CHAPTER-III

IMPORTANT THEORETICAL DISTRIBUTIONS

We now discuss some generalized distributions which are very useful and relevant to the policy makers. These distributions are based on certain assumptions. Using these distributions, predictions can be made on theoretical grounds.

3.1 Discrete Probability Distributions:

Discrete Uniform Distribution: The probability distribution in which the random variable assumes all its values with equal probabilities, is known as discrete uniform distribution. If the random variable X assumes the value x_1, $x_{2,}\cdots\cdots,x_k$ with equal probabilities, then its probability function is given by:

$$P(X = x_i) = \frac{1}{k} \text{ for } i = 1,2,.....,k$$

Bernoulli Distribution: A random variable X which takes only two values 1 and 0 (called as success and failure) with respective probabilities p and q such that P(X=1) = p, and P(X = 0) = q and p + q = 1, is called a Bernoulli Variate and its distribution is called Bernoulli Distribution.

Binomial Distribution: This distribution was given by James Bernoulli (1713) and is based on the assumptions of Bernoulli trials.

Bernoulli Trials: A series of independent trials which can result in one of the two mutually exclusive outcomes called success or failure such that the probability of success (or failure) in each trial is constant, then such repeated independent trials are called Bernoulli trials.

If we perform a series of n Bernoulli trials such that for each trial, p is the probability of success and q is the probability of failure (p + q = 1), then probability of x successes in a series of n independent trials is given by

$p(x) = {}^nC_x\, p^x\, q^{n-x}$ where $x = 0, 1, 2,\cdots\cdots,n$

This is known as binomial distribution and the probabilities $p(0) = q^n$; $p(1) = {}^nC_1 pq^{n-1}$; $p(2) = {}^nC_2 p^2 q^{n-2},\cdots\cdots$, $p(n) = q^n$ of 0 success, 1 success, 2 successes $\cdots\cdots$n successes are nothing but the first, the second, the thirdé é the $(n+1)^{th}$ terms in the

CHAPTER-IV

STATISTICAL INFERENCE

Statistical Inference is a branch of statistics which deals with drawing conclusions about the population on the basis of few observations (sample) and sampling is a process of selecting a small fraction (sample) from the population so that it possesses the characteristics of the population.

In applied investigations especially in agricultural and allied sciences, it may not be possible to study the whole population and thus the investigator is forced to draw inferences about the population on the basis of the information obtained from the sample data (due to high operational cost and time consideration), which is called as statistical inference. For example, it is out of question to harvest and record the produce from all the fields growing the wheat crop which constitute the population under study.

4.1 Some Basic Concepts:

Population: The word population in Statistics is used to refer to any aggregate collection of individuals possessing a specified characteristic e.g. population of teachers in HAU, plants in a field etc. A population containing a limited number of individuals or members is called a finite population, whereas a population with unlimited number of individuals or members is known as infinite population. The population of books in a library, population of Indian students in UK are examples of finite population, whereas fish population in Pacific Ocean and the number of stars in the sky are infinite **populations.**

Sample: A portion or a small section selected from the population by some sampling procedure is called a sample.

Census: The recording of all the units of a population for a certain characteristic is known as census or complete enumeration.

Parameter: Parameter are the numerical constants of the population, e.g. population mean (μ), population variance (σ^2) etc.

Statistic: Any function of sample observations is called a statistic. Its value may vary from sample to sample, e.g. sample mean (x) and, sample variance (s^2) are the statistics.

Some commonly used Sampling Techniques: For drawing a sample from the population, several sampling designs are used, some of which are discussed as under:

Probability and Non-Probability Sampling

Probability Sampling: This is a method of selecting a sample according to certain laws of probability in which each unit of population is assigned some definite probability of being selected in the sample. The followings are some such sampling methods:

Simple Random Sampling (SRS): It is the simplest and most commonly used method in which the sample is drawn unit by unit with equal probability of selection at each draw for each unit. Hence simple random sampling is a method of selecting ón' units out of a population of size óN' by assigning equal probability to all units. It is a sampling procedure in which all possible combinations of ón' units that may be formed from the population of óN' units have the same probability of selection. The procedure, where a selected unit is replaced in the population and ón' units are drawn successively is called simple random sampling with replacement (SRSWR). If sample is drawn without replacing the units selected at each draw, it is called simple random sampling without replacement (SRSWOR). In SRSWR, there are N^n possible samples of size n each with probability of selection $\frac{1}{N^n}$ that can be drawn from a population of size N, while in SRSWOR, there are ${}^{N}C_{n}$ possible samples each with probability of selection as $1 / {}^{N}C_{n}$.

Methods of Drawing a Random Sample:

i) Lottery Method: Suppose we wish to draw a random sample of size n from a finite population of N units. We take N pieces of paper of the same size and shape and number them from 1 to N such that each unit in the population corresponds to one piece of paper. These pieces are then put in a container and mixed up thoroughly and a sample of n pieces is drawn either one by one or in a single stroke. The sampling units bearing the numbers on the selected pieces will constitute the desired random sample. Lottery method cannot be applied if the study population is infinite.

ii) Random Number Table Method: Suppose a random sample of n units is to be taken from a population of N units. We mark all the units serially from 1 to N, and take any page of random number table. Starting from anywhere either row-wise or column-wise, random numbers are selected in the sample ignoring the values greater than N. In this method all the digits greater than N are rejected.

Sampling with and without Replacement: In lottery method, while drawing a piece of paper from the container, we may have the choice of replacing or not replacing the selected piece into the container before the next draw is made. In the first case the

numbers can come up again and again, while in the second case they can come up only once. The sampling in which each member of a population may be chosen more than once is called sampling with replacement, whereas the sampling in which no member can be chosen more than once is called sampling without replacement.

Stratified Sampling: Stratified sampling technique is generally followed when the population is heterogeneous and where it is possible to divide it into certain homogeneous sub-populations, say k, called strata. The strata differ from one another but each is homogeneous within itself. The units are selected at random from each of these strata. The number of units selected from different strata may vary according to their relative importance in the population. The sample, which is the aggregate of the sampled units from various strata, is called a stratified random sample and the technique of drawing such a sample is known as stratified sampling or stratified random sampling.

Advantages of Stratified Sampling over Random Sampling:

i) The cost per observation in the survey may be reduced

ii) Estimates of the population parameters may be obtained for each sub-population

iii) Accuracy at given cost is increased

iv) Administrative control is much better as compared to simple random sampling

Systematic Sampling: If the sampling units are arranged in a systematic manner and then a sample is drawn not at random but by taking sampling units systematically at equally spaced intervals along some order. The sample obtained in this manner is called a systematic sample and the technique is called the systematic sampling.

Cluster Sampling: In some situations the elementary units are in the form of groups, composed of smaller units. A group of elementary units is called a cluster. The procedure in which sampling is done by selecting a sample of clusters and then carrying out the complete enumeration of clusters is called cluster sampling. For example in taking a sample of households we select a few villages and then enumerate them completely. Cluster sampling is typically used when the researchers cannot get a complete list of the members of a population they wish to study but can get a complete list of groups or 'clusters' of the population. It is also used when a random sample would produce a list of individuals so widely scattered that surveying them would prove to be much expensive. This sampling technique may be more practical and/or economical than simple random

sampling or stratified sampling. The systematic sampling may also be taken as the cluster sampling in which a sample of one cluster is taken and then it is completely investigated.

Multistage Sampling:

The cluster sampling is more economical but the method restricts the spread of the sample over the population which increases the variance of the estimator. The method of sampling which consists in first selecting the clusters and then selecting specified elements from each cluster is known as two stage sampling. Here clusters which form the units of sampling at the 1^{st} stage are called first stage units (fsu) or primary stage units (psu) and the elements within clusters are called second stage unit (ssu). This procedure can be generalized for more than two stages which is termed as multistage sampling. For example, in crop survey, district may be fsu, blocks as ssu and village may be considered as third stage units or ultimate stage units (usu).

Non-Probability Sampling: It is a sampling procedure in which the sample units are selected not according to law of chance but according to some prior judgement. This procedure is adopted when we wish to collect some confidential information or quick information at low cost. Popular non-probability sampling schemes are purposive, judgement sampling and quota sampling

4.2 Sampling Distributions:

If all possible samples of size n are drawn from a given population and for each sample the value of a statistic, such as the mean, variance etc. is calculated. The value of the statistic will vary from sample to sample and resulting distribution of the statistic is called its sampling distribution. If the particular statistic is the sample mean, the distribution is called the sampling distribution of the mean and so on. The standard deviation of a sampling distribution of a statistic is often called its standard error.

Sampling Distribution of Sample Mean ($\overline{X}$):

Suppose X_1, X_2,é .., X_n is a random sample from normal population N (μ, σ^2). Then sample mean is $\overline{X} = \frac{1}{n}$Ł$X_i$

We assume that the variance σ^2 is known. Since $\overline{X}$ is a linear combination of normal variates, therefore, distribution of $\overline{X}$ is also normal having mean $\varepsilon_{\bar{x}}$ and variance $\mathring{u}^2_{\bar{x}}$ where.

$$\varepsilon_{\bar{x}} = E[\bar{X}] = E\left[\frac{X_1 + X_2, ..., X_n}{n}\right] = \frac{1}{n}\left[E(X_1) + E(X_2) + ... + E(X_n)\right]$$

$$= \frac{1}{n}\left[\varepsilon + \varepsilon + ... + \varepsilon\right] = \frac{n\varepsilon}{n} = \varepsilon$$

and $\sigma^2 = V(\bar{X}) = V\left[\frac{X_1 + X_2 + ..., + X_n}{n}\right] = \frac{1}{n^2}\left[V(X_1) + V(X_2) + ..., + V(X_n)\right]$

$$= \frac{1}{n^2}\left[\mathring{u}^2 + \mathring{u}^2 + ... + \mathring{u}^2\right] = \frac{n\mathring{u}^2}{n^2} = \frac{\mathring{u}^2}{n}$$

Thus, sampling distribution of $\bar{X}$ is also normal with mean μ and variance σ^2/n.

or $\bar{X} \sim N\left(\varepsilon, \frac{\mathring{u}^2}{n}\right)$ and $Z = \frac{\bar{X} - \varepsilon_{\bar{X}}}{\mathring{u}/\sqrt{n}} \sim N(0, 1)$

When parent population is not normal, the sampling distribution will depend on the form of the parent population. However, as n gets large, the form of the sampling distribution will become more and more like a normal distribution, no matter what the parent distribution is. This is stated in the following theorem, which is popularly known as central limit theorem.

Central Limit Theorem: If random samples of size n are taken from any distribution with mean μ and variance σ^2, the sample mean $\bar{X}$ will have a distribution approximately normal with mean μ and variance σ^2/n. The approximation becomes better as n increases.

If σ^2 is unknown, then an estimate of σ^2 is obtained from the sample, which is given by $s^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$. Further if n < 30, then $t = \frac{\bar{X} - \varepsilon_{\bar{X}}}{s/\sqrt{n}}$ follows a student's t-distribution with (n-1) degrees of freedom and when n > 30, $Z = \frac{\bar{X} - \varepsilon_{\bar{X}}}{s/\sqrt{n}}$ follows a standard normal distribution i.e. N (0, 1). Also, the standard error of the statistic sample mean ($\bar{X}$) is given by SE ($\bar{X}$) $= \frac{s}{\sqrt{n}}$.

Sampling Distribution of Difference of Means: Let $x_1, x_2, \cdots\cdots, x_{n1}$ and $y_1, y_2, \cdots\cdots, y_{n2}$ be two independent random samples of sizes n_1 and n_2 from N (μ_1, σ_1^2) and N (μ_2, σ_2^2), respectively.

Then the statistics $\overline{X} = \frac{1}{n_1}\sum X_i$ and $\overline{Y} = \frac{1}{n_2}\sum Y_i$ have sampling distributions $N\left(\varepsilon_1, \frac{\sigma_1^2}{n_1}\right)$ and $N\left(\varepsilon_2, \frac{\sigma_2^2}{n_2}\right)$ respectively provided the population variances σ_1^2 and σ_2^2 are known. The sampling distribution of the difference of means $\overline{X} - \overline{Y}$ is normal with mean $\varepsilon_{\bar{x}-\bar{y}}$ and variance $\sigma^2_{\bar{x}-\bar{y}}$ where $\varepsilon_{\bar{x}-\bar{y}} = \mu_1 - \mu_2$ and $\sigma^2_{\bar{x}-\bar{y}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Also standard error of the difference of means $\overline{X} - \overline{Y}$

$$SE_d(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \text{ and thus the statistic } Z = \frac{(\overline{X} - \overline{Y}) - (\varepsilon_1 - \varepsilon_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

If σ_1^2 and σ_2^2 are unknown but equal and samples are small, then the pooled estimate s_p^2 of common variance σ^2 is given by:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \text{where } s_1^2 = \frac{1}{n_1 - 1}\sum_i (x_i - \bar{x})^2; s_2^2 = \frac{1}{n_2 - 1}\sum_i (y_i - \bar{y})^2$$

Then $SE(\overline{X} - \overline{Y}) = s_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ and the statistic $t = \frac{(\bar{x} - \bar{y}) - (\varepsilon_1 - \varepsilon_2)}{s_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ follows

student's t-distribution with $(n_1 + n_2 - 2)$ d.f. In case of large samples the statistic

$$Z = \frac{(\overline{X} - \overline{Y}) - (\varepsilon_1 - \varepsilon_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$$

Remarks:

1) For Simple Random Sampling without Replacement (SRSWOR) from a finite population, we have

   $E(\bar{x}) = \mu$ and $S.E_m = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N - n}{N - 1}}$ where N is the population size.

2) If the original population is not normal then also for large samples ($n \geq 30$), the sample mean ($\overline{X}$) is approximately normally distributed

   Thus if $X \sim (\mu, \sigma^2)$, then for large samples $\overline{X} \sim N(\mu, \sigma^2/n)$ approximately.

Sampling Distribution of Sample Proportion: Suppose a population is divided into two non-overlapping classes C and C' i.e. individuals possessing a characteristic are put in class C and those not possessing the characteristic in class C' e.g. Smokers and Non-Smokers, Defectives and Non-Defectives, Males and Females etc.

Let A be the number of individuals possessing a particular characteristic in a population of size N and let a be the number of individuals in a sample of size n possessing the characteristic C. Then

$P = \frac{A}{N}$ denotes proportion of units in the population possessing characteristic C

and $p = \frac{a}{n}$ denotes proportion of units in the sample possessing characteristic C

Then the sampling distribution of sample proportion is as follows:

If all possible samples of size n are drawn from a population of size N, then sample proportion (p) is distributed with mean P and variance $\frac{PQ}{n}$ i.e. $p \sim \left(P, \frac{PQ}{n}\right)$ and for large samples (n>30), the distribution of p is approximately normal i.e. $p \sim N\left(P, \frac{PQ}{n}\right)$; where Q = 1 ï P or $Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$

The Z statistic obtained above is used for testing the null hypothesis that the population proportion (P) is equal to some specified value P_0 i.e. H_0: $P = P_0$ for large samples.

Sampling Distribution of difference between two Sample Proportions:

Let p_1 and p_2 be two sample proportions obtained from samples of sizes n_1 and n_2 from two populations with population proportions P_1 and P_2, respectively. If all possible samples of sizes n_1 and n_2 are drawn from two populations with population proportions P_1 and P_2, respectively, then the difference between sample proportions p_1 and p_2 i.e. p_1 ï p_2 is distribution with mean P_1 ï P_2 and variance $\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}$

For large samples i.e. n_1>30 and n_2>30, the distribution is approximately normal i.e. $p_1 - p_2 \sim N\left(P_1 - P_2, \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)$ which implies

$$Z = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1)$$

The Z statistic obtained above is used for testing the significance of difference between two population proportions for large samples.

The mean and standard error of some of the important statistics are given below:

Statistic	Parameter	S.E.	Estimated S.E.	Remarks
Sample mean ($\overline{X}$)	μ	$\mathring{u}/\sqrt{n}$	$s/\sqrt{n}$	
Sample proportion (p)	P	$\sqrt{PQ/n}$	$\sqrt{pq/n}$	P is the population proportion and Q = 1-P
Difference of two sample means ($\overline{X} - \overline{Y}$)	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	σ_1^2 and σ_2^2 are variances of two populations
Difference of two sample proportions (p_1-p_2)	P_1-P_2	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$	P_1 and P_2 are population proportions and $Q_1 = 1-P_1$; $Q_2 = 1-P_2$

Uses of Standard Error:

i) The reciprocal of the S.E. gives an estimate of the reliability or precision of the statistic.

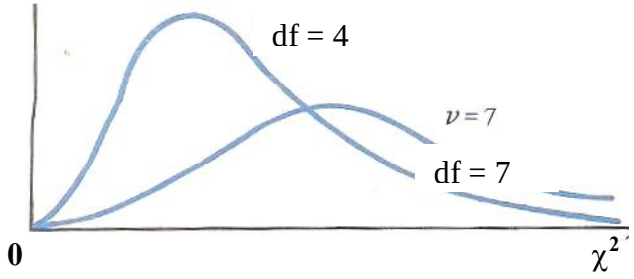
ii) S.E. enables us to determine the confidence limits, which are expected to contain the population parameter.

iii) If t is any statistic, then for large samples $Z = \frac{t - E(t)}{S.E(t)}$ is a standard normal variate having mean zero and variance unity (in case the population is normal, then there is no restriction on the sample size). This Z value forms the basis for testing of hypothesis.

Chi-square (χ^2) Distribution: A continuous random variable X is said to follow χ^2 distribution if its probability density function is:

$$f(x) = ke^{-x/2}(x)^{n/2-1}$$

where k is a constant such that the area under probability density curve is unity. The only parameter (positive integer) of the χ^2 (pronounced as Ky) distribution is ónô which is called the ónumber of degrees of freedomô. Range of χ^2 is from 0 to ∞ i.e. $0 \le \chi^2 \le \infty$.



Shape of the Chi-square distribution curve

Theorem: If $Z_1, Z_2, \ldots, Z_n$ are ónô independent standard normal variates i.e. $Z_i = \frac{x_i - \varepsilon}{\sigma}$ then $Z_1^2 + Z_2^2 + \ldots\ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$ follows a χ^2 distribution with n degrees of freedom. However, if Z_i is taken as $\frac{X_i - \bar{x}}{\sigma}$ where $\bar{x}$ is the sample mean instead of μ then $\sum Z_i^2$ follows χ^2 distribution with (n-1) degrees of freedom.

Properties of χ^2 Distribution:

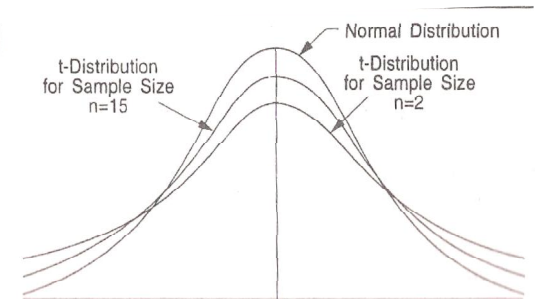
i) The mean and S.D. of χ^2 distribution with n d.f. are n and $\sqrt{2n}$ respectively

ii) χ^2 distribution is positively skewed but with an increased degree of freedom, the χ^2 curve approaches more and more close to the normal curve and for $n \geq 30$ i.e. in the limiting case, χ^2 curve/distribution can be approximated by the normal curve/distribution

iii) Sum of independent χ^2- variates is also a χ^2- variate

Student's 't' distribution: The form of the probability density function for ótô distribution is given by

$$f(t) = K\left(1 + \frac{t^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}; (-\infty < t < \infty)$$



Normal Distribution and t-Distribution curves for sample sizes n = 2 and n = 15

where k is a constant such that total area under the curve is unity. The only parameter n (a positive integer) is called the ñnumber of degrees of freedomò. This distribution was found out by W.S. Gosset (1908) who used the pen-name ñStudentò and hence it is known as studentôs átô distribution. A variable which follows Studentôs distribution is denoted by átô.

If a random sample of size n is drawn from a normal population with mean μ and S.D. σ (unknown), then $\frac{\bar{x} - \varepsilon}{s/\sqrt{n}}$ follows átô distribution with (n-1) d.f. where $\bar{x}$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_1 - \bar{x})^2$ denote sample mean and sample variance respectively.

Properties:

i) The curve is leptokurtic with $\beta_2 > 3$ but curve approaches more and more close to the normal curve and for n > 30. t-curve approaches to the normal curve.

ii) t-curve is symmetrical about t = 0 hence mean = mode = median = 0. Its standard deviation is $\sqrt{\frac{n}{n-2}}, (n > 2)$. Also the t-curve extends from -∞ to +∞ like the normal curve.

iv) The degrees of freedom n is the only parameter of the t-distribution.

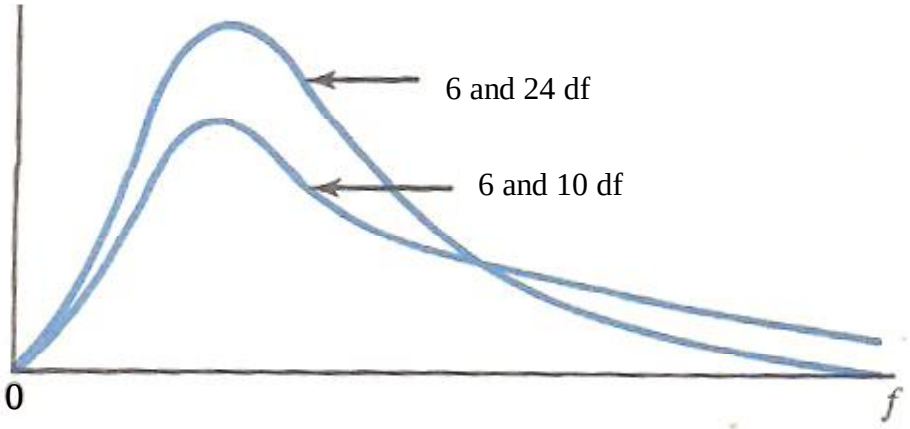
Snedcor's F-Distribution:

F-distribution is defined as the ratio of two independent chi-square variates, each divided by their respective degrees of freedom. Let χ_1^2 and χ_2^2 are independent random variables having Chi-square distributions with degrees of freedom v_1 and v_2 respectively, then distribution of the ratio $F = \frac{\chi_1^2/v_1}{\chi_2^2/v_2}$ follows F-distribution with v_1 and v_2 degrees of freedom.

If we have independent random samples of sizes n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2 respectively, then $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ has F distribution with degrees of freedom (n_1-1) and $(n_2 - 1)$. F-distribution is also known as the variance ratio distribution. This distribution is primarily applied in the analysis of variance, where

we wish to test the equality of several means simultaneously. F-distribution is also used to make inferences concerning the variance of two normal populations.

Definition: A random variable follows F distribution if the probability density function is defined by $f(f) = Kf^{\left(\frac{v_1}{2}\right)-1}(3_2 + 3_1 f)^{-\left(\frac{3_1+3_2}{2}\right)}$ where ófô ranges between 0 to ∞ and K is constant. The parameters 3_1 and 3_2 gives the degrees of freedom (3_1, 3_2) of the distribution. The distribution was proposed by Snedecor and named óFô in honour of the distinguished statistician Sir R.A. Fisher.



F-distribution curve

Properties:

i) The F-distribution is positively skewed with ϭ$_1$ but for g_1 and g_2 both greater than 30, F-curve/distribution can be approximated by the normal curve

ii) Mean of F-distribution is $\frac{\upsilon_2}{\upsilon_2 - 1}$ and variance is $\frac{2\upsilon_2^{\,2}(\upsilon_1 + \upsilon_2 - 2)}{\upsilon_1(\upsilon_2 - 2)^2(\upsilon_2 - 4)}$

iii) If $X \sim Fg_1, g_2$ then $\frac{1}{X} \sim Fg_2, v_1$

iv) υ_1 and υ_2 are the only two parameters of the F-distribution

4.3 Point and Interval Estimation:

The use of sample statistic such as sample mean ($\overline{X}$) sample variance (s^2), sample proportion (p) etc. to draw conclusions about the population parameters such as mean (μ), variance (σ^2), proportion (P) is of fundamental importance in statistical inference.

The following two concepts are used for drawing valid inferences about the unknown population parameter based upon random samples

i) **Estimation:** A procedure of using a sample statistic to approximate a population parameter is called estimation. A statistic used to estimate a population parameter value is called an **estimator** and the value taken by the estimator for a particular sample is called an **estimate.**

A few examples are given below for illustration:

a) A breeder needs to know the average yield of a newly released variety V_1

b) A manufacturer needs to know the proportion of second quality items of his product.

c) A social worker needs to know the proportion of families having three or more children in a particular region.

ii) **Testing of Hypothesis:** It is a procedure to test the claim or belief about an unknown parameter value and will be discussed in the next chapter.

There are two types of estimators: (i) Point estimator and (ii) confidence interval estimator.

Point Estimation: When the estimator of unknown population parameter is given by a single value or point value then it is called **Point Estimator**. For example, sample mean $(\overline{X})$ and sample variance (s^2) are point estimators of the population mean (μ) and population variance (σ^2) respectively.

There are several alternative estimators which might be used for estimating the same parameter. For example, the population mean is a measure of central tendency of the population values and sample mean, sample median and sample mode may be considered as the possible estimators of the population mean. Now the question arises which sample statistic should be used as the estimator of the population parameter. The best estimator is one that is more suitable to a given problem, and has same desirable properties like unbiasedness, consistency, efficiency and sufficiency.

Properties of a Good Estimator: A good estimator is one which is close to the parameter being estimated. Some of the desirable properties of a estimator are discussed below:

i) **Unbiasedness:** An estimator is a random variable and it is always a function of sample observations. If the expected value of an estimator is equal to the population parameter, then it is called an **unbiased estimator**. Thus an estimator

(of a parameter) is said to be unbiased for parameter θ, if E(t) = θ; if E(t) ≠ θ, then it is biased estimator of θ and Bias of the estimator is given by B(t) = E(t) ï θ. For example, sample mean is an unbiased for population mean and sample variance computed with the divisor (n-1) is also unbiased for σ^2.

ii) Consistency: It is a limiting property of an estimator i.e. it concerns with the behavior of the estimator for large sample sizes. If the difference between estimator and the corresponding population parameter continues to become smaller and smaller as the sample size increases, i.e. the estimator coverages in probability to the population parameter then it is called consistent estimator of that parameter. Symbolically, if t_n is an estimator computed from a sample of size n and θ is the parameter being estimated and Pr $\left[\left|t_n - \theta\right| < \in\right] \to 1$, as $n \to \infty$ for any positive ε, however small, then t_n is said to be consistent estimator of θ. It is true for $\overline{X}$ and s^2 which are consistent estimators of μ and σ^2 respectively.

iii) Efficiency: An estimator t_1 is said to be more efficient than t_2 for parameter θ if MSE (t_1) < MSE (t_2) where MSE (t) = E $(t-\theta)^2$ is the mean square error of the estimator t. It can be proved that sample mean is more efficient estimator of population mean μ then the sample median.

iv) Sufficiency: A sufficient estimator is one that uses all the information about the population parameter contained in the sample. It ensures that all information that a sample can furnish with respect to the estimation of a parameter is utilized. It may be noted that sample mean $\overline{X}$ and sample proportion (p) are sufficient estimators of corresponding parameters since all the information in the sample is used in their computation. On the other hand, sample mid-range is not a sufficient estimator since it is computed by averaging only the highest and lowest values in the sample.

Interval Estimation:

Point Estimates provides no information regarding the reliability of estimates i.e. how close an estimate is to the true population parameter. Thus, point estimators are rarely used alone to estimate the population parameters. It is always better to construct an

interval estimator which contains the population parameter so that the reliability of the estimator can be measured. This is the purpose of interval estimation.

When the estimator of a parameter is given in the form an interval (A, B) with a specified level of confidence instead of a single value, it is called **Confidence Interval Estimator**.

Definition: The interval (A, B) is said to be (1-α)x100% Confidence Interval for the population parameter θ if P (A ≤ θ ≤ B) = 1-α.

The quantities A and B are called **Confidence Limits** or **Fiducial Limits** while (1-Ŭ) 100% is called **level of confidence**.

Remarks:

i) By putting α = 0.05 and 0.01 we have 95% and 99% confidence intervals.

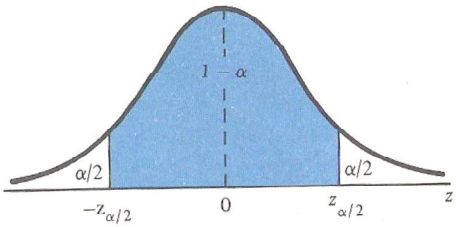
ii) By 95% confidence interval, we mean that the interval will contain the parameter in atleast 95% cases if the experiment is repeated with different samples.

Construction of C.I. for Population Mean μ:

Case-1: When population variance σ^2 is known or sample size n ≥ 30

Let $x_1, x_2, \cdots, x_n$ be a random sample of size n from N (μ, σ^2), then $Z = \frac{\overline{X} - \varepsilon}{\mathring{u}/\sqrt{n}} \sim N(0,1)$ is a standard normal variate. The (1-α) 100% C.I. for μ is given by

$$P\left[-z_{\check{U}/2} \le \frac{\overline{X} - \varepsilon}{\mathring{u}/\sqrt{n}} \le z_{\check{U}/2} \right] = 1 - \check{U}$$



where $z_{\alpha/2}$ is the Z-value for which the area on the right tail of standard normal curve is α/2

or $$P\left[\overline{X} - z_{\check{U}/2} \frac{\mathring{u}}{\sqrt{n}} \le \varepsilon \le \overline{X} + z_{\check{U}/2} \frac{\mathring{u}}{\sqrt{n}} \right] = 1 - \check{U}$$

Thus (1-α) 100% upper and lower confidence limits for μ are $\overline{X} \pm z_{\check{U}/2} \frac{\mathring{u}}{\sqrt{n}}$ and the quantity $z_{\check{U}/2} \frac{\mathring{u}}{\sqrt{n}}$ is called margin of error whereas $\frac{\mathring{u}}{\sqrt{n}}$ is called standard error of sample mean.

Values of the Standard Normal Variate

Confidence level (1-α) 100%	**90%**	**95%**	**98%**	**99%**
Level of significance α	0.10	0.05	0.02	0.01
z_α (for one tailed test)	1.28	1.64	2.05	2.33
$z_{\alpha/2}$ (for two tailed test)	1.64	1.96	2.33	2.58

Thus the 95% and 99% confidence limits for μ are $\overline{X} \pm 1.96$ ů/$\sqrt{n}$ and $\overline{X} \pm 2.58$ ů/$\sqrt{n}$ respectively.

Remarks:

i) Even if σ^2 is unknown but for large sample size ($n \geq 30$), we can substitute sample standard deviation s in place of σ and the interval estimator for μ is given by $\overline{X} \pm z_{\alpha/2} s/\sqrt{n}$.

ii) The confidence limits as given above are applicable for sampling from infinite population or sampling with replacement from finite population. For sampling without replacement from finite population, standard error will be multiplied by the factor $\sqrt{(N-n)/(N-1)}$ known as **finite population correction factor**.

Case-2: When σ^2 is unknown and n<30: For this situation, the statistic $\frac{\overline{x} - \varepsilon}{s/\sqrt{n}}$ follows student's t-distribution with (n-1) d.f. In this case Z-values are replaced by t-distribution values with (n-1) d.f. Thus the (1-α) 100% confidence limits for μ are $\overline{X} \pm t_{\alpha/2,(n-1)}\ s/\sqrt{n}$ where $t_{\alpha/2,(n-1)}$ is the tabulated value of t leaving an area α/2 on the right tail of t- distribution curve with (n-1) d.f. and $s^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2$ is the sample variance.

	Confidence interval for μ (summary)	
Sample size	**σ known**	**σ unknown**
Large	$\overline{X} \pm z_{\alpha/2}\ \sigma/\sqrt{n}$	$\overline{X} \pm z_{\alpha/2} s/\sqrt{n}$
Small	-do-	$\overline{X} \pm t_{\alpha/2,(n-1)}\ s/\sqrt{n}$

Confidence interval for population proportion (P):

For large sample, the confidence interval for P with (1-α)100% level of confidence is given by $p \pm z_{\alpha/2}\sqrt{PQ/n}$ where Q = 1-P. Since P and Q are unknown and

they are estimated by sample statistics p and q respectively and the confidence limits for P can be taken as $p \pm z_{\alpha/2}\sqrt{pq/n}$.

Example-1: 400 labourers were selected from a certain district and their mean income was found to be Rs. 700 per week with a standard deviation of Rs. 140. Set up 95% and 98% confidence limits for the mean income of the labour community of the district.

Solution: Given $\bar{x} = 700$; $s = 140$ and $n = 400$

$SE_m = s/\sqrt{n} = 140/20 = 7$

The confidence limits for population mean (average weekly income)

$= \bar{x} \pm z_{\alpha/2}\ s/\sqrt{n}$

Thus 95% confidence limits are: $700 \pm 1.96(7) = 700 \pm 13.72 = (686.28, 713.72)$

and 98% confidence limits are: $700 \pm 2.33(7) = 700 \pm 16.31 = (683.69, 716.31)$

Example-2: A random sample of 144 families shows that 48 families have two or more children. Construct a 90% confidence interval for the proportion of families having two or more children.

Solution: Sample proportion $p = 48/144 = 1/3$

The confidence limits for population proportion P are given by $p \pm z_{Ŭ2}\sqrt{\frac{pq}{n}}$

where $p = 1/3$, $q = 2/3$, $z_{0.05} = 1.645$, $n = 144$

$SE_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(1/3)(2/3)}{144}} = 0.0393$

Thus the 90% confidence interval is

$\frac{1}{3} \pm 1.645(0.0393) = 0.333 \pm 0.065 = (0.268, 0.398)$

4.4 Testing of Hypothesis:

One of the prime objectives of experimentation whether it is in field or laboratory, is the comparison of treatments means and variances under study. A researcher is usually interested in the following comparisons for drawing logical conclusions of the study undertaken by him. The statistical tests which are applicable in different situations are also given:

i) Comparison of a treatment mean with its hypothetical value (one sample Z-test and one sample t-test)

ii) Comparison of two treatment means (two sample Z-test, two sample t-test and paired t-test)

iii) Comparison of two population variances (F-test)

Some Basic Concepts to Hypotheses Testing:

Any statement or assumption about the population or the parameters of the population is called as **statistical hypothesis**.

The truth or falsity of a statistical hypothesis is never known with certainty unless we examine the entire population. This, of course, would be impractical in most situations. Instead, we take a random sample from the population of interest and use the information contained in this sample to decide whether the hypothesis is likely to be true or false. Evidence from the sample if is inconsistent with the stated hypothesis leads to the rejection of the hypothesis, whereas evidence supporting the hypothesis leads to its acceptance. The investigator should always state his hypothesis in a manner so as to test it for possible rejection. If the investigator is interested in a new vaccine, he should assume that the new vaccine is not better than the vaccine now in the market and then set out to reject this contention. Similarly, to test if the new ploughing technique is superior to old one, we test the hypothesis that there is no difference between these two techniques.

The hypothesis which is being tested for possible rejection is referred to as **Null hypothesis** and is represented by H_0 where as the hypothesis complementary to the null hypothesis is referred to as **Alternative hypothesis** and is represented by H_1.

These hypotheses are constructed such that the acceptance (or rejection) of one leads to the rejection (or acceptance) of the other. Thus if we state the null hypothesis as H_0 : μ (yield of c.v. WH-542) = 65 q/ha, then the alternative hypothesis might be H_1 : $\mu \neq 65$ q/ha or $\mu > 65$ q/ha or $\mu < 65$ q/ha.

Two Types of Errors in Hypothesis Testing: In order to make any decision to accept or to reject the null hypothesis we have to draw a random sample $x_1, x_2, ..é , x_n$ of size n from the population under study and on the basis of the information contained in the

sample we have to decide whether to accept or reject the null hypothesis. Because of the random nature of the sample the above decision could lead to two types of errors.

	Decision	
	Accept	**Reject**
H_0 is true	Correct decision (no error)	Type ï I error
H_0 is false	Type ï II error	Correct decision (no error)

Type I error occurs when we reject a true null hypothesis, i.e. we reject the null hypothesis when it should be accepted. **Type II error** is committed when we accept a false null hypothesis, i.e. we accept the null hypothesis when it should be rejected.

The relative importance of these two types of errors depends upon the individual problem under study. For instance, in the above example, it is expensive to replace the existing variety V_1 and so one should be very careful about the type I error. Whatever may be the relative importance of these errors, it is preferable to choose a test for which the probability of both types of error is as small as possible. Unfortunately, when the sample size n is fixed in advance, it is not possible to control simultaneously both types of errors. What is possible is to choose a test that keeps the probability of one type of error a minimum when the probability of other type is fixed. It is customary to fix type I error and to choose a test that minimize the probability of type II error.

Level of significance is the probability of committing a type I error i.e. it is the risk of rejecting a true null hypothesis. It is denoted by the symbol Ŭ

On the other hand, the probability of committing a type II error is denoted by the symbol ɓ and consequently (1- ɓ) is called the **power of the test**. There is no hard and fast rule for the choice of Ŭ, it is customary to choose Ŭ equal to 0.05 or 0.01. A test is said to be significant if H_0 is rejected at Ŭ= 0.05 and is considered as highly significant if H_0 is rejected at Ŭ= 0.01.

P-value: It indicates the strength of evidence for rejecting the null hypothesis H_0, rather than simply concluding 'reject H_0' or 'do not reject H_0'. Small p-values suggest that the null hypothesis is unlikely to be true. The smaller, it is the more convincing, is the rejection of the null hypothesis.

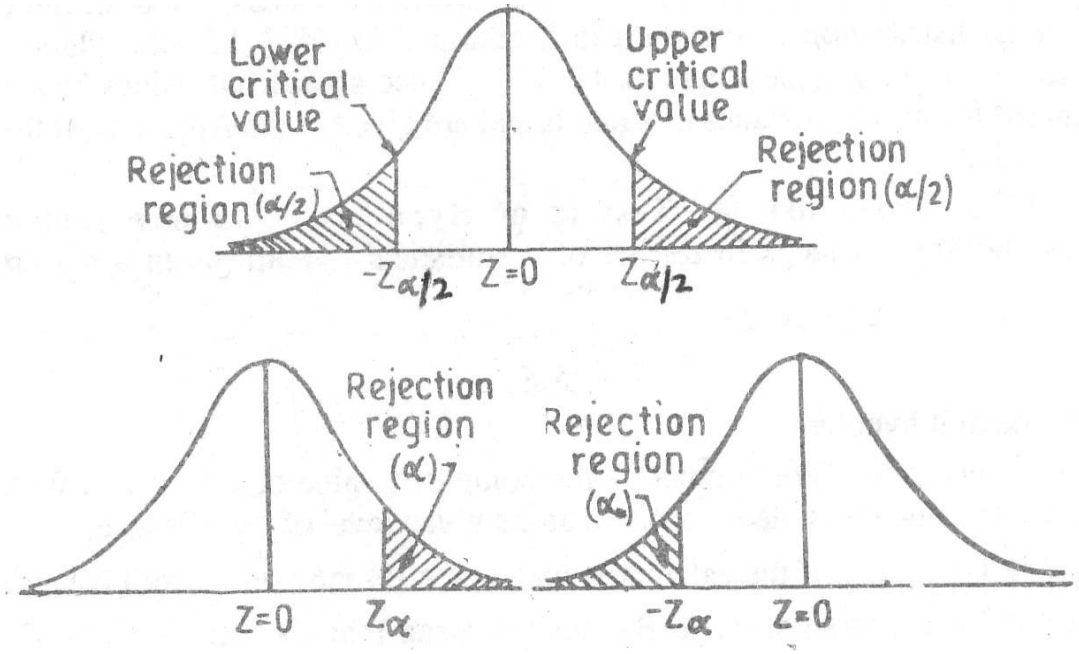
Test statistic: It is the statistic whose value is calculated from the sample data and then compared with critical or table value to decide whether to reject or accept H_0.

The procedure of testing any hypothesis consists of partitioning the total sample space in two regions. One is referred to as **region of rejection** or the **critical region** and other as region of acceptance. If the test statistic on which we base our decision falls in the critical region, then we reject H_0. If it falls in the acceptance region, we accept H_0.

One tailed and two tailed tests:

A test of any statistical hypothesis where the alternative is one sided (right sided or left sided) is called a **one tailed test**. For example, a test for testing the mean of a population H_0: $\mu = \mu_0$ against H_1: $\mu > \mu_0$ (right tailed) or H_1: $\mu < \mu_0$ (left tailed) is a one tailed test. A test of statistical hypothesis where the alternative is two sided such as: H_0: $\mu = \mu_0$ against H_1 : $\mu \neq \mu_0$ ($\mu > \mu_0$ or $\mu < \mu_0$), is known as two tailed test. The critical or table value of Z for one tailed test at level of significance α is the same as the critical value of Z for a two tailed test at level of significance 2α as shown in the figure.

Two-tailed Test



Right-tailed Test **Left-tailed Test**

4.5 One Sample Tests for Mean:

Here we will discuss tests for determining whether we should reject or accept H_0 about the population mean μ.

Case (i): Population S.D. (σ) is known (One Sample Z-test)

Let a random sample $x_1, x_{2,}$.é ., x_n of size n be drawn from a normal population whose SD(σ) is known. We want to test the null hypothesis that the population mean is equal to the specified mean μ_0 against the alternative hypothesis that the population mean is not equal to μ_0.

Assumptions:

i) Population is normal

ii) The sample is drawn at random

iii) Population S.D. (σ) is known

Procedure:

1. Formulate the null hypothesis $H_0: \mu = \mu_0$
2. Formulate the alternative hypothesis $H_1: \mu \neq \mu_0$

   Situation (i) $H_1: \mu \neq \mu_0$ (Two tailed test)

   Situation (ii) $H_1: \mu > \mu_0$ (Right tailed test)

   Situation (iii) $H_1: \mu < \mu_0$ (Left tailed test)
3. Choose the level of significance $\alpha = 0.05$ or 0.01
4. Compute the test statistic value

$$Z_{cal} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

5. **Conclusion:**

   For two tailed test if $|Z_{cal}| \geq z_{\alpha/2}$, reject H_0 otherwise accept H_0.

   For right tailed test if $Z_{cal} \geq z_{\alpha}$, reject H_0 otherwise accept H_0.

   For left tailed test if $Z_{cal} \leq - z_{\alpha}$, reject H_0 otherwise accept H_0.

Example-3: The average number of mango fruit per tree in a particular region is known from a past experience as 520 with a standard deviation 4. A sample of 20 trees gives an average number of fruit 450 per tree. Test whether the average number of fruit selected in the sample is in agreement with the average production in that region.

Solution: A stepwise solution is as follows:

1. $H_0 : \mu = \mu_0 = 520$ fruit
2. $H_1 : \mu \neq$ 520 fruit (Two tailed test)

3. $\alpha = 0.05$

4. $\bar{x} = 450$, $\mathring{u} = 4$ and $n = 20$

$$|Z_{cal}| = \left|\frac{\bar{x} - \varepsilon_0}{\mathring{u}/\sqrt{n}}\right| = \left|\frac{450 - 520}{4/\sqrt{20}}\right| = 78.26$$

Conclusion: $|Z_{cal}| > Z_{tab}$ i.e. 1.96 at $\alpha = 0.05$. Therefore, we reject H_0 and conclude that average number of fruit per tree in the sample is not in agreement with the average production in the region.

Case (ii): If the population S.D. (σ) is not known but sample size is large (say > 30). Still we can use the one sample Z-test.

Assumptions:

i) Population is normal

ii) The sample is drawn at random

iii) Population S.D. (σ) is unknown

iv) Sample size is large

Test statistic, we can use sample S.D. (s) in place of (σ), then

$$Z_{cal} = \frac{\bar{x} - \varepsilon_0}{s/\sqrt{n}}; \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Conclusion: same as in case (i).

Example-4: Certain gram variety tested on 64 plots gave an average yield as 985 kg/ha, and variance 1600 kg^2/ha. Test at 5% level of significance that the experiment agreed with the breeders claim that the average yield of the variety is 1000 kg/ha. Also construct 95% confidence interval for population mean.

Solution: Here $n = 64$, $\bar{X} = 985$ kg/ha and $s^2 = 1600$ kg^2/ha or $s = 40$ kg/ha

$H_0 : \mu_0 = 1000$ kg/ha

$H_1 : \mu_0 \neq 1000$ kg/ha

Level of significance $\alpha = 0.05$

Population variance is unknown and sample is large so, Z-test is used

$$Z_{cal} = \frac{\bar{X} - \varepsilon_0}{s/\sqrt{n}} = \frac{985 - 1000}{40/\sqrt{64}} = \frac{985 - 1000}{40/8} = \frac{985 - 1000}{5}$$

$$= \frac{-15}{5} = -3 \text{ or } \quad |Z_{cal}| = 3.0$$

because $|Z_{cal}| > 1.96$ so, we reject the null hypothesis at 5% level of significance. Hence it can be concluded that experiment does not confirm breederôs claim that average yield of variety is 1000 kg/ha.

95% confidence interval for mean (μ) = $\bar{x} \pm z_{\ddot{y}/2} \frac{s}{\sqrt{n}} = 985 \pm \frac{40}{\sqrt{64}} \times 1.96$

$= 985 \pm 5 \text{ x } 1.96 = (975.2, 994.8)$

Example-5: A sample of 121 tyres is taken from a lot. The mean life of tyres is found to be 39350 kms with a standard deviation of 3267 kms. Could the sample come from a population with mean life of 40000 kms considering $\alpha = 0.02$?

Solution:

H_0 : Mean life of tyres in the population (μ) = 40,000 kms

H_1 : $\mu \neq 40000$ kms (Two tailed test)

Since the sample size (n = 121) is large, therefore, we apply Z-test

$Z = \frac{\overline{X} - \varepsilon_0}{s/\sqrt{n}}$, since σ is not known and therefore, it is replaced by sample standard deviation (s) = 3267 kms

Thus $Z_{cal} = \frac{39350 - 40000}{3267/\sqrt{121}} = \frac{650}{297} = 2.19$

$Z_{\alpha/2} = Z_{005} = 2.33$

Since $|Z_{cal}| < 2.33$, therefore, H_0 cannot be rejected. Hence we conclude that mean life time of tyres in the population is equal to 40000 kms or the sample has been drawn a population with life time equal to 40000 kms.

Case (iii): Population SD (σ) is unknown and sample size is small i.e. < 30 (one sample t-test) this case is important in the sense that it is always feasible and less expensive to have a small sample size.

Let $X_1, X_2, .é ., X_n$ be a random sample of size n from a normal population with SD σ (unknown) and we want to test the null hypothesis that the population mean μ is

equal to a specified value μ_0 against the alternative hypothesis. The stepwise testing procedure is as follows:

Assumptions:

i) Population is normal

ii) The sample is drawn at random

iii) Population SD (σ) is unknown and sample size is small.

Procedure:

1. $H_o : \mu = \mu_0$

2. Situation (i) $H_1 : \mu \neq \mu_0$ (Two tailed test)

   Situation (ii) $H_1 : \mu > \mu_0$ (Right tailed test)

   Situation (iii) $H_1 : \mu < \mu_0$ (Left tailed test)

3. Choose the level of significance $\alpha = 0.05$ or 0.01

4. Test Statistic

   Obtain sample mean $\bar{x}$ and sample SD's

   $\bar{x} = \frac{1}{n}\sum x$ and $s^2 = \frac{1}{n-1}\sum (x_i - x)^2$ and finally compute $t = \frac{\bar{x} - \varepsilon_0}{s/\sqrt{n}}$

5. Obtain tabulated value of 't' distribution with (n-1) df at the level of significance α

6. **Conclusion:**

   For two tailed test if $|t_{cal}| \geq t_{\alpha/2, n-1}$ reject H_0 otherwise accept H_0.

   For right tailed test if $t_{cal} \geq t_{\alpha, n-1}$ reject H_0 otherwise accept H_0.

   For left tailed test if $t_{cal} \leq - t_{\alpha, n-1}$ reject H_0 otherwise accept H_0.

Example-6: The life time (000ô hours) of a random sample of 11 electric bulbs from a large consignment are 4.2, 3.6, 3.9, 3.1, 4.2, 3.8, 3.9, 4.3, 4.4, 4.6 and 4.0. Can we accept the hypothesis at 5% level of significance that the average life time is more than 3.75.

Solution: H_0 : Average life time (μ) = 3.75

i) $H_1 : \mu > 3.75$ (Right tailed test)

Test Statistic

$$t = \frac{\bar{x} - \varepsilon_0}{s/\sqrt{n}}$$

Calculation of $\bar{x}$ and s

Here n = 11; $\Sigma x = 44$ and $\bar{x} = 4$

x_i	4.2	3.6	3.9	3.1	4.2	3.8	3.9	4.3	4.4	4.6	4.0	**Total**
$x-\bar{x}$	0.2	-0.4	-0.1	-0.9	0.2	-0.2	-0.1	0.3	0.4	0.6	0	
$(x-\bar{x})^2$	0.04	0.16	.01	.81	.04	.04	.01	.09	.16	.36	0	**1.72**

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} = \sqrt{0.172} = 0.415$$

$$t_{cal} = \frac{4 - 3.75}{0.415}\sqrt{11} = 2.00$$

ii) Since $t_{cal} > t_{0.05,\ 10} = 1.812$, therefore H_0 is rejected and it can be concluded that the average life time of bulbs is more than 3750 hours.

Example-7: Ten individuals are chosen at random from a normal population and their heights are found as follows: 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71 inches, respectively. Test whether the mean height is 69.6 inches in the population (use α = 0.05). Also construct the 95% confidence interval for population mean (μ).

Solution: From the given data, we obtain n = 10, $\Sigma x = 678$, $\Sigma x^2 = 46050$, $\bar{x} = 67.8$

$H_0 : \mu_0 = 69.6$

$H_1 : \mu_0 \neq 69.6$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$$

$$= \frac{1}{9}\left[46050 - \frac{(678)^2}{10}\right] = \frac{1}{9}[46050 - 45968.4] = \frac{81.6}{9} = 9.07$$

Test statistic

$$t_{cal} = \frac{\bar{x} - \varepsilon}{\sqrt{s^2/n}} = \frac{67.8 - 69.6}{\sqrt{\frac{9.07}{10}}} = \frac{-1.8}{\sqrt{0.907}} = -1.89$$

$|t_{cal}| = 1.89$ and $t_{0.05}(9)$ df = 2.26

As $|t_{cal}|$ is less than table t-value at Ŭ= 0.05 at 9 df, the test provides no evidence against null hypothesis and we conclude that the mean of the population is 69.6ò.

95% confidence interval for population mean

$$= \bar{x} \pm t_{\ddot{U}/2} \frac{s}{\sqrt{n}} = 67.8 \pm \frac{3.01}{\sqrt{10}} \times 2.26 = 67.8 \pm \frac{3.01}{3.16} \times 2.26$$

$$= 67.8 \pm 0.95 \text{ x } 2.26 = 67.8 \pm 2.15 = (65.65, 69.95)$$

Example-8: A new feed was given to 25 animals and it was found that the average gain in weight was 7.18 kg with a standard deviation 0.45 kg in a month. Can the new feed be regarded having similar performance as that of the standard feed, which has the average gain weight 7.0 kg?

Solution:

$H_0 : \mu = 7.0$ kg

$H_1 : \mu \neq 7.0$ kg $\alpha = 0.05$ (Two tailed test)

Here $\bar{x} = 7.18$ kg and $s = 0.45$ kg and $n = 25$

$$t_{cal} = \frac{7.18 - 7.0}{0.45/\sqrt{25}} = 2.0$$ t_{tab} at $\alpha = 0.05$ with 24 d.f. = 2.06

Conclusion:

As $|t_{cal}| < t_{tab}$, we accept H_0 at 5% level of significance therefore, we conclude that new feed do not differ in performance than the existing feed.

4.6 Two Sample Tests for Means:

The Comparison of a sample mean with its hypothetical value is not a problem of frequent occurrence. A more common problem is the comparison of two population means. For example, we may wish to compare two training methods or two diets to see their effect on the increase in weight. Here we will like to test the null hypothesis whether the two population means are same (H_0: $\mu_1 = \mu_2$) against the alternative hypothesis that the two population means are different.

Case (i): Population SD's are known (Two sample Z test)

Let X_1, X_2,é .., X_{n1} and Y_1, Y_2,é ..,Y_{n2} be the two independent random samples of sizes n_1 and n_2 from the two normal populations with known standard deviations σ_1 and σ_2 and we want to test the null hypothesis $H_0 : \mu_1 = \mu_2$.

Assumptions:

i) Populations are normal.

ii) Samples are drawn independently and at random.

iii) Population SD's are known.

Stepwise procedure is as follows:

1. Formulate the null hypothesis H_0: $\mu_1 = \mu_2$

2. Formulate the alternative hypothesis that the two means are not equal

   Situation (i) $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

   Situation (ii) $H_1 : \mu_1 > \mu_2$ (Right tailed test)

   Situation (iii) $H_1 : \mu_1 < \mu_2$ (Left tailed test)

3. Choose the level of significance $\alpha = 0.05$ or 0.01

4. Test statistic

Obtain $\overline{X}$ and $\overline{Y}$ from the two independent random samples of size n_1 and n_2 respectively and compute.

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\mathring{u}_1^2}{n_1} + \frac{\mathring{u}_2^2}{n_2}}}$$

Conclusion:

For two tailed test if $|Z_{cal}| \geq z_{\alpha/2}$, reject H_0 otherwise accept H_0.

For right tailed test if $Z_{cal} \geq z_{\alpha}$, reject H_0 otherwise accept H_0.

For left tailed test if $Z_{cal} \leq - z_{\alpha}$, reject H_0 otherwise accept H_0.

Case (ii): Population SD's are unknown but the samples sizes are large (Two sample Z-test).

If the sample sizes are large, then we can replace the population SD's with corresponding sample values s_1 and s_2.

Assumptions:

i) Populations are normal.

ii) Samples are drawn independently and at random.

iii) Population SD's are unknown.

iv) Sizes of the samples are large.

Procedure is same as in case (i) except the test statistic

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ where } s_1^2 = \frac{1}{n_1}\sum (x_i - \bar{x})^2; \qquad s_1^2 = \frac{1}{n_2}\sum (y_i - \bar{y})^2$$

Conclusion: same as in case (i)

Example-9: A random sample of the heights in inches of adult males living in two different countries gave the following results

$n_1 = 640$ $\bar{X} = 67.35$ò and $s_1 = 2.56$ò

$n_2 = 160$ $\bar{Y} = 68.56$ and $s_2 = 2.52$

Test at 0.01 level of significance whether the average height of males in the two countries differ significantly?

Solution: Given

	Country – I	**Country – II**
No. of observation	640	160
Mean	67.35	68.56
s.d	2.56	2.52

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$$Z_{cal} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{67.85 - 68.56}{\sqrt{\frac{(2.56)^2}{640} + \frac{(2.52)^2}{160}}} = \frac{-0.7}{\sqrt{0.01024 + 0.03969}}$$

$$= \frac{-0.7}{\sqrt{0.04993}} = \frac{-0.7}{0.22} = -3.18$$

$|Z_{cal}| = 3.18$

$|Z_{cal}| > Z_{tab} = 2.58$, H_0 is rejected. Hence we conclude that the average heights of males in two countries are different.

Example-10: I.Q. Test of two groups of boys and girls gave the following results

Boys: $\bar{x} = 80$, SD = 10, $n_1 = 30$

Girls: $\bar{y} = 75$, SD = 13, $n_2 = 70$

Is there a significant difference in the mean scores of boys and girls at 5% level of significance?

Solution: H_0 : $\mu_1 = \mu_2$ i.e. mean scores of boys and girls is same.

H_1 : μ_1 Í μ_2 (Two tailed test)

Since both the sample sizes are large, therefore, two sample Z-test is applicable.

$$Z_{cal} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{80 - 75}{\sqrt{\frac{100}{30} + \frac{169}{70}}} = \frac{5}{\sqrt{5.75}} = 2.08$$

Since $|Z_{cal}| > Z_{\alpha/2}$ ($\alpha = 0.05$) = 1.96, therefore, H_0 is rejected and it is concluded that there is significant difference in the mean scores of boys and girls.

Example-11: A random sample of 90 birds of one breed gave on average production of 240 eggs per bird per year with a SD of 18 eggs. Another random sample of 60 birds of another breed gave an average production of 195 eggs per bird/year with a SD of 15 eggs. Is there any significant difference between the two breeds with respect to their egg production?

Solution:

Stepwise solution is as follows:

1. $H_0 : \mu_1 = \mu_2$
2. $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)
3. $\alpha = 0.05$
4. $\bar{x} = 240$ $n_1 = 90$ $s_1 = 18$

   $\bar{y} = 195$ $n_2 = 60$ $s_2 = 15$

   Given test statistic

$$Z_{cal} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{240 - 195}{\sqrt{\frac{(18)^2}{90} + \frac{(15)^2}{60}}} = 16.61$$

Conclusion:

Since $|Z_{cal}| > Z_{tab} = 1.96$ at 5% level of significance, therefore, we reject H_0 and conclude that there is a significant difference between the two breeds of birds with respect to egg production.

Case (iii): Population SD's are unknown but assumed same and sample sizes are small (Two sample t-test)

Let $X_1, X_{2,}\cdots\cdots, X_{n1}$ and $Y_1, Y_{2,}\cdots\cdots, Y_{n2}$ in be two independent random samples of sizes n_1 and n_2 (small) from two normal populations with unknown but equal standard deviations σ_1 and σ_2. Here we want to test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative that the population means are different.

Assumptions:

1. Populations are normal.
2. Samples are drawn independently and at random.
3. Population SD's are unknown but assumes to be the same.
4. Sample sizes are small

We proceed by the following steps:

1. $H_0 : \mu_1 = \mu_2$
2. Situation (i) $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

   Situation (ii) $H_1 : \mu_1 > \mu_2$ (Right tailed test)

   Situation (iii) $H_1 : \mu_1 < \mu_2$ (Left tailed test)
3. $\alpha = 0.05$ or 0.01
4. **Test statistic**

   Let $\overline{X}$, $\overline{Y}$ denote the sample means and s_1^2 and s_2^2 be sample variances , then the statistic $t = \dfrac{\overline{X} - \overline{Y}}{s}\sqrt{\dfrac{n_1 n_2}{n_1 + n_2}}$ follows Studentôs t-distribution with $(n_1 + n_2 ï 2)$ d.f.

   where $s^2 = \dfrac{(n_1 - 1)\ s_1^2 + (n_2 - 1)\ s_2^2}{n_1 + n_2 - 2}$ is called pooled variance.
5. Obtain t_{tab} at $(n_1 + n_2 ï 2)$ at α level of significance.
6. **Conclusion:**

   For two tailed test if $|t_{cal}| \geq t_{Ï2, n_1+n_2-2}$ reject H_0 otherwise accept H_0.

   For right tailed test if $t_{cal} \geq t_{Ï, n_1+n_2-2}$ reject H_0 otherwise accept H_0.

   For left tailed test if $t_{cal} \leq -\ t_{Ï, n_1+n_2-2}$ reject H_0 otherwise accept H_0.

Example-12: Two types of drugs X and Y were tried on certain persons for increasing weight, 5 persons were given drug X and 7 persons were given drug Y. The increase in weight is given below

Drug X	**:**	7	11	12	8	2	
Drug Y	:	12	10	14	17	8	10	13

Do the two drugs differ significantly with regard to their effect in increasing weight?

Solution: H_0: $\mu_1 = \mu_2$ i.e. there is no significant difference in the efficacy of two drugs.

H_1: $\mu_1 \neq \mu_2$ (Two tailed test)

Since population variances are unknown and sample sizes are small (we assume the population variances to be equal).

Therefore, applying two sample t-test

$$t_{cal} = \frac{\bar{x} - \bar{y}}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Calculation of $\bar{x}, \bar{y}$ and s

	x	x- $\bar{x}$	$(x-\bar{x})^2$	y	y- $\bar{y}$	$(y-\bar{y})^2$
	7	-1	1	12	0	0
	11	3	9	10	-2	4
	12	4	16	14	2	4
	8	0	0	17	5	25
	2	-6	36	8	-4	16
				10	-2	4
				13	1	1
Total	**40**	**0**	**62**	**84**	**0**	**54**

$\bar{x} = 40/5 = 8$; $\bar{y} = 84/7 = 12$

$$s = \sqrt{\frac{\sum (x-\bar{x})^2 + \sum (y-\bar{y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{62+54}{5+7-2}} = 3.406$$

$$t_{cal} = \frac{8-12}{3.406}\sqrt{\frac{5x7}{5+7}} = -2.0$$

Since $|t_{cal}| < t_{\alpha/2}$ ($\alpha = 0.05$, df = 10) = 2.23, therefore, we accept H_0 and conclude that there is no significant difference in the efficacy of two drugs in increasing the weight.

Example-13: An experiment was conducted to compare the effectiveness of two sources of nitrogen, namely ammonium chloride and urea, on grain yield of paddy. The results on the grain yield of paddy (kg/plot) under the two treatments are given below.

Ammonium chloride (x): 13.4, 10.9, 11.2, 11.8, 14.0, 15.3, 14.2, 12.6, 17.0, 16.2, 16.5, 15.7

Urea(y): 12.0, 11.7, 10.7, 11.2, 14.8, 14.4, 13.9, 13.7, 13.7, 16.9, 16.0, 15.6, 16.0

Which source of nitrogen is better for paddy?

Solution: $H_0 : \mu_1 = \mu_2$ The effect of the two source of nitrogen on paddy yield are same

$H_1 : \mu_1 \neq \mu_2$ The effects of two sources of nitrogen on paddy yield are not same.

Let $\alpha = 0.05$

Ammonimum Chloride

$n_1 = 12$

$\Sigma x = 168.8$

$\Sigma x^2 = 2423.72$

$\bar{x} = 14.07$

$$s_1^2 = \frac{1}{n_1 - 1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$

$$= \frac{1}{11}\left[2423.72 - \frac{(168.8)^2}{12}\right]$$

$$= \frac{1}{11}[2423.72 - 2374.45]$$

$$= \frac{1}{11}[49.27]$$

$= 4.48$

Urea

$n_2 = 12$

$\Sigma y = 166.8$

$\Sigma y^2 = 2369.09$

$\bar{y} = 13.91$

$$s_2^2 = \frac{1}{n_2 - 1}\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right]$$

$$= \frac{1}{11}\left[2369.09 - \frac{(166.9)^2}{12}\right]$$

$$= \frac{1}{11}[2369.09 - 2321.30]$$

$$= \frac{1}{11}[47.79]$$

$= 4.34$

Before applying two-sample t-test, it is required to test the equality of variability in populations, first use F-test.

$$F_{cal} = \frac{s_1^2}{s_2^2} = \frac{4.48}{4.34} = 1.03$$

Table F value at (11, 11) degrees of freedom and $\alpha = 0.05$ level of significance is 2.82.

Here $F_{cal} < F_{tab}$., F is not significant. Therefore, the variances are equal, and we can pool them. The pooled variance is

$$s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{11 \text{ x } 4.48+11 \text{ x } 4.34}{12+12-2} = \frac{11\ (4.48+4.34)}{22} = \frac{97.02}{22} = 4.41$$

$$t_{cal} = \frac{\bar{x}-\bar{y}}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{14.07-13.91}{\sqrt{4.41\left(\frac{1}{12}+\frac{1}{12}\right)}} = \frac{0.16}{\sqrt{0.735}} = \frac{0.16}{0.857} = 0.186$$

Table t-value for 22 df at $\alpha = 0.05$ is 2.074. Since $|t_{cal}|$ is less than table t-value, we accept the null hypothesis and conclude that both the sources of nitrogen have similar effect on the grain yield of paddy.

Example-14: Descriptive summary for samples obtained from two electric bulb manufacturing companies is as under:

	Company A	**Company B**
Mean life (in hours)	1234	1136
Standard deviation (in hours)	36	40
Sample size	8	7

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

Solution:

H_0 : $\mu_1 = \mu_2$ i.e. there is no significant difference in the mean life of two brands of bulbs.

H_1 : $\mu_1 \neq \mu_2$ (Two tailed test)

Applying two sample t-test

$$t = \frac{\bar{x}-\bar{y}}{s}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

$\bar{x} = 1234$; $\bar{y} = 1136$; $n_1 = 8$, $n_2 = 7$, $s_1 = 36$, $s_2 = 40$

$$s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{(8-1)(36)^2+(7-1)(40)^2}{8+7-2}} = \sqrt{\frac{9072+9600}{13}} = 37.9$$

$$t_{cal} = \frac{1234-1136}{37.9}\sqrt{\frac{8x7}{8+7}} = 4.99$$

Since $|t_{cal}| > t_{\alpha/2}$ ($\alpha = 0.05$, df = 13) = 2.16, therefore, we reject H_0. Thus bulbs of brand A should be purchased as their mean life time is significantly greater than that of B.

Example-15: For a random sample of 10 animals fed on diet A, the increase in weight in kg in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 kg and for another random sample of 12 animals of the same species fed on diet B, the increase in weights for same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 kg.

i) Test whether diet A and B differs significantly as regards the effect on increase in weight is concerned.

ii) How will we modify the testing procedure if the population variances are known to be 5 and 9 Kg^2

Solution:

1. $H_0 : \mu_1 = \mu_2$

2. $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

3. $\alpha = 0.05$

4. $$\bar{x} = \frac{\sum x}{n_1} = \frac{120}{10} = 12$$

   $$\bar{y} = \frac{\sum y}{n_2} = \frac{180}{12} = 15$$

   $(n_1-1)\, s_1^2 = \Sigma(x-x)^2 = 120$

   $(n_2-1)s_2^2 = \Sigma \quad (y-y)^2 = 314$

   $$s^2 = \frac{120+314}{10+2-2} = 21.1$$

   $$t_{cal} = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} + \frac{12-15}{\sqrt{21.1\left(\frac{1}{10}+\frac{1}{12}\right)}} = -1.6$$

5. t_{tab} at 5% level of significance with 20 d.f. is 2.086.

6. As $|t_{cal}| < 2.086$, we accept H_0 and conclude that the two diets do not differ significantly.

iii) Here the population variances are known, therefore, we can apply two sample Z-test [case (i)] where $\mathring{u}_1^2 = 5$ and $\mathring{u}_2^2 = 9$

1. $H_0 : \mu_1 = \mu_2$

2. $H_1 : \mu_1 \neq \mu_2$

3. $\alpha = 0.05$

4. Test Statistic

$$Z_{cal} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\mathring{u}_1^2}{n_1} + \frac{\mathring{u}_2^2}{n_2}}} = \frac{12 - 15}{\sqrt{\frac{5}{10} + \frac{9}{12}}} = -2.68$$

Conclusion: As $|Z_{cal}| > Z_{tab} = 1.96$ at 5% level of significance, therefore, we reject H_0 and conclude that diet A differs significantly from diet B as far as increase in weight is concerned.

Case (iv): Population variances are unknown but different

For testing the significance of the differences between two means, we have made the assumption that the variances of two populations are same. Before applying the t-test, it is desirable to test this assumption by F-test by comparing variance ratio s_1^2/s_2^2 ($s_1^2 > s_2^2$) against F distribution with (n_1-1, n_2-1) degrees of freedom. If the two variances are different then in this case we find out t_{cal} as:

$$t_{cal} = \frac{\bar{x} - \bar{y} - (\varepsilon_1 - \varepsilon_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

and compare it with t_{tab} with v d.f. where

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

Case (v) Test for the paired samples (paired t-test): In above tests, we have assumed that the two random samples are independent, but some times in practice we find that two random samples may be correlated. For instance due to the shortage of material, the experiment have to be carried out on same set of units on two different occasions or two varieties/fertilizers are tested on adjacent plots. Other examples where paired t-test may be used are to see (i) effect of coaching in securing good marks (ii) the effect of medical practice in changing the blood pressure etc.

Let there be n pairs and the observations be denoted by (x_1, y_1), (x_2, y_2), é .., (x_n, y_n) and let d_1, d_2,é .., d_n represents the differences of n related pairs of measurements, where $d_i = x_i$ ï y_i, i = 1, 2,·····,n.

Assumptions:

1. Populations are normal.
2. Samples are dependent and taken at random.
3. Population SD's are unknown but equal
4. Sizes of the samples are small.

Procedure:

1. H_0: $\mu_d = 0$ i.e. mean differences in the population are zero.
2. Situation (i) H_1: $\mu_d \neq 0$ (Two tailed test)

   Situation (ii) H_1: $\mu_d > 0$ (Right tailed test)

   Situation (iii) H_1: $\mu_d < 0$ (Left tailed test)
3. Choose the level of significance α
4. Test Statistic Computation

   Compute $\bar{d} = \frac{\sum d_i}{n}$, $s_d^2 = \frac{\sum (d_i - \bar{d})^2}{n-1}$ and find out $t_{cal} = \frac{\sqrt{n}\,\bar{d}}{s_d}$
5. Obtain t_{tab} at α level of significance with (n-1) d.f.
6. **Conclusion:**

   For two tailed test if $|t_{cal}| \geq t_{\alpha/2, n-1}$ reject H_0 otherwise accept H_0.

   For right tailed test if $t_{cal} \geq t_{\alpha, n-1}$ reject H_0 otherwise accept H_0.

   For left tailed test if $t_{cal} \leq - t_{\alpha, n-1}$ reject H_0 otherwise accept H_0.

Example-16: Eleven employees of a company were given training for accounts test. The marks obtained by them before and after training are given below:

Before Training	32	29	28	30	27	29	27	26	32	25	28
After Training	34	29	31	28	31	33	30	30	33	30	37

Test at 5% level of significance that the training has contributed towards increase in marks.

Solution: H_0: $\mu_d = 0$ i.e. the training has no effect on marks of employees

H_1: $\mu_d > 0$ (Right tailed test)

Test statistic:

$t = \frac{\bar{d}\sqrt{n}}{s_d}$, where $\bar{d}$ = mean increase in marks after training

Calculation of $\bar{d}$ and s

												Total
d	2	0	3	-2	4	4	3	4	1	5	9	$\Sigma d=33$
d^2	4	0	9	4	16	16	9	16	1	25	81	$\Sigma d^2=181$

$\bar{d} = 33/11 = 3$ and $s_d = \sqrt{\frac{1}{n-1}\left[\sum d^2 - n(\bar{d})^2\right]} = \sqrt{\frac{1}{10}(181 - 11(3)^2} = 2.863$

$t_{cal} = \frac{3\sqrt{11}}{2.863} = 3.475$

Since $t_{cal} > t_{0.05,10} = 1.812$, therefore, we reject H_0 and conclude that training has contributed significantly towards increase in marks.

Example-17: A drug is given to 10 patients, and the increments in their blood pressure were recorded to be 3, 6, -2, 4, -3, 4, 6, 0, 0, 2. Is it reasonable to believe that the drug has no effect on blood pressure? Use $\alpha = 0.05$.

Solution: H_0: $\mu_d = 0$ i.e. drug has no effect on the blood pressure of patients

H_1: μ_d Í 0 (Two tailed test)

Applying paired t-test

$t = \frac{\bar{d}\sqrt{n}}{s_d}$

Calculation of $\bar{d}$ and s_d

											Total
d	3	6	-2	4	-3	4	6	0	0	2	$\Sigma d=20$
d^2	9	36	4	16	9	16	36	0	0	4	$\Sigma d^2=130$

$\bar{d} = 20/10 = 2$

$s_d = \sqrt{\frac{1}{9}(130 - 10(2)^2} = 3.162$

$$t_{cal} = \frac{2\sqrt{10}}{3.162} = 2$$

Since $|t_{cal}| < t_{0.025,\ 9\ df} = 2.262$, therefore, we do not reject H_0 and conclude that drug has no significant effect on the blood pressure of patients.

Testing of Hypothesis about a Population Proportion:

If all possible samples of size n are drawn from a population of size N, then sample proportion (p) is distributed with mean P and variance PQ/n and for large samples (n Ó30), p is approximately normally distributed, i.e. p ~ N (P, PQ/n) where Q = 1 ï P

or $$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$$

The Z-statistic obtained above is used for testing H_0: $P = P_0$ for large samples.

Testing Hypothesis about Difference of Proportions:

Let p_1, p_2 be two sample proportions obtained from independent samples of sizes n_1 and n_2 from two populations with population proportions P_1 and P_2 respectively.

Here H_0: $P_1 = P_2$ and H_1: P_1 Í P_2

For large samples (i.e. n_1 Ó 30 and n_2 Ó 30), the distribution of p_1 ï p_2 is approximately normal

i.e. p_1 ï $p_2 \sim N(P_1$ ï $P_2, P_1Q_1/n_1 + P_2Q_2/n_2)$

or $$Z = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1)$$

Under H_0, p_1 and p_2 are independent unbiased estimators of the same parameter $P_1 = P_2 = P$. Thus we use the weighted mean of p_1 and p_2 as estimator of P. i.e.

$$p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} \text{ and } q = 1 - p$$

thus $$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

and is used for testing H_0: $P_1 = P_2$ for large samples.

Example-18: A coin is tossed 900 times and heads appears 480 times. Does this result support the hypothesis that the coin is unbiased at (i) $\alpha = 0.05$ (ii) $\alpha = 0.01$.

Solution:

H_0: Coin is unbiased i.e. the proportion of heads (P) = 0.5

H_1: P Í 0.5 (Two tailed test)

Sample proportion (p) = 480/900 = 0.533

Standard error of (p) = $\sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.5x0.5}{900}} = 0.0167$

Test statistic

$$Z_{cal} = \frac{p - P}{SE(p)} = \frac{0.533 - 0.500}{0.0167} = 1.98$$

i) Since $|Z_{cal}| > z_{\alpha/2}$ (α=0.05) = 1.96, therefore, H_0 is rejected and it is concluded that the coin is biased.

ii) Since $|Z_{cal}| < z_{\alpha/2}$ (α=0.01) = 2.58, therefore H_0 is accepted and it is concluded that the coin is unbiased.

Example-19: A sales clerk in the departmental store claims that 60% of the customers entering the store leave without buying anything. A random sample of 50 customers showed that 35 of them left without making any purchase. Test the claim of the sales clerk at 5% level of significance.

Solution:

H_0: P = 0.60

H_1: P Í 0.60 (Two tailed test)

Sample proportion (p) = 35/50 = 0.70

Standard error of (p) = $\sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.60 \text{ x } 0.4}{50}} = 0.0693$

Test statistic

$$Z_{cal} = \frac{p - P}{SE(p)} = \frac{0.70 - 0.60}{0.0693} = 1.44$$

Since $|Z_{cal}| < z_{0.025} = 1.96$, therefore, H_0 is not rejected and it supports the claim of sales clerk.

Example-20: In a random sample of 100 persons taken from a village A, 60 are found consuming tea. In another sample of 200 persons taken from village B, 100 persons are found consuming tea. Do the data reveal significant difference between the two villages so far as the habit of taking tea is concerned?

Solution: H_0: $P_1 = P_2$ Tea habit in the two villages is same

H_1: $P_1 \neq P_2$ (Two tailed test)

$n_1 = 100$ $p_1 = 60/100 = 0.6$

$n_2 = 200$ $p_2 = 100/200 = 0.5$

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = (60 + 100)/(100 + 200) = 0.53$$

$$Z_{cal} = \frac{0.6 - 0.5}{\sqrt{(0.53)(0.47)\left(\frac{1}{100} + \frac{1}{200}\right)}} = \frac{0.1}{\sqrt{(0.53)(0.47)(0.015)}} = 1.64$$

Since $|Z_{cal}| < z_{\alpha/2}$ (α=0.05) = 1.96 therefore H_0 is accepted and it is concluded that there is no significant difference in the habit of taking tea in the two villages A and B.

4.7 Chi-square (χ^2) Test and its Applications:

The Chi-square test (written as χ^2-test) is one of the simplest and most widely used non-parametric test which was given by Karl Pearson (1900).

Assumption:

i) Totals of observed and expected frequencies are same i.e. $\Sigma O_i = \Sigma E_i = N$ and $N > 50$ and

ii) No observed frequency should be less than 5. If any frequency is less than 5, then for the application of Chi-square test it to be pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjusted for the degrees of freedom lost in pooling.

Here, we shall discuss following applications of χ^2 test

i) Test of goodness of fit

ii) Test of independence

iii) Test for the population variance

iv) Test for the homogeneity of several population variances (Bartlettôs test)

Applications of Chi-square:

(1) Test of Goodness of Fit of a Distribution:

H_0 : observed and expected frequencies are in complete agreement i.e. fit is good.

H_1 : observed and expected frequencies are not in agreement.

The goodness of fit of any set of data to a probability distribution can be tested by a chi-square test. For carrying out the goodness of fit test, we calculate expected frequencies E_i corresponding to the observed frequencies O_i on the basis of given distribution i =1, 2,·····, k and compute the value of the test statistic:

$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$, which follows Chi-square distribution with (k-1) d.f. and gives the magnitude of discrepancy between expected and observed patterns.

Conclusion: Reject H_0 if χ^2_{cal} is greater than table value of χ^2 at α level of significance with (k-1) d.f. and conclude that fit is not good i.e. the observed frequencies are not in agreement with expected frequencies.

Example-21: The data relate to the distribution of printing mistakes. Fit the Poisson distribution and test for goodness of fit.

x :	0	1	2	3	4	5	6
f :	275	72	30	7	5	2	1

Solution:

H_0: Printing mistakes follow Poisson law

H_0: Printing mistakes does not follow Poisson law

The mean of the distribution $\bar{X} = \frac{\Sigma fx}{N}$ where N=Σf, determines the estimate of λ value equal to $\bar{X}$, since in Poisson distribution mean is equal to λ.

i) Compute the expected frequencies corresponding to observed frequencies by the Poisson distribution as explained in Example-14 (Chapter-3).

ii) Compute $\chi^2_{cal} = \sum_{i=1}^{7} (O_i - E_i)^2 / E_i$ and compare with tabulated value of χ^2.

X	f (Observed frequency)	fx	Expected frequency	Expected Frequency after rounding
0	275	0	242.10	242
1	72	72	116.69	117
2	30	60	28.12	28
3	7	21	4.52	5
4	5	20	0.54 =5.12	
5	2	10	0.05	
6	1	6	0.01	
Total	**392**	**189**		

$$\overline{X} = \frac{189}{392} = 0.482$$

Note that last four classes have expected frequency less than 5. Hence to maintain the continuity of χ^2 distribution, the last four classes are pooled so that the expected frequency of the last class becomes (4.52 + 0.54 + 0.05 + 0.01) = 5.12 $\simeq$ 5. In this way, the number of classes reduced to four and d.f. will be reduced to 3.

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

= (275 - 242)^2/242 + (72 ï 117)^2 /117 + (30 - 28)^2/28 + (15 ï 5)^2 /5

= 4.5 + 17.31 + 0.14 + 20 = 41.95

Tabulated value $\chi^2_{3,\,0.05} = 7.82$

Since $\chi^2_{cal} > \chi^2_{tab}$, we reject H_0 and conclude that the Poisson distribution did not fit well to the given data.

Example-22: The following figures shows the distribution of digits in numbers chosen at random from a telephone directory

Digit	0	1	2	3	4	5	6	7	8	9	**Total**
Frequency	180	200	190	230	210	160	250	220	210	150	**2000**

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

Solution: H_0: Digits occur equally frequently in the directory

H_1: Digits do not occur equally frequently in the directory i.e. fit is not good

The expected frequency (E) for each digit

0, 1, 2,·····,9 is 2000/10 = 200

We will now use χ^2 ï test of goodness of fit

$\chi^2 = \Sigma(O ï E)^2/E$

O	E	$(O ï E)^2/E$
180	200	2.0
200	200	0
190	200	0.5
230	200	4.5
210	200	0.5
160	200	8.0
250	200	12.5
220	200	2.0
200	200	0
150	200	12.5

$\chi^2_{cal} = \sum(O-E)^2/E = 42.5$

χ^2_{tab} (for k-1 = 9 df at 5% level of significance) is 16.9. Thus $\chi^2_{cal} > \chi^2_{tab}$ and thus H_0 is rejected. Thus it can be concluded that digits are not uniformly distributed in the directory.

Example-23: The following table gives the number of road accidents that occurred during the various days of the week. Find whether the accidents are uniformly distributed over the week?

Days	**Sunday**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Total**
No. of accidents	16	18	10	13	13	11	17	98

Solution: H_0 : Road accidents are uniformly distributed over the week.

H_1 : Road accidents are not uniformly distributed over the week.

Under H_0, expected frequency = 98/7 = 14

Let O and E represent the observed and expected frequencies, then

Day	**O**	**E**	**O^2**	**O^2/E**
Sunday	16	14	256	18.29
Monday	18	14	324	23.14
Tuesday	10	14	100	7.14
Wednesday	13	14	169	12.07
Thursday	13	14	169	12.07
Friday	11	14	121	8.64
Saturday	17	14	289	20.64
Total	**98**	**98**		**101.99**

$N = \Sigma O_i = \Sigma E_i = 98$

$$\chi^2_{cal} = \sum \frac{O_i^{\ 2}}{E_i} - N = 101.99 \text{ ï } 98 = 3.99$$

Table value of χ^2 at $\alpha = 0.05$ and (7-1) = 6 degree of freedom is 12.6. Since $\chi^2_{cal} < \chi^2_{tab}$ so we do not reject the null hypothesis and conclude the accidents are uniformly distributed over the week.

Example-24: In experiments on pea breeding, Mendal obtained the following frequencies of seeds : 315 round and yellow, 101 wrinkled and yellow; 108 round and green; 32 wrinkled and green, total 556. Theory predicts that the frequencies should be in the ration 9 : 3 : 3 : 1. Find χ^2 and examine correspondence between theory and experiment.

Solution: H_0: The data follow the ratio 9 : 3 : 3 : 1 (or the fit is good)

H_1: The data does not follow the ratio 9 : 3 : 3 : 1 (or the fit is not good)

Expected frequencies for group I = $\frac{9}{16} \times 556 = 313$

Expected frequencies for group II = $\frac{3}{16} \times 556 = 104$

Expected frequencies for group III = $\frac{3}{16} \times 556 = 104$

Expected frequencies for group IV = $\frac{1}{16} \times 556 = 35$

Observed and expected frequencies	**I**	**II**	**III**	**IV**	**Total**
O :	315	101	108	32	**556**
E :	313	104	104	35	**556**

$$\chi^2_{cal} = \frac{(315-313)^2}{313} + \frac{(101-104)^2}{104} + \frac{(108-104)^2}{104} + \frac{(32-35)^2}{35} = 0.51$$

The table value $\chi^2_{3,\,0.05}$ = 7.82

The calculated value is less than 7.82, hence we do not reject H_0. We conclude that there is a correspondence between the theory and experiment or the data follows the ratio 9 : 3 : 3 : 1

(ii) Test of Independence of Attributes in Contingency Tables:

Another application of the Chi-square test is in testing independence of attributes A and B in a m x n contingency table, which contains mn cell frequencies in m rows and n columns, where m and n are the categories of the attributes A and B respectively. For testing independence of row and column classifications, we define the null and alternative hypothesis as follows

H_0 : Attributes A and B are independent

H_1 : Attributes A and B are not independent

Let O_{ij} denote the observed frequency in the (i, j) cell and E_{ij} be the expected frequency under the null hypothesis.

When H_0 is true $E_{ij} = \frac{R_i \text{ x } C_j}{N}$ where R_i is the i^{th} row total C_j the j^{th} column total and N is the total frequency.

Test Statistic:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{O_{ij}^2}{E_{ij}} - N, \text{ is distributed as } \chi^2 \text{ with (m ï 1) (n ï 1) d.f.}$$

Conclusion: Reject H_0 if $\chi^2_{cal} > \chi^2_{\alpha,\,(m-1)(n-1)}$ with (m-1)(n-1) d.f. at α per cent level of significance, otherwise we do not have sufficient evidence for rejection of H_0 and hence accept H_0.

Yates' Correction:

If any cell frequency is <5, then Yatesô correction of continuity is to be applied and we get modified χ^2 as:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left[\,|\, O_{ij} - E_{ij} \,|\, - 0.5\right]^2}{E_{ij}}$$

Alternately, we can merge the nearby classes for attributes A or B or both so that no cell frequency in the modified table remains less than 5. Compute the value of χ^2 for the modified table and adjust the d.f. as per new dimensions.

Example-25: Show that the conditions at home have a bearing on the condition of the child on the basis of the following observed table:

	Condition at home		
Conditions of child	**Clean**	**Not clean**	**Total**
Clean	75	40	115
Fairly clean	35	15	50
Dirty	25	45	70
Total	**135**	**100**	**235**

Solution: H_0 : Condition of child is independent of condition at home. H_1: condition of the child depends on condition at home:

The expected frequencies are computed as follows:

E_{11} = (115x135)/235 = 66.01

E_{21} = (50x135)/235 = 28.7 and so on

Expected frequency table

	Clean	**Not clean**	**Total**
Clean	66.1	48.9	115
Fairly clean	28.7	21.3	50
Dirty	40.2	29.8	70
Total	135	100	235

$$G^2_{cal} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ji})^2}{E_{ij}}$$

$$= \frac{(75 - 66.1)^2}{66.1} + \frac{(35 - 28.7)^2}{28.7} + \frac{(25 - 40.2)^2}{40.2} + \frac{(40 - 48.9)^2}{48.9}$$

$$+ \frac{(15 - 21.3)^2}{21.3} + \frac{(45 - 29.8)^2}{29.8} = 18.95$$

The table value of χ^2 at 2 d.f. and at 5% level of significance ($G_2^2, 0.05$) = 5.99. The calculated value is more than the table value, hence the null-hypothesis is rejected. Our decision is that the condition at home has a bearing on the condition of the child.

Example-26: The data relate to the sample of married women according to their level of education and marriage adjustment score.

Level of education	**Marriage adjustment score**				
	Very low	**Low**	**High**	**Very high**	**Total**
Post Graduate	24	97	62	58	**241**
Matriculate	22	28	30	41	**121**
Illiterate	32	10	11	20	**73**
Total:	78	135	103	119	**435**

Can you say that two attributes are independent?

Solution: Here null and alternative hypotheses are

H_0 : The two attributes are independent i.e. marriage adjustment is independent of education level.

H_1 : The two attributes are associated i.e. adjustment in marriage is a function of education.

We compute the expected frequency corresponding to observed ones using formula: $E_{ij} = \frac{R_i \text{ x } C_j}{N}$ which is given as follows:

$$E_{11} = \frac{78 \text{ x } 241}{435} = 43.2 \quad E_{12} = \frac{135 \text{ x } 241}{435} = 74.8 \quad E_{13} = \frac{103 \text{ x } 241}{435} = 57.1 \quad E_{14} = \frac{119 \text{ x } 241}{435} = 65.9$$

$$E_{21} = \frac{78 \text{ x } 121}{435} = 21.7 \quad E_{22} = \frac{135 \text{ x } 121}{435} = 37.6 \quad E_{23} = \frac{103 \text{ x } 121}{435} = 28.7 \quad E_{24} = \frac{119 \text{ x } 121}{435} = 33.1$$

$$E_{31} = \frac{78 \text{ x } 73}{435} = 13.1 \quad E_{32} = \frac{135 \text{ x } 73}{435} = 22.7 \quad E_{33} = \frac{103 \text{ x } 73}{435} = 17.3 \quad E_{34} = \frac{119 \text{ x } 73}{435} = 20.1$$

Compute $\chi^2_{cal} = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{(O_{ij} - E_j)^2}{E_{ij}} = 57.26$

also χ^2_{tab} (at (4-1) (3-1) = 6 d.f.) = 12.59

Since calculated value of χ^2 is greater than tabulated value, hence we reject H_0 and say that the two attributes i.e., level of education and marriage adjustment score are related to each other. That is, higher the level of education, greater is the adjustment in marriage.

Fisher Exact Test for 2 x 2 Contingency Table:

If two attributes are divided into only two classes to form a 2 x 2 contingency table

Attribute A	**Attribute B**		**Total**
	B_1	**B_2**	
A1	a	b	a + b
A2	c	d	c + d
Total	**a + c**	**b + d**	**N = a + b + c + d**

In this case, the value of χ^2 can be calculated directly from the observed frequencies by the formula:

$$G^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{N(ad-bc)^2}{R_1R_2C_1C_2}, \text{ where } N = (a+b+c+d)$$

which follows chi-square distribution with one degree of freedom

Yates Correction for 2 x 2 Contingency Table:

If some of the cell frequencies in 2 x 2 contingency table are less than 5, the continuity of χ^2 distribution is not maintained. So, Yatesô correction should be used to remove this discrepancy. For applying Chi-square test Yates suggested that add 0.5 in the frequency which are less than 5 and add or 0.5 to the remaining cell frequencies in such a way that the marginal totals remain the same. Specially for 2 x 2 contingency table the value of χ^2 under Yatesô correction can be obtained from the formula $G^2 = \frac{N\left[\,|\,ad-bc\,|\,-N/2\right]^2}{(a+b)(c+d)(a+c)(b+d)}$ which follows χ^2ï distribution with 1 d.f.

Example-27: The following data relate to the height of fathers and their first sons at the age of 35 years.

		Height of Fathers		**Total**
Height of Sons	Tall	8	2	10
	Short	7	6	13
	Total	15	8	23

Test whether the height of sons is independent of the height of the fathers.

Solution: Since one of the cell frequency is < 5, therefore, the value of χ^2 is calculated using the formula of Yatesôcorrection i.e.

$$\chi^2 = \frac{N\left[|ad - bc| - N/2\right]^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{23\left[|48 - 14| - 11.5\right]^2}{10 \times 13 \times 15 \times 8} = 0.746$$

and $\chi^2_{0.05}$ at 1 df = 3.84

Since the calculated value of χ^2 is less than tabulated value of χ^2, we do not reject H_0 and conclude that the height of sons is independent of the height of their fathers.

Example-28: From the following data test at 5% level of significance if literacy depends upon the region.

Education	**Region**		**Total**
	Rural	**Urban**	
Literature	10	46	56
Illiterate	40	4	44
Total	**50**	**50**	**100**

Solution: H_0 : Literacy is independent of region.

H_1 : Literacy is not independent of region.

Since frequency in one cell is less than 5, so using Yatesô Corrected Formula for chi-square

$$\chi^2_{cal} = \frac{\left[|ad - bc| - \frac{N}{2}\right]^2 N}{R_1 R_2 C_1 C_2} = \frac{\left[|10 \times 4 - 40 \times 46| - \frac{100}{2}\right]^2 100}{50 \times 50 \times 56 \times 44}$$

$$\chi^2_{cal} = \frac{\left[|40 - 1840| - 50\right]^2}{25 \times 56 \times 44} = \frac{(1750)^2}{25 \times 56 \times 44} = \frac{3062500}{61600} = 49.72$$

Table χ^2_{tab} at 1 df at 5% level of significance is 3.84. Since $\chi^2_{cal} > \chi^2_{tab}$, therefore we reject the null hypothesis and conclude that literacy is dependent on region.

Example-29: From the following results regarding eye colour of mother and son, test at Ŭ = 0.05 if the colour of sonĝs eyes is associated with that of mother?

Mother's eye colour	**Son's eye colour**		
	Light blue	**Not light blue**	**Total**
Light blue	47	16	63
Not light blue	4	33	37
Total	**51**	**49**	**100**

Solution:

H_0: Colours of motherôs and sonôs are independent i.e. not associated

H_1: Colours of motherôs and sonôs are not independent (i.e. associated)

Since one cell frequency is less than 5, therefore, applying Yatesô correction for 2 x 2 contingency table, we get:

$$\chi^2_{cal} = \frac{N\left(|ad-bc|-\frac{N}{2}\right)^2}{R_1R_2C_1C_2} = \frac{100(|1551-64|-50)^2}{63x37x51x49} = 35.45$$

Since $\chi^2_{cal} > \chi^2_{0.05,1} = 3.84$, therefore, we reject H_0 and conclude that the attributes are not independent i.e. the colour of sonôs eye depend on the colour of motherôs eye.

Chi-square test for the Population Variance:

This test is used to test the null hypothesis that whether the sample has been drawn from population with the specified variance σ_0^2.

Let $x_1, x_2,$ é x_n be n independent observations from N (ε, σ^2)

1. Formulate H_0 and H_1 as:

$$H_0 : \sigma^2 = \sigma_0^2 \quad Vs \quad H_1 : \sigma^2 > \sigma_0^2$$

2. Choose Ǔ

3. Compute χ^2 statistic

If H_0 is true then:

$$\chi^2 = \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma_0^2} \text{ which follows } \chi^2 \text{ with n -1 d.f.}$$

where, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is the sample variance

4. Hence if $\chi^2_{cal} \geq \chi^2_{\alpha,n-1}$, reject H_0 and not otherwise; where $\chi^2_{\alpha,n-1}$ = tabulated value of χ^2 at 5% level of significance with n ï 1 d.f.

Approximation of χ^2-distribution for large sample size (n > 30):

If the sample size n is large (>30), then we can use Fisherôs approximation i.e. $\sqrt{2\chi^2}$ follows normal distribution with mean $\sqrt{2n-1}$ and variance 1.

i.e. $\sqrt{2\chi^2} \sim N(\sqrt{2n-1}, 1)$

$$\text{Thus } Z = \sqrt{2\chi^2} - (\sqrt{2n-1}) \sim N(0,1)$$

And usual Z (standard normal) test can be applied.

Example-30: The variability in the yield of a crop variety by the conventional method (measures in terms of standard deviation) for a random sample of size 30 was 2.8 qha^{-1}. Can it be concluded at $\alpha = 0.05$ that it is not more than that of standard method which is believed to be equal to 2.2 q ha^{-1}.

Suppose:

1. $H_0 : \sigma = 2.2$ q ha^{-1}
2. $H_1 : \sigma > 2.2$ q ha^{-1}
3. $\alpha = 0.05$
4. Test statistic: Here s = 2.8 $\sigma_0 = 2.2$ n = 30

$$\chi^2_{cal} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{29(2.8)^2}{(2.2)^2} = \frac{227.36}{4.84} = 46.98$$

Since $\chi^2_{cal} > \chi^2_{tab}$ (at $\alpha = 0.05$) for 29 d.f. = 42.56, therefore H_0 is rejected. Thus, it is concluded that variability in yields by conventional method is more than that of standard method.

By using Fisherôs approximation for large sample size:

Test Statistic:

$$Z_{cal} = \sqrt{2\chi^2} - (\sqrt{2n-1})$$

$$= \sqrt{2 \text{ x } 46.98} - (\sqrt{60-1})$$

= 9.693 ï 7.681 = 2.012

Since $Z_{cal} > Z_{tab} = 1.645$, therefore H_0 is rejected.

Snedcor's F-test: It is used

i) As a test for the equality of two population variances i.e. whether the two samples may be regarded as drawn from the normal populations having the same variance.

ii) As a test for the equality of several population means.

Testing the equality of two population variances:

The F-test may be used to test the equality of two population variances. Let x_1, x_2, é ..xn_1 and y_1, y_2,·····,yn_2 be two independent samples drawn randomly from two normal

populations with variances σ_1^2 and σ_2^2. Let s_1^2 and s_2^2 be the estimates of population variances. We want to test the null hypothesis that the population variances are equal.

Assumptions:

i) Populations are normal.

ii) Samples are drawn independently and at random

Stepwise testing procedure is as follows

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

2. $H_1: \sigma_1^2 > \sigma^2_2$

3. Choose level of significant $\alpha = 0.05$ or 0.01

4. Test statistic $F_{cal} = \frac{s_x^2}{s_y^2}$ where $s_x^2 = \frac{1}{n_1 - 1}\sum\left(x_i - \bar{x}\right)^2$ $s_y^2 = \frac{1}{n_2 - 1}\sum(y_i - \bar{y})^2$ are unbiased estimators of σ_1^2 and σ_2^2 respectively. Here numerator corresponds to greater variance

If F_{cal}>F-tabulated at ($v_1 = n_1$-1, $v_2 = n_2$-1) d.f. and at α level of significance, then we reject H_0 and conclude that the population variances are significantly different, otherwise we accept H_0.

Example-31: Test the assumption for equality of two population variances in the example 15 on two sample t-test.

Solution: $n_1 = 10$ $\bar{x} = 12$ $s_1^2 = \frac{1}{n_1 - 1}\sum(x_i - \bar{x})^2 = 120/9 = 13.33$

$n_2 = 12$ $\bar{y} = 15$ $s_2^2 = \frac{1}{n_2 - 1}\sum(y_i - \bar{y})^2 = 314/11 = 28.55$

$H_0: \sigma_1^2 = \sigma_2^2$; $H_1: \sigma_2^2 > \sigma_1^2$ (since the larger variance is in numerator)

$$F_{cal} = \frac{s_2^2}{s_1^2} = \frac{28.55}{13.33} = 2.14$$

Since $F_{cal} < F_{(11, 9)} = 3.59$ at $\alpha = 0.05$, therefore, we do not reject H_0 and conclude that population variances are equal. Thus, the usual two sample t-test can be applied.

Example-32: The life times for random samples of batteries (type A and B) were recorded and the following results were obtained.

Type of Battery	No. of Batteries	Mean life (hours)	Sum of squares of deviations from mean
A	10	500	1800
B	12	555	2160

Test if there is any significant difference between the life times of two types of batteries.

Solution: Equality of means will be tested by applying two sample t-test where we assume that $\sigma_1^2 = \sigma_2^2$, therefore, we first apply F-test for the equality of two population variances.

1. $H_0 : \sigma_1^2 = \sigma_2^2$
2. $H_1 : \sigma_1^2 > \sigma_2^2$
3. $\alpha = 0.05$ or 0.01
4. Test statistic $F_{cal} = \frac{s_1^2}{s_2^2}$

$n_1 = 10,\ \bar{x} = 500 \qquad \Sigma (x_i - \bar{x})^2 = 1800 \qquad s_1^2 = \frac{1800}{9} = 200$

$n_2 = 12,\ \bar{y} = 555 \qquad \Sigma (y_i - \bar{y})^2 = 2160 \qquad s_1^2 = \frac{2160}{11} = 196.3$

$$Fcal = \frac{200}{196.3} = 1.02$$

Tabulated $F_{(9, 11)}$ at $\alpha = 0.05$ is equal to 3.92

Since $F_{cal} < F_{tab}$, therefore, we do not reject H_0 and conclude that the population variances are not significantly different.

Two sample t-test: After testing the null hypothesis of equality of population variances, we now apply the t-test.

1. $H_0 : \mu_1 = \mu_2$
2. $H_0 : \mu_1 \neq \mu_2$
3. $\alpha = 0.05$ or 0.01
4. Test Statistic

$$t_{cal} = \frac{\bar{x} - \bar{y}}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x_i - \bar{x})^2 + (y_i - \bar{y})^2\right] = \frac{1}{20}[1800 + 2160] = 198$$

$$t_{cal} = \frac{500 - 555}{\sqrt{198}} \sqrt{\frac{10 \times 12}{10 + 12}} = -21.32 \qquad \Rightarrow |t_{cal}| = 21.32$$

Conclusion: Since $|t_{cal}| > t_{0.025,\ 20} = 2.08$ therefore we reject H_0 and conclude that there is significant difference in the life times of two types of batteries.

Example-33: In a test given to two groups of students, the marks obtained are as follows:

First Group	18	20	36	50	49	36	34	49	41
Second Group	36	31	29	35	37	30	40		

Examine the significance of difference in the mean marks secured by students of two groups.

Solution:

$H_0 : \mu_1 = \mu_2$ There is no significant difference between the mean marks

$H_1 : \mu_1 \neq \mu_2$

$\alpha = 0.05$

Calculation of $\bar{X}, \bar{Y}$ and s_1, s_2

First group X	**$(X - \bar{X})$ = X – 37**	**$(X - \bar{X})^2$**	**Second Group X**	**$(Y - \bar{Y})$ = Y – 34**	**$(Y - \bar{Y})^2$**
18	-19	361	36	2	4
20	-17	289	31	-3	9
36	-1	1	29	-5	25
50	13	169	35	1	1
49	12	144	37	3	9
36	-1	1	30	-4	16
34	-3	9	40	6	36
49	12	144			
41	4	16			
333	0	1134	238	0	100

$$\overline{X} = \frac{333}{9} = 37 \qquad \overline{Y} = \frac{238}{7} = 34$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum (X_i - \overline{X})^2 = \frac{1134}{8} = 141.75 \qquad s_2^2 = \frac{1}{n_2 - 1}\sum (Y_i - \overline{Y})^2 = \frac{100}{6} = 16.66$$

For testing the equality of population variances, use F-test:

$$F_{cal} = \frac{s_1^2}{s_2^2} = \frac{141.75}{16.66} = 8.51$$

F_{tab} for (8, 6) d.f. at Ŭ= 0.05 = 3.58

Here $F_{cal} > F_{tab}$, therefore, we conclude that population variances are significantly different.

Remark: Since population variances $\mathring{u}_1^2$ and $\mathring{u}_2^2$ are unknown and differ significantly, therefore, it falls under case (iv).

Test Statistic: $$t_{cal} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{37 - 34}{\sqrt{\frac{141.75}{9} + \frac{16.66}{7}}} = \frac{3}{\sqrt{18.13}} = 0.704$$

$$\text{d.f.} = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_1^2/n_1\right)^2}{n_2 - 1}} = \frac{(18.13)^2}{31.01 + 0.94} = \frac{328.70}{31.95} = 10 \text{ (after rounding off)}$$

t_{tab} for 10 d.f. at Ŭ= 0.05 (Two tailed test) = 2.228

Here $|t_{cal}| < t_{tab}$, thus H_0 is not rejected and it is concluded that mean marks secured by two groups do not differ significantly.

Testing the equality of several Population Means:

The F-test can also be used to test the equality of several population means in the analysis of variance technique. The F-test has wider application as it provides an overall test for the equality of several population means where as t-test may be used to test the equality of only two population means.

Here we shall discuss ANOVA for i) Completely Randomized Design ii) Randomized Block Design. The null hypotheses we want to test here

H_0 : All treatment effects (means) are equal

H_1 : Atleast two treatments means differ

Completely Randomized Design (CRD):

It is a simplest design that uses only two basic principles of experimental design. In this design, total experimental area is divided into a number of experimental units and the treatments are allotted to units entirely at random. CRD is used when there is homogenous experimental material e.g. in agriculture field experiments when all the plots have same soil fertility, soil texture and uniform agronomical practices and in animal science experiments when the animals are of same breed, grown in similar conditions etc. CRD is used in laboratory, pot and green house experiments etc.

It has easy layout, its analysis is the simplest one, it has flexibility with respect to number of treatments and number of replications, it allows maximum number of degree of freedom in error. In CRD missing data can be handled easily. It is suited for only small number of treatments because large number of treatments needs large material in which variation increases, so we opt for other designs.

Layout:

The layout for CRD is the simplest one. The whole experimental area is divided into a no. of units $N = \sum_{i=1}^{t} r_i$ and all the treatments are allotted randomly to all the units. Assign t treatments randomly to N units such that T_i is allotted to r_i units. Suppose we have 5 treatments T_1, T_2, T_3, T_4 and T_5 with replication 4, 3, 4, 4 and 5 respectively. The whole experimental area is to be divided into

$\sum r_i = 4 + 3 + 4 + 4 + 5 = 20$ plots

T_1	T_4	T_2	T_5	T_3
T_3	T_5	T_1	T_3	T_4
T_2	T_5	T_3	T_2	T_1
T_1	T_4	T_5	T_4	T_5

Model for CRD

Let Y_{ij} is jth unit in ith treatment.

$Y_{ij} = \mu + t_i + e_{ij}$

where μ = General Mean

$t_i = i^{th}$ treatment effect

e_{ij} = random error ~ $N(0, \sigma^2)$

For the analysis purpose the data is written systematically

	Treatments					
Replications	1	2	é é é é é .	i	é é é .	t
	Y_{11}	Y_{21}	é é é é ..	Y_{i1}	é é é	Y_{t1}
	Y_{12}	Y_{22}	é é é é ..	Y_{i2}	é é é	Y_{t2}
	.					
	.					
	Y_{1r_1}	Y_{1r_2}	é é é é ..	Y_{ir_i}	é é é	Y_{tr_t}
Total	$\mathbf{T_1}$	$\mathbf{T_2}$	…….	$\mathbf{T_i}$	……..	$\mathbf{T_t}$
Mean	$\bar{Y}_1$	$\bar{Y}_2$	…….	$\bar{Y}_3$	……	$\bar{Y}_t$

$N = \sum_{i=1}^{t} r_i$ Let $G = \sum_i T_i = \sum_i \sum_j y_{ij}$ be the grand total.

i) Correction factor = G^2/N

ii) Total sum of squares = $\sum\sum Y_{ij}^2$ ï CF

iii) Treatment sum of squares = $\sum \frac{T_i^2}{r_i} - CF$

iv) Error sum of squares = Total SS ï treatments SS

ANOVA

Source	d.f.	SS	M.S.S.	F_{cal}
Treatments	t-1	$\sum \frac{T_i^2}{r_i} - CF = SS_T$	$SS_T/(t-1) = T$	T/E
Error	N-t	Total SS- Treat. SS = SS_E	$SS_E/(N-t) = E$	
Total	N-1	$\sum\sum Y_{ij}^2$ ï CF = Total SS		

It the F_{cal} value is greater than F_{tab} at (t-1, N-t) d.f. and at given α, we conclude that the treatments are significantly different.

SE (mean of T_i) = $\sqrt{\frac{E}{r_i}}$

SE (d) = $\sqrt{E\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$ where r_i and r_j are replications of T_i and T_j. If $r_i = r_j = r$

SE (d) = $\sqrt{\frac{2E}{r}}$ we go for CD only when the treatment effects are significant

CD = SE x $t_{\text{error d.f. i .e. (N-t) d.f.}}$

If $\left|\overline{Y}_i - \overline{Y}_j\right| \geq$ C.D. we say T_i is significant different from T_j.

Example-34: Given below are the weight gain of baby chicks (gms) under 4 different feeds, analyze the data using CRD.

	Treatments				
Observations	**t_1**	**t_2**	**t_3**	**t_4**	
1	55	61	42	169	
2	49	112	97	137	
3	42	30	91	169	
4	21	89	95	85	
5	52	63	92	154	
Total	**219**	**355**	**407**	**714**	**1695**
Mean	**43.8**	**71**	**81.4**	**142.8**	

CF = $\frac{G^2}{N} = \frac{(1695)^2}{20} = 143651$ General Mean (GM) = $\frac{1695}{20} = 84.75$

Total SS = $\sum\sum Y_{ij}^2$ ï CF = $55^2 + 61^2 +$ é é .. 154^2 ï CF = 37794

SS due to treatments = $\frac{(219)^2 + (355)^2 + (407)^2 + (714)^2}{5} - \text{CF} = 26235.2$

Error SS = Total SS ï treat SS = 37794.0 ï 26235.2 = 11558.8

ANOVA

S.V.	d.f.	SS	MSS	F_{cal}
Treatment	3	26235.2	8745.1	12.1*
Error	16	11558.8	722.4	
Total	19	37794.0		

$F_{3,16(0.05)} = 3.25$

Since $F_{cal} \geq F_{tab}$ so treatment effects are significantly different

CD 5% = $\sqrt{\frac{2E}{r}}$ x $t_{16,\,0.025}$ = $\sqrt{\frac{2x722.4}{4}}$ x 2.12 = 40.3

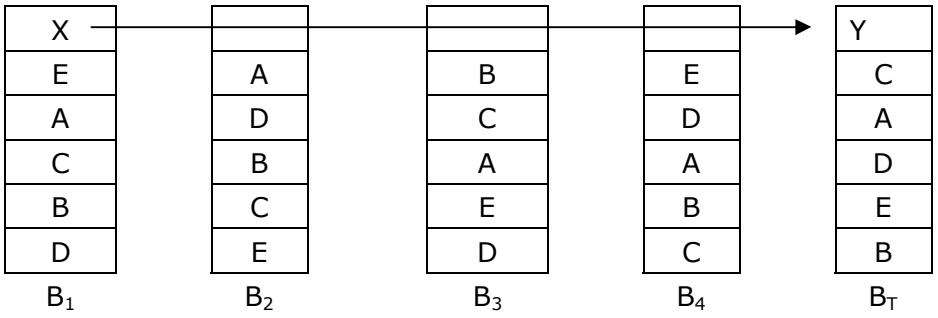
C.V. (Coefficient of Variation) = $\frac{\sqrt{E}}{G.M.}$ x 100 = $\frac{\sqrt{722}}{84.75}$ x 100 = 31.7%

Conclusion: t_1, t_2 and t_3 do not differ significantly in respect of weight gain but these treatments differ significantly from treatment t_4.

Randomized Block Design (R.B.D.)

If the whole experimental area is not homogeneous and the fertility gradient is in one direction only, then it is possible to divide the whole area into homogeneous blocks perpendicular to the direction of fertility gradient. The treatments are randomly allocated separately to each of these blocks, and the result is a randomized block design.

Layout of R.B.D.: The plots within block must be homogeneous as far as possible. Thus, if the direction called fertility gradient in which fertility changes are maximum is known, we proceed as follows.



Suppose the fertility of the field might be having a slope from X to Y, it would be advantageous to place the blocks one after another along the gradient XY

1. Whole experimental material is divided into blocks or groups such that each treatment occurs once and only once in each block. The number of blocks to be formed is equal to the number of replications.
2. Each of these groups is further divided into a number of experimental units (plots). The number of plots within a block should be equal to the number of treatments.
3. The number of treatments within each block are applied by a random procedure.

Why Randomized Block Design is used:

1. Sensitiveness: This design removes the variation between the blocks and from that within blocks which generally results in decrease of experimental error and thus sensitivity is increased. Cochran has shown that experimental error of a R.B.D. is 60 per cent of a C.R.D.

2. Flexibility: This design allows any number of treatments and replications and the only restriction is that number of replications is equal to the number of blocks.

3. Ease of analysis: The statistical analysis is easy even in the case of missing values.

Demerits:

i) It cannot control the variation in the experimental material from two sources and in such cases is not an efficient design.

ii) If the number of treatments is large then size of blocks will increase and thus heterogeneity within the blocks will increase.

Analysis:

For analysis we use the linear additive model

$Y_{ij} = \mu + t_i + b_j + e_{ij}$ where Y_{ij} is the value of the unit for the i^{th} treatment in the j^{th} block (i= 1, 2, é .., t; j = 1, 2,é ., r)

μ is the general mean effect, t_i is the effect due to i^{th} treatment, b_j is the effect due to j^{th} block, e_{ij} is random error which is assumed to be independently and normally distributed with mean zero and variance σ_e^2.

Let there be t treatments, each treatment being replicated r times (equal to number of blocks)

Let $T_i = \sum_j Y_{ij}$; $R_j = \sum_i Y_{ij}$

Treatments/Blocks	1	2	r	Totals
1	Y_{11}	Y_{12} - - - - - - - - - -	Y_{2r}	T_1
2 . . . t	Y_{21} Y_{t1}	Y_{22} - - - - - - - - - - Y_{t2} - - - - - - - - - -	Y_{22} Y_{tr}	T_2 T_t
	R_1	R_2 - - - - - - - - - -	R_r	G

$$\text{C.F.} = \frac{(GT)^2}{N} = \frac{G^2}{rt}$$

$$\text{Total S.S.} = \sum_i \sum_j Y_{ij}^2 - \text{C.F.} = S \text{ (Say)}$$

Sum of squares due to treatments = $\sum_i \frac{T_i^2}{r} - C.F. = S_1$

Sum of square due to blocks = $\sum_j \frac{R_j^2}{t} - C.F. = S_2$

S.S. due to error = Total S.S. ï S.S. due to treatments ï S.S. due to blocks

ANOVA

Source	d.f.	SS	MSS	F_{cal}
Blocks	(r-1)	S_1	B	$F_{(r-1)(r-1)(t-1)} = B/E$
Treatment	(t-1)	S_2	T	$F_{(t-1)(r-1)(t-1)} = T/E$
Error	(r-1)(t-1)	S_3	E	-
Total	rt-1	S		

$H_0 : \mu_1 = \mu_2$ é .. = μ_t

Against alternative H_1 that treatment means are not equal. If F_{cal} (treatments) come out to be significant at a specified Ŭ, then we compare the treatment means with the C.D.

$SE_m = \sqrt{\frac{E}{r}}$

$SE_d = \sqrt{\frac{2E}{r}}$

C.D. at Ŭ= 0.005 = SE(d) x t value at error d.f. at 5%; C.D. 1% = SE(d) x t value at error d.f. at 1%

Example-35: Five varieties of cotton A, B, C, D and E were tried in RBD with five replications and following yields were obtained

	B	E	C	A	D
Block-1	6.87	4.82	7.87	5.94	9.60
	E	D	B	C	A
Block-2	I6.66	8.46	8.91	6.69	6.84
	C	A	D	B	E
Block-3	6.65	8.02	6.78	8.44	5.32
	A	C	E	D	B
Block-4	6.65	8.02	6.78	8.44	5.32
	D	B	A	E	C
Block-5	5.27	8.95	6.12	4.46	5.98

Analyze the data and draw the conclusions.

Solution:

Replications / Treatments	R_1	R_2	R_3	R_4	R_5	Total	Mean
A	5.94	6.84	8.02	7.24	6.12	34.16	6.83
B	6.87	8.91	8.44	7.59	8.98	40.79	8.16
C	7.87	6.69	6.65	7.98	5.98	35.17	7.03
D	9.60	8.46	6.78	7.50	5.27	37.61	7.52
E	4.82	6.66	5.32	5.79	4.46	27.05	5.41
Total:	**35.20**	**37.56**	**35.21**	**36.10**	**30.81**	**174.88**	

G = 174.88, r = 5 and t = 5

C.F. = $(174.88)^2/25$ = 1223.06

Total sum of squares = $5.94^2 + 6.84^2 +$ é é é é é é é é é . 4.46^2 ï C.F. = 42.072

$$\text{Block S.S.} = \frac{R_1^2 + R_5^2 + \ldots\ldots\ldots + R_5^2}{5} - \text{C.F.}$$

$$= \frac{35.20^2 + 37.66^2 + \ldots\ldots\ldots + 30.81^2}{5} - 1223.06 = 4.135$$

$$\text{Treat S.S.} = \frac{T_1^2 + T_5^2 + \ldots\ldots\ldots + T_5^2}{5} - \text{C.F.}$$

$$= \frac{34.16^2 + 40.79^2 + \ldots\ldots\ldots + 27.05^2}{5} - 1223.06 = 21.547$$

Error S.S. = Total S.S. ï Block S.S. ï Treat S.S. = 16.380

ANOVA

Source	d.f.	SS	MSS	F_{cal}
Blocks	4	4.135		
Treatment	4	21.457	5.387	4.90
Error	16	16.380	1.026	
Total	**24**	**42.072**		

SE (m) = $\sqrt{\frac{1.026}{5}}$ = 0.453; SE (d) = $\sqrt{\frac{2\,x\,1.026}{5}}$ = 0.640

$CD_{5\%}$ = SE(d) x t_{16} at Ŭ= 0.05 = 0.640 x 2.120 = 1.357

$CD_{5\%}$ = SE(d) x t_{16} at Ŭ= 0.01 = 0.640 x 2.721 = 1.870

Bartlett's Test of Homogeneity of Variances:

Sometimes the question arises whether the two or more variances obtained from different samples differ significantly from one another or not. In case of two variances, the answer can be obtained by the F test. But in case of more than two variances, Bartlettôs test of homogeneity is adequate.

Let the number of samples be k and their variances are s_1^2, s_2^2,é .,s_k^2 and their corresponding d.f. is $ɜ_1, ɜ_2, ɜ_3$,é .., $ɜ_k$ ($ɜ_i = n_i$ ï 1) where n_i is the size of i^{th} sample.

Now the following steps are needed for the test:

i) Calculate the total degrees of freedom

$N = ɜ_1 + ɜ_2 + ɜ_3$ é .. + $ɜ_k$

ii) Calculate $\bar{s}^2$, weighted average of the variances:

$$\bar{s}^2 = \frac{ɜ_1 s_1^2 + ɜ_2 s_2^2 + \ldots\ldots + ɜ_k s_k^2}{ɜ_1 + ɜ_2 + \ldots\ldots + ɜ_k} = \frac{\sum_{i=1}^{k} \nu_i s_i^2}{N} \qquad N = \sum_{i=1}^{k} \nu_i$$

iii) Calculate $ɢ^{2'}$ and C, the correction factor:

$$ɢ^{2'} = N \log_e \bar{s}^2 - \sum_{i=1}^{k} ɜ_i \log_e s_i^2$$

$$\text{and } C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\left(\frac{1}{\nu_1}\right) - \frac{1}{N}\right]$$

iv) Test Statistic

Calculate $ɢ^2 = ɢ^{2N}/C$ which follows $ɢ^2$ distribution with (k-1) d.f.

Conclusion:

If $ɢ^2_{cal} \geq ɢ^2_{tab}$ with (k-1) d.f. at a specified Ŭ, then we reject H_0 and conclude that population variances are not homogeneous.

Example-36: The sample variances 1.27, 2.58 and 3.75 based on 9, 13 and 12 degrees of freedom are obtained from three different samples. Apply Bartlettôs test for testing the homogeneity of three population variances.

Sample	**Sample Variances s^2**	**Degree of Freedom ν**	$\frac{1}{\nu}$	**νs^2**	**$\log_e s^2 = 2.3026 \log_{10} s^2$**	**$\nu \log_e s^2$**
1	1.27	9	0.11111	11.43	0.2390	2.1510
2	2.58	13	0.07692	33.54	0.9478	12.3214
3	3.75	12	0.08333	45.00	1.3218	15.8616
Total	**7.60**	**N=34**	**0.27136**	**89.97**	**2.5086**	**30.3340**

Here $\bar{s}^2 = \frac{\sum \nu_i s_i^2}{N} = \frac{89.97}{34} = 2.646$ $\qquad N = \sum_{i=1}^{k} \nu_1$

And $\chi^{2Nj} = N \log_e s^{-2} - \sum \nu_i \log_e s_i^2$

$= 34 \log_e 2.646 - 30.3340 = 2.7514$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum\left(\frac{1}{\nu_1}\right) - \frac{1}{N}\right] = 1 + \frac{1}{3 \times 2}[0.27136 - 0.02941] = 1.04033$$

$$\therefore \text{Corrected } \chi^2_{cal} = \frac{\chi'^2}{C} = \frac{2.7514}{1.04033} = 2.645$$

χ^2_{tab} (with k - 1) = 2 df at Ŭ= 0.05 = 5.991

Since the $\chi^2_{cal} < \chi^2_{tab}$, hence we conclude that the population variances are homogenous.

Example-37: Certain gram variety tested on 64 plots gave an average yield as 985 kg/ha, and variance 1600 kg^2/ha. Test at 5% level of significance that the experiment agreed with the breeders claim that the average yield of the variety is 1000 kg/ha. Also construct 95% confidence interval for population mean.

Solution: Here n = 64, $\bar{X}$ = 985 kg/ha and s^2 = 1600 kg^2/ha or s = 40 kg/ha

$H_0 : \mu_0 = 1000$ kg/ha

$H_1 : \mu_0 \neq 1000$ kg/ha

Level of significance $\alpha = 0.05$

Population variance is unknown and sample is large so, Z-test is used

$$Z_{cal} = \frac{\overline{X} - \varepsilon_0}{s/\sqrt{n}} = \frac{985-1000}{40/\sqrt{64}} = \frac{985-1000}{40/8} = \frac{985-1000}{5}$$

$$= \frac{-15}{5} = -3 \text{ or } \quad |Z_{cal}| = 3.0$$

because $|Z_{cal}| > 1.96$ so, we reject the null hypothesis at 5% level of significance. Hence it can be concluded that experiment does not confirm breederôs claim that average yield of variety is 1000 kg/ha.

95% confidence interval for mean (μ) = $\overline{x} \pm z_{\ddot{y}/2} \frac{s}{\sqrt{n}} = 985 \pm \frac{40}{\sqrt{64}} \times 1.96$

$= 985 \pm 5 \text{ x } 1.96 = (975.2, 994.8)$

Summary Table for Various Tests of Hypotheses

H_o	Test Statistic	H_1	Critical Region
$\mu = \mu_o$	$Z = \dfrac{\overline{X} - \varepsilon_0}{\sigma/\sqrt{n}}$; σ known	$\mu \neq \mu_o$ $\mu > \mu_o$ $\mu < \mu_o$	$\lvert Z\rvert \geq z_{\alpha/2}$ $Z > z_{\alpha}$ $Z < -z_{\alpha}$
$\mu = \mu_o$	$t = \dfrac{\overline{X} - \varepsilon_0}{s/\sqrt{n}}$; $\nu = n - 1$, σ known	$\mu \neq \mu_o$ $\mu > \mu_o$ $\mu < \mu_o$	$\lvert t\rvert \geq t_{\alpha/2,\, n-1}$ $t > t_{\alpha,\, n-1}$ $t < -t_{\alpha,\, n-1}$
$\mu_1 - \mu_2 = d_o$	$Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - d_0}{\sqrt{(\sigma_1^2 n_1)} + \sqrt{(\sigma_2^2 n_2)}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 \neq d_o$ $\mu_1 - \mu_2 < d_o$ $\mu_1 - \mu_2 > d_o$	$\lvert Z\rvert \geq z_{\alpha/2}$ $Z > z_{\alpha}$ $Z < -z_{\alpha}$
$\mu_1 - \mu_2 = d_o$	$t = \dfrac{(\overline{X}_1 - \overline{X}_2) - d_0}{s_p\sqrt{(1/n_1)} + \sqrt{(1/n_2)}}$; $\nu = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$;	$\mu_1 - \mu_2 \neq d_o$ $\mu_1 - \mu_2 < d_o$ $\mu_1 - \mu_2 > d_o$	$\lvert t\rvert \geq t_{\alpha/2,\, n_1+n_2-2}$ $t < -t_{\alpha} n_1+n_2-2$ $t > t_{\alpha,\, n_1+n_2-2}$
$\mu_1 - \mu_2 = d_o$	$t' = \dfrac{(\overline{X}_1 - \overline{X}_2) - d_0}{\sqrt{(s_1^2 n_1)} + \sqrt{(s_2^2 n_2)}}$; $\nu = \dfrac{(s_1^2/n_1 + (s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$; $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 \neq d_o$ $\mu_1 - \mu_2 < d_o$ $\mu_1 - \mu_2 > d_o$	$\lvert t\rvert \geq t_{\alpha/2}$ $t > t_{\alpha}$ $t < -t_{\alpha}$
$\mu_D = d_o$	$t = \dfrac{\overline{D} - d_0}{sd/\sqrt{n}}$; $\nu = n - 1$, paired observations	$\mu_D \neq d_o$ $\mu_D > d_o$ $\mu_D < d_o$	$\lvert t\rvert \geq t_{\alpha/2,\, n-1}$ $t > t_{\alpha,\, n-1}$ $t < -t_{\alpha,\, n-1}$
$\sigma^2 = \sigma_0^2$	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$; $\nu = n - 1$	$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi_{\alpha}^2$
$\sigma_1^2 = \sigma_2^2$	$F = \dfrac{s_1^2}{s_2^2}$; $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$	$\sigma_1^2 > \sigma_2^2$	$F > F_{\alpha(\nu_1, \nu_2)}$

EXERCISES

1. A sample of 400 male adultôs from Haryana is found to have a mean height of 171.38 cms. Can it be reasonably regarded as a sample from a large population of mean height 171.17 cms and standard deviation of 3.30 cms? [Hint: use one sample Z-test]

2. Ten specimens of copper wires drawn from a large lot have the breaking strengths (in kg. wt) equal to 578, 572, 570, 568, 512, 578, 570, 572, 569, 548. Test whether the mean breaking strength of the lot may be taken to be 578 kg wt. [Hint: use one sample t-test]

3. A manufacturer claimed that atleast 90% of the tools which he supplied were upto the standard quality. A random sample of 200 tools showed that only 164 were upto the standard. Test his claim at 1% level of significance. [Hint: use Z-test for single proportion]

4. You are working as a purchase manager for a company. The following information has been supplied to you by two manufacturers of electric bulbs

	Company A	**Company B**
Mean life (in hours)	1300	1250
Standard deviation (in hours)	82	93
Sample size	100	80

Is brand A of bulbs is superior in respect to higher mean life at a risk of 5%.

[Hint: Since sample size are large, therefore, use two samples Z-test].

5. A company is interested to know if there is any difference in the average salary received by the managers of two divisions. Accordingly samples of 12 managers in the first division and 10 in the second division were selected at random and results are given below:

	First Division	**Second Division**
Sample size	12	10
Average monthly salary (Rs.)	25000	22400
Standard deviation (Rs)	640	960

Apply two sample t-test to find out whether there is a significant difference in the average salary.

6. Given below is the contingency table for production in three shifts the number of defective goods turn over. Use chi-square test to test whether the number of defective goods depends on the shift run by the factory.

No. of Defective goods			
Shifts	**1st week**	**2nd week**	**3rd week**
1	15	5	20
2	20	10	20
3	25	15	

7. Ten individuals are chosen at random from a population and their heights are found to be in inches,

   64, 65, 65, 66, 68, 69, 70, 70, 71, 71. In the light of these data, discuss the suggestion that the mean height in the population is 66 inches (Hint: One sample t-test).

8. Two independent random samples were taken from two populations:

Sample I:	12	14	10	8	16	5	3	9	11	
Sample II:	21	18	14	20	11	19	8	12	13	15

   Assuming a normal distribution for the population, test significance of difference between the population means (Hint: Two sample t-test)

9. 10 women were given an injection to induce blood pressure. Their blood pressures before and after the injection were as follows:

S. No.	**Before Injection**	**After Injection**
1	70	87
2	86	93
3	84	94
4	88	90
5	96	95
6	70	72
7	99	102
8	94	97
9	72	89
10	98	101

(a) Do you think mean blood pressure before injection is the same as mean blood pressure after injection?

(b) Give a 95% confidence interval for the mean change in blood pressure.

10. A random sample of 16 values from a normal population should mean 41.5 cm, and sum of squares of deviation from mean is 135 cm^2. Construct a 95% confidence interval for population mean.

11. A random sample of 300 were taken from a population of 9000 buffaloes in a region. 90 buffaloes were found suffering from a disease. Construct 95% confidence interval for the (a) proportion and (b) total number of buffaloes suffering from disease in the whole region.

12. Random sample of 64 men from a population has mean height equal to 68.8 inches and standard deviation of height equal to 2.4 inches. Find the 90% confidence interval for μ the mean height of men in the population.

13. Three treatments A, B & C are compared in a completely randomized design with 6 replications for each. The layout and wheat yield in Kg./plot are given in the following table ï

A 17	B 19	A 29	C 33	B 23	B 21
B 15	A 25	A 17	C 35	C 29	B 23
A 34	C 25	B 19	C 37	A 23	C 27

Analyze the experimental yields and state your conclusions?

14. The plants of wheat of 6 varieties were selected at random and the heights of their shoots were measured in cms.

Varieties			**Height in cms**						
1	85	90	89	93	84	87	-	-	-
2	88	87	94	95	91	-	-	-	-
3	95	93	87	89	91	95	92	93	-
4	83	89	90	84	85	85	-	-	-
5	90	89	61	93	88	89	90	-	-
6	93	89	87	88	88	89	90	87	90

Do the data indicate that there is no significant difference between the mean height of the plants of the different varieties?