \frac{(B_j)}{N}$$

$$= \frac{(A_i)(B_j)}{N}$$

Thus, we can find and expected frequencies of all the cells. From observed frequencies o_{ij} and expected frequencies e_{ij}; the value of χ^2 can be obtained by the following formula.

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

The number of independent cells in a $r \times c$ contingency table is $(r - 1)(c - 1)$. Hence the degrees of freedom in a $r \times c$ table is $(r - 1)(c - 1)$.

For testing the hypothesis of independence of two attributes A and B, the value of χ^2 is found out and it is compared with the table value of χ^2 on $(r - 1)(c - 1)$ d.f. and at a required level of significance. If calculated χ^2 is greater than the table value of χ^2, the hypothesis may be rejected i.e the two attributes may not be regarded independent. If calculated χ^2 is less than the table value of χ^2, the hypothesis that the attributes are independent may be accepted.

Illustration 6 : In an industry, 200 workers employed for a specific job were classified according to their performance and training received / not received. Test independence of training and performance. The data are summarised as follows.

	Performance		
	Good	Not Good	Total
Trained	100	50	150
Untrained	20	30	50
	120	80	200

(C.A., foundation, May, 1997)

Ans.: H_0 : Performance is independent of training.

	Performance		
	Good	Not Good	Total
Trained	100 (90)	50 (60)	150
Untrained	20 (30)	30 (20)	50
Total	120	80	200

Expected frequency of cell (1, 1) $= \frac{150 \times 120}{200} = 90$

The expected frequencies of different cells are indicated in brackets in the cells.

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(100-90)^2}{90} + \frac{(50-60)^2}{60} + \frac{(20-30)^2}{30} + \frac{(30-20)^2}{20}$$

$$= 1.11 + 1.67 + 3.33 + 5$$

$$= 11.11$$

D.f. $= (r-1)(c-1) = (2-1)(2-1) = 1$

On 1 d.f. and at 5% level of significance table value of $\chi^2 = 3.84$

$\chi^2_{cal} > \chi^2_{tab}$

$\therefore$ Ho may be rejected.

$\therefore$ Performance depends upon training.

Illustration 7 : The result in the last examination of a sample of 100 students is given below :

	Ist class	2nd class	3rd class	Total
Boys	10	28	12	50
Girls	20	22	8	50
Total	30	50	20	100

Can it be said that the performance in the examination depends upon sex.

Ans. : H_0 : Sex and performance in the examination are independent.

	1st class	2nd class	3rd class	Total
Boys	10 (15)	28 (25)	12 (10)	50
Girls	20 (15)	22 (25)	8 (10)	50
Total	30	50	20	100

Expected frequency of cell (1,1) $= \frac{50 \times 30}{100} = 15$

Expected frequency of cell (1,2) $= \frac{50 \times 50}{100} = 25$

Thus, expected frequencies of different cells are found out and are shown in brackets in the cells.

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(10-15)^2}{15} + \frac{(28-25)^2}{25} + \frac{(12-10)^2}{10} + \frac{(20-15)^2}{15}$$

$$+ \frac{(22-25)^2}{25} + \frac{(8-10)^2}{10}$$

$$= 1.67 + 0.36 + 0.4 + 1.67 + 0.36 + 0.4$$

$$= 4.86$$

D.f. $= (r-1)(c-1) = (2-1)(3-1) = 2$

on 2 d.f. and at 5 % level of significance table value of $\chi^2 = 5.99$

$\chi^2_{cal} < \chi^2_{tab}$

$\therefore$ H_0 may be accepted.

$\therefore$ Sex and performance in the examination may be regarded independent.

Illustration 8 : In a certain sample of 2000 families, 1400 families are consumers of tea. Out of 1800 Hindu families 1236 families consume tea. Use χ^2 test and state whether there is any significant difference between consumption of tea among Hindu and Non–Hindu families.

(C.A., May, 1978)

Ans. :

First of all we shall prepare a 2×2 contingency table as follows :

	Consumers of Tea	Not consumers of Tea	Total
Hindu Families	1236 (1260)	564 (540)	1800
Non-Hindu Families	164 (140)	36 (60)	200
Total	1400	600	2000

Ho : Consumption of tea does not depend upon Hindu and Non-Hindu families.

Expected frequency of cell (1,1) $= \frac{1800 \times 1400}{2000} = 1260$

Expected frequencies of different cells are found out and are shown in the cells in brackets.

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(1236 - 1260)^2}{1260} + \frac{(564 - 540)^2}{540} + \frac{(164 - 140)^2}{140} + \frac{(36 - 60)^2}{60}$$

$$= \frac{576}{1260} + \frac{576}{540} + \frac{576}{140} + \frac{576}{60}$$

$$= 0.46 + 1.07 + 4.11 + 9.60$$

$$= 15.24$$

D.f. $= (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$

Table value of χ^2 on 1 d.f and at 5% level of significance = 3.84

$\chi^2_{cal} > \chi^2_{tab}$

$\therefore$ H_o may be rejected.

$\therefore$ We may conclude that there is a significant difference between consumption of tea among Hindu and Non–Hindu families.

6. Yate's Correction :

χ^2 is a continuous distribution and it fails to maintain its characteristic of continuity if any of the expected frequency is less than 5. So in a 2×2 contingency table with cell frequencies $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$ F. Yate's in 1934 suggested