Sampling Theory and Tests of Significance

Every day, people in the ordinary business of life make judgements and take action based on samples. We come across people making assessment of the population through samples. The importance of the theory of sampling lies in the fact that for a large population it is neither practical nor necessary to collect data for each and every member of the population. For example, to collect an information about the economic conditions of the rural population of Kerala, it would require a large establishment to collect data, and then processing, tabulation etc. in order to calculate the parameters from it. But it would be quite sufficient if a village is selected and, if that village is adequately representative of the population, we shall be able to arrive at valid conclusions. People have been using sampling theory without knowing about it. For example, a grain merchant does not examine each grain of wheat that he purchases. The merchant simply takes out a handful and from it gets an idea about the quality of the whole consignment. In some cases, only sample method of study is possible. For example, if the universe is infinite or is hypothetical, then the census type of enquiry is not possible and only sample studies can be conducted. Again, for example, if some bombs are to be tested for effectiveness, inspection would destroy the bombs themselves. Similarly, if lives of bulbs, engines, crackers etc. are to be estimated, inspection will destroy the items themselves. Only a small sample study is enough to draw inferences. Larger group from which the sample is drawn is called the universe or population. The meaning of the term 'population' or 'universe' in statistical sampling differs from the ordinary meaning of this term. In Statistics, population is not restricted to mean a number of persons. That is, in statistics, it is the totality of persons, objects, items or anything conceivable pertaining to certain characteristics.

When we are interested to examine a set of data which is large in respect of a particular character, much time and energy is needed for the analysis. Therefore, from the big mass, a part, which is representative of the whole, is selected ; this part is known as sample and the process of selection is known as sampling. The mass is known as population or parent universe or universe from where samples are

taken. W.A. Spur states, "The process by which we draw a conclusion about some measure of a population is based on sample value. The measure might be a variable, such as the average or mean. The purpose of sampling is to estimate some characteristics for the population from which the sample is selected." In almost all the fields, the statistical enquiry deals with the drawing of conclusions concerning a population, from a sample selected from it. A major theory of sampling is the basis on which the estimation of the different characteristics of the population is evaluated. This means, the results of the samples can be generalised to the population and verified how far these results stand with confidence. Because of practical difficulties, the census method of enquiry is not possible. Therefore, inferences of the population have to be analysed on the basis of the information contained in the samples. Statistical sample plays a part in almost all statistical studies on which decisions for future action are to be used. Sometimes, the population is infinite and can never be studied in its entirety ; therefore sample study is adopted and becomes necessary. As mentioned earlier, even if the population is finite, studying them through sample is easier, economical and time-saving. Sampling theory is a study of relationship between a population and a sample drawn from it and estimation of the population parameters. Thus the statistical inference is divided into two heads.

1. Estimation Theory.
2. Testing of Hypothesis.

Estimation Theory

Statistical estimation is a method by which population parameters are estimated from the sample information. As already pointed out, the estimation can be obtained either by census or by sample method. Due to infinite population or limitation of money, time, etc., sample is adopted. Statistical estimation helps in estimating the mean, standard deviation, etc., of population quantities.

When estimating parameters of the population, the following two types of the estimates are possible.

(*a*) Point estimate

(*b*) Interval estimate

(*a*) *Point estimate.* An estimate by a single value of statistic, used to approximate the parameter of an unknown population, is known as point estimate or estimater of the parameter. For *e.g.,* sample mean, used for estimating the population mean, is an estimator.

(*b*) *Interval estimate.* While point estimate is a single value of statistics used as an estimate of the population parameter, Interval estimate means the population parameter given by two numbers between which the parameter is considered. Generally, point

estimation does not confidently tie down our information. Therefore, two values are computed in such a way that the interval lies between the two values containing the parameter. An interval so obtained is called interval estimate or confidence interval. For instance, studying a sample, we estimate that the average salary of a factory worker is Rs. 600 ; it is a point estimate. At the same time we may estimate through a sample study that an average salary of factory workers can lie between Rs. 600 and Rs. 700 ; this is an interval estimate.

Both methods of estimation have their own advantages and limitations. But in Statistics, the exact value of the parameter is not necessary. As such, generally interval estimate is adopted in practice. A good estimator is one which is very close to the value of the parameter. It must possess the following properties:

1. *Unbiasedness.* An estimator is expected to be unbiased if the value is equal to the parameter. For instance, when r becomes sufficiently large, many estimators are there and thus biases can be reduced to negligible value.

2. *Consistency.* When the sample size increases, the difference between sample statistics and parameters becomes smaller and smaller. If sufficient estimators are available, then the value of the parameter can be closely estimated.

3. *Efficiency.* Smaller the variance of an estimator, the distribution of the estimator is better, because its value will be closer to the parameter value. The efficiency depends upon the minimum variance.

4. *Sufficiency.* If the estimator possesses all the information regarding the parameter, the estimator is said to be sufficient. In other words, the estimator contains or conveys as much information as possible about the parameter.

Testing of Hypothesis

A sample investigation produces results ; and with these results, decisions are made on the population. But such decisions involve an element of uncertainty causing wrong decisions. Hypothesis is an assumption which may or may not be true about a population parameter. For example, tossing a coin 300 times, one may get 190 heads and 110 tails. At this instance, we are interested in testing whether the coin is unbiased or not. Therefore, we may conduct a test to judge significance whether the difference is due to sampling. The procedure of carrying out a significance test is as follows :

The word PARAMETER is used to indicate various statistical measures like, mean, standard deviation, correlation etc, in the universe. As against this the term STATISTIC refers to the statistical measures relating to the sample. That is parameters are functions of the population values while statistics are functions of the sample observations.

1. Laying Down of Hypothesis

To verify our assumption, which is based on sample study, we collect data and find out the difference between the sample value and the population value. If there is no difference or if the difference is very small, then our hypothesized value is correct. Generally two hypotheses must be constructed ; and if one hypothesis is correct, the other one is rejected.

(a) Null Hypothesis

It is a very useful tool to test the significance of difference. Any hypothesis concerning a population is called a statistical hypothesis. In the process of statistical test, the hypothesis is rejected or accepted, based on sample drawn from population. The statistician tests the hypothesis through observation and gives a probability statement. The simple hypothesis reveals that the value of sample and the value of the population under study do not show any difference. The hypothesis we have assumed is said to be null hypothesis ; it means that the true difference between the mean of sample and the mean of population is nil ; the least difference found is unimportant or due to sampling error. The rejection of null hypothesis ; it means that the true difference between the mean of sample and the mean of population is nil ; the least difference found is unimportant or due to sampling error. The rejection of null hypothesis reveals that the decision is correct.

For *e.g.*,

(*i*) The average height of the students of a university is 155 cms.

(*ii*) The average daily sales of a firm is Rs. 1500.

(*iii*) The average income of a mean of a particular village is Rs. 100.

All these statements will have to be verified on the basis of sample tests. Generally a hypothesis states that there is no difference between the mean of sample and the population. A statistical hypothesis is a null hypothesis if it is accepted. A null hypothesis is denoted by H_0.

(b) Alternative hypothesis

Rejection of H_0 leads to the acceptance of alternative hypothesis, which is denoted by H_1. For *e.g.*

$$H_0 = \mu = 155 \text{ (Null hypothesis)}$$

$$H_1 = \mu \neq 155 \text{ i.e., } \mu > 155 \text{ or } \mu < 155$$

(Alternative hypothesis)

When there are two hypotheses set up, the acceptance or rejection of a null hypothesis is based on a sample study. Thus it leads

to two wrong conclusions. *i.e.*, (*i*) Rejecting H_0, when H_0 is true (μ) Accepting H_0, when H_0 is false. This can be expressed in the following table.

	Decision from sample	
	Accept H_0	*Reject* H_0
H_0 true	correct	wrong (Type I error)
H_0 false (H_1 true)	wrong (Type II error)	correct

By rewriting

Reject H_0 when it is true (Type I error) $= \alpha$

Accept H_0 when it is false (Type II error) $= \beta$

Accept H_0 when it is true (Correct decision)

Reject H_0 when it is false (Correct decision)

2. Level of Significance

The maximum probability of committing type I error, which we specified in a test, is known as the level of significance. Generally 5% level of significance is fixed in statistical tests. This implies that we can have 95% confidence in accepting a hypothesis or we could be wrong 5%.

3. Critical Region

The range of variation has two regions—acceptance region and critical region or rejection region. If the sample statistics falls in critical region we have to reject the hypothesis, as it leads to false decision. We go for H_1, if the computed value of sample statistic falls in rejecting region.

4. One-tailed and two-tailed tests

The critical region under a normal curve, as stated earlier can be divided into two ways. (*a*) two sides under a curve (*b*) one side under a curve ; and both are either at the right tail or at the left tail.

5. Making a Decision or Conclusion

Finally we come to a conclusion either to accept or reject the null hypothesis. The decision is on the basis of computed value whether it lies in the acceptance region or rejected region.

If the computed value of the test statistic is less than the critical value, the computed value of the test statistic falls in the acceptance region and the null hypothesis is accepted. If the computed value of the test statistic is greater than the critical value, the computed value of the statistic falls in the rejection region and the null hypothesis is rejected.

Standard Error

If we select a number of independent random samples of a definite size from a given population and calculate some statistics, like mean, standard deviation etc., from each sample, we shall get a series of values of these statistics. These values obtained from the different samples can be put in the form of a frequency distribution. If we calculate the mean of the sampling distribution it could be deemed to be the Mean of the universe. Similarly, the standard deviation of the Sampling Distribution would be called the Standard Error. If 10 samples have been taken from a universe and if $X_1, X_2, X_3 ..... X_n$ represent their mean values, then the mean of these mean values would be close to the mean of the universe and the standard deviation of these values or the standard deviation of the Sampling Distribution of mean values would be called the Standard Error. The formula for this is $\frac{\sigma}{r}$.

Utility

1. It is a useful instrument in testing the hypothesis. We may test the hypothesis at 5% level of significance, which means, if the difference between observed and expected mean is more than 1.96 S.E., the hypothesis is not accepted and one has to go in for the alternative hypothesis. The level of significance can be 1%. Generally, the hypothesis is accepted if the difference is less than 3 S.E.; 5% level is popular.
2. Reliablity of a sample can be known.
3. The value of the parameters can be determined along with limits.

Now we discuss the various tests of significance to be applied on various situations under the following heads :

1. Tests of Significance for Attributes
2. Tests of Significance for Variables (Large Samples)
3. Tests of Significance for Variables (Small Samples)

Tests of Significance for Attributes

The sampling of attributes may be regarded as the drawing of samples from a population whose members consist of presence or

absence of a particular characteristics. For example, in the study of attribute blind, a sample may be drawn and its members are classified as blind and not blind. The presence of attribute may be represented by P and the absence of attribute may be represented by q. Thus, of 1000 people, 25 are blind and remaining are not blind. In other words $p = \frac{25}{1000}$ or 0.025 and $q = 0.975$. The various types of significance test may be studied under the following heads :

(A) Test for Number of Successes

(B) Test for Proportion of Successes, and

(C) Test for Difference between Proportions.

(A) Test for Number of Successes

The sampling distribution of the number of successes follows a binomial probability distribution. Hence its standard error is given by the formula :

S.E. of No. of successes $= \sqrt{npq}$

n = Size of sample
p = Probability of success in each trial
q = (1-p) *i.e.*, probability of failure

Illustration : 1

In 600 throws of a six faced dice, odd points appeared 360 times. Would you say that the dice is fair at 5% level of significance ?

(MBA, Kurukshetra)

Solution :

Let us take the hypothesis that the dice is fair.

In a fair dice we would expect 300 odd points in 600 throws.

$$S.E. = \sqrt{npq} = \sqrt{600 x \frac{1}{2} x \frac{1}{2}} = 12.247$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{360 - 300}{12.247} = 4.89$$

Since the difference is more than 1.96 at 5% level of significance, the hypothesis is rejected. Hence we cannot say that the dice is fair at 5% level of significance.

Illustration : 2

A person throws 10 dice 500 times and obtains 2560 times 4, 5 or 6. Can this be attributed to fluctuations of sampling.

(M.Com., Osmania)

Solution :

$$\text{S.E.} = \sqrt{npq} = \sqrt{5000 \times \frac{1}{2} \times \frac{1}{2}} = 35.355$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{2560 - 2500}{35.355} = 1.697 \text{ or } 1.7$$

Since the difference is less than 1.96 S.E. at 5% level of significance the hypothesis holds true. Hence the difference could be attributed to fluctuations in sampling.

Illustration : 3

In a sample of 400 population from a village 230 are found to be eaters of vegetarian items and the rest non-vegetarian items. Can we assume that both vegetarian and non-vegetarian food are equally popular ? *(B.Com., Madurai)*

(Given that for 5% level of significance, the S.E is 1·96)

Solution : Let us take the hypothesis that both type of food are equally popular.

$$n = 400; p = \frac{1}{2} \text{ and } q = \frac{1}{2}$$

Standard Error of No. of vegetarian $= \sqrt{npq}$

$$= \sqrt{400 \times \frac{1}{2} \times \frac{1}{2}} = 10$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{230 - 200}{10} = 3$$

Since the difference observed and expected number of vegetable eaters is more than 1.96 S.E. at 5% level of significance, the result of the experiment does not support hypothesis and thus vegetarian and non-vegetarian food are not equally popular.

Illustration : 4

In a sample of 500 people from a village in Rajasthan, 280 are found to be rice eaters and the rest wheat eaters. Can we assume that both the food articles are equally popular ? *(M.A. Ecos. Madras)*

Solution :

We take the hypothesis that the food articles are equally popular.

Then, the expected frequency of wheat eaters and rice eaters are 250 : 250.

$$\text{S.E.} = \sqrt{npq} = \sqrt{500 \times \tfrac{1}{2} \times \tfrac{1}{2}} = 11.18$$

Difference between actual and observed $= 280 - 250 = 30$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{30}{11.18} = 2.68$$

The difference is more than 2.58 S.E. at 1% level. It is not because of sampling fluctuations. Therefore, we may assume that both the food articles are not equally popular.

Illustration : 5

A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin may be an unbiased one, and explain briefly the theoretical principles you would use for this purpose. *(B. Com. Madurai)*

Solution :

An unbiased coin turns up head $= \frac{1}{2}$

The expected number of heads in tosses of 400 $= 200$

But the observed number of heads $= 216$

$$\text{S.E.} = \sqrt{npq}$$

$$= \sqrt{400 \times \tfrac{1}{2} \times \tfrac{1}{2}} = \sqrt{100} = 10$$

Deviation from actual $= 216 - 200 = 16$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{16}{10} = 1.6$$

Since the observed deviation is 1.6 times the S.E., which is less than 1.96 S.E. (5% level), it can be concluded that the hypothesis is accepted. Therefore, the coin is an unbiased one.

(B) Test for Proportion of Succeses

Instead of taking the number of success in each sample, a proportion of success *i.e.*, $\frac{1}{n}$ is recorded. Formula :

$$\text{S.E.} = \sqrt{\frac{pq}{n}}$$

Illustration : 6

A random sample of 500 pineapples were taken from a large consignment and 65 were found to be bad. Show that the standard error of the population of bad ones in a sample of the size is 0.015, and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.

Solution :

Here $p = \frac{65}{500} = 0.13$,

$q = 1 - 0.13 = 0.87$

$$\text{S.E.} = \sqrt{\frac{pq}{n}} = \frac{0.13 \times 0.87}{500}$$

$$= \sqrt{\frac{0.1131}{500}}$$

$$= \sqrt{.000226} = 0.015$$

The limits of percentage of bad pineapples in the consignment are :

$$(0.13 \pm 3\text{S.E.}) \times 100 = (0.13 \pm 3 \times 0.015)100 = (.13 \pm .045)100$$

$$= (13 \pm 4.5) = 17.5 \text{ and } 8.5$$

Note : (*i*) 3 S.E. limits are "almost certain".

Illustration : 7

A wholesaler in apples claims that only 4% of the apples supplied by him are defective. A random sample of 600 apples contained 36 defective apples. Test the claim of the wholesaler. (*B.A. Ecos. Delhi*)

Solution :

$$\text{S.E.} = \sqrt{\frac{pq}{n}}$$

$$= \sqrt{\frac{0.96 \times 0.04}{600}} = 0.008$$

95% confidence limit $= P \pm 1.96$ S.E

$= P \pm 1.96 \times 0.008$

$= .96 \pm 0.01568$

$= 0.94432$ to $= .97568$

Out of 600 apples, good apples may be between $.94432 \times 600 = 566.59$ to $.97568 \times 600 = 585.4$ or 567 to 585.

Therefore the number of defective is expected between 15 to 33 apples. His claim is that 4% of apples are defective. But actual number is 36 defectives. Hence his claim cannot be accepted.

Illustration : 8

A cultivator of bananas claims that only 3 out of 100 supplied by him are defective. A random sample of 700 bananas contained 45 defective bananas. Test whether the claim of cultivator is correct.

Solution :

The cultivator claims only 3% of bananas are defective. Hence the 95% confidence limit given by $X \pm 1.96$ S.E.

$$p = \frac{3}{100} \text{(Defective bananas)} (0.03)$$

$$q = 1 - \frac{3}{100} = \frac{97}{100} (0.97)$$

$n = 700$

$$\text{S.E.} = \sqrt{\frac{pq}{n}}$$

$$= \sqrt{\frac{0.03 \times 0.97}{700}}$$

$= 0.006$

95% confidence limit $= X \pm 1.96$ S.E.

$= 0.97 \pm 1.96 \times 0.006$

$= 0.97 \pm 0.01176$

$= .95824$ to 0.98176

Out of 700 bananas, good bananas may lie between 0.98176×700 or 687.23 and 0.95824×700 or 670.76 or 671. That is good bananas lie between 671 and 687. Thus the number of defectives expected to lie between 13 and 29. Since the actual defective is 45. Therefore, the cultivator's claim that only 3% defective cannot be accepted.

Illustration : 9

A sample size of 600 persons selected at random from a large city shows that the percentage of males in the sample is 53. It is believed that the ratio of males to total population in the city is $\frac{1}{2}$. Test whether this belief is confirmed by the observation.

(*M. Com. Calcutta*)

Solution :

Let the null hypothesis be that the number of males to total population is $\frac{1}{2}$ or 0.5.

The observed value = 0.53

$$\text{S.E.} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{600}} = \sqrt{\frac{\frac{1}{4}}{600}} = \sqrt{\frac{1}{2400}} = 0.02$$

$$\text{S.E.} = \frac{0.53 - 0.05}{\text{S.E.}} = \frac{(0.53 - 0.5)}{0.02} = 1.5$$

Since Z is less than 1.96, the difference is not significant at 5% level of confidence and could have arisen because of sampling fluctuations. Therefore the null hypothesis cannot be rejected. The belief is confirmed.

(C) Test for Difference in Proportions

We draw two samples from different populations and verify whether the proportion of success is significant or not.

Formula :

$$S.E.\ (P_1-P_2) = \sqrt{pq\frac{1}{n_1}+\frac{1}{n_2}}$$

where p = the pooled estimate of the actual proportion in the population. The value of p is obtained by :

$$p = \frac{n_1P_1+n_2P_2}{n_1+n_2}$$

$$q = 1-p$$

If S.E. $(P_1-P_2)<1.96$

The difference is regarded as due to random sampling fluctuations.

Illustration : 10

One thousand articles from a factory are examined and found to be 3% defective. Fifteen hundred similar articles from a second factory are found to be only 2% defective. Can it reasonably be concluded that the product of the first factory is inferior to the second ?

(M. Com.)

Solution :

Let us set up the null hypothesis $H_o : P_1 = P_2$

$$P_1 = \frac{30}{1000} = 0.03$$

$$P_2 = \frac{30}{1500} = 0.02$$

$$SE\ (P_1-P_2) = \sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$P = \frac{(1000\times0.03)+(1500\times0.02)}{1000+1500} = \frac{(30+30)}{2500} = 0.024$$

$$= \sqrt{0.024\times0.976\left(\frac{1}{1000}+\frac{1}{1500}\right)}$$

$$= 0.006$$

$$Z = \frac{0.03-0.02}{0.006} = 1.67$$

At 95% level of confidence, z = 1.96, the difference is not significant. The null hypothesis that is $P_1=P_2$ is accepted.

Illustration : 11

A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved ? *(M. Com. Nagpur)*

Solution :

$$P_1 = \frac{16}{500} = 0.032 \text{ (in the first sample)}$$

$$P_2 = \frac{3}{100} = 0.03 \text{ (in the second sample)}$$

Let us assume that the machine has not improved after overhauling.

$$S.E.\ (P_1-P_2) = \sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$P = \frac{n_1P_1+n_2P_2}{n_1+n_2}$$

$$P = \frac{500\times0.032+100\times0.3}{500+100} = \frac{16+3}{600} = 0.03$$

$$q = 1-0.03 = 0.97$$

$$S.E.\ (P_1-P_2) = \sqrt{Pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= \sqrt{(0.03)(0.97)\left(\frac{1}{500}\right)+\left(\frac{1}{100}\right)}$$

$$= \sqrt{(0.03)(0.97)[0.002+0.01]}$$

$$= 0.0187$$

$$Z = \frac{0.032-0.03}{0.0187} = \frac{0.002}{0.0187} = 0.106$$

Since the difference is less than 2.58 S.E (1% level), the result of the experiment supports the hypothesis. Therefore, we conclude that the machine has not improved after overhauling.

Illustration : 12

In a random sample of 1000 persons from town A, 400 are found to be consumers of wheat. In a sample of 800 from town B, 400 are found to be consumers of wheat. Do these data reveal a significant difference between town A and town B so far as the proportion of wheat consumers is concerned ?

(M.A. Punjab, Madurai)

Solution :

Let us assume the hypothesis that the two towns do not differ so far as proportion of wheat consumption. Ho : $P_1 = P_2$

$$P_1 = \frac{400}{1000} = 0.4 \; P_2 = \frac{400}{800} = 0.5$$

$$P = \frac{(1000 \times 0.4) + (800 \times 0.5)}{1000 + 800} = \frac{4}{9} \therefore \; q = \frac{5}{9}$$

$$\text{S.E.}\,(P_1 - P_2) = \sqrt{\frac{4}{9} \times \frac{5}{9}\left(\frac{1}{1000} + \frac{1}{800}\right)}$$

$$= \sqrt{\frac{20}{81} \times \frac{9}{4000}} = 0.024$$

$$P_1 - P_2 = 0.4 - 0.5 = 0.1$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{0.1}{0.024} = 4.17$$

Since the difference is more than 2.58 SE (1% level) it could not have arisen due to fluctuations of sampling . Hence, the data reveal a significant difference between town *A* and town *B* so far as the proportion of wheat consumers is concerned.

Illustration : 13

In a village 'A' out of random sample of 1000 persons 100 were found to be vegetarians while in another village 'B' out of 1500 persons 180 were found to be vegetarians. Do you find a significant difference in the food habits of the people of the two villages ?

(B.Com., Madurai)

Solution :

Let us take the hypothesis that there is no significant difference in the food habits of the people of two villages

$$\text{S.E.}\,(P_1 - P_2) = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

P_1 = Village A : percentage of vegetarians $= \frac{100 \times 100}{1000} = 10$

P_2 = Village B : percentage of Vegetarians $= \frac{180}{1500} \times 100 = 12$

$$P = \frac{x_1}{n_1} + \frac{x_2}{n_2} = \frac{100}{1000} + \frac{180}{1500} = \frac{280}{2500} = 0.112 \text{ or } 11.2\%$$

$q = 100 - 11.2 = 88.8\%$

$$\text{S.E.}\,(P_1 - P_2) = \sqrt{11.2 \times 88.8\left(\frac{1}{1000} + \frac{1}{1500}\right)}$$

$$= \sqrt{994.56 \times \frac{5}{3000}}$$

$$= 1.288$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{12 - 10}{1.288} = 1.55$$

Since the calculated value (1.55) is less than 1.96 at 5% level of significance the hypothesis is accepted. Hence, there is no significant difference in the food habits of people of the two villages 'A' and 'B'.

Illustration : 14

In a simple random sample of 600 men taken from a big city, 400 are found to be smokers. In another simple random sample 900 men taken from another city 450 are smokers. Do the data indicate that there is a significant difference in the habit of smoking in the two cities.

(M.Com., Rajasthan)

Solution :

Let us assume that there is no significant difference in the habit of smoking in the two cities.

$$P_1 = \frac{400}{600} = 0.667; P_2 = \frac{450}{900} = 0.5$$

$$P = \frac{X_1 + X_2}{n_1 + n_2} = \frac{400 + 450}{600 + 900} = \frac{17}{30}$$

$$q = 1 - p = 1 - \frac{17}{30} = \frac{13}{30}$$

$$\text{S.E.}\,(P_1 - P_2) = \sqrt{\frac{17}{30} \times \frac{13}{30}\left(\frac{1}{600} + \frac{1}{900}\right)} = 0.026$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{0.667 - 0.5}{0.026} = 6.42$$

Since the difference is more than 2.58 S.E. at 1% level of significance, the hypothesis is rejected. Hence there is a significant difference in the habit of smoking of the two cities.

Tests of Significance for Large Samples

It is very difficult to distinguish between large and small samples. If the sample size is greater than 30 *i.e.*, if $n > 30$, then those samples may be regarded as large samples. There is difference between large and small samples in using the test of significance, because the

assumption we make for the two samples are also not the same. The assumptions made for large samples are :

1. The random sampling distribution of statistics is approximately normal.
2. Sampling values are sufficiently close to the population value and can be used for the calculation of standard error of estimate.

In the case of large samples, when we are testing the significance of statistic, the concept of standard error is used. The following are the formula for finding out the standard error for different statistics.

1. The standard error of mean

It measures only sampling errors. Sampling errors are involved in estimating a population parameter from a sample, instead of including all the essential information in the population.

(i) When standard deviation of the population is known, the formula is

$$S.E.\ \bar{X} = \frac{\sigma p}{\sqrt{n}}$$

S.E. $\bar{X}$ = The standard error of the mean

σp = Standard deviation of the population

n = Number of observations in the sample

(ii) When standard deviation of population is not known, we have to use the standard deviation of the sample in calculating standard error of mean. The formula is

$$S.E.\ \bar{X} = \frac{\sigma(\text{Sample})}{\sqrt{n}}$$

σ = standard deviation of the sample

If standard deviation of sample and population are available, then for the calculation of standard error of mean, we must use standard deviation of the population.

Illustration : 15

A sample of 1000 students from Bombay University was taken and their average weight was found to be 112 lbs with a standard deviation of 20 lbs. Could the mean weight of students in the population be 120 pounds ?

(M.A. Ecos.)

Solution :

Let us take the hypothesis that there is no significant difference between the sample mean and the hypothetical population mean.

$$S.E.\ \bar{X} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{1000}} = \frac{20}{31.623} = 0.632$$

$$\frac{\text{Difference}}{S.E.\bar{X}} = \frac{120-112}{0.632} = \frac{8}{0.632} = 12.66$$

Since the difference is more than 2.58 S.E. (1% level) it could not have arisen due to fluctuations of sampling. Hence the mean weight of students in the population could not be 120 lbs.

Illustration : 16

A company manufacturing electric light bulbs claims that the average life of its bulbs is 1600 hours. The average life and standard deviation of a random sample of 100 such bulbs were 1570 hours and 120 hours respectively. Should we accept the claim of the company ?

(B.A. Hons. Delhi)

Solution :

$$S.E.\ \bar{X} = \frac{\sigma}{\sqrt{n}} = \frac{120}{\sqrt{100}} = 12$$

$$\frac{\text{Difference}}{S.E.\bar{X}} = \frac{1600-1570}{12} = \frac{30}{12} = 2.5$$

2.5 > 1.96 S.E. at 5% level of significance, the hypothesis cannot be accepted. We cannot accept the claim of the company.

Illustration 17

Calculate standard error of mean from the following data, showing the amount paid by 100 firms in Calcutta on the occasion of Durga Puja.

Mid value (Rs.)	39	49	59	69	79	89	99
No. of firms	2	3	11	20	32	25	7

(M.B.A. Rohtak,)

Solution :

$$S.E.\bar{X} = \frac{\sigma}{\sqrt{n}}.$$

Computation of Standard Deviation

Mid value m	f	$\frac{m-69}{10} = d'$	fd'	fd'^2
39	2	-3	-6	18
49	3	-2	-6	12
59	11	-1	-11	11
69	20	0	0	0
79	32	+1	+32	32
89	25	+2	+50	100
99	7	+3	+21	63
	N = 100		$\sum fd' = 80$	$\sum fd'^2 = 236$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times C$$

$$= \sqrt{\frac{236}{100} - \left(\frac{80}{100}\right)^2} \times 10$$

$$= \sqrt{2.36 - 0.64} \times 10 = \sqrt{1.72} \times 10$$

$$= 1.311 \times 10 = 13.11$$

$$\text{S.E.}_{\bar{X}} = \frac{13.11}{\sqrt{100}} = \frac{13.11}{10} = 1.311$$

Testing the difference between means of Two Samples

(A) If two independent random samples with n_1 and n_2 respectively are drawn from the same population of standard deviation 61, the standard error of the difference between the same means is given :

S.E. of the difference between sample means =

$$= \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

(B) If two random samples with $\bar{X}_1, \sigma_1, n_1$ and $\bar{X}_2, \sigma_2 n_2$ respectively are drawn from different populations, then the S.E. of the difference between the mean is given :

$$= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If σ_1 and σ_2 are unknown, S.E. of the difference between mean

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

where S_1 and S_2 represent standard deviations of the two samples.

Illustration : 18

A test on two groups of boys and girls gave the following results :

	Mean	S.D.	N
Boys	45	3	150
Girls	75	5	100

Do you conclude that the marks scored by girls are more than boys at 0.05 level of significance.

(B.Com., Madurai)

Solution :

Let us take the hypothesis that there is no significant difference in the marks scored between boys and girls.

$$\text{S.E.}\,(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} =$$

$\sigma_1 = 3,\ \sigma_2 = 5,\ n_1 = 150,\ n_2 = 100$

Substituting the Values :

$$\text{S.E.}\,(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{3^2}{150} + \frac{5^2}{100}} = \sqrt{\frac{9}{150} + \frac{25}{100}}$$

$$= \sqrt{0.31} = 0.56$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{45 - 75}{0.56} = 53.57$$

Since the calculated value *i.e.*, 53.57 is greater than 1.96 at 5% level of significance, we reject the hypothesis. We conclude that the girl students have scored more than the boys.

Illustration : 19

An examination was given to two classes consisting of 40 and 50 students respectively. In the first class the mean mark was 74 with a standard deviation of 8, while in the second class the mean mark was 78 with a standard deviation of 7. Is there a significant difference between the performance of the two classes at a level of significance of 0.05 ?

(BA (Hons.,) Delhi)

Solution :

Let us take the hypothesis that there is no significant difference in the mean marks of two classes.

$$\text{S.E.}\,(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$\sigma_1 = 8; n_1 = 40; \sigma_2 = 7; n_2 = 50$

$$\therefore \text{SE} = \sqrt{\frac{8^2}{40} + \frac{(7)^2}{50}}$$

$$= \sqrt{1.6 + 0.98} = 1.606$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{78 - 74}{1.606} = 2.49$$

Since the difference is more than 1.96 SE (5% level of significance), the hypothesis is rejected. Hence there is a significant difference in the performance of the two classes.

Illustration : 20

A college conducted both day and evening classes intended to be identical. A sample of 100 day students yields examination results as under :

$$\bar{X}_1 = 72.4 \quad \sigma_1 = 14.8$$

A sample of 200 evening students yields examination results as under :

$$\bar{X}_2 = 73.9 \quad \sigma_2 = 17.9$$

Are the two means statistically equal at 1% level ?

(M.Com., Sukhadia)

Solution :

Let us take hypothesis that the two means are statistically equal.

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

$$= \sqrt{\frac{(14.8)^2}{100} + \frac{(17.9)^2}{200}}$$

$$= \sqrt{2.1904 + 1.602}$$

$$= 1.947$$

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{72.4 - 73.9}{1.947} = 0.77$$

Since the difference is less than 2.58 SEC 1% level of significance, the hypothesis holds true. Hence the two means may be regarded statistically equal.

Standard Error of the Difference between two Standard Deviations

In case of two large random samples, each drawn from a normally distributed population, the S.E. of the difference between the standard deviations is given by :

$$\text{S.E.}(\sigma_1 - \sigma_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

Illustration : 21

Productivity test of two food articles – paddy and wheat gives the following results :

	Mean yield (tonnes)	S.D.	No. of hectares
Paddy	80	10	120
Wheat	75	12	90

Is the difference between standard deviation is significant ?

Solution :

Let us take the hypothesis that there is no significant difference in the Standard Deviation of productivity of paddy and wheat.

$$\text{S.E.}(\sigma_1 - \sigma_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

$$\sigma_1 = 10;\ \sigma_2 = 12 \quad n_1 = 120 \quad n_2 = 90$$

$$= \sqrt{\frac{10^2}{2 \times 120} + \frac{12^2}{2 \times 90}}$$

$$= \sqrt{\frac{100}{240} + \frac{144}{180}}$$

$$= \sqrt{1.22}$$

$$= 1.10$$

$$Z = \frac{\text{Difference}}{\text{SE}} = \frac{10 - 12}{1.10} = 1.82$$

Since the difference is less than 1.96 at 5% level of significance, the given data support the hypothesis. Thus we conclude that there is no significant difference in standard deviation of productivity between paddy and wheat.

Test of Significance for Small Samples

If the sample size is less than 30 *i.e.*, $n < 30$, then those samples may be regarded as small samples. As a rule, the methods and the theory of small samples are applicable to large samples ; but the methods and the theory of large samples are not applicable to small samples. The small samples are used in testing a given hypothesis, to find out the observed values, which could have arisen by sampling fluctuations from some values given in advance. For example, if a sample of 12 gives a correlation coefficient of +0.5, we can test whether the value is significant of correlation in the parent population.

In a small sample, the investigator's estimate will vary widely from sample to sample. An inference drawn from a smaller sample result is less precise than the inference drawn from a large sample result.

Students' t – Distribution

The greatest contribution to the theory of small samples was made by Sir William Gossett and R.A. Fisher. Gossett published his discovery in 1905 under the pen name 'students' and it is popularly known as t-test or students' t-distribution or students' distribution.

When the sample size is 30 or less and the population standard deviation is unknown, we can use the t-distribution.