MODULE 5

Statistical Techniques–III

5.1 POPULATION OR UNIVERSE

An aggregate of objects (animate or inanimate) under study is called **population or universe.** It is thus a collection of individuals or of their attributes (qualities) or of results of operations which can be numerically specified.

A universe containing a finite number of individuals or members is called a **finite inverse.** For example, the universe of the weights of students in a particular class.

A universe with infinite number of members is known as an **infinite universe.** For example, the universe of pressures at various points in the atmosphere.

In some cases, we may be even ignorant whether or not a particular universe is infinite, *e.g.*, the universe of stars.

The universe of concrete objects is an **existent universe.** The collection of all possible ways in which a specified event can happen is called a **hypothetical universe.** The universe of heads and tails obtained by tossing a coin an infinite number of times (provided that it does not wear out) is a hypothetical one.

5.2 SAMPLING

The statistician is often confronted with the problem of discussing universe of which he cannot examine every member *i.e.*, of which complete enumeration is impracticable. For example, if we want to have an idea of the average per capita income of the people of India, enumeration of every earning individual in the country is a very difficult task. Naturally, the question arises : What can be said about a universe of which we can examine only a limited number of members ? This question is the origin of the Theory of Sampling.

A finite subset of a universe is called a **sample**. A sample is thus a small portion of the universe. The number of individuals in a sample is called the **sample size**. The process of selecting a sample from a universe is called **sampling**.

The theory of sampling is a study of relationship existing between a population and samples drawn from the population. The fundamental objective of sampling is to get as much information as possible of the whole universe by examining only a part of it. An attempt is thus made through sampling to give the maximum information about the parent universe with the minimum effort.

Sampling is quite often used in our day-to-day practical life. For example, in a shop we assess the quality of sugar, rice or any other commodity by taking only a handful of it from the bag and then decide whether to purchase it or not. A housewife normally tests the cooked

products to find if they are properly cooked and contain the proper quantity of salt or sugar, by taking a spoonful of it.

5.3 SAMPLING METHODOLOGIES

Sampling methodologies are classified under two general categories:

1. Probability sampling and 2. Non-probability sampling

In the former, the researcher knows the exact possibility of selecting each member of the population while in the latter, the chance of being included in the sample is not known. A probability sample tends to be more difficult and costly to conduct. However, probability samples are the only type of samples where the results can be generalized from the sample to the population. In addition, probability samples allow the researcher to calculate the precision of the estimates obtained from the sample and to specify the sampling error.

Non-probability samples, incontrast, donot allow the study's findings to be generalized from the sample to the population. When discussing the results of a non-probability sample, the researchers must limit his/her findings to the persons or elements sampled.

This procedure also does not allow the researcher to calculate sampling statistics that provide information about the precision of the results. The advantage of non-probability sampling is the case in which it can be administered.

Non-probability samples tend to be less complicated and less time consuming than probability samples. If the researcher has no intention of generalizing beyond the sample, one of the non-probability sampling methodologies will provide the desired information.

5.4 NON-PROBABILITY SAMPLING

The three common types of non-probability samples are:

(*i*) **Convenience Sampling:** As the name implies, convenience sampling involves choosing respondents at the convenience of the researcher. Examples of convenience sampling include people-in-the street interviews—the sampling of people to which the researcher has easy access, such as a class of students and studies that use people who have volunteered to be questioned as a result of an advertisement or another type of promotion. A drawback to this methodology is the lack of sampling accuracy. Because the probability of inclusion in the sample is unknown for each respondent, none of the reliability or sampling precision statistics can be calculated. Convenience samples, however, are employed by researchers because the time and cost of collecting information can be reduced.

(*ii*) **Quota Sampling:** Quota sampling is often confused with stratified and cluster sampling—two probability sampling methodologies. All of these methodologies sample a population that has been subdivided into classes or categories.

The primary differences between the methodologies is that with stratified and cluster sampling, the classes are mutually exclusive and are isolated prior to sampling. Thus, the probability of being selected is known and members of the population selected to be sampled are not arbitrarily disqualified from being included in the results. In quota sampling, the classes cannot be isolated prior to sampling and respondents are categorized into the classes as the survey proceeds. As each class fills or reaches its quota, additional respondents that would have fallen into these classes are rejected or excluded from the results.

An example of a quota sample would be a survey in which the researcher desires to obtain a certain number of respondents from various income categories. Generally, researchers donot know the income of the persons they are sampling until they ask about income.

population. If this assumption is not valid, then systematic sampling will be less precise than simple random sampling. In conducting systematic sampling, it is also essential that the researcher does not introduce bias into the sample by selecting an inappropriate sampling interval. For instance, when conducting a sample of financial records, or other items that follow a calendar schedule, the researcher would not want to select "7" as the sampling interval because the sample would then be comprised of observations that were all on the same day of the week. Day-of-the-week influences may cause contamination of the sample, giving the researcher biased results.

(*v*) **Multi-Stage Sampling:** Multi-stage sampling is like cluster sampling, but involves selecting a sample within each chosen cluster, rather than including all units in the cluster. Thus, multi-stage sampling involves selecting a sample in at least two stages. In the first stage, large groups or clusters are selected. These clusters are designed to contain more population units than are required for the final sample.

In the second stage, population units are chosen from selected clusters to derive a final sample. If more than two stages are used, the process of choosing population units within clusters continues until the final sample is achieved.

An example of multi-stage sampling is where, firstly, electoral sub-divisions (clusters) are sampled from a city or state. Secondly, blocks of houses are selected from within the electoral sub-divisions and, thirdly, individual houses are selected from within the selected blocks of houses.

The advantages of multi-stage sampling are convenience, economy and efficiency. Multi-stage sampling does not require a complete list of members in the target population, which greatly reduces sample preparation cost. The list of members is required only for those clusters used in the final stage. The main disadvantage of multi-stage sampling is the same as for cluster sampling: lower accuracy due to higher sampling error.

5.6 PARAMETERS OF STATISTICS

The statistical constants of the population such as mean, the variance etc. are known as the parameters. The statistical concepts of the sample from the members of the sample to estimate the parameters of the population from which the sample has been drawn is known as *statistic*.

Population mean and variance are denoted by μ and σ^2, while those of the samples are given by $\bar{x}$, s^2.

5.7 STANDARD ERROR

The standard deviation of the sampling distribution of a statistic is known as the **standard error (S.E.)**. It plays an important role in the theory of large samples and it forms a basis of the testing of hypothesis. If t is any statistic, for large sample. $z = \dfrac{t - \mathrm{E}(t)}{\mathrm{S.E.}(t)}$ is normally distributed with mean 0 and variance unity.

For large sample, the standard errors of some of the well known statistic are listed below:

n—sample size; σ^2—population variance; s^2—sample variance; p—population proportion; $Q = 1 - p$; n_1, n_2—are sizes of two independent random samples.

S. No.	Statistic	Standard error
1.	$\bar{x}$	$\sigma/\sqrt{n}$
2.	s	$\sqrt{\sigma^2/2n}$
3.	Difference of two sample means $\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
4.	Difference of two sample standard deviation $s_1 - s_2$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
5.	Difference of two sample proportions $p_1 - p_2$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$
6.	Observed sample proportion p	$\sqrt{PQ/n}$

5.8 TEST OF SIGNIFICANCE

An important aspect of the sampling theory is to study the test of significance which will enable us to decide, on the basis of the results of the sample, whether

(*i*) *the deviation between the observed sample statistic and the hypothetical parameter value or*

(*ii*) *the deviation between two sample statistics is significant or might be attributed due to chance or the fluctuations of the sampling.*

5.9 TESTING OF STATISTICAL HYPOTHESIS

Step 1. Null hypothesis

For applying the tests of significance, we first set up a hypothesis which is a definite statement about the population parameter called Null Hypothesis. It is denoted by H_0.

Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true. First, we set up H_0 in clear terms.

Step 2. Alternative hypothesis

Any hypothesis which is complementary to the null hypothesis (H_0) is called an alternative hypothesis. It is denoted by H_1.

For example, if we want to test the null hypothesis that the population has a specified mean μ_0 then we have

$$H_0 : \mu = \mu_0$$

then the alternative hypothesis will be

(*i*) $H_1 : \mu \neq \mu_0$ (Two tailed alternative hypothesis)

(*ii*) $H_1 : \mu > \mu_0$ (right tailed alternative hypothesis (or) single tailed)

(*iii*) $H_1 : \mu < \mu_0$ (left tailed alternative hypothesis (or) single tailed)

Hence alternative hypothesis helps to know whether the test is two tailed test or one tailed test. Therefore, we set up H_1 for this decision.

Step 3. Level of significance

The probability of the value of the variate falling in the critical region is known as level of significance. A region corresponding to a statistic t in the sample space S which amounts to rejection of the null hypothesis H_0 is called as **critical region** or region of rejection while which amounts to acceptance of H_0 is called **acceptance region.** The probability α that a random value of the statistic t belongs to the critical region is known as the **level of significance.**

$$P(t \in w/H_0) = \alpha$$

i.e., the level of significance is the size of the type I error (refer art. 5.7) or the maximum producer's risk.

We select the appropriate level of significance in advance depending on the reliability of the estimates.

Step 4. Test statistic (or test criterion): We compute the test statistic z under the null hypothesis. For larger samples corresponding to the statistic t, the variable $z = \dfrac{t - E(t)}{S.E.(t)}$ is normally distributed with mean 0 and variance 1. The value of z given above under the null hypothesis is known as **test statistic.**

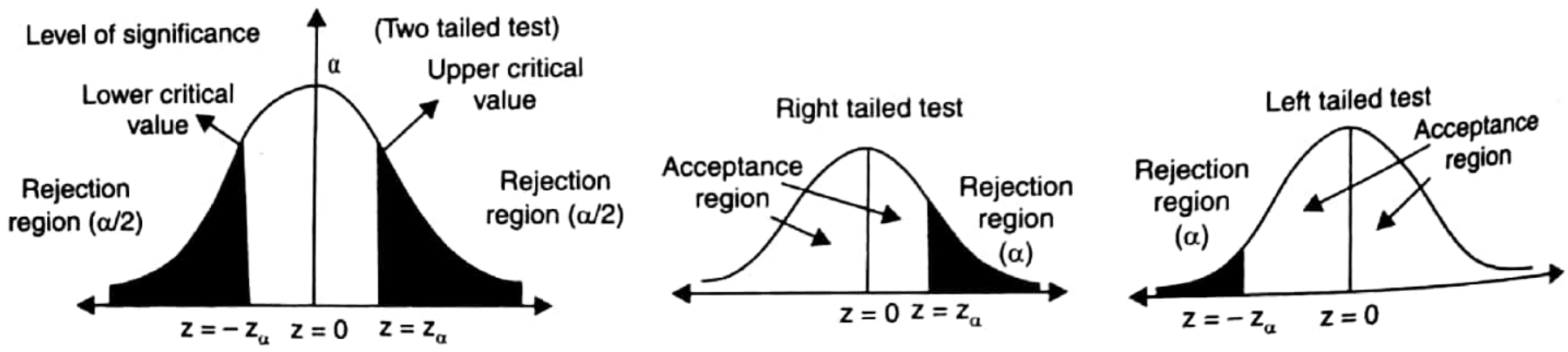
Step 5. Conclusion: We compare the computed value of z with the critical value z_α at level of significance (α). The critical value of z_α of the test statistic at level of significance α for a two tailed test is given by

$$p(|z| > z_\alpha) = \alpha \qquad \ldots(1)$$

i.e., z_α is the value of z so that the total area of the critical region on both tails is α. Since the normal curve is symmetrical, from equation (1), we get

$$p(z > z_\alpha) + p(z < -z_\alpha) = \alpha;\ i.e.\ 2p(z > z_\alpha) = \alpha;\ i.e.,\ p(z > z_\alpha) = \alpha/2$$

i.e., the area of each tail is $\alpha/2$.



The critical value z_α is that value such that the area to the right of z_α is $\alpha/2$ and the area to the left of $-z_\alpha$ is $\alpha/2$.

In the case of one tailed test,

$p(z > z_\alpha) = \alpha$ if it is right tailed; $p(z < -z_\alpha) = \alpha$ if it is left tailed.

The critical value of z for a single tailed test (right or left) at level of significance α is same as the critical value of z for two tailed test at level of significance 2α.

Using the equation, also using the normal tables, the critical value of z at different levels of significance (α) for both single tailed and two tailed test are calculated and listed below. The equations are

$$p(|z| > z_\alpha) = \alpha;\ p(z > z_\alpha) = \alpha;\ p(z < -z_\alpha) = \alpha$$

	Level of significance		
	1% (0.01)	5% (0.05)	10% (0.1)
Two tailed test	$\mid z_\alpha \mid = 2.58$	$\mid z_\alpha \mid = 1.96$	$\mid z_\alpha \mid = 1.645$
Right tailed test	$z_\alpha = 2.33$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
Left tailed test	$z_\alpha = -2.33$	$z_\alpha = -1.645$	$z_\alpha = -1.28$

If $|z| > z_\alpha$, we reject H_0 and conclude that there is significant difference. If $|z| < z_\alpha$, we accept H_0 and conclude that there is no significant difference.

5.10 ERRORS IN SAMPLING

The main aim of the sampling theory is to draw a valid conclusion about the population parameters on the basis of the sample results. In doing this we may commit the following two types of errors:

Type I Error

When H_0 is true, we may reject it. P(Reject H_0 when it is true) = P(Reject H_0/H_0) = α. α is called the size of the type I error, also referred to as **producer's risk.**

Type II Error: When H_0 is wrong, we may accept it. P(Accept H_0 when it is wrong) = P(Accept H_0/H_1) = β. β is called the size of the type II error, also referred to as **consumer's risk.**

Note. The values of the test statistic which separates the critical region and acceptance region are called the **critical values** or **significant values.** This value is dependent on (*i*) the level of significance used and (*ii*) the alternative hypothesis, whether it is one tailed or two tailed.

5.11 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

If the sample size $n > 30$, the sample is taken as large sample. For such sample we apply distributions, as Binomial, Poisson, which are closely approximated by normal distributions assuming the population as normal.

Under large sample test, the following are the important tests to test the significance:

1. *Testing of significance for single proportion.*
2. *Testing of significance for difference of proportions.*
3. *Testing of significance for single mean.*
4. *Testing of significance for difference of means.*
5. *Testing of significance for difference of standard deviations.*

5.11.1 Testing of Significance for Single Proportion

This test is used to find the significant difference between proportion of the sample and the population. Let X be the number of successes in n independent trials with constant probability P of success for each trial.

$E(X) = nP$; $V(X) = nPQ$; $Q = 1 - P$ = Probability of failure.

Let $p = X/n$ called the observed proportion of success.

$$E(p) = E(X/n) = \frac{1}{n}E(X) = \frac{nP}{n} = P$$

$$V(p) = V(X/n) = \frac{1}{n^2}V(X) = \frac{nPQ}{n^2} = PQ/n$$

$$\text{S.E.}(p) = \sqrt{\frac{PQ}{n}} \; ; z = \frac{p - E(p)}{\text{S.E.}(p)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$$

This z is called test statistic which is used to test the significant difference of sample and population proportion.

Note. 1. The probable limits for the observed proportion of successes are $P \pm 3\sqrt{PQ/n}$.

2. If p is not known, the probable limits for the proportion in the population are $p \pm z_\alpha \sqrt{pq/n}$, $q = 1 - p$, where sample proportion, p is taken as an estimate of p and z_α is the significant value of z at level of significance α.

ILLUSTRATIVE EXAMPLES

Example 1. *A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased.*

Sol. Null hypothesis:

H_0: The coin is unbiased *i.e.*, $P = 0.5$

Alternative hypothesis:

H_1: The coin is biased *i.e.*, $P \neq 0.5$

Hence we use **two tailed test.**

Here, $n = 400$, X = no. of success = 216

$$\therefore \quad p = \text{proportion of success in the sample} = \frac{X}{n} = \frac{216}{400} = 0.54$$

P = population proportion = 0.5, $\quad Q = 1 - P = 0.5$

Test Statistic:

Under H_0, test statistic $z = \dfrac{p - P}{\sqrt{PQ / n}}$

$$|z| = \left| \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{400}}} \right| = 1.6$$

Conclusion:

Since $|z| = 1.6 < 1.96$ *i.e.*, $|z| < z_\alpha$ where z_α is the significant value of z at 5% level of significance.

Hence we accept H_0 and conclude that the coin is unbiased.

Example 2. *A machine is producing bolts of which a certain fraction is defective. A random sample of 400 is taken from a large batch and is found to contain 30 defective bolts. Does this indicate that the proportion of defectives is larger than that claimed by the manufacturer where the manufacturer claims that only 5% of his product are defective. Find 95% confidence limits of the proportion of defective bolts in batch.*

Sol. Null hypothesis H_0: The manufacturer claim is accepted *i.e.*, $P = \frac{5}{100} = 0.05$

$$Q = 1 - P = 1 - 0.05 = 0.95$$

Alternative hypothesis: $P > 0.05$

Hence we use **Right tailed test.**

p = observed proportion of sample $= \frac{30}{400} = 0.075$

Test statistic

Under H_0, the test statistic $\quad z = \frac{p - P}{\sqrt{PQ/n}} \quad \therefore \quad z = \frac{0.075 - 0.05}{\sqrt{\frac{0.05 \times 0.95}{400}}} = 2.2941.$

Conclusion: The tabulated value of z at 5% level of significance for right tailed test is $z_\alpha = 1.645$. Since $|z| = 2.2941 > 1.645$, H_0 is rejected at 5% level of significance. *i.e.*, the proportion of defective is larger than the manufacturer claim.

To find 95% confidence limits of the proportion.

It is given by $P \pm z_\alpha \sqrt{PQ/n}$

i.e., $$0.05 \pm 1.96 \sqrt{\frac{0.05 \times 0.95}{400}} = 0.05 \pm 0.02135 = 0.07136, 0.02865$$

Hence 95% confidence limits for the proportion of defective bolts are (0.07136, 0.02865).

TEST YOUR KNOWLEDGE

1. In a hospital 475 female and 525 male babies were born in a week. Do these figures confirm the hypothesis that males and females are born in equal number?
2. In a city a sample of 1000 people were taken and out of them 540 are vegetarian and the rest are non-vegetarian. Can we say that the both habits of eating (vegetarian or non-vegetarian) are equally popular in the city at (*i*) 1% level of significance (*ii*) 5% level of significance?
3. 325 men out of 600 men chosen from a big city were found to be smokers. Does this information support the conclusion that the majority of men in the city are smokers?

Answers

1. H_0 accepted at 5% level
2. H_0 rejected at 5% level, accepted at 1% level
3. H_0 rejected at 5% level.

5.11.2 Testing of Significance for Difference of Proportions

Consider two samples X_1 and X_2 of sizes n_1 and n_2 respectively taken from two different populations. To test the significance of the difference between the sample proportions p_1 and p_2, the test statistic under the null hypothesis H_0, that there is no significant difference between the two sample proportion, is

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{where} \quad P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad \text{and} \quad Q = 1 - P.$$

ILLUSTRATIVE EXAMPLES

Example 1. *Before an increase in excise duty on tea, 800 people out of a sample of 1000 persons were found to be tea drinkers. After an increase in the duty, 800 persons were known to be tea drinkers in a sample of 1200 people. Do you think that there has been a significant decrease in the consumption of tea after the increase in the excise duty?*

Sol. Here, $n_1 = 800, n_2 = 1200$

$$p_1 = \frac{X_1}{n_1} = \frac{800}{1000} = \frac{4}{5}; p_2 = \frac{X_2}{n_2} = \frac{800}{1200} = \frac{2}{3}$$

$$P = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{8}{11}; Q = \frac{3}{11}$$

Null hypothesis H_0: $p_1 = p_2$ *i.e.*, there is no significant difference in the consumption of tea before and after increase of excise duty.

Alternative hypothesis $H_1 : p_1 > p_2$

Hence we use **right tailed test.**

Test statistic:

The test statistic is $$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.8 - 0.6666}{\sqrt{\frac{8}{11} \times \frac{3}{11}\left(\frac{1}{1000} + \frac{1}{1200}\right)}} = 6.842.$$

Conclusion: Since the calculated value of $|z| > 1.645$ and also $|z| > 2.33$, both the significant values of z at 5% and 1% level of significance, hence H_0 is rejected *i.e.*, there is a significant decrease in the consumption of tea due to increase in excise duty.

Example 2. *A machine produced 16 defective articles in a batch of 500. After overhauling it produced 3 defectives in a batch of 100. Has the machine improved?*

Sol. $$p_1 = \frac{16}{500} = 0.032, \quad n_1 = 500$$

$$p_2 = \frac{3}{100} = 0.03, \quad n_2 = 100$$

Null hypothesis:

H_0: The machine has not improved due to overhauling, *i.e.*, $p_1 = p_2$.

Alternative hypothesis:

H_1: $p_1 > p_2$

Hence we use **right tailed test.**

$\therefore$ $$P = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{19}{600} \cong 0.032$$

Test Statistic:

Under H_0, the test statistic

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 0.104$$

Conclusion: The calculated value of $|z| < 1.645$ which is the significant value of z at 5% level of significance, H_0 is accepted *i.e.*, the machine has not improved due to overhauling.

TEST YOUR KNOWLEDGE

1. Random sample of 400 men and 600 women were asked whether they would like to have a school near their residence. 200 men and 325 women were in favour of proposal. Test the hypothesis that the proportion of men and women in favour of the proposal are same at 5% level of significance.
2. In a town A, there were 956 births of which 52.5% were males while in towns A and B combined, this proportion in total of 1406 births was 0.496. Is there any significant difference in the proportion of male births in the two towns ?

Answers

1. H_0 : Accepted **2.** H_0 : Rejected.

5.11.3 Testing of Significance for Single Mean

To test whether the difference between sample mean and population mean is significant or not.

Let $X_1, X_2, \ldots\ldots, X_n$ be a random sample of size n from a large population $X_1, X_2, \ldots\ldots, X_N$ of size N with mean μ and variance σ^2.

$\therefore$ the standard error of mean of a random sample of size n from a population with variance σ^2 is $\sigma/\sqrt{n}$.

To test whether the given sample of size n has been drawn from a population with mean μ *i.e.*, to test whether the difference between the sample mean and population mean is significant or not under the null hypothesis that there is no difference between the sample mean and population mean.

The test statistic is $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, where σ is the standard deviation of the population.

If σ is not known, we use the test statistic $z = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$, where s is the standard deviation of the sample.

Note. If the level of significance is α and z_α is the critical value $-z_\alpha < |\, z \,| = \left| \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right| < z_\alpha$.

The limits of the population mean μ are given by $\bar{x} - z_\alpha \dfrac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_\alpha \dfrac{\sigma}{\sqrt{n}}$.

At 5% level of significance, 95% confidence limits are $\bar{x} - 1.96 \dfrac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \dfrac{\sigma}{\sqrt{n}}$.

At 1% level of significance, 99% confidence limits are $\bar{x} - 2.58 \dfrac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2.58 \dfrac{\sigma}{\sqrt{n}}$.

These limits are called **confidence limits** or **fiducial limits.**

ILLUSTRATIVE EXAMPLES

Example 1. *A random sample of 900 members has a mean 3.4 cms. Can it be reasonably regarded as a sample from a large population of mean 3.2 cms and S.D. 2.3 cms?*

Sol. Here $n = 900$, $\bar{x} = 3.4$, $\mu = 3.2$, $\sigma = 2.3$

Null hypothesis:

H_0 : Assume that the sample is drawn from a large population with mean 3.2 and S.D. 2.3

Alternative hypothesis:

$H_1 : \mu \neq 3.2$ (**two tailed test**)

Test statistic:

$$\text{Under } H_0 ; z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{3.4 - 3.2}{2.3/\sqrt{900}} = 0.261.$$

Conclusion: As the calculated value of $| z | = 0.261 < 1.96$, the significant value of z at 5% level of significance, H_0 is accepted *i.e.,* the sample is drawn from the population with mean 3.2 and S.D. 2.3.

Example 2. *The mean weight obtained from a random sample of size 100 is 64 gms. The S.D. of the weight distribution of the population is 3 gms. Test the statement that the mean weight of the population is 67 gms at 5% level of significance. Also set up 99% confidence limits of the mean weight of the population.*

Sol. Here $n = 100$, $\mu = 67$, $\bar{x} = 64$, $\sigma = 3$

Null hypothesis:

H_0: There is no significant difference between sample and population mean.

i.e., $\mu = 67$, the sample is drawn from the population with $\mu = 67$.

Alternative hypothesis:

H_1: $\mu \neq 67$ **(Two tailed test).**

Test statistic:

$$\text{Under } H_0, z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{64 - 67}{3/\sqrt{100}} = -10 \quad \therefore \quad | z | = 10.$$

Conclusion: Since the calculated value of $| z | > 1.96$, the significant value of z at 5% level of significance, H_0 is rejected *i.e.,* the sample is not drawn from the population with mean 67.

To find 99% confidence limits. It is given by $\bar{x} \pm 2.58\ \sigma/\sqrt{n} = 64 \pm 2.58(3/\sqrt{100}) =$ 64.774, 63.226.

TEST YOUR KNOWLEDGE

1. A sample of 1000 students from a university was taken and their average weight was found to be 112 pounds with a S.D. of 20 pounds. Could the mean weight of students in the population be 120 pounds?
2. A sample of 400 male students is found to have a mean height of 160 cms. Can it be reasonably regarded as a sample from a large population with mean height 162.5 cms and standard deviation 4.5 cms?

3. A random sample of 200 measurements from a large population gave a mean value of 50 and a S.D. of 9. Determine 95% confidence interval for the mean of population.

Answers

1. H_0 is rejected 2. H_0 accepted 3. 48.8 and 51.2.

5.11.4 Test of Significance for Difference of Means of Two Large Samples

Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1, and variance σ_1^2. Let $\bar{x}_2$ be the mean of an independent sample of size n_2 from another population with mean μ_2 and variance σ_2^2. The test statistic is given by $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$.

Under the null hypothesis that the samples are drawn from the same population where $\sigma_1 = \sigma_2 = \sigma$ *i.e.*, $\mu_1 = \mu_2$ the test statistic is given by $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$.

Note 1. If σ_1, σ_2 are not known and $\sigma_1 \neq \sigma_2$ the test statistic in this case is $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2 + s_2^2}{n_1 + n_2}}}$.

Note 2. If σ is not known and $\sigma_1 = \sigma_2$. We use $\sigma^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$ to calculate σ.

ILLUSTRATIVE EXAMPLES

Example 1. *The average income of persons was ₹ 210 with a S.D. of ₹ 10 in sample of 100 people of a city. For another sample of 150 persons, the average income was ₹ 220 with S.D. of ₹ 12. The S.D. of incomes of the people of the city was ₹ 11. Test whether there is any significant difference between the average incomes of the localities.*

Sol. Here $n_1 = 100$, $n_2 = 150$, $\bar{x}_1 = 210$, $\bar{x}_2 = 220$, $s_1 = 10$, $s_2 = 12$.

Null hypothesis: The difference is not significant. *i.e.*, there is no difference between the incomes of the localities. H_0: $\bar{x}_1 = \bar{x}_2$

Alternative hypothesis

$H_1 : \bar{x}_1 \neq \bar{x}_2$ **(two tailed test)**

Test statistic:

Under H_0, $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{210 - 220}{\sqrt{\dfrac{10^2}{100} + \dfrac{12^2}{150}}} = -7.1428 \quad \therefore \quad |z| = 7.1428.$

Conclusion: As the calculated value of $|z| > 1.96$, the significant value of z at 5% level of significance, H_0 is rejected *i.e.*, there is significant difference between the average incomes of the localities.

Example 2. *Intelligence tests were given to two groups of boys and girls.*

	Mean	*S.D.*	*Size*
Girls	*75*	*8*	*60*
Boys	*73*	*10*	*100*

Examine if the difference between mean scores is significant.

Sol. Null hypothesis H_0: There is no significant difference between mean scores *i.e.*, $\bar{x}_1 = \bar{x}_2$.

Alternative hypothesis H_1: $\bar{x}_1 \neq \bar{x}_2$ **(two tailed test)**

Test statistic: Under the null hypothesis, $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{75 - 73}{\sqrt{\dfrac{8^2}{60} + \dfrac{10^2}{100}}} = 1.3912.$

Conclusion: As the calculated value of $|z| < 1.96$, the significant value of z at 5% level of significance, H_0 is accepted *i.e.*, there is no significant difference between mean scores.

TEST YOUR KNOWLEDGE

1. Two random samples of sizes 1000 and 2000 farms gave an average yield of 2000 kg and 2050 kg respectively. The variance of wheat farms in the country may be taken as 100 kg. Examine whether the two samples differ significantly in yield.
2. The means of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms respectively. Can the samples be regarded as drawn from the same population of standard deviation 6.25 cms?
3. In a survey of buying habits, 400 women shoppers are chosen at random in supermarket A. Their average weekly food expenditure is ₹ 250 with a S.D. of ₹ 40. For 500 women shoppers chosen at supermarket B, the average weekly food expenditure is ₹ 220 with a S.D. of ₹ 45. Test at 1% level of significance whether the average food expenditures of the two groups are equal.

Answers

1. Highly significant
2. Not significant
3. Highly significant.

5.11.5 Test of Significance for the Difference of Standard Deviations

If s_1 and s_2 are the standard deviations of two independent samples, then under the null hypothesis $H_0 : \sigma_1 = \sigma_2$, *i.e.*, the sample standard deviations don't differ significantly, the statistic

$$z = \frac{s_1 - s_2}{\sqrt{\dfrac{\sigma_1^2}{2n_1} + \dfrac{\sigma_2^2}{2n_2}}},$$ where σ_1 and σ_2 are population standard deviations.

When population standard deviations are not known, then $z = \dfrac{s_1 - s_2}{\sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}}}$.

Example. *Random samples drawn from two countries gave the following data relating to the heights of adult males:*

	Country A	*Country B*
Mean height (in inches)	*67.42*	*67.25*
Standard deviation	*2.58*	*2.50*
Number in samples	*1000*	*1200*

(*i*) *Is the difference between the means significant?*

(*ii*) *Is the difference between the standard deviations significant?*

Sol. Given: $n_1 = 1000$, $n_2 = 1200$, $\bar{x}_1 = 67.42$; $\bar{x}_2 = 67.25$, $s_1 = 2.58$, $s_2 = 2.50$

Since the samples size are large we can take $\sigma_1 = s_1 = 2.58$; $\sigma_2 = s_2 = 2.50$.

(*i*) **Null hypothesis** $H_0 : \mu_1 = \mu_2$ *i.e.*, sample means do not differ significantly.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ **(two tailed test)**

Test statistic:
$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{67.42 - 67.25}{\sqrt{\frac{(2.58)^2}{1000} + \frac{(2.50)^2}{1200}}} = 1.56.$$

Conclusion:

Since $|z| < 1.96$ we accept the null hypothesis at 5% level of significance.

(*ii*) **Null hypothesis:**

H_0: $\sigma_1 = \sigma_2$ *i.e.*, the sample S.D.'s do not differ significantly.

Alternative hypothesis H_1: $\sigma_1 \neq \sigma_2$ **(two tailed test)**

Test statistic:
$$z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} = \frac{s_1 - s_1}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = 1.0387.$$

Since $|z| < 1.96$ we accept the null hypothesis at 5% level of significance.

TEST YOUR KNOWLEDGE

1. The mean yield of two sets of plots and their variability are as given. Examine
   (*i*) whether the difference in the mean yield of the two sets of plots is significant.
   (*ii*) whether the difference in the variability in yields is significant.

	Set of 40 plots	*Set of 60 plots*
Mean yield per plot	1258 lb	1243 lb
S.D. per plot	34	28

2. The yield of wheat in a random sample of 1000 farms in a certain area has a S.D. of 192 kg. Another random sample of 1000 farms gives a S.D. of 224 kg. Are the S.D's significantly different?

Answers

1. $z = 2.315$, Difference significant at 5% level ; $z = 1.31$, Difference not significant at 5% level
2. $z = 4.851$. The S.D's are significantly different.

5.12 TEST OF SIGNIFICANCE OF SMALL SAMPLES

When the size of the sample is less than 30, then the sample is called small sample. For such sample it will not be possible for us to assume that the random sampling distribution of a statistic is approximately normal and the values given by the sample data are sufficiently close to the population values and can be used in their place for the calculation of the standard error of the estimate.

5.13 STUDENT'S t-DISTRIBUTION (t-Test)

[G.B.T.U. (MBA) 2011 ; G.B.T.U. (MCA) 2010]

This t-distribution is used when sample size is ≤ 30 and the population standard deviation is unknown.

t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \qquad \text{where,} \qquad S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

$\bar{x}$ is the mean of sample, μ is population mean. S is the standard deviation of population and n is sample size.

If the standard deviation of the sample 's' is given then t-statistic is defined as

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

Note. The relation between s and S is $ns^2 = (n-1)S^2$.

5.13.1 The *t*-Table

The t-table given at the end is the probability integral of t-distribution. The t-distribution has different values for each degrees of freedom and when the degrees of freedom are infinitely large, the t-distribution is equivalent to normal distribution and the probabilities shown in the normal distribution tables are applicable.

5.13.2 Applications of *t*-Distribution

[G.B.T.U. (MBA) 2011]

Some of the applications of t-distribution are given below:

1. *To test if the sample mean ($\bar{x}$) differs significantly from the hypothetical value μ of the population mean.*
2. *To test the significance between two sample means.*
3. *To test the significance of observed partial and multiple correlation coefficients.*

5.13.3 Critical Value of *t*

The critical value or significant value of t at level of significance α, degrees of freedom γ for two tailed test is given by

$$P[|\, t \,| > t_\gamma(\alpha)] = \alpha$$
$$P[|\, t \,| \leq t_\gamma(\alpha)] = 1 - \alpha$$

The significant value of t at level of significance α, for a single tailed test can be got from those of two tailed test by referring to the values at 2α.

5.14 TEST I: t-TEST OF SIGNIFICANCE OF THE MEAN OF A RANDOM SAMPLE

To test whether the mean of a sample drawn from a normal population deviates significantly from a stated value when variance of the population is unknown.

H_0: There is no significant difference between the sample mean $\bar{x}$ and the population mean μ, *i.e.*, we use the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \qquad \text{where} \quad S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

with degree of freedom $n - 1$.

At given level of significance α and degrees of freedom $(n - 1)$, we refer to t-table t_α (two tailed or one tailed). If calculated t value is such that $| t | < t_\alpha$, the null hypothesis is accepted. If $| t | > t_\alpha$, H_0 is rejected.

5.14.1 Fiducial Limits of Population Mean

If t_α is the value of t at level of significance α at $(n - 1)$ degrees of freedom then,

$$\left| \frac{\bar{x} - \mu}{S/\sqrt{n}} \right| < t_\alpha \text{ for acceptance of } H_0.$$

$$\bar{x} - t_\alpha S/\sqrt{n} < \mu < \bar{x} + t_\alpha S/\sqrt{n}$$

95% confidence limits (level of significance 5%) are $\bar{x} \pm t_{0.05} S/\sqrt{n}$.

99% confidence limits (level of significance 1%) are $\bar{x} \pm t_{0.01} S/\sqrt{n}$.

ILLUSTRATIVE EXAMPLES

Example 1. *A random sample of size 16 has 53 as mean. The sum of squares of the deviation' from mean is 135. Can this sample be regarded as taken from the population having 56 as mean? Obtain 95% and 99% confidence limits of the mean of the population.*

Sol. Null hypothesis, H_0: There is no significant difference between the sample mean and hypothetical population mean *i.e.*, $\mu = 56$.

Alternative hypothesis, H_1: $\mu \neq 56$ **(Two tailed test)**

Test statistic Under H_0, test statistic is $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$

Given : $\bar{x} = 53$, $\mu = 56$, $n = 16$, $\Sigma(x - \bar{x})^2 = 135$

$$S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{135}{15}} = 3$$

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{53 - 56}{3/\sqrt{16}} = -4$$

$| t | = 4$

$d.fv. = 16 - 1 = 15.$

Conclusion: Since $| t | = 4 > t_{0.05} = 2.13$ *i.e.*, the calculated value of t is more than the tabulated value, the null hypothesis is rejected. Hence, the sample mean has not come from a population having 56 as mean.

95% confidence limits of the population mean

$$= \bar{x} \pm \frac{S}{\sqrt{n}} t_{0.05} = 53 \pm \frac{3}{\sqrt{16}} (2.13) = 51.4025, 54.5975$$

99% confidence limits of the population mean

$$= \bar{x} \pm \frac{S}{\sqrt{n}} t_{0.01,} = 53 \pm \frac{3}{\sqrt{16}} (2.95) = 50.7875, 55.2125.$$

Example 2. *The lifetime of electric bulbs for a random sample of 10 from a large consignment gave the following data:*

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Life in '000 hrs.	*4.2*	*4.6*	*3.9*	*4.1*	*5.2*	*3.8*	*3.9*	*4.3*	*4.4*	*5.6*

Can we accept the hypothesis that the average lifetime of bulb is 4000 hrs?

Sol. Null hypothesis H_0 : There is no significant difference in the sample mean and population mean. *i.e.*, $\mu = 4000$ hrs.

Alternative hypothesis: $\mu \neq 4000$ hrs (**Two tailed test**)

Test statistic: Under H_0, the test statistic is $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$

x	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6
$x - \bar{x}$	– 0.2	0.2	– 0.5	– 0.3	0.8	– 0.6	– 0.5	– 0.1	0	1.2
$(x - \bar{x})^2$	0.04	0.04	0.25	0.09	0.64	0.36	0.25	0.01	0	1.44

$$\bar{x} = \frac{\Sigma x}{n} = \frac{44}{10} = 4.4, \qquad \Sigma (x - \bar{x})^2 = 3.12$$

$$S = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}} = 0.589$$

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{4.4 - 4}{\left(\frac{0.589}{\sqrt{10}}\right)} = 2.123$$

For $\gamma = 9$, $t_{0.05} = 2.26$.

Conclusion: Since the calculated value of t is less than the tabulated value of t at 5% level of significance.

$\therefore$ The null hypothesis $\mu = 4000$ hrs is accepted *i.e.*, the average lifetime of bulbs could be 4000 hrs.

Example 3. *A sample of 20 items has mean 42 units and S.D. 5 units. Test the hypothesis that it is a random sample from a normal population with mean 45 units.*

Sol. Null hypothesis H_0: There is no significant difference between the sample mean and the population mean. *i.e.*, $\mu = 45$ units

Alternative hypothesis, H_1: $\mu \neq 45$ **(Two tailed test)**

Given: $n = 20$, $\bar{x} = 42$, $s = 5$; $\gamma = 19$ d.f.

Test statistic: Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{42 - 45}{5/\sqrt{19}} = -2.615$$

$\therefore$ $\quad | t | = 2.615$

The tabulated value of t at 5% level for 19 d.f. is $t_{0.05} = 2.09$.

Conclusion: Since the calculated value $| t |$ is greater than the tabulated value of t at 5% level of significance, the null hypothesis H_0 is rejected. *i.e.,* there is significant difference between the sample mean and population mean.

i.e., the sample could not have come from this population.

Example 4. *The 9 items of a sample have the following values:*

45, 47, 50, 52, 48, 47, 49, 53, 51.

Does the mean of these values differ significantly from the assumed mean 47.5?

Sol. Here, $n = 9$, $\mu = 47.5$, $\bar{x} = \frac{\Sigma x}{n} = 49.1$

x	45	47	50	52	48	47	49	53	51
$x - \bar{x}$	– 4.1	– 2.1	0.9	2.9	– 1.1	– 2.1	– 0.1	3.9	1.9
$(x - \bar{x})^2$	16.81	4.41	0.81	8.41	1.21	4.41	0.01	15.21	3.61

$$\Sigma(x - \bar{x})^2 = 54.89,$$

$$S^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1} = 6.86$$

$\therefore$ $\quad S = 2.619$

Null hypothesis:

$$H_0: \mu = 47.5$$

i.e., there is no significant difference between the sample and population means.

Alternative hypothesis:

$$H_1: \mu \neq 47.5$$

Hence we apply **two-tailed test.**

Test statistic: Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{49.1 - 47.5}{(2.619/\sqrt{9})} = 1.8327$$

$$t_{0.05} = 2.31 \text{ for } \gamma = 8$$

Conclusion: Since $|t|_{\text{calculated}} < t_{\text{tabulated}}$ at 5% level of significance, the null hypothesis H_0 is accepted *i.e.,* there is no significant difference between their means.

TEST YOUR KNOWLEDGE

1. Ten individuals are chosen at random from a normal population of students and their marks are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. In the light of these data, discuss the suggestion that mean mark of the population of students is 66.
2. The following values gives the lengths of 12 samples of Egyptian cotton taken from a consignment: 48, 46, 49, 46, 52, 45, 43, 47, 47, 46, 45, 50. Test if the mean length of the consignment can be taken as 46.
3. A sample of 18 items has a mean 24 units and standard deviation 3 units. Test the hypothesis that it is a random sample from a normal population with mean 27 units.
4. A random sample of 10 boys had the I.Q.'s 70, 120, 110, 101, 88, 83, 95, 98, 107 and 100. Do these data support the assumption of a population mean I.Q. of 160?

5.15 TEST II: t-TEST FOR DIFFERENCE OF MEANS OF TWO SMALL SAMPLES (from a Normal Population)

This test is used to test whether the two samples $x_1, x_2, \ldots\ldots, x_{n_1}, y_1, y_2, \ldots\ldots, y_{n_2}$ of sizes n_1, n_2 have been drawn from two normal populations with mean μ_1 and μ_2 respectively under the assumption that the population variance are equal ($\sigma_1 = \sigma_2 = \sigma$).

H_0: The samples have been drawn from the normal population with means μ_1 and μ_2 *i.e.*, $H_0: \mu_1 \neq \mu_2$.

Let $\bar{x}$, $\bar{y}$ be their means of the two samples.

Under this H_0 the test statistic t is given by $t = \dfrac{(\bar{x} - \bar{y})}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$.

Degree of freedom is $n_1 + n_2 - 2$.

Note 1. If the two sample's standard deviations s_1, s_2 are given then we have $S^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$.

Note 2. If s_1, s_2 are not given then $S^2 = \dfrac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$.

ILLUSTRATIVE EXAMPLES

Example 1. *Two samples of sodium vapour bulbs were tested for length of life and the following results were got:*

	Size	*Sample mean*	*Sample S.D.*
Type I	*8*	*1234 hrs*	*36 hrs*
Type II	*7*	*1036 hrs*	*40 hrs*

Is the difference in the means significant to generalise that Type I is superior to Type II regarding length of life?

Sol. Null hypothesis:

$H_0 : \mu_1 = \mu_2$ *i.e.,* two types of bulbs have same lifetime.

Alternative hypothesis:

$H_1 : \mu_1 > \mu_2$ *i.e.,* type I is superior to Type II.

Hence we use **right tailed test.**

$$S^2 = \frac{n_1 s_1^{\,2} + n_2 s_2^{\,2}}{n_1 + n_2 - 2} = \frac{8(36)^2 + 7(40)^2}{8 + 7 - 2} = 1659.076$$

$\therefore$ $S = 40.7317$

Test statistic: Under H_0, the test statistic t is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1234 - 1036}{40.7317\sqrt{\frac{1}{8} + \frac{1}{7}}} = 18.1480$$

$t_{0.05}$ at d.f. $\gamma = n_1 + n_2 - 2 = 13$ is 1.77.

Conclusion: Since calculated $| t | > t_{\text{tabulated}}$ at 5% level of significance, H_0 is rejected.

$\therefore$ Type I is definitely superior to Type II.

Example 2. *Samples of sizes 10 and 14 were taken from two normal populations with S.D. 3.5 and 5.2. The sample means were found to be 20.3 and 18.6. Test whether the means of the two populations are the same at 5% level.*

Sol. We have, $\bar{x}_1 = 20.3$, $\bar{x}_2 = 18.6$, $n_1 = 10$, $n_2 = 14$, $s_1 = 3.5$, $s_2 = 5.2$

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = 22.775$$

$\therefore$ $S = 4.772$

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ *i.e.,* the means of the two populations are the same.

Alternative hypothesis:

$H_1 : \mu_1 \neq \mu_2$

Test statistic: Under H_0, the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{20.3 - 18.6}{4.772\sqrt{\frac{1}{10} + \frac{1}{14}}} = 0.8604$$

The tabulated value of t at 5% level of significance for 22 *d.f.* is $t_{0.05} = 2.0739$

Conclusion:

Since $t = 0.8604 < t_{0.05}$, the null hypothesis H_0 is accepted; *i.e.,* there is no significant difference between their means.

Example 3. *The height of 6 randomly chosen sailors in inches are 63, 65, 68, 69, 71 and 72. Those of 9 randomly chosen soldiers are 61, 62, 65, 66, 69, 70, 71, 72 and 73. Test whether the sailors are on the average taller than soldiers.*

Sol. Let X_1 and X_2 be the two samples denoting the heights of sailors and soldiers.

$$n_1 = 6, n_2 = 9$$

Null hypothesis H_0: $\mu_1 = \mu_2$.

i.e., the mean of both the population are the same.

Alternative hypothesis H_1: $\mu_1 > \mu_2$ (one tailed test)

Calculation of two sample means :

X_1	63	65	68	69	71	72
$X_1 - \bar{X}_1$	– 5	– 3	0	1	3	4
$(X_1 - \bar{X}_1)^2$	25	9	0	1	9	16

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = 68; \Sigma(X_1 - \bar{X}_1)^2 = 60$$

X_2	61	62	65	66	69	70	71	72	73
$X_2 - \bar{X}_2$	– 6.66	– 5.66	– 2.66	1.66	1.34	2.34	3.34	4.34	5.34
$(X_2 - \bar{X}_2)^2$	44.36	32.035	7.0756	2.7556	1.7956	5.4756	11.1556	18.8356	28.5156

$$\bar{X}_2 = \frac{\Sigma X_2}{n_2} = 67.66; \Sigma(X_2 - \bar{X}_2)^2 = 152.0002$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}[\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2] = 16.3077$$

$\therefore$ $S = 4.038$

Test statistic:

Under H_0, $$t = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{68 - 67.666}{4.038\sqrt{\frac{1}{6} + \frac{1}{9}}} = 0.1569$$

The value of t at 5% level of significance for 13 d.f. is 1.77. (d.f. $= n_1 + n_2 - 2$)

Conclusion: Since $t_{calculated} < t_{0.05} = 1.77$, the null hypothesis H_0 is accepted.

i.e., there is no significant difference between their average.

i.e., the sailors are not on the average taller than the soldiers.

TEST YOUR KNOWLEDGE

1. The mean life of 10 electric motors was found to be 1450 hrs with S.D. of 423 hrs. A second sample of 17 motors chosen from a different batch showed a mean life of 1280 hrs with a S.D. of 398 hrs. Is there a significant difference between means of the two samples?
2. The marks obtained by a group of 9 regular course students and another group of 11 part time course students in a test are given below:

Regular :	56	62	63	54	60	51	67	69	58		
Part time :	62	70	71	62	60	56	75	64	72	68	66

Examine whether the marks obtained by regular students and part time students differ significantly at 5% and 1% level of significance.

3. A group of 5 patients treated with the medicine A weigh 42, 39, 48, 60 and 41 kgs. A second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 kg. Do you agree with the claim that medicine B increases the weight significantly? It is given that the value of t at 10% level of significance for 10 degree of freedom is 1.81.

[G.B.T.U. (B. Pharm.) 2010]

4. Two independent samples of sizes 7 and 9 have the following values:

Sample A :	10	12	10	13	14	11	10		
Sample B :	10	13	15	12	10	14	11	12	11

Test whether the difference between the mean is significant.

5. The average number of articles produced by two machines per day are 200 and 250 with standard deviation 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 5% level of significance?

5.16 SNEDECOR'S VARIANCE RATIO TEST OR F-TEST

In testing the significance of the difference of two means of two samples, we assumed that the two samples came from the same population or population with equal variance. The object of the F-test is to discover whether two independent estimates of population variance differ significantly or whether the two samples may be regarded as drawn from the normal populations having the same variance. Hence before applying the t-test for the significance of the difference of two means, we have to test for the equality of population variance by using F-test.

Let n_1 and n_2 be the sizes of two samples with variance s_1^2 and s_2^2. The estimate of the population variance based on these samples are $s_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$ and $s_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$. The degrees of freedom of these estimates are $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$.

To test whether these estimates s_1^2 and s_2^2 are significantly different or if the samples may be regarded as drawn from the same population or from two populations with same variance σ^2, we set-up the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$.

i.e., the independent estimates of the common population do not differ significantly.

To carry out the test of significance of the difference of the variances we calculate the test statistic $F = \frac{s_1^2}{s_2^2}$, the Numerator is greater than the Denominator. *i.e.*, $s_1^2 > s_2^2$.

Conclusion: If the calculated value of F exceeds $F_{0.05}$ for $(n_1 - 1)$, $(n_2 - 1)$ degrees of freedom given in table we conclude that the ratio is significant at 5% level.

i.e., we conclude that the sample could have come from two normal population with same variance.

The assumptions on which F-test is based are:

1. *The populations for each sample must be normally distributed.*
2. *The samples must be random and independent.*
3. *The ratio of σ_1^2 to σ_2^2 should be equal to 1 or greater than 1. That is why we take the larger variance in the Numerator of the ratio.*

5.16.1 Applications: *F-test is used to test*

(i) whether two independent samples have been drawn from the normal populations with the same variance σ^2.

(ii) Whether the two independent estimates of the population variance are homogeneous or not.

ILLUSTRATIVE EXAMPLES

Example 1. *Two random samples drawn from 2 normal populations are as follows:*

A	17	27	18	25	27	29	13	17
B	16	16	20	27	26	25	21	

Test whether the samples are drawn from the same normal population.

Sol. To test if two independent samples have been drawn from the same population we have to test (*i*) equality of the means by applying the *t*-test and (*ii*) equality of population variance by applying F-test.

Since the *t*-test assumes that the sample variances are equal, we shall first apply the F-test.

F-test: Null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ *i.e.*, the population variance do not differ significantly.

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

Test statistic: $F = \dfrac{s_1^2}{s_2^2}$, (if $s_1^2 > s_2^2$)

Computations for s_1^2 and s_2^2

X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$	X_2	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
17	– 4.625	21.39	16	– 2.714	7.365
27	5.735	28.89	16	– 2.714	7.365
18	– 3.625	13.14	20	1.286	1.653
25	3.375	11.39	27	8.286	68.657
27	5.735	28.89	26	7.286	53.085
29	7.735	54.39	25	6.286	39.513
13	– 8.625	74.39	21	2.286	5.226
17	– 4.625	21.39			

$\bar{X}_1 = 21.625$; $n_1 = 8$; $\Sigma(X_1 - \bar{X}_1)^2 = 253.87$

$\bar{X}_2 = 18.714$; $n_2 = 7$; $\Sigma(X_2 - \bar{X}_2)^2 = 182.859$

$$s_1^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{253.87}{7} = 36.267; \quad s_2^2 = \frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{182.859}{6} = 30.47$$

$$F = \frac{s_1^2}{s_2^2} = \frac{36.267}{30.47} = 1.190.$$

Conclusion: The table value of F for $v_1 = 7$ and $v_2 = 6$ degrees of freedom at 5% level is 4.21. The calculated value of F is less than the tabulated value of F. $\therefore$ H_0 is accepted. Hence we conclude that the variability in two populations is same.

t-test: Null hypothesis H_0: $\mu_1 = \mu_2$ *i.e.,* the population means are equal.

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$

Test of statistic

$$s^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{253.87 + 182.859}{8 + 7 - 2} = 33.594 \quad \therefore \quad s = 5.796$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{21.625 - 18.714}{5.796\sqrt{\frac{1}{8} + \frac{1}{7}}} = 0.9704 \sim t(n_1 + n_2 - 2) \text{ d.f.}$$

Conclusion: The tabulated value of t at 5% level of significance for 13 d.f. is 2.16.

The calculated value of t is less than the tabulated value. H_0 is accepted *i.e.,* there is no significant difference between the population mean. *i.e.,* $\mu_1 = \mu_2$. $\therefore$ We conclude that the two samples have been drawn from the same normal population.

Example 2. *Two independent sample of sizes 7 and 6 had the following values:*

Sample A	28	30	32	33	31	29	34
Sample B	29	30	30	24	27	28	

Examine whether the samples have been drawn from normal populations having the same variance.

Sol. H_0: The variance are equal. *i.e.,* $\sigma_1^2 = \sigma_2^2$

i.e., the samples have been drawn from normal populations with same variance.

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

Under null hypothesis, the test statistic $F = \frac{s_1^2}{s_2^2} (s_1^2 > s_2^2)$

Computations for s_1^2 and s_2^2

X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$	X_2	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
28	– 3	9	29	1	1
30	– 1	1	30	2	4
32	1	1	30	2	4
33	2	4	24	– 4	16
31	0	0	27	– 1	1
29	– 2	4	28	0	0
34	3	9			
		28			26

$$\bar{X}_1 = 31, \quad n_1 = 7; \quad \Sigma(X_1 - \bar{X}_1)^2 = 28$$

$$\bar{X}_2 = 28, \quad n_2 = 6; \quad \Sigma(X_2 - \bar{X}_2)^2 = 26$$

$$s_1^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{28}{6} = 4.666; \quad s_2^2 = \frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{26}{5} = 5.2$$

$$F = \frac{s_2^2}{s_1^2} = \frac{5.2}{4.666} = 1.1158. \qquad (\because \ s_2^2 > s_1^2)$$

Conclusion: The tabulated value of F at $v_1 = 6 - 1$ and $v_2 = 7 - 1$ d.f. for 5% level of significance is 4.39. Since the tabulated value of F is less than the calculated value, H_0 is accepted *i.e.*, there is no significant difference between the variance. *i.e.*, the samples have been drawn from the normal population with same variance.

Example 3. *The two random samples reveal the following data:*

Sample no.	*Size*	*Mean*	*Variance*
I	*16*	*440*	*40*
II	*25*	*460*	*42*

Test whether the samples come from the same normal population.

Sol. A normal population has two parameters namely the mean μ and the variance σ^2. To test whether the two independent samples have been drawn from the same normal population, we have to test

(*i*) the equality of means (*ii*) the equality of variance.

Since the *t*-test assumes that the sample variance are equal, we first apply F-test.

F-test: Null hypothesis: $\sigma_1^2 = \sigma_2^2$

The population variance do not differ significantly.

Alternative hypothesis: $\sigma_1^2 \neq \sigma_2^2$

Under the null hypothesis the test statistic is given by $F = \frac{s_1^2}{s_2^2}$, $(s_1^2 > s_2^2)$

Given: $n_1 = 16$, $n_2 = 25$; $s_1^2 = 40$, $s_2^2 = 42$

$$\therefore \quad F = \frac{s_1^2}{s_2^2} = \frac{\frac{n_1 s_1^2}{n_1 - 1}}{\frac{n_2 s_2^2}{n_2 - 1}} = \frac{16 \times 40}{15} \times \frac{24}{25 \times 42} = 0.9752.$$

Conclusion: The calculated value of F is 0.9752. The tabulated value of F at 16 – 1, 25 – 1 d.f. for 5% level of significance is 2.11.

Since the calculated value is less than that of the tabulated value, H_0 is accepted. *i.e.*, the population variance are equal.

t-test: Null hypothesis H_0 : $\mu_1 = \mu_2$ *i.e.*, the population means are equal.

Alternative hypothesis H_1 : $\mu_1 \neq \mu_2$

Given: $n_1 = 16$, $n_2 = 25$, $\bar{X}_1 = 440$, $\bar{X}_2 = 460$

$$s^2 = \frac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2} = \frac{16 \times 40 + 25 \times 42}{16 + 25 - 2} = 43.333 \quad \therefore \quad s = 6.582$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{440 - 460}{6.582\sqrt{\frac{1}{16} + \frac{1}{25}}} = -9.490 \text{ for } (n_1 + n_2 - 2) \text{ d.f.}$$

Conclusion: The calculated value of $|t|$ is 9.490. The tabulated value of t at 39 d.f. for 5% level of significance is 1.96.

Since the calculated value is greater than the tabulated value, H_0 is rejected.

i.e., there is significant difference between means. *i.e.*, $\mu_1 \neq \mu_2$.

Since there is significant difference between means, and no significant difference between variance, we conclude that the samples do not come from the same normal population.

TEST YOUR KNOWLEDGE

1. From the following two sample values, find out whether they have come from the same population:

Sample 1	17	27	18	25	27	29	27	23	17
Sample 2	16	16	20	16	20	17	15	21	

2. The daily wages in Rupees of skilled workers in two cities are as follows:

	Size of sample of workers	*S.D. of wages in the sample*
City A	16	25
City B	13	32

3. The standard deviation calculated from two random samples of sizes 9 and 13 are 2.1 and 1.8 respectively. Can the samples be regarded as drawn from normal populations with the same standard deviation?

4. Two independent samples of size 8 and 9 had the following values of the variables:

Sample I	20	30	23	25	21	22	23	24	
Sample II	30	31	32	34	35	29	28	27	26

Do the estimates of the population variance differ significantly?

Answers

1. rejected **2.** accepted **3.** accepted **4.** accepted.

5.17 CHI-SQUARE (χ^2) TEST

[G.B.T.U. 2010]

When a coin is tossed 200 times, the theoretical considerations lead us to expect 100 heads and 100 tails. But in practice, these results are rarely achieved. The quantity χ^2 (a Greek letter, pronounced as chi-square) describes the magnitude of discrepancy between theory and observation. If $\chi^2 = 0$, the observed and expected frequencies completely coincide. The greater the discrepancy between the observed and expected frequencies, the greater is the value of χ^2. Thus χ^2 **affords a measure of the correspondence between theory and observation.**

If O_i $(i = 1, 2, ..., n)$ is a set of observed (experimental) frequencies and E_i $(i = 1, 2, ..., n)$ is the corresponding set of expected (theoretical or hypothetical) frequencies, then, χ^2 **is defined as**

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

where $\Sigma O_i = \Sigma E_i = N$ (total frequency) and degrees of freedom (d.f.) $= (n - 1)$.

Note. (*i*) If $\chi^2 = 0$, the observed and theoretical frequencies agree exactly.

(*ii*) If $\chi^2 > 0$, they do not agree exactly.

5.18 DEGREES OF FREEDOM

While comparing the calculated value of χ^2 with the tabular value, we have to determine the degrees of freedom.

If we have to choose any four numbers whose sum is 50, we can exercise our independent choice for any three numbers only, the fourth being 50 minus the total of the three numbers selected. Thus, though we were to choose any four numbers, our choice was reduced to three because of one condition imposed. There was only one restraint on our freedom and our degrees of freedom were $4 - 1 = 3$. If two restrictions are imposed, our freedom to choose will be further curtailed and degrees of freedom will be $4 - 2 = 2$.

In general, the number of degrees of freedom is the total number of observations less the number of independent constraints imposed on the observations. Degrees of freedom (d.f.) are usually denoted by ν.

Thus, $\nu = n - k$, where k is the number of independent constraints in a set of data of n observations.

Note. (*i*) For a $p \times q$ contingency table (p columns and q rows), $\nu = (p - 1)(q - 1)$

(*ii*) In the case of a contingency table, the expected frequency of any class

$$= \frac{\text{Total of rows in which it occurs} \times \text{Total of columns in which it occurs}}{\text{Total number of observations}}$$

5.19 APPLICATIONS OF CHI-SQUARE TEST

χ^2 test is one of the simplest and the most general test known. It is applicable to a very large number of problems in practice which can be summed up under the following heads:

(*i*) as a test of goodness of fit.

(*ii*) as a test of independence of attributes.

(*iii*) as a test of homogeneity of independent estimates of the population variance.

(*iv*) as a test of the hypothetical value of the population variance σ^2.

(*v*) as a list to the homogeneity of independent estimates of the population correlation coefficient.

5.20 CONDITIONS FOR APPLYING χ^2 TEST

χ^2 test is an approximate test for large values of n. For the validity of chi-square test of goodness of fit between theory and experiment, the following conditions must be satisfied.

(*a*) The sample observations should be independent.

(*b*) The constraints on the cell frequencies, if any, should be linear *e.g.*, $\Sigma n_i = \Sigma \lambda_i$ or $\Sigma O_i = \Sigma E_i$.

(*c*) N, the total number of frequencies should be reasonably large. It is difficult to say what constitutes largeness, but as an arbitrary figure, we may say that **N should be at least 50,** however, few the cells.

(*d*) No theoretical cell-frequency should be small. Here again, it is difficult to say what constitutes smallness, but 5 should be regarded as the very minimum and **10 is better.** If small theoretical frequencies occur (*i.e.*, < 10), the difficulty is overcome by grouping two or more classes together before calculating (O – E). **It is important to remember that the number of degrees of freedom is determined with the number of classes after regrouping.**

Note 1. If any one of the theoretical frequency is less than 5, then we apply a correction given by F Yates, which is usually known as 'Yates correction for continuity', we add 0.5 to the cell frequency which is less than 5 and adjust the remaining cell frequency suitably so that the marginal total is not changed.

Note 2. It may be noted that the χ^2 test depends only on the set of observed and expected frequencies and on degrees of freedom (d.f.). It does not make any assumption regarding the parent population from which the observations are taken. Since χ^2 does not involve any population parameters, it is termed as a statistic and the test is known as Non-parametric test or Distribution-free test.

5.21 THE χ^2 DISTRIBUTION

For large sample sizes, the sampling distribution of χ^2 can be closely approximated by a continuous curve known as the chi-square distribution. The probability function of χ^2 distribution is given by

$$f(\chi^2) = c(\chi^2)^{(\nu/2-1)}\, e^{-x^2/2}$$

where $e = 2.71828$, ν = number of degrees of freedom; c = a constant depending only on ν.

Symbolically, the degrees of freedom are denoted by the symbol ν or by d.f. and are obtained by the rule $\nu = n - k$, where k refers to the number of independent constraints.

In general, when we fit a binomial distribution the number of degrees of freedom is one less than the number of classes; when we fit a Poisson distribution the degrees of freedom are 2 less than the number of classes, because we use the total frequency and the arithmetic mean to get the parameter of the Poisson distribution. When we fit a normal curve the number of degrees of freedom are 3 less than the number of classes, because in this fitting we use the total frequency, mean and standard deviation.

If the data is given in a series of "n" numbers then degrees of freedom = $n - 1$.
In the case of Binomial distribution d.f. = $n - 1$
In the case of Poisson distribution d.f. = $n - 2$
In the case of Normal distribution d.f. = $n - 3$.

5.22 χ^2 TEST AS A TEST OF GOODNESS OF FIT

χ^2 test enables us to ascertain how well the theoretical distributions such as Binomial, Poisson or Normal etc., fit empirical distributions, *i.e.*, distributions obtained from sample data. If the **calculated value of χ^2 is less than the tabular value** at a specified level (generally 5%) of significance, the **fit is considered to be good** *i.e.*, the divergence between actual and expected frequencies is attributed to fluctuations of simple sampling. If the calculated value of χ^2 is greater than the tabular value, the fit is considered to be poor.

ILLUSTRATIVE EXAMPLES

Example 1. *In experiments on pea breeding, the following frequencies of seeds were obtained:*

Round and yellow	*Wrinkled and yellow*	*Round and green*	*Wrinkled and green*	*Total*
315	*101*	*108*	*32*	*556*

Theory predicts that the frequencies should be in proportions 9 : 3 :: 3 : 1. Examine the correspondence between theory and experiment.

Sol. Null hypothesis

H_0 : The experimental result support the theory *i.e.,* there is no significant difference between the observed and theoretical frequency.

Under H_0, The theoretical (expected) frequencies can be calculated as follows:

$$E_1 = \frac{556 \times 9}{16} = 312.75$$

$$E_2 = \frac{556 \times 3}{16} = 104.25$$

$$E_3 = \frac{556 \times 3}{16} = 104.25$$

$$E_4 = \frac{556 \times 1}{16} = 34.75$$

To calculate the value of χ^2:

Observed frequency O_i	315	101	108	32
Expected Frequency E_i	312.75	104.25	104.25	34.75
$\frac{(O_i - E_i)^2}{E_i}$	0.016187	0.101319	0.134892	0.217626

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = 0.470024$$

Tabular value of χ^2 at 5% level of significance for $n - 1 = 3$ *d.f.* is 7.815 *i.e.,* $\chi^2_{0.05} = 7.815$.

Conclusion: Since the calculated value of χ^2 is less than that of the tabulated value, hence H_0 is accepted. Therefore, the experimental results support the theory.

Example 2. *The following table gives the number of accidents that took place in an industry during various days of the week. Test if accidents are uniformly distributed over the week.*

Day	*Mon*	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*
No. of accidents	*14*	*18*	*12*	*11*	*15*	*14*

Sol. Null hypothesis H_0 : The accidents are uniformly distributed over the week.

Under this H_0, the expected frequencies of the accidents on each of these days $= \frac{84}{6} = 14$

Observed frequency O_i	14	18	12	11	15	14
Expected frequency E_i	14	14	14	14	14	14
$(O_i - E_i)^2$	0	16	4	9	1	0

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = \frac{\Sigma(O_i - E_i)^2}{E_i} = \frac{30}{14} = 2.1428.$$

Tabular value of χ^2 at 5% level for $(6 - 1 = 5$ d.f.) is 11.09.

Conclusion: Since the calculated value of χ^2 is less than the tabulated value, H_0 is accepted *i.e.*, the accidents are uniformly distributed over the week.

Example 3. *A die is thrown 276 times and the results of these throws are given below:*

No. appeared on the die	*1*	*2*	*3*	*4*	*5*	*6*
Frequency	*40*	*32*	*29*	*59*	*57*	*59*

Test whether the die is biased or not. (A.K.T.U. 2019)

Sol. Null hypothesis H_0: Die is unbiased.

Under this H_0, the expected frequencies for each digit is $\frac{276}{6} = 46$.

To find the value of χ^2

O_i	40	32	29	59	57	59
E_i	46	46	46	46	46	46
$(O_i - E_i)^2$	36	196	289	169	121	169

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = \frac{\Sigma(O_i - E_i)^2}{E_i} = \frac{980}{46} = 21.30.$$

Tabulated value of χ^2 at 5% level of significance for $(6 - 1 = 5)$ d.f. is 11.09.

Conclusion: Since the calculated value of $\chi^2 = 21.30 > 11.07$ the tabulated value, H_0 is rejected. *i.e.*, die is not unbiased or die is biased.

Example 4. *Records taken of the number of male and female births in 800 families having four children are as follows:* [U.P.T.U. (MCA) 2009]

No. of male births	*0*	*1*	*2*	*3*	*4*
No. of female births	*4*	*3*	*2*	*1*	*0*
No. of families	*32*	*178*	*290*	*236*	*64*

Test whether the data are consistent with the hypothesis that the Binomial law holds and the chance of male birth is equal to that of female birth, namely $p = q = 1/2$.

Sol. Null hypothesis H_0: The data are consistent with the hypothesis of equal probability for male and female births. *i.e.*, $p = q = 1/2$.

We use Binomial distribution to calculate theoretical frequency given by:

$$N(r) = N \times P(X = r) = N \times {}^nC_r\, p^r q^{n-r}$$

where N is the total frequency, N(r) is the number of families with r male children.

p and q are probabilities of male and female births respectively, n is the number of children.

$$N(0) = 800 \times {}^4C_0 \left(\frac{1}{2}\right)^4 = 50,$$

$N(1) = 200$, $N(2) = 300$, $N(3) = 200$ and $N(4) = 50$

Observed frequency O_i	32	178	290	236	64
Expected frequency E_i	50	200	300	200	50
$(O_i - E_i)^2$	324	484	100	1296	196
$\frac{(O_i - E_i)^2}{E_i}$	6.48	2.42	0.333	6.48	3.92

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = 19.633.$$

Tabular value of χ^2 at 5% level of significance for $5 - 1 = 4$ d.f. is 9.49.

Conclusion: Since the calculated value of χ^2 is greater than the tabulated value, H_0 is rejected. *i.e.*, the data are not consistent with the hypothesis that the Binomial law holds and that the chance of a male birth is not equal to that of a female birth.

Example 5. *The theory predicts the proportion of beans in the four groups, G_1, G_2, G_3, G_4 should be in the ratio 9 : 3 : 3 : 1. In an experiment with 1600 beans the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory?*

Sol. Null hypothesis H_0 : The experimental result support the theory. *i.e.*, there is no significant difference between the observed and theoretical frequency.

Under H_0, the theoretical frequency can be calculated as follows:

$$E(G_1) = \frac{1600 \times 9}{16} = 900; \; E(G_2) = \frac{1600 \times 3}{16} = 300;$$

$$E(G_3) = \frac{1600 \times 3}{16} = 300; \; E(G_4) = \frac{1600 \times 1}{16} = 100$$

To calculate the value of χ^2.

Observed frequency O_i	882	313	287	118
Expected frequency E_i	900	300	300	100
$\frac{(O_i - E_i)^2}{E_i}$	0.36	0.5633	0.5633	3.24

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] = 4.7266.$$

Tablular value of χ^2 at 5% level of significance for 3 d.f. is 7.815.

Conclusion. Since the calculated value of χ^2 is less than that of the tabulated value, hence H_0 is accepted. *i.e.,* the experimental results support the theory.

Example 6. *The following table shows the distribution of digits in numbers chosen at random from a telephone directory:*

Digits	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Frequency	*1026*	*1107*	*997*	*966*	*1075*	*933*	*1107*	*972*	*964*	*853*

Test whether the digits may be taken to occur equally frequently in the directory.

[G.B.T.U. (MCA) 2011]

Sol. Null hypothesis H_0 : The digits taken in the directory occur equally frequently *i.e.,* there is no significant difference between the observed and expected frequency.

Under H_0, the expected frequency $= \frac{10000}{10} = 1000$

Calculation of χ^2

O_i	1026	1107	997	966	1075	933	1107	972	964	853
E_i	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
$(O_i - E_i)^2$	676	11449	9	1156	5625	4489	11449	784	1296	21609

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] = \frac{58542}{1000} = 58.542$$

The tabulated value of χ^2 at 5% level of significance for 9 d.f. is 16.919.

Conclusion: Since $\chi^2_{\text{calculated}} > \chi^2_{\text{tabulated}}$, H_0 is rejected *i.e.,* there is significant difference between the observed and theoretical frequencies. Therefore, the digits taken in the directory donot occur equally frequently.

Example 7. *When the first proof of 392 pages of a book of 1200 pages were read, the distribution of printing mistakes were found to be as follows:*

No. of mistakes in a page (x) :	*0*	*1*	*2*	*3*	*4*	*5*	*6*
No. of pages (f) :	*275*	*72*	*30*	*7*	*5*	*2*	*1*

Fit a poisson distribution to the above data and test the goodness of fit.

Sol. Null Hypothesis H_0: Poisson distribution is a good fit to the data.

$$\text{Mean } (\lambda) = \frac{\sum fx}{\sum f} = \frac{189}{392} = 0.4821$$

The frequency of x mistakes per page is given by the poisson law as follows:

$$N(x) = N \cdot P(x)$$

$$= 392\left[\frac{e^{-0.4821}(0.4821)^x}{x!}\right] = \frac{242.05(0.4821)^x}{x!}\ ;\ 0 \le x \le 6$$

Under H_0, expected frequencies are,

$N(0) = 242.05$, $N(1) = 116.69$, $N(2) = 28.13$, $N(3) = 4.52$

$N(4) = 0.54$, $N(5) = 0.052$, $N(6) = 0.0042$

The χ^2-table is as follows:

Mistakes per page (x)	*Observed frequency (O_i)*	*Expected frequency (E_i) (correct to one place of decimal)*	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	275	242.1	1082.41	4.471
1	72	116.7	1998.09	17.121
2	30	28.1	3.61	0.128
3	7	4.5		
4	5 } 15	0.5 } 5.1	98.01	19.217
5	2	0.1		
6	1	0		
Total	392	392		40.937

$$\chi^2_{\text{cal.}} = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] = 40.937$$

d.f. $= 7 - 1 - 1 - 3 = 2$

One d.f. is lost because of linear constraint $\Sigma O_i = \Sigma E_i$. One d.f. is lost because the parameter λ has been estimated from the given data and is then used for computing the expected frequencies. 3 d.f. are lost because of grouping the last four expected cell frequencies which were less than 5.

Tabulated value of χ^2 for 2 d.f. at 5% level of significance is 5.991.

Conclusion: Since $\chi^2_{\text{cal.}} > \chi^2_{\text{tab.}}$, the null hypothesis is rejected at 5% level of significance. Hence, we conclude that poisson distribution is not a good fit to the given data.

Example 8. *Fit a Poisson distribution to the following data and best the goodness of fit:*

x:	0	1	2	3	4
f:	109	65	22	3	1

Sol. Null hypothesis, H_0 : Poisson distribution is a good fit to the data.

$$\text{Mean } (\lambda) = \frac{\Sigma fx}{\Sigma f} = \frac{122}{200} = 0.61$$

$$N(x) = N \cdot P(x) = (200)\ \frac{e^{-0.61}\,(0.61)^x}{x\,!} = \frac{(108.67)\,(0.61)^x}{x\,!}$$

Under H_0, expected frequencies are

$N(0) = 108.67 \approx 109$, $N(1) = 66.29 \approx 66$, $N(2) = 20.22 \approx 20$

$N(3) = 4.11 \approx 4$, $N(4) = 0.63 \approx 1$

The χ^2-table is as follows:

x	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	109	109	0	0
1	65	66	1	0.01515
2	22	20	4	0.2
3	3 } 4	4 } 5	1	0.2
4	1 }	1 }		
Total	200	200		0.41515

$$\chi^2_{cal.} = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = 0.41515$$

d.f. $= 5 - 1 - 1 - 1 = 2$

Tabulated value of χ^2 for 2 d.f. at 5% level of significance is 5.991.

Conclusion: Since $\chi^2_{cal.} < \chi^2_{tab.}$, the null hypothesis H_0 is accepted at 5% level of significance. Hence we conclude that Poisson distribution is a good fit to the given data.

TEST YOUT KNOWLDEDGE

1. A sample analysis of examination results of 500 students, it was found that 220 students have failed, 170 have secured a third class, 90 have secured a second class and the rest, a first class. Do these figures support the general belief that above categories are in the ratio 4 : 3 : 2 : 1 respectively? (The tabular value of χ^2 for d.f. 3 at 5% level of significance is 7.81)

2. What is χ^2–test? *[G.B.T.U. 2010]*

A die is thrown 90 times with the following results:

Face :	1	2	3	4	5	6	Total
Frequency :	10	12	16	14	18	20	90

Use χ^2 test to test whether these data are consistent with the hypothesis that die is unbiased.

Given $\chi^2_{0.05} = 11.07$ for 5 degrees of freedom.

3. (*i*) A survey of 320 families with 5 children shows the following distribution:

No. of boys & girls :	5 boys & 0 girl	4 boys & 1 girl	3 boys & 2 girls	2 boys & 3 girls	1 boy & 4 girls	0 boy & 5 girls	Total
No. of families :	18	56	110	88	40	8	320

Given that values of χ^2 for 5 degrees of freedom are 11.1 and 15.1 at 0.05 and 0.01 significance level respectively, test the hypothesis that male and female births are equally probable.

(G.B.T.U. 2010)

(*ii*) A survery of 240 families with 4 children shows the following distribution:

No. of boys :	1	2	3	4	5	6	Total
No. of families :	10	12	16	14	18	20	90

(Given: $\chi^2_{0.05} = 9.49$ and 11.1 for 4 d.f. and 5 d.f., respectively)

4. A chemical extraction plant processes sea water to collect sodium chloride and magnesium. It is known that sea water contains sodium chloride, magnesium and other elements in the ratio 62 : 4 : 34. A sample of 200 tonnes of sea water has resulted in 130 tonnes of sodium chloride and 6 tonnes of magnesium. Are these data consistent with the known composition of sea water at 5% level of significance? (Given that the tabular value of χ^2 is 5.991 for 2 degree of freedom).

5. The demand for a particular spare part in a factory was found to vary from day-to-day. In a sample study, the following information was obtained:

Days :	Mon	Tue	Wed	Thurs	Fri	Sat
No. of parts demanded :	1124	1125	1110	1120	1126	1115

Test the hypothesis that the number of parts demanded does not depend on the day of the week.

[**Given.** The values of chi-square significance at 5, 6, 7 d.f. are respectively 11.07, 12.59, 14.07 at 5% level of significance]

(G.B.T.U. 2011)

6. The sales in a supermarket during a week are given below. Test the hypothesis that the sales donot depend on the day of the week using a significant level of 0.05.

Days	*Mon*	*Tue*	*Wed*	*Thurs*	*Fri*	*Sat*
Sales (in 1000 ₹)	65	54	60	56	71	84

7. 4 coins were tossed at a time and this operation is repeated 160 times. It is found that 4 heads occur 6 times, 3 heads occur 43 times, 2 heads occur 69 times, one head occur 34 times. Discuss whether the coin may be regarded as unbiased?

8. 200 digits are chosen at random from a set of tables. The frequencies of the digits were:

Digits :	0	1	2	3	4	5	6	7	8	9
Frequency :	18	19	23	21	16	25	22	20	21	15

Use χ^2-test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the table, given that the value of χ^2 are respectively 16.9, 18.3 and 19.7 for 9, 10 and 11 degrees of freedom at 5% level of significance.

9. A genetical law says that children having one parent of blood group M and the other parent of blood group N will always be one of the three blood groups M, MN, N and that the average no. of children in these groups will be in the ratio 1 : 2 : 1. The report on an experiment states as follows:

"Of 162 children having one M parent and one N parent, 28.4% were found to be of group M, 42% of group MN and the rest of the group N." Do the data in the report conform to the expected genetic ratio 1 : 2 : 1?

10. Every clinical thermometer is classified into one of the four categories A, B, C and D on the basis of inspection and test. From past experience, it is knwon that thermometers produced by a certain manufacturer are distributed among the four categories in the following proportions:

Category :	A	B	C	D
Proportion :	0.87	0.09	0.03	0.01

A new lot of 1336 thermometers is submitted by the manufacturer for inspection and test and the following distribution into four categories results:

Category :	A	B	C	D
No. of thermometers reported :	1188	91	47	10

Does this new lot of thermometers differ from the previous experience with regards to proportion of thermometers in each category?

11. Test for goodness of fit of a poisson distribution at 5% level of significance to the following frequency distribution:

(*i*)

x :	0	1	2	3	4	5	6	7	8
f :	52	151	130	102	45	12	5	1	2

[**Hint.** Group the last three frequencies]

(*ii*)

x :	0	1	2	3	4	5	6	7	8	9	10	11	12	13
f :	3	15	47	76	68	74	46	39	15	9	5	2	0	1

[**Hint.** Group the first two and last four frequencies]

(*iii*)

x :	0	1	2	3	4	5
f :	275	138	75	7	4	1

[**Hint.** Club the last three frequencies]

(*iv*)

x :	0	1	2	3	4
f :	419	352	154	56	19

12. (*i*) Fit a binomial distribution to the data and test for goodness of fit at 5% level of significance.

x :	0	1	2	3	4	5
y :	38	144	342	287	164	25

(*ii*) A random number table of 250 digits showed the following distribution of digits 0, 1, 2, ..., 9.

Digit :	0	1	2	3	4	5	6	7	8	9
Observed Frequency :	17	31	29	18	14	20	35	30	20	36
Expected Frequency :	25	25	25	25	25	25	25	25	25	25

Does the observed distribution differ significantly from expected distributions using a significance level of 0.01? Given that $\chi^2_{0.99}$ for 9 degrees of freedom is 21.7 [*G.B.T.U. MCA (SUM) 2010*]

Answers

1. No **2.** Yes

3. H_0 accepted at 1% level of significance and rejected at 5% level of significance

4. $\chi^2 = 1.025$, Yes **5.** H_0 accepted at 5% level **6.** H_0 accepted

7. Coin is unbiased **8.** H_0 accepted at 5% level **9.** H_0 accepted at 5% level

10. H_0 rejected at 5% level

11. (*i*) H_0 accepted at 5% level, (*ii*) H_0 accepted at 5% level, (*iii*) H_0 rejected at 5% level, not a good fit. (*iv*) H_0 accepted at 5% level

12. (*i*) H_0 accepted at 5% level, provides a good fit. (*ii*) Yes.

5.23 χ^2 TEST AS A TEST OF INDEPENDENCE

With the help of χ^2 test, we can find whether or not, two attributes are associated. We take the null hypothesis that there is no association between the attributes under study, *i.e.*, **we assume that the two attributes are independent. If the calculated value of χ^2 is less than the table value** at a specified level (generally 5%) of significance, the hypothesis holds good, *i.e.*, **the attributes are independent** and do not bear any association. On the other hand, if the calculated value of χ^2 is greater than the table value at a specified level of significance, we say that the results of the experiment do not support the hypothesis. In other words, the attributes are associated. Thus a very useful application of χ^2 test is to investigate the relationship between trials or attributes which can be classified into two or more categories.

The sample data set out into two-way table, called **contingency table.**

Let us consider two attributes A and B divided into r classes $A_1, A_2, A_3, \ldots\ldots, A_r$ and B divided into s classes $B_1, B_2, B_3, \ldots\ldots, B_s$. If (A_i), (B_j) represents the number of persons possessing the attributes A_i, B_j respectively, $(i = 1, 2, \ldots\ldots, r, j = 1, 2, \ldots\ldots, s)$ and $(A_i\ B_j)$ represent the number of persons possessing attributes A_i and B_j. Also we have $\sum_{i=1}^{r} A_i = \sum_{j=1}^{s} B_j = N$, where N is the total frequency. The contingency table for $r \times s$ is given as follows:

B \ A	A_1	A_2	A_3	$\ldots A_r$	*Total*
B_1	(A_1B_1)	(A_2B_1)	(A_3B_1)	$\ldots\ldots(A_rB_1)$	B_1
B_2	(A_1B_2)	(A_2B_2)	(A_3B_2)	$\ldots\ldots(A_rB_2)$	B_2
B_3	(A_1B_3)	(A_2B_3)	(A_3B_3)	$\ldots\ldots(A_rB_3)$	B_3
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
B_s	(A_1B_s)	(A_2B_s)	(A_3B_s)	$\ldots\ldots(A_rB_s)$	(B_s)
Total	(A_1)	(A_2)	(A_3)	$\ldots\ldots(A_r)$	N

H_0 : Both the attributes are independent, *i.e.*, A and B are independent under the null hypothesis, we calculate the expected frequency as follows:

$$P(A_i) = \text{Probability that a person possesses the attribute } A_i = \frac{(A_i)}{N} \quad i = 1, 2, \ldots\ldots, r$$

$$P(B_j) = \text{Probability that a person possesses the attribute } B_j = \frac{(B_j)}{N}$$

$$P(A_iB_j) = \text{Probability that a person possesses both attributes } A_i \text{ and } B_j = \frac{(A_iB_j)}{N}$$

If $(A_iB_j)_0$ is the expected number of persons possessing both the attributes A_i and B_j

$$(A_iB_j)_0 = NP(A_iB_j) = NP(A_i)(B_j)$$

$$= N \frac{(A_i)}{N} \frac{(B_j)}{N} = \frac{(A_i)(B_j)}{N} \qquad (\because \text{ A and B are independent})$$

Hence, $$\boxed{\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \left[\frac{[(A_iB_j) - (A_iB_j)_0]^2}{(A_iB_j)_0} \right]}$$

which is distributed as a χ^2 variate with $(r-1)(s-1)$ degrees of freedom.

Note 1. For a 2 × 2 contingency table where the frequencies are $\frac{a|b}{c|d}$, χ^2 can be calculated from independent frequencies as $\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$.

Note 2. If the contingency table is not 2 × 2, then the above formula for calculating χ^2 can't be used. Hence, we have another formula for calculating the expected frequency $(A_iB_j)_0 = \frac{(A_i)(B_j)}{N}$

i.e., expected frequency in each cell is $= \frac{\text{Product of column total and row total}}{\text{whole total}}$.

Note 3. If $\frac{a|b}{c|d}$ is the 2 × 2 contingency table with two attributes, $Q = \frac{ad-bc}{ad+bc}$ is called the coefficient of association. If the attributes are independent then $\frac{a}{b} = \frac{c}{d}$.

Note 4. Yates's Correction. In a 2 × 2 table, if the frequencies of a cell is small, we make Yates's correction to make χ^2 continuous. Decrease by $\frac{1}{2}$ those cell frequencies which are greater than expected frequencies, and increase by $\frac{1}{2}$ those which are less than expectation. This will not affect the marginal columns. This correction is known as Yates's correction to continuity. After Yates's correction

$$\chi^2 = \frac{N\left(bc - ad - \frac{1}{2}N\right)^2}{(a+c)(b+d)(c+d)(a+b)} \quad \text{when} \quad ad - bc < 0$$

and

$$\chi^2 = \frac{N\left(ad - bc - \frac{1}{2}N\right)^2}{(a+c)(b+d)(c+d)(a+b)} \quad \text{when} \quad ad - bc > 0.$$

ILLUSTRATIVE EXAMPLES

Example 1. *What are the expected frequencies of 2 × 2 contingency tables given below:*

(*i*)

a	*b*
c	*d*

(*ii*)

2	10
6	6

Sol. Observed frequencies → Expected frequencies

(*i*)

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d=\mathrm{N}$

→

$\frac{(a+c)(a+b)}{a+b+c+d}$	$\frac{(b+d)(a+b)}{a+b+c+d}$
$\frac{(a+c)(c+d)}{a+b+c+d}$	$\frac{(b+d)(c+d)}{a+b+c+d}$

Observed frequencies → Expected frequencies

(*ii*)

2	10	12
6	6	12
8	16	24

→

$\frac{8\times12}{24}=4$	$\frac{16\times12}{24}=8$
$\frac{8\times12}{24}=4$	$\frac{16\times12}{24}=8$

Example 2. *From the following table regarding the colour of eyes of father and son, test if the colour of son's eye is associated with that of the father.*

	Eye colour of son	
	Light	*Not light*
Light	*471*	*51*
Not light	*148*	*230*

Eye colour of father

Sol. Null hypothesis $\mathrm{H_0}$: The colour of son's eye is not associated with that of the father, *i.e.*, they are independent.

Under $\mathrm{H_0}$, we calculate

$$\text{the expected frequency in each cell} = \frac{\text{Product of column total and row total}}{\text{whole total}}$$

Expected frequencies are:

Eye colour of father \ *Eye colour of son*	*Light*	*Not light*	*Total*
Light	$\frac{619\times522}{900}=359.02$	$\frac{289\times522}{900}=167.62$	522
Not light	$\frac{619\times378}{900}=259.98$	$\frac{289\times378}{900}=121.38$	378
Total	619	289	900

$$\chi^2=\frac{(471-359.02)^2}{359.02}+\frac{(51-167.62)^2}{167.62}+\frac{(148-259.98)^2}{259.98}+\frac{(230-121.38)^2}{121.38}=261.498.$$

Tabulated value of χ^2 at 5% level for 1 d.f. is 3.841.

Conclusion. Since the calculated value of χ^2 > tabulated value of χ^2, H_0 is rejected. They are dependent *i.e.*, the colour of son's eye is associated with that of the father.

Example 3. *The following table gives the number of good and bad parts produced by each of the three shifts in a factory:*

	Good parts	*Bad parts*	*Total*
Day shift	*960*	*40*	*1000*
Evening shift	*940*	*50*	*990*
Night shift	*950*	*45*	*995*
Total	*2850*	*135*	*2985*

Test whether or not the production of bad parts is independent of the shift on which they were produced.

Sol. Null hypothesis H_0: The production of bad parts is independent of the shift on which they were produced. *i.e.,* the two attributes, production and shifts are independent.

Under H_0,
$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\left[\frac{[(A_iB_j)_0 - (A_iB_j)]^2}{(A_iB_j)_0}\right]$$

Calculation of expected frequencies

Let A and B be the two attributes namely production and shifts. A is divided into two classes A_1, A_2 and B is divided into three classes B_1, B_2, B_3.

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{(2850)\times(1000)}{2985} = 954.77$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{(2850)\times(990)}{2985} = 945.226$$

$$(A_1B_3)_0 = \frac{(A_1)(B_3)}{N} = \frac{(2850)\times(995)}{2985} = 950$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{(135)\times(1000)}{2985} = 45.27$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{(135)\times(990)}{2985} = 44.773$$

$$(A_2B_3)_0 = \frac{(A_2)(B_3)}{N} = \frac{(135)\times(995)}{2985} = 45.$$

To calculate the value of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
(A_1B_1)	960	954.77	27.3529	0.02864
(A_1B_2)	940	945.226	27.3110	0.02889
(A_1B_3)	950	950	0	0
(A_2B_1)	40	45.27	27.7729	0.61349
(A_2B_2)	50	44.773	27.3215	0.61022
(A_2B_3)	45	45	0	0
				1.28126

The tabulated value of χ^2 at 5% level of significance for 2 degrees of freedom $(r-1)(s-1)$ is 5.991.

Conclusion: Since the calculated value of χ^2 is less than the tabulated value, we accept H_0. *i.e.*, the production of bad parts is independent of the shift on which they were produced.

Example 4. *From the following data, find whether hair colour and sex are associated.*

Sex \ Colour	*Fair*	*Red*	*Medium*	*Dark*	*Black*	*Total*
Boys	*592*	*849*	*504*	*119*	*36*	*2100*
Girls	*544*	*677*	*451*	*97*	*14*	*1783*
Total	*1136*	*1526*	*955*	*216*	*50*	*3883*

Sol. Null hypothesis H_0: The two attributes hair colour and sex are not associated. *i.e.*, they are independent.

Let A and B be the attributes hair colour and sex respectively. A is divides into 5 classes $(r = 5)$. B is divided into 2 classes $(s = 2)$.

$\therefore$ Degrees of freedom $= (r-1)(s-1) = (5-1)(2-1) = 4$

Under H_0, we calculate $\chi^2 = \sum_{i=1}^{5}\sum_{j=1}^{2} \dfrac{[(A_iB_j)_0 - (A_iB_j)]^2}{(A_iB_j)_0}$

To calculate the expected frequency $(A_i B_j)_0$ as follows:

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{1136 \times 2100}{3883} = 614.37$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{1136 \times 1783}{3883} = 521.629$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{1526 \times 2100}{3883} = 852.289$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{1526 \times 1783}{3883} = 700.71$$

$$(A_3B_1)_0 = \frac{(A_3)(B_1)}{N} = \frac{955 \times 2100}{3883} = 516.482$$

$$(A_3B_2)_0 = \frac{(A_3)(B_2)}{N} = \frac{955 \times 1783}{3883} = 483.517$$

$$(A_4B_1)_0 = \frac{(A_4)(B_1)}{N} = \frac{216 \times 2100}{3883} = 116.816$$

$$(A_4B_2)_0 = \frac{(A_4)(B_2)}{N} = \frac{216 \times 1783}{3883} = 99.183$$

$$(A_5B_1)_0 = \frac{(A_5)(B_1)}{N} = \frac{50 \times 2100}{3883} = 27.04$$

$$(A_5B_2)_0 = \frac{(A_5)(B_2)}{N} = \frac{50 \times 1783}{3883} = 22.959$$

Calculation of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
A_1B_1	592	614.37	500.416	0.8145
A_1B_2	544	521.629	500.462	0.959
A_2B_1	849	852.289	10.8175	0.0127
A_2B_2	677	700.71	562.1641	0.8023
A_3B_1	504	516.482	155.800	0.3016
A_3B_2	451	438.517	155.825	0.3553
A_4B_1	119	116.816	4.7698	0.0408
A_4B_2	97	99.183	4.7654	0.0480
A_5B_1	36	27.04	80.2816	2.9689
A_5B_2	14	22.959	80.2636	3.495
				9.79975

$$\chi^2_{\text{cal.}} = 9.799.$$

Tabular value of χ^2 at 5% level of significance for 4 d.f. is 9.488.

Conclusion. Since the calculated value of χ^2 < tabulated value, H_0 is rejected. *i.e.,* the two attributes are not independent. *i.e.,* the hair colour and sex are associated.

Example 5. *Can vaccination be regarded as preventive measure of small pox as evidenced by the following data of 1482 persons exposed to smallpox in a locality. 368 in all were attacked of these 1482 persons and 343 were vaccinated and of these only 35 were attacked.*

Sol. For the given data we form the contingency table. Let the two attributes be vaccination and exposed to small pox. Each attributes is divided into two classes.

Vaccination A / *Disease small-pox B*	*Vaccinated*	*Not*	*Total*
Attacked	35	333	368
Not	308	806	1114
Total	343	1139	1482

Null hypothesis H_0: The two attributes are independent *i.e.*, vaccination can't be regarded as preventive measure of small pox.

Degrees of freedom $\nu = (r-1)(s-1) = (2-1)(2-1) = 1$

Under H_0, $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{[(A_iB_j)_0 - (A_iB_j)]^2}{(A_iB_j)_0}$

Calculation of expected frequency

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{343 \times 368}{1482} = 85.1713$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{343 \times 1114}{1482} = 257.828$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{1139 \times 368}{1482} = 282.828$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{1139 \times 1114}{1482} = 856.171$$

Calculation of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
(A_1B_1)	35	85.1713	2517.159	29.554
(A_1B_2)	308	257.828	2517.229	8.1728
(A_2B_1)	333	282.828	2517.2295	7.5592
(A_2B_2)	806	856.171	2517.1292	2.9399
				48.2261

Calculated value of $\chi^2 = 48.2261$.

Tabulated value of χ^2 at 5% level of significance for 1 d.f. is 3.841.

Conclusion: Since the calculated value of χ^2 > tabulated value, H_0 is rejected.

i.e., the two attributes are not independent. *i.e.*, the vaccination can be regarded as preventive measure of small pox.

Example 6. *To test the effectiveness of inoculation against cholera, the following table was obtained:*

	Attacked	*Not attacked*	*Total*
Inoculated	*30*	*160*	*190*
Not inoculated	*140*	*460*	*600*
Total	*170*	*620*	*790*

(The figures represent the number of persons.)

Use χ^2-test to defend or refute the statement that the inoculation prevents attack from cholera. **(A.K.T.U. 2009, 2018)**

Sol. Null hypothesis H_0: The inoculation does not prevent attack from cholera.
Under H_0, we calculate the expected frequencies as:

	Attacked	*Not attacked*
Inoculated	$\frac{190 \times 170}{790} = 40.886$	$\frac{190 \times 620}{790} = 149.11$
Not inoculated	$\frac{600 \times 170}{790} = 129.11$	$\frac{600 \times 620}{790} = 470.89$

Calculation of χ^2

O_i	30	160	140	460
E_i	40.886	149.11	129.11	470.89
$\frac{(O_i - E_i)^2}{E_i}$	2.898	0.795	0.918	0.252

$$\chi^2_{cal.} = \sum \frac{(O_i - E_i)^2}{E_i} = 4.863$$

Tabulated value of χ^2 at 5% level of significance for 1 d.f. is 3.841.

Conclusion: Since $\chi^2_{cal.} > \chi^2_{tab.}$ at 5% level of significance, null hypothesis H_0 is rejected. Hence we defend the statement that inoculation prevents attack from cholera.

TEST YOUR KNOWLEDGE

1. In a locality 100 persons were randomly selected and asked about their educational achievements. The results are given below:

		Education		
		Middle	*High school*	*College*
Sex	Male	10	15	25
	Female	25	10	15

Based on this information can you say that the education depends on sex.

2. The following data is collected on two characters:

	Smokers	*Non smokers*
Literate	83	57
Illiterate	45	68

Based on this information can you say that there is no relation between habit of smoking and literacy.

3. 500 students at school were graded according to their intelligences and economic conditions of their homes. Examine whether there is any association between economic condition and intelligence, from the following data:

Economic conditions	*Intelligence*	
	Good	*Bad*
Rich	85	75
Poor	165	175

4. In an experiment on the immunisation of goats from anthrox, the following results were obtained. Derive your inferences on the efficiency of the vaccine.

	Died anthrox	*Survived*
Inoculated with vaccine	2	10
Not inoculated	6	6

5. By using χ^2-test, find out whether there is any association between income level and type of schooling:

Income	*Public School*	*Govt. School*
Low	200	400
High	1000	400

(Given for degree of freedom 1, $\chi^2_{0.05} = 3.84$) *[G.B.T.U. (MBA) 2011]*

6. Examine by any suitable method, whether the nature of area is related to voting preference in the election for which the data are tabulated below:

Votes for Area	*A*	*B*	*Total*
Rural	620	480	1100
Urban	380	520	900
Total	1000	1000	2000

7. The groups of 100 people each were taken for testing the use of a vaccine. 15 persons contracted the disease out of the inoculated persons, whicle 25 contracted the diesease in the other group. Test the efficiency of the vaccine using Chi-square test. (The value of χ^2 for one degree of freedom at 5% level of significance is 3.84).

8. By using χ^2-test, find out whether there is any association between income level and type of schooling:

Social Status Health	*Poor*	*Rich*	*Total*
Below Normal	130	20	150
Normal	102	108	210
Above Normal	24	96	120
Total	256	224	480

Answers

1. Yes	**2.** No	**3.** No	**4.** Not effective
5. Yes	**6.** Yes	**7.** Not, associated	**8.** Yes, associated

5.24 ANALYSIS OF VARIANCE (ANOVA) *[G.B.T.U. (MCA) 2011]*

ANOVA is a statistical technique which can be used to make comparisons among more than two groups. For example, suppose we want to see the effects of three different fertilisers on the yield of a crop. We select 100 identical plots of which 25 are control plots and 75 are experimental plots. Plants are grown in all the 100 plots in the same manner except that no fertiliser is applied in the control plots, while the three fertilisers are applied in the 75 experimental plots. The 75 experimental plots are divided into 3 groups of 25 plots each. Each of these groups is treated with different kinds of fertiliser. When the crop is harvested, the yield of each individual plot is recorded in four groups. We calculate the mean yield of all the 100 plots. Then we apply the technique (analysis of variance) to find out if there is any significant difference between the overall mean yield and the mean yields of the individual groups.

Similarly, suppose we want to test the effectiveness of various doses of a drug on blood pressure. We take a large number of individuals and divide them into several groups depending on the number of doses. For example, if we want to test four different doses, we make five groups (one control group and four experimental groups). Each experimental group is meant for a given dose of the drug.

One way ANOVA is used to see the effects of an independent variable on a given dependent variable. Two way ANOVA is used to see the effects of two independent variables on a given dependent variable.

In ANOVA, we compare the variations existing in the observed values of the dependent variable in the classified groups. We calculate the variance of the total population, variance between the groups and the variance within the groups. Then we examine whether there is any significant difference between the calculated variances. If we find, by ANOVA, that the variance between the groups is not significantly different from the variance within the groups, we conclude that there is no significant difference between the means of the groups, and hence the independent variable (factor) has no effect on the dependent variable.

5.24.1 ANOVA Table

The technique of analysis of variance is referred to as ANOVA. A table showing the source of variation, the sum of squares, degrees of freedom, mean square (variance) and the formula for the F-ratio is known as ANOVA table.

5.24.2 Computation of Test Statistic

The actual analysis of variance is carried out on the basis of ratio between the variances. This ratio forms the test statistic known as *F-statistic*, *given by*

$$\text{F-statistic} = \frac{\text{Variance between the samples}}{\text{Variance within the samples}}$$

5.24.3 Assumptions

Following are the assumptions for the study of analysis of variance:

(*i*) Each of the samples is a simple random sample.

(*ii*) Population, from which the samples are selected, are normally distributed.

(*iii*) Each of the samples is independent of the other samples.

(*iv*) Each of the population has the same variations and identical means.

(*v*) The effects of various components are additive.

5.25 ONE WAY CLASSIFICATION

(*a*) **Null hypothesis H_0:** $\mu_1 = \mu_2 = ... = \mu_p$, *i.e.*, the means of the populations from which p samples are drawn are equal to one another.

(*b*) **Alternative Hypothesis H_i:** At least two of the means of the populations are unequal, or *all the* μ_i*'s are not equal.*

(*c*) **Computation of test statistic:** To obtain the test statistic known as F-value, we define certain terms as follows:

Sample observation x_{ij} : A sample observation x_{ij} has two subscripts. The subscript i denotes the row or *sample observation*, while the subscript j denotes the column or *population from which the observation came*. Thus x_{ij} is the i^{th} observation from the j^{th} population.

The total observations in the above matrix A is $m \times n$.

The various sums of squares involved in the computation of F-statistics are:

(*i*) **Sum of the Squares of Variations Amongst the Columns (SSC):** It is the sum of the squares of deviation between the column or group means and the grand mean, *i.e.*,

$$SSC = r\,\Sigma(\bar{x}_j - \bar{x})^2 ,$$

where $\bar{x}_j$ = mean of the j^{th} sample, $\bar{x}$ = mean of the sample (column) means, r = number of rows or size of each sample.

Variance amongst columns = $\boxed{MSC = \dfrac{SSC}{c-1}}$, where c is number of columns.

It indicates the degree of explained variance due to sampling variations.

(*ii*) **Sum of the Squares of Variations within Columns (SSE):** It is the sum of the squares of variations between individual items and the column means, *i.e.*,

$$SSE = \sum_i \sum_j (x_{ij} - \bar{x}_j)^2$$

where $x_{ij} = i^{th}$ observation in j^{th} column, $\bar{x}_j$ = mean of j^{th} column

Mean of the square of Column Errors = $\boxed{MSE = \dfrac{SSE}{c(r-1)}}$

where c is number of columns and r is number of rows.

This is called *Unexplained Variance* as it indicates only the chance variation which cannot be explained in terms of variation in population.

(*iii*) **Total Sum of Squares of Variation (SST).**

$$SST = \sum_j \sum_i x_{ij}^2 - C,$$

where C = Correction factor = $\dfrac{T^2}{rc}$, T = grand total of the values in all samples, r = number of rows, c = number of columns.

It is the sum of squares of observations between the individual values and the grand mean $\bar{x}$.

Also, **SST = SSC + SSE.**

$$\text{Total Variance} = \frac{\text{SST}}{n-1},$$

where $n = r \times c$ = total number of observations in all the samples and $(n-1)$ is the degree of freedom.

Test Statistic: The test statistic is the **F-value** or **F-statistic** and is given by

$$\textbf{F-statistic or F-value} = \frac{\textbf{Explained variance}}{\textbf{Unexplained variance}} = \frac{\textbf{MSC}}{\textbf{MSE}}.$$

$$\Rightarrow \quad \mathbf{F} = \frac{\mathbf{MSC}}{\mathbf{MSE}} \text{ with } (c-1) \text{ and } c\,(r-1) \text{ degrees of freedom.}$$

Thus we have

ANOVA Table for one-way classification for equal sample

Source of variation	*Sum of squares*	*Degree of freedom*	*Mean square*	*F*
Between samples (Column Means)	SSC	$c-1$	$\text{MSC} = \frac{\text{SSC}}{c-1}$	
Within Samples	SSE	$c\,(r-1)$	$\text{MSE} = \frac{\text{SSE}}{c(r-1)}$	$\text{F} = \frac{\text{MSC}}{\text{MSE}}$
Total	SST	$cr-1$		

(*d*) **Level of significance:** It is denoted by α and the value is taken generally as $\alpha = 0.05$.

(*e*) **Conclusion:** If $F_{computed} < F_{tabular}$ at α level of significance then the null hypothesis H_0 is accepted and if $F_{computed} > F_{tabular}$ at α level of significance then the null hypothesis H_0 is rejected and H_1 is accepted.

5.26 PROCEDURE TO PREPARE ANOVA TABLE

The basic structure of ANOVA table is

Source of variation	*Sum of squares*	*Degree of freedom*	*Mean sum of squares*	*Variance ratio or F-statistic*
Between samples				
Within samples				
Total				

(*i*) Find the sum of the values of all the items of all the samples. We call it **Grand Total (G.T.)**

$$\text{G.T.} = \Sigma A + \Sigma B + \Sigma C + \ldots\ldots$$

(*ii*) Calculate the correction factor (C.F.)

$$\text{C.F.} = \frac{(\text{G. T.})^2}{n}$$

where n is the total number of items in all the samples.

(*iii*) Find the total sum of squares. It is obtained as $(\Sigma A^2 + \Sigma B^2 + \Sigma C^2 + ...) - \text{C.F.}$

(*iv*) Find the sum of squares between samples.

It is obtained as

$$\left[\frac{(\Sigma A)^2}{n_1} + \frac{(\Sigma B)^2}{n_2} + \frac{(\Sigma C)^2}{n_3} + ...\right] - \text{C.F.}$$

(*v*) Calculate the sum of squares within samples.

Sum of squares within samples = (Total sum of squares) – (Sum of squares between samples)

(*vi*) Find total degree of freedom and degree of freedom between samples. Degree of freedom within samples is obtained by subtraction.

(*vii*) Mean sum of squares is obtained by dividing sum of squares by degree of freedom.

(*viii*) Variance ratio is obtained by dividing the corresponding mean sum of squares.

This completes the preparation of ANOVA table.

5.27 CHANGE OF ORIGIN METHOD (Coding Method)

F remains unchanged if all the figures are multiplied or divided by a common factor or if a common factor is added to or subtracted from each figure. This property helps in calculating F-ratio when the figures are large. When such simplifications are done, it is said that data have been coded to simplify calculations.

ILLUSTRATIVE EXAMPLES

Example 1. *It is desired to compare three hospitals with regards to the number of deaths per month. A sample of death records were selected from the records of each hospital and the number of deaths was as given below. From these data, suggest a difference in the no. of the deaths per months among three hospitals:*

Hospitals

A	*B*	*C*
3	*6*	*7*
4	*3*	*3*
3	*3*	*4*
5	*4*	*6*
0	*4*	*5*

Sol. Null hypothesis H_0: There is no difference in the no. of deaths per months among three hospitals.

Alternate hypothesis H_1: There is a significant difference in the no. of deaths per months among three hospitals.

Level of significance: We use 5% level of significance.

Test statistic: To find the variance ratio, F, we set up an ANOVA table as follows:
Sample totals:

$$\Sigma y_A = 3 + 4 + 3 + 5 + 0 = 15$$
$$\Sigma y_B = 6 + 3 + 3 + 4 + 4 = 20$$
$$\Sigma y_C = 7 + 3 + 4 + 6 + 5 = 25$$
$$\text{Grand total (G.T.)} = \Sigma y_A + \Sigma y_B + \Sigma y_C = 60$$

$$\text{Correction factor (C.F.)} = \frac{(\text{G.T})^2}{n} = \frac{(60)^2}{15} = 240$$

Sum of squares of samples:

$$\Sigma y_A^2 = 3^2 + 4^2 + 3^2 + 5^2 + 0^2 = 59$$
$$\Sigma y_B^2 = 6^2 + 3^2 + 3^2 + 4^2 + 4^2 = 86$$
$$\Sigma y_C^2 = 7^2 + 3^2 + 4^2 + 6^2 + 5^2 = 135$$

Total sum of squares $= \Sigma y_A^2 + \Sigma y_B^2 + \Sigma y_C^2 - \text{C.F.}$
$= 59 + 86 + 135 - 240 = 40$

Sum of squares between samples

$$= \frac{(\Sigma y_A)^2}{n_1} + \frac{(\Sigma y_B)^2}{n_2} + \frac{(\Sigma y_C)^2}{n_3} - \text{C.F.}$$

$$= \frac{(15)^2}{5} + \frac{(20)^2}{5} + \frac{(25)^2}{5} - 240 = 10$$

Sum of squares within samples
= Total sum of squares – Sum of squares between samples
$= 40 - 10 = 30$

Degrees of freedom for total sum of squares $= n - 1 = 15 - 1 = 14$
Degrees of freedom for Hospitals $= k - 1 = 3 - 1 = 2$
Degrees of freedom for error $= n - k = 15 - 3 = 12$

ANOVA TABLE

Source of variation	*Sum of squares*	*Degree of freedom*	*Mean sum of squares*	*Variance ratio or F*
Between samples	10	2	5	$F_{2, 12} = \frac{5}{2.5} = 2$
Within samples	30	12	2.5	
Total	40	14	—	—

The tabular value of F at 5% level of significance with $\nu_1 = 2$, $\nu_2 = 12$ is 3.89

Conclusion: Since $F_{cal.} < F_{tab.}$, the difference is insignificant and we conclude that data donot suggest a difference in the number of deaths per month among the three hospitals.

Example 2. *A manufacturing company purchased three new machines of different makes and wishes to determine whether one of them is faster than the others in producing a certain output. Five hourly production figures are observed at random from each machine and results are given below:*

Observations	A_1	A_2	A_3
1	25	31	24
2	30	39	30
3	36	38	28
4	38	42	25
5	31	35	28

Use ANOVA and determine whether the machines are significantly different in their mean speed. (Given: at 5% level, $F_{2,12} = 3.89$) **Sol. Null hypothesis H_0:** Machines are not significantly different in their mean speeds *i.e.*, $\mu_1 = \mu_2 = \mu_3$

Alternate hypothesis H_1: Machines are significantly different in their mean speed.

Level of significance: We use 5% level of significance.

Test statistic: To find the variance ratio F, we set up an ANOVA table as follows:

let us shift the origin at 30 *i.e.*, reduce each observation by 30.

Now,

Observations	A_1	A_2	A_3
1	– 5	1	– 6
2	0	9	0
3	6	8	– 2
4	8	12	– 5
5	1	5	– 2

Machine totals:

$$\Sigma A_1 = -5 + 0 + 6 + 8 + 1 = 10$$

$$\Sigma A_2 = 1 + 9 + 8 + 12 + 5 = 35$$

$$\Sigma A_3 = -6 + 0 - 2 - 5 - 2 = -15$$

$$\text{Grand total (G.T.)} = \Sigma A_1 + \Sigma A_2 + \Sigma A_3 = 30$$

$$\text{Correction factor (C.F)} = \frac{(30)^2}{15} = 60$$

Sum of squares of samples

$$\Sigma A_1^2 = (-5)^2 + 0^2 + 6^2 + 8^2 + 1^2 = 126$$

Similarly, $\Sigma A_2^2 = 315$, $\Sigma A_3^2 = 69$

$$\text{Total sum of squares} = 126 + 315 + 69 - (\text{C.F.}) = 510 - 60 = 450$$

$$\text{Machine sum of squares (between samples)} = \frac{(10)^2}{5} + \frac{(35)^2}{5} + \frac{(-15)^2}{5} - \text{C.F.}$$

$$= 310 - 60 = 250$$

Error sum of squares (within samples)

= Total sum of squares – Machine sum of squares

= 450 – 250 = 200

Degrees of freedom for total sum of squares = $n - 1 = 15 - 1 = 14$

Degrees of freedom for Machines = $k - 1 = 3 - 1 = 2$

Degrees of freedom for Error = $n - k = 15 - 3 = 12$

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean sum of squares*	*Variance ratio of F*
Machine (between samples)	250	2	$\frac{250}{2} = 125$	$F_{2, 12} = \frac{125}{16.67}$
Error (within samples)	200	12	$\frac{200}{12} = 16.67$	$= 7.498$
Total	450	14	—	—

The tabular value of F at 5% level of significance with $\nu_1 = 2$, $\nu_2 = 12$ is 3.89 (given).

Conclusion: Since $F_{calculated} > F_{tabulated}$, the null hypothesis is rejected and the difference is significant and we conclude that there is a significant difference in the mean speed of machines.

TEST YOUR KNOWLEDGE

1. To test the significance of the variations of the retail prices in the commodity in three principal cities : Mumbai, Bangalore and Chennai, the four shops were chosen at random in each city and prices observed in rupees were as follows:

Mumbai	16	8	12	14
Bangalore	14	10	10	6
Chennai	4	10	8	8

Do the data indicate that the prices in the three cities are significantly different?

2. Below are given the yield kg for four varieties of tablets. Prepare ANOVA table and test that varieties differ significantly.

A	*B*	*C*	*D*
20	25	24	23
19	23	20	20
21	21	22	20

[**Hint.** Shift the origin at 20]

3. The following table gives the yields on 15 sample plots under three varieties of seeds:

Variety I:	20	21	23	16	20
Variety II:	18	20	17	15	25
Variety III:	25	28	22	28	32

Show that the seed varieties show variations more than could be covered by sampling variations. Given the tabulated value of F for $v_1 = 2$ and $v_2 = 12$ at 5% level of significance is 3.88.

4. The following figures relate to the production in kg of three varieties of wheat A, B and C used on 15 plots.

Wheat Variety	*Yields (kg)*					
RR 21–A	14	17	16	16		
K.68–B	15	11	13	15	13	14
Sonalika–C	18	16	18	19	15	

Test whether there is any significant difference in the production of three varieties.

[**Hint.** Shift the origin at 16]

5. The following figures relate to the production in kg of three varieties I, II, III of wheat shown in 12 plots:

Variety I:	14	16	18		
Variety II:	14	13	15	22	
Variety III:	18	16	19	19	20

Is there any significant difference in the production of three varieties? Given the tabulated value of F for $\nu_1 = 2$ and $\nu_2 = 9$ at 5% level of significance is 4.26.

6. A trucking company wishes to test the average life of each of the four brands of tyres. The company uses all brands on randomly selected trucks. The records showing the lives (thousands of miles) of tyres are as given in the table given below:

Brand I:	20	23	18	17	
Brand II:	19	15	17	20	16
Brand III:	21	19	20	17	16
Brand IV:	15	17	16	18	

Test the hypothesis that the average life for each brand of tyres is the same. Assume $\alpha = 0.01$.

7. What is analysis of variance? Mention its importance and uses.

8. Write short notes on:

(*i*) ANOVA [*G.B.T.U. 2011*]

(*ii*) One-way classification.

9. Under what conditions do we require to use one way Analysis of Variance ? Give examples and explain the method involved by taking a hypothetical example.

Answers

1. No **2.** Varieties donot differ significantly.

4. The difference is significant. **5.** No

6. No significant difference between the average lives.

5.28 STATISTICAL QUALITY CONTROL [*G.B.T.U. (MCA) 2010*]

A quality control system performs inspection, testing and analysis to ensure that the quality of the products produced is as per the laid down quality standards. It is called

"Statistical Quality Control" when statistical techniques are employed to control, improve and maintain quality or to solve quality problems. Building an information system to satisfy the concept of prevention and control and improving upon product quality requires statistical thinking.

Statistical quality control (S.Q.C.) is systematic as compared to guess-work of haphazard process inspection and the mathematical statistical approach neutralizes personal bias and uncovers poor judgement. S.Q.C. consists of three general activities:

(1) Systematic collection and graphic recording of accurate data

(2) Analyzing the data

(3) Practical engineering or management action if the information obtained indicates significant deviations from the specified limits.

Modern techniques of statistical quality control and acceptance sampling have an important part to play in the improvement of quality, enhancement of productivity, creation of consumer confidence and development of industrial economy of the country.

Following statistical tools are generally used for the above purposes :

(*i*) **Frequency distribution:** Frequency distribution is a tabulation of the number of times a given quality characteristic occurs within the samples. Graphic representation of frequency distribution will show :

(*a*) Average quality (*b*) Spread of quality

(*c*) Comparison with specific requirements (*d*) Process capability.

(*ii*) **Control chart:** Control chart is a graphical representation of quality characteristics, which indicates whether the process is under control or not.

(*iii*) **Acceptance sampling:** Acceptance sampling is the process of evaluating a portion of the product/material in a lot for the purpose of accepting or rejecting the lot on the basis of conforming to a quality specification.

It reduces the time and cost of inspection and exerts more effective pressure on quality improvement than it is possible by 100% inspection.

It is used when assurance is desired for the quality of materials/products either produced or received.

(*iv*) **Analysis of data:** It includes analysis of tolerances, correlation, analysis of variance, analysis for engineering design, problem solving technique to eliminate cause to troubles. Statistical methods can be used in arriving at proper specification limits of product, in designing the product, in purchase of raw-material, semi-finished and finished products, manufacturing processes, inspection, packaging, sales and also after sales service.

5.29 ADVANTAGES OF STATISTICAL QUALITY CONTROL

1. Efficiency: The use of statistical quality control ensures rapid and efficient inspection at a minimum cost. It eliminates the need of 100% inspection of finished products because the acceptance sampling in S.Q.C. exerts more effective pressure for quality improvement than it.

2. Reduction of scrap: It uncovers the cause of excessive variability in manufactured products forecasting trouble before rejections occur and reducing the amount of spoiled work.

3. Easy detection of faults: In statistical quality control after plotting the control, charts ($\bar{X}$, R, P, C, U) etc. when the points fall above the upper control limits or below the lower control limit, it gives an indication of deterioration in quality. Necessary corrective action may be then taken immediately.

4. Adherence to specifications: So long as a statistical control continues, specifications can be accurately predicted for future by which it is possible to assess whether the production processes are capable of producing the products with the given set of specifications.

5. Increases output and reduces wasted machine and man hours.

6. Efficient utilization of personnel, machines and materials results in higher productivity.

7. Creates quality awareness in employees. However, it should be noted that statistical quality control is not a panacea for assuring product quality.

8. Provides a common language that may be used by all three groups designers, production personnel and inspectors in arriving at a rational solution of mutual problems.

9. Points out when and where 100% inspection, sorting or screening is required.

10. Elimination of bottlenecks in the process of manufacturing.

It simply furnishes 'perspective facts' upon which intelligent management and engineering action can be based. Without such action, the method is ineffective.

Even the application of standard procedures is also very dangerous without adequate study of the process.

5.30 REASONS FOR VARIATIONS IN THE QUALITY OF A PRODUCT

Two extremely similar things are rarely obtained in nature. This fact holds good for production processes as well. No production process is good enough to produce all items of products exactly alike. The variations are due to two main reasons:

(*i*) **Chance or random causes:** Variations due to chance causes are inevitable in any process or product. They are difficult to trace and to control also even under best conditions of production.

These variations may be due to some inherent characteristic of the process or machine which functions at random.

If the variations are due to chance factors alone, the observations will follow a "normal curve". The knowledge of the behaviour of chance variation is the foundation on which control chart analysis rests. The conditions which produce these variations are accordingly said to be **"under control"**. On the other hand, if the variations in the data do not conform to a pattern that might reasonably be produced by chance causes, then in this case, conditions producing the variations are said to be **"out of control"** as it may be concluded that one or more assignable causes are at work.

(*ii*) **Assignable causes:** The variations due to assignable causes possess greater magnitude as compared to those due to chance causes and can be easily traced or detected. The power of the shewhart control chart lies in its ability to separate out these assignable causes of quality variations *e.g.* in length thickness, weight or diameter of a component.

The variations due to assignable causes may be because of following factors :

(*i*) Differences among machines

(*ii*) Differences among workers

(*iii*) Differences among materials

(*iv*) Differences in each of these factors over time

(*v*) Differences in their relationship to one another.

These variations may also be caused due to change in working conditions, mistake on the part of operator etc.

5.31 TECHNIQUES OF STATISTICAL QUALITY CONTROL

To control the quality characteristics of the product, there are two main techniques :

1. Process Control: It is a process of monitoring and measuring variability in the performance of a process or a machine through the interpretation of statistical techniques and it is employed to manage in-process quality. This technique ensures the production of requisite standard product and makes use of control charts.

2. Product control: This technique is concerned with inspection of already produced goods to ascertain whether they are fit to be despatched or not. To achieve the objectives, it makes use of sampling inspection plans.

5.32 CONTROL CHART

A control chart is a graphical representation of the collected information. It detects the variation in processing and warns if there is any departure from the specified tolerance limits. In other words, control charts is a device which specifies the state of statistical control or is a device for attaining quality control or is a device to judge whether the statistical control has been attained.

The control limits on the chart are so placed as to disclose the presence or absence of the assignable causes of quality variation which makes the diagnosis possible and brings substantial improvements in product quality and reduction of spoilage and rework.

Moreover, by identifying chance variations, the control chart tells when to leave the process alone and thus prevents unnecessarily frequent adjustments that tend to increase the variability of the process rather than to decrease it.

There are many types of control charts designed for different control situations. Most commonly used control charts are

(*i*) **Control charts for variables:** They are useful to measure quality characteristics and to control fully automatic process. It includes $\bar{X}$ and R-charts and charts for $\bar{X}$ and σ.

(*ii*) **Control charts for attributes:** It includes P-chart for fraction defective. A fraction defective control chart discloses erratic fluctuations in the quality of inspection which may result in improvement in inspection practice and inspection standards.

It also includes C-chart for number of defects per unit.

5.33 OBJECTIVES OF CONTROL CHARTS

[G.B.T.U. (C.O.) 2010)]

Control charts are based on statistical techniques.

1. $\bar{X}$ and R or $\bar{X}$ and σ charts are used in combination for control process. $\bar{X}$-chart shows the variation in the averages of samples. It is the most commonly used variables chart. R-chart shows the uniformity or consistency of the process *i.e.*, it shows the variations in the ranges of samples. It is a chart for measure of spread. σ-chart shows the variation of process.

2. To determine whether a given process can meet the existing specifications without a fundamental change in the production line or to tell whether the process is in control and if so, at what dispersion.

3. To secure information to be used in establishing or changing production procedures.

4. To secure information when it is necessary to widen the tolerances.

5. To provide a basis for current decisions or acceptance or rejection of manufactured or purchased product.

6. To secure information to be used in establishing or changing inspection procedure or acceptance procedure or both.

5.34 CONSTRUCTION OF CONTROL CHARTS FOR VARIABLES

First of all, a random sample of size n is taken during a manufacturing process over a period of time and quality measurements $x_1, x_2, \ldots\ldots, x_n$ are noted

Sample mean $$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample range $$R = x_{max.} - x_{min.}$$

If the process is found stable, k consecutive samples are selected and for each sample, $\bar{x}$ and R are calculated. Then we find $\bar{\bar{x}}$ and $\bar{R}$ as

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \ldots\ldots + \bar{x}_k}{k} = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i$$

and $$\bar{R} = \frac{R_1 + R_2 + \ldots\ldots + R_k}{k} = \frac{1}{k}\sum_{i=1}^{k} R_i$$

For $\bar{X}$-chart

$$\text{Central line} = \left.\begin{array}{l} \bar{\bar{x}}, \text{ when tolerance limits are not given} \\ \mu, \text{ when tolerance limits are given} \end{array}\right\}$$

where $$\mu = \frac{1}{2}[\text{LCL} + \text{UCL}]$$

LCL is lower control limit and UCL is upper control limit

Now, LCL (for $\bar{X}$-chart) $= \bar{\bar{x}} - A_2\bar{R}$ and UCL (for $\bar{X}$-chart) $= \bar{\bar{x}} + A_2\bar{R}$ are set.

A_2 depends on sample size n and can be found from the following table :

Sample size (n)	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A_2	1.88	1.02	0.73	0.58	0.48	0.42	0.37	0.34	0.31	0.29	0.27	0.25	0.24	0.22	0.21	0.20	0.19	0.19	0.18

For R-chart Central line (CL) $= \bar{R}$

Now, LCL (for R-chart) $= D_3\bar{R}$ UCL (for R-chart) $= D_4\bar{R}$ are set.

where D_3 and D_4 depend on sample size and are found from the following table :

Sample size (n)	D_3	D_4	d_2
2	0	3.27	1.13
3	0	2.57	1.69
4	0	2.28	2.06
5	0	2.11	2.33
6	0	2.00	2.53
7	0.08	1.92	2.70
8	0.14	1.86	2.85
9	0.18	1.82	2.97
10	0.22	1.78	3.08
11	0.26	1.74	3.17
12	0.28	1.72	3.26
13	0.31	1.69	3.34
14	0.33	1.67	3.41
15	0.35	1.65	3.47
16	0.36	1.64	3.53
17	0.38	1.62	3.59
18	0.39	1.61	3.64
19	0.40	1.60	3.69
20	0.41	1.59	3.74

To compute upper and lower process tolerance limits for the values of x, we have

$$\text{LTL} = \bar{\bar{x}} - \frac{3\overline{\text{R}}}{d_2} \qquad \text{UTL} = \bar{\bar{x}} + \frac{3\overline{\text{R}}}{d_2}$$

where d_2 is found from the above table.

Moreover, the process capability is given by $6\sigma = 6\,\dfrac{\overline{\text{R}}}{d_2}$ where σ is standard deviation.

While plotting the $\overline{\text{X}}$-chart the central line on the $\overline{\text{X}}$ chart should be drawn as a solid horizontal line at $\overline{\text{X}}$. The upper and lower control limits for $\overline{\text{X}}$ chart should be drawn as dotted horizontal lines at the computed values.

Similarly, for R-chart, the central line should be drawn as a solid horizontal line at $\overline{\text{R}}$. The upper control limit should be drawn as dotted horizontal line at the computed value of UCL_R. If the subgroup size is 7 or more, the lower control limit should be drawn as dotted horizontal line at LCL_R. However, if the subgroup size is ≤ 6, the lower control limit for R is zero.

Plot the averages of subgroups in $\overline{\text{X}}$-chart, in the order collected and ranges in R-chart which should be below the $\overline{\text{X}}$-chart so that the subgroups correspond to one-another in both the charts. Points outside the control limits are indicated with cross (×) on $\overline{\text{X}}$-chart and the points outside the limits on R chart by a circle (⊙).

5.35 CONTROL CHARTS FOR ATTRIBUTES

Following control charts will be discussed here

(*i*) P chart (*ii*) *np* chart (*iii*) C chart

As an alternative to $\overline{X}$ and R chart and as a substitute when characteristic is measured only by attribute, a control chart based on fraction defective $\overline{p}$ is used, called P-chart.

$$\overline{p} = \frac{\text{No. of defective articles found in any inspection}}{\text{Total no. of articles actually inspected}}.$$

(*i*) **Control limits (3σ limits) on p-chart:** We know that for binomial distribution, the mean value of total number of defectives in a sample n is np and standard deviation is $\sqrt{npq}$ or $\sqrt{np(1-p)}$.

$\therefore$ Mean value of fraction defective is $\overline{p}$ and S.D.

$$\sigma_p = \frac{1}{n}\sqrt{n\overline{p}(1-\overline{p})} = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

$\therefore$ $\quad CL = \overline{p}$

The upper and lower limits for P-chart are, $\quad UCL_P = \overline{p} + 3\sigma_p = \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$

and

$$LCL_P = \overline{p} - 3\sigma_p = \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}.$$

Due to the lower inspection and maintenance costs of P-charts, they usually have a greater area of economical applications.

(*ii*) **Control limits for *np* chart:** Whenever subgroup size is variable, P-chart is used but if it is constant, the chart for actual number of defectives called *np* chart is used.

$$CL = n\overline{p} \quad \text{where} \quad \overline{p} = \frac{\Sigma np}{\Sigma n}$$

$$UCL_{np} = n\overline{p} + 3\sigma_{np} = n\overline{p} + 3\sqrt{n\overline{p}(1-\overline{p})} \qquad (\text{where } \sigma_{np} = n\sigma_p)$$

and

$$LCL_{np} = n\overline{p} - 3\sqrt{n\overline{p}(1-\overline{p})}.$$

Note. In case of $\overline{X}$ and R chart, it may not be necessary to draw lines connecting the points which represent the successive subgroups. But incase of P-chart, a line connecting the points is usually helpful in interpretation of the chart. Such a line assists in the interpretation of trends.

(*iii*) **Control limits for C chart**

(*a*) **Difference between a defect and defective**

An item is called defective if it fails to conform to the specifications in any of the characteristics. Each characteristic that does not meet the specifications is a defect. An item is defective if it contains atleast one defect. The *np* chart applies to the number of defectives in subgroups of constant size while C chart applies to the number of defects in a subgroup of constant size.

(*b*) **Basis for control limits on C chart**

Control limits on C chart are based on **Poisson distribution:** Hence two conditions must be satisfied. The first condition specifies that the area of opportunity for occurrence of

defects should be fairly constant from period to period. Second condition specifies that opportunities for defects are large while the chances of a defect occurring in any one spot are small.

(*c*) **Calculation of control limits on C chart**

Standard deviation $\sigma_c = \sqrt{\overline{C}}$

Thus 3σ limits on a C chart are $UCL_c = \overline{C} + 3\sqrt{\overline{C}}$ and $LCL_c = \overline{C} - 3\sqrt{\overline{C}}$

and central line $CL = \overline{C}$

where $$\overline{C} = \frac{\text{Number of defects in all samples}}{\text{Total number of samples}}.$$

ILLUSTRATIVE EXAMPLES

Example 1. *The following are the mean lengths and ranges of lengths of a finished product from 10 samples each of size 5. The specification limits for length are 200 ± 5 cm. Construct $\overline{X}$ and R-chart and examine whether the process is under control and state your recommendations.*

Sample No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Mean ($\overline{X}$)	*201*	*198*	*202*	*200*	*203*	*204*	*199*	*196*	*199*	*201*
Range (R)	*5*	*0*	*7*	*3*	*3*	*7*	*2*	*8*	*5*	*6*

Assume for $n = 5$, $A_2 = 0.58$, $D_4 = 2.11$ *and* $D_3 = 0$.

Sol. (*i*) **Control limits for $\overline{X}$ chart:**

Central limit $CL = 200$ | $\because$ Tolerance / specification limits are given, $\therefore$ $\mu = 200$

$$UCL_{\bar{x}} = \overline{\bar{x}} + A_2\overline{R} = \mu + A_2\overline{R}$$

$$LCL_{\bar{x}} = \overline{\bar{x}} - A_2\overline{R} = \mu - A_2\overline{R}$$

where $$\overline{R} = \frac{R_1 + R_2 + \ldots\ldots + R_{10}}{10} = \frac{46}{10} = 4.6$$

Then, $UCL_{\overline{X}} = 200 + (0.58 \times 4.6) = 202.668$

$$LCL_{\overline{X}} = 200 - (0.58 \times 4.6) = 197.332.$$

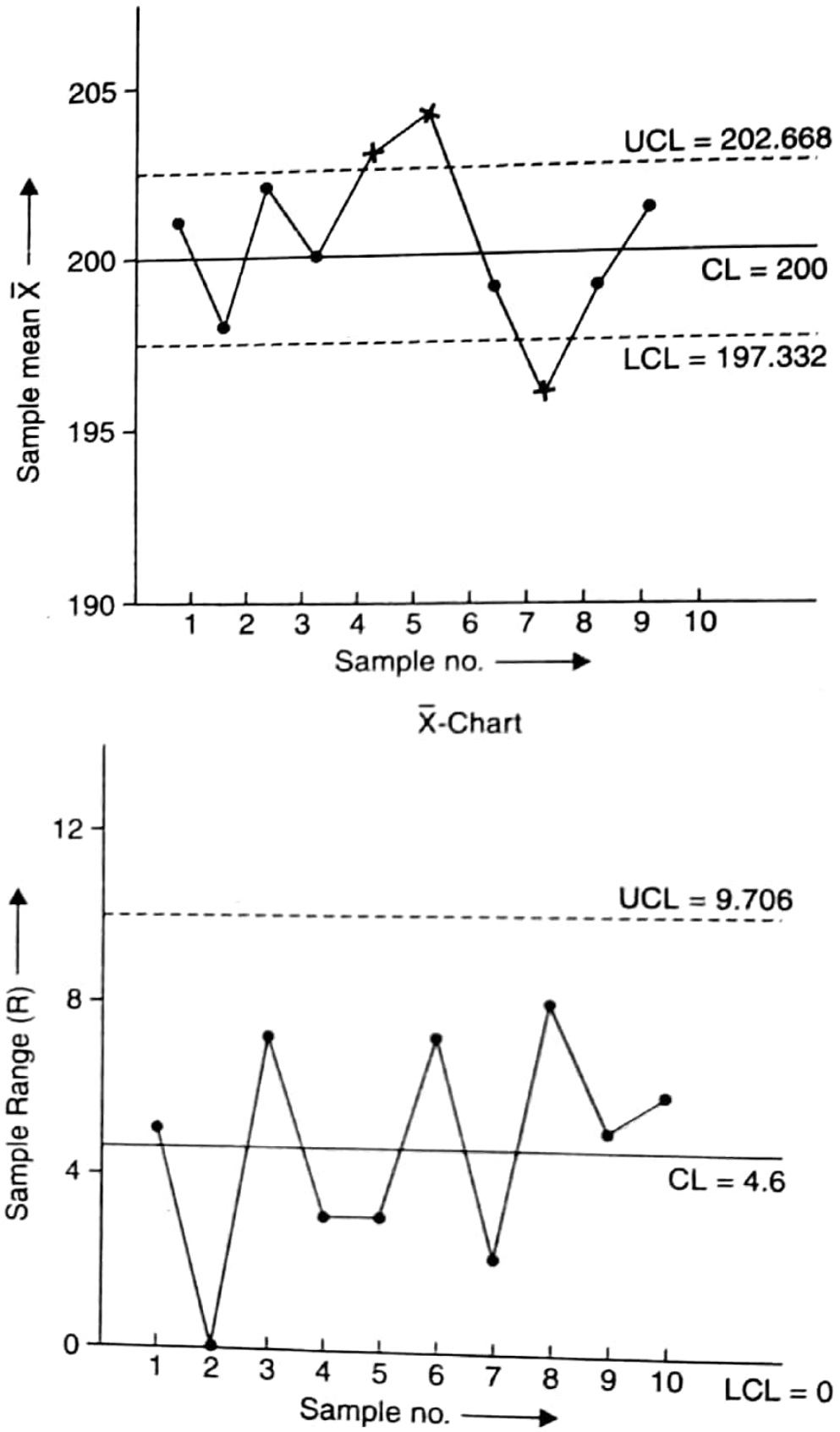
(*ii*) **Control limits for R chart:**

Control limit $CL = \overline{R} = 4.6$

$$UCL_R = D_4\overline{R} = 2.11 \times 4.6 = 9.706$$

$$LCL_R = D_3\overline{R} = 0 \times 4.6 = 0$$

The $\overline{X}$ and R-charts are drawn below:



It is noted that all points lie within the control limits on R chart. Hence the process variability is under control. But in X-chart, points corresponding to sample number 5, 6 and 8 lie outside the control limits. Therefore the **process is not in statistical control.** The process should be halted and it is recommended to check for any assignable causes. Fluctuation will remain until these causes, if found, are removed.

Example 2. *A drilling machine bores holes with a mean diameter of 0.5230 cm and a standard deviation of 0.0032 cm. Calculate the 2-sigma and 3-sigma upper and lower control limits for means of sample of 4.*

[G.B.T.U. (C.O) 2010]

Sol. Mean diameter $\bar{\bar{x}} = 0.5230$ cm

S.D. $\sigma = 0.0032$ cm

$n = 4$

(*i*) 2-sigma limits are as follows:

$$\text{CL} = \bar{\bar{x}} = 0.5230 \text{ cm}$$

$$\text{UCL} = \bar{\bar{x}} + 2\frac{\sigma}{\sqrt{n}} = 0.5230 + 2 \times \frac{0.0032}{\sqrt{4}} = 0.5262 \text{ cm}$$

$$\text{LCL} = \bar{\bar{x}} - 2\frac{\sigma}{\sqrt{n}} = 0.5230 - 2 \times \frac{0.0032}{\sqrt{4}} = 0.5198 \text{ cm}.$$

(*ii*) 3-sigma limits are as follows:

$$\text{CL} = \bar{\bar{x}} = 0.5230 \text{ cm}$$

$$\text{UCL} = \bar{\bar{x}} + 3\frac{\sigma}{\sqrt{n}} = 0.5230 + 3 \times \frac{0.0032}{\sqrt{4}} = 0.5278 \text{ cm}$$

$$\text{LCL} = \bar{\bar{x}} - 3\frac{\sigma}{\sqrt{n}} = 0.5230 - 3 \times \frac{0.0032}{\sqrt{4}} = 0.5182 \text{ cm}.$$

Example 3. *In a blade manufacturing factory, 1000 blades are examined daily. Draw the np chart for the following table and examine whether the process is under control?*

Date	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of defective blades	:	9	10	12	8	7	15	10	12	10	8	7	13	14	15	16

[G.B.T.U. (C.O) 2010]

Sol. Here, $n = 1000$

$\sum np$ = total number of defectives = 166

$\sum n$ = total number inspected = 1000×15

$$\therefore \quad \bar{p} = \frac{\sum np}{\sum n} = \frac{166}{1000 \times 15} = 0.011$$

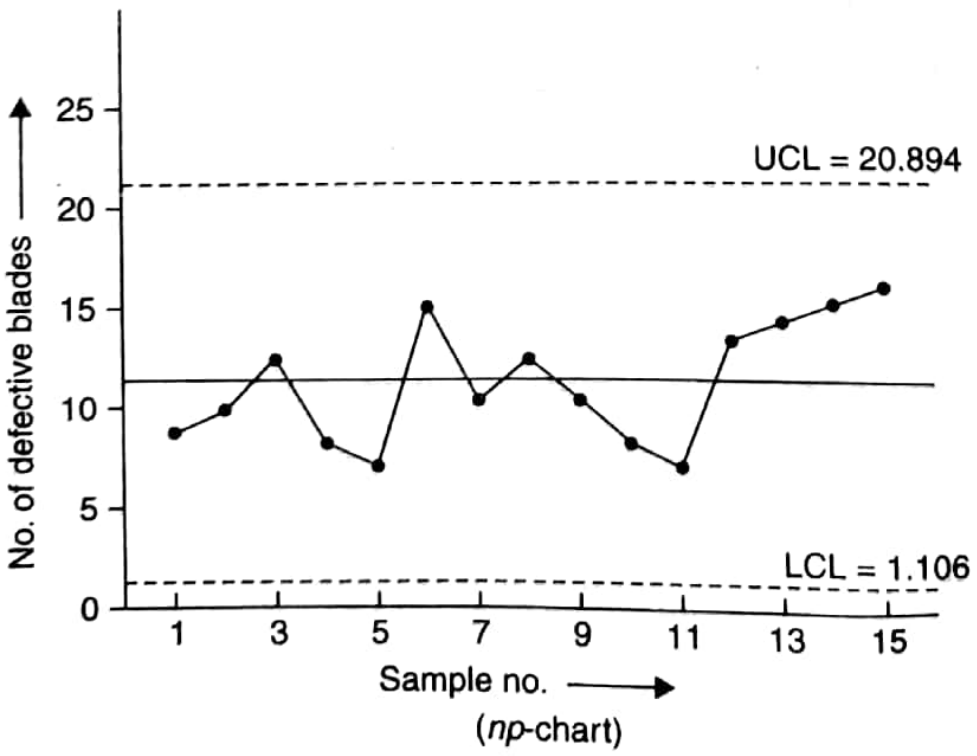
$$\therefore \quad n\bar{p} = 1000 \times 0.011 = 11$$

Control limits are $\text{CL} = n\bar{p} = 11$

$$\text{UCL}_{np} = n\bar{p} + 3\sqrt{n\bar{p}\,(1-\bar{p})} = 11 + 3\sqrt{11(1-0.011)} = 20.894$$

$$\text{LCL}_{np} = n\bar{p} - 3\sqrt{n\bar{p}\,(1-\bar{p})} = 11 - 3\sqrt{11(1-0.011)} = 1.106$$

The *np* chart is drawn in the figure. Since all the points lie within the control limits, the process is under control.



Example 4. *In a manufacturing process, the number of defectives found in the inspection of 20 lots of 100 samples is given below:*

Lot no.	No. of defectives	Lot no.	No. of defectives
1	5	11	7
2	4	12	6
3	3	13	3
4	5	14	5
5	4	15	4
6	6	16	2
7	9	17	8
8	15	18	7
9	11	19	6
10	6	20	4

(i) Determine the control limits of p-chart and state whether the process is in control.

(ii) Determine the new value of mean fraction defective if some points are out of control. Compute the corresponding control limits and state whether the process is still in control or not.

(iii) Determine the sample size when a quality limit not worse than 9% is desirable and a 10% bad product will not be permitted more than three times in thousand.

Sol. (*i*)
$$\bar{p} = \frac{\text{Total number of defectives}}{\text{Total number of items inspected}} = \frac{120}{20 \times 100} = 0.06$$

$$\text{UCL}_\text{P} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 + 3\sqrt{\frac{0.06\,(1-0.06)}{100}} = 0.13095$$

$$\text{LCL}_\text{P} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 - 3\sqrt{\frac{0.06\,(1-0.06)}{100}} = -0.01095$$

Since the fraction defective cannot be (–) ve

$\therefore$ $\text{LCL}_\text{P} = 0$

After observing the values of defectives in the given example, it is clear that only 8th lot having fraction defective $\frac{15}{100} = 0.15$ will go above UCL_P.

(*ii*) After eliminating the 8th lot,

$$\text{Revised value of } \bar{p} = \frac{120 - 15}{100 \times 19} = 0.056$$

Revised control limits will be

$$\text{UCL}_\text{P} = 0.056 + 3\sqrt{\frac{0.056\,(1-0.056)}{100}} = 0.125$$

$$\text{LCL}_\text{P} = 0.056 - 3\sqrt{\frac{0.056\,(1-0.056)}{100}} = -0.013 \textit{ i.e.,} \text{ zero.}$$

It is clear that all the points are within control limits.

$\therefore$ Revised quality level $\bar{p} = 0.056$

(*iii*) Since a probability that a defective worse than 9% defective quality will not be permitted is more than 3 times in thousand (0.3%) is corresponding 36 limits

$$\therefore \qquad \bar{p} + 3p = 0.09$$

$$0.056 + 3\sqrt{\frac{0.056\,(1-0.056)}{n}} = 0.09 \quad \Rightarrow \quad \sqrt{\frac{0.056 \times 0.944}{n}} = \frac{0.034}{3}$$

Squaring, $$\frac{0.056 \times 0.944}{n} = \left(\frac{0.034}{3}\right)^2 = (0.01133)^2$$

$$n = \frac{0.056 \times 0.944}{0.01133 \times 0.01133} = 333.$$

Example 5. *Determine the control limits for $\bar{X}$ and R charts if $\Sigma\bar{X} = 357.50$, $\Sigma R = 9.90$, number of subgroups = 20. It is given that $A_2 = 0.18$, $D_3 = 0.41$, $D_4 = 1.59$ and $d_2 = 3.736$. Also, find the process capability.*

Sol. $$\bar{\mathrm{X}} = \frac{\Sigma\bar{\mathrm{X}}}{\mathrm{N}} = \frac{357.50}{20} = 17.875, \quad \bar{\mathrm{R}} = \frac{\Sigma \mathrm{R}}{\mathrm{N}} = \frac{9.90}{20} = 0.495$$

$$\mathrm{UCL}_{\bar{\mathrm{X}}} = \bar{\bar{\mathrm{X}}} + \mathrm{A}_2\bar{\mathrm{R}} = 17.875 + (0.18 \times 0.495) = 17.9641$$

$$\mathrm{LCL}_{\bar{\mathrm{X}}} = \bar{\bar{\mathrm{X}}} - \mathrm{A}_2\bar{\mathrm{R}} = 17.875 - (0.18 \times 0.495) = 17.7859$$

$$\mathrm{UCL}_{\mathrm{R}} = \mathrm{D}_4\bar{\mathrm{R}} = 1.59 \times 0.495 = 0.78705$$

$$\mathrm{LCL}_{\mathrm{R}} = \mathrm{D}_3\bar{\mathrm{R}} = 0.41 \times 0.495 = 0.20295$$

$$\sigma = \frac{\bar{\mathrm{R}}}{d_2} = \frac{0.495}{3.735} = 0.13253$$

$\therefore$ Process capability = $6\sigma = 6 \times 0.13253 = 0.79518$.

Example 6. *If the average fraction defective of a large sample of a product is 0.1537, Calculate the control limits given that sub-group size is 2000.*

Sol. Average fraction defective

$$\bar{p} = 0.1537$$

Sub-group size is 2000

$$\therefore \qquad n = 2000$$

Central line $$\mathrm{CL} = n\bar{p} = 2000 \times 0.1537 = 307.4$$

$$\mathrm{UCL}_{np} = n\bar{p} + 3\sigma_{np} = n\bar{p} + 3\sqrt{n\bar{p}\,(1-\bar{p})}$$

$$= 307.4 + 3\sqrt{307.4\,(1-0.1537)}$$

$$= 355.787742$$

and $$\mathrm{LCL}_{np} = n\bar{p} - 3\sqrt{n\bar{p}\,(1-\bar{p})} = 307.4 - 48.38774204$$

$$= 259.012258$$

TEST YOUR KNOWLEDGE

1. A company manufactures screws to a nominal diameter 0.500 ± 0.030 cm. Five samples were taken randomly from the manufactured lots and 3 measurements were taken on each sample at different lengths. Following are the readings:

Sample no.	*Measurement per sample x(in cm)*		
	1	*2*	*3*
1	0.488	0.489	0.505
2	0.494	0.495	0.499
3	0.498	0.515	0.487
4	0.492	0.509	0.514
5	0.490	0.508	0.499

Calculate the control limits of $\overline{X}$ and R charts. Draw $\overline{X}$ and R charts and examine whether the process is in statistical control? [Take $A_2 = 1.02$, $D_4 = 2.57$, $D_3 = 0$ for $n = 3$]

2. Discuss how control charts can be used in quality control of industrial products. The average percentage of defectives in 27 samples of size 1500 each was found to be 13.7%. Construct P-chart for this situation. Explain how the control chart can be used to control quality.

[**Hint.** $\overline{p} = 0.137$]

(G.B.T.U. 2010)

3. The following data shows the value of sample mean $\overline{X}$ and range R for 10 samples of size 5 each.

Calculate the values for central line and control limits for $\overline{X}$-chart and R chart and determine whether the process is under control.

Sample no. :	1	2	3	4	5	6	7	8	9	10
Mean $\overline{X}$:	11.2	11.8	10.8	11.6	11	9.6	10.4	9.6	10.6	10
Range R :	7	4	8	5	7	4	8	4	7	9

Assume for $n = 5$, $A_2 = 0.577$, $D_3 = 0$ and $D_4 = 2.115$.

4. It was found that when a manufacturing process is under control, the average number of defectives per sample batch of 10 is 1.2. What limits would you set in a quality control chart based on the examination of defectives in sample batches of 10?

[**Hint.** $\overline{p} = 0.12$, $n\overline{p} = 1.2$]

5. In a manufacturing process, the number of defective items found in the inspection of 15 lots of 400 items each are given below:

Lot No. :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of defectives :	2	5	0	14	3	0	1	0	18	8	6	0	3	0	6

6. (*i*) What are statistical quality control techniques? Discuss the objectives and advantages of statistical quality control.

(*ii*) Give statistical quality control methods. Explain one of them with example.

7. Write short notes on the following :

(*i*) Use of statistical techniques/methods in quality control *[G.B.T.U. (MCA) 2010]*

(*ii*) P-chart, *np*-chart and *C*-chart *(G.B.T.U. 2010)*

8. Distinguish between the *np*-chart and *p*-chart. Following is the data of defectives of 10 samples of size 100 each. Construct *np*-chart and give your comments.

Sample no. :	1	2	3	4	5	6	7	8	9	10
No. of difectives :	6	9	12	5	12	8	8	16	13	7

(G.B.T.U. 2010)

9. In a factory producing spark plug, the number of defectives found in inspection of 20 lots of 100 each is given below. *(G.B.T.U. 2011)*

Lot No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
No. of defectives	5	10	12	8	6	4	6	3	3	5	4	7	8	3	3	4	5	8	6	10

Construct *p*-chart and state whether the process is in statistical control.

10. The data below given is the number of defective bearing in samples of size 150. Construct *np*-chart for these data. If any point(s) lie outside the control limits, assume that assignable cause can be found and determine the revised control limits :

Sample No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
No. of defectives	12	7	5	4	1	5	9	0	15	6	7	4	1	3	6	8	10	5	2	7

Determine metrial control limits for *np* chart and state whether the proven is in control.

11. The following table shows the number of missing rivets observed at the time of inspection of 12 aircrafts. Find the control limits for the number of defects chart and comment on the state of control.

Air craft number :	1	2	3	4	5	6	7	8	9	10	11	12
No. of missing rivets :	7	15	13	18	10	14	13	10	20	11	22	15

12. The number of customer complaints received daily by an organization is given below :

Day :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Complaints :	2	3	0	1	9	2	0	0	4	2	0	7	0	2	4

Does it mean that the number of complaints is under statistical control? Establish a control scheme for the future.

Answers

1. $CL_{\overline{X}} = 0.4988$, $UCL_{\overline{X}} = 0.5172$, $LCL_{\overline{X}} = 0.4804$, $CL_R = 0.018$, $UCL_R = 0.0463$, $LCL_R = 0$. The process is in control.

2. $CL_P = 0.137$, $UCL_P = 0.164$, $LCL_P = 0.110$

3. $CL_{\overline{X}} = 10.66$, $UCL_{\overline{X}} = 14.295$, $LCL_{\overline{X}} = 7.025$, $CL_R = 0.3$, $UCL_R = 13.32$, $LCL_R = 0$; The process is under control

4. $CL_{np} = 1.2$, $UCL_{np} = 2.175$, $LCL_{np} = 0.225$

5. $p = 0.011$, CL = 4.4, UCLnp = 10.6581, LCLnp = 0; No

8. Process is in statistical control.

9. Process is in statistical control. $UCL_p = 0.131$, $LCL_p = 0$

10. $CL_{np} = 5.36842$, $UCL_{np} = 12.19385$, $LCL_{np} = 0$

11. $UCL_C = 25.23$, $LCL_C = 2.77$. The process is in control.

12. $CL_C = 2.4$, $UCL_C = 7.05$, $LCL_C = 0$, the process is not under control.