DIRECTORATE OF DISTANCE EDUCATION

UNIVERSITY OF JAMMU

JAMMU

SELF LEARNING MATERIAL

B.A. SEMESTER III

SUBJECT : STATISTICS	**UNIT I–V**
COURSE CODE : ST 301	**LESSON NO. 1–19**

Dr. HINA S. ABROL
COURSE CO-ORDINATOR

http:/www.distanceeducationju.in

Printed and Published on behalf of the Directorate of Distance Education, University of Jammu, Jammu by the Director, DDE, University of Jammu, Jammu.

STATISTICAL INFERENCE

*Content Editing and
Format Editing*

PROF. MOHINDER PAL

- The script writer shall be responsible for the lesson / script submitted to the DDE and any plagiarism shall be his/her entire responsibility.

Printed by : Classic / 20 / 50

STATISTICS

III Semester (B.A./B.Sc)

Paper Code : ST 301 (Practical) **Title : Statistical Computing-III**

Maximum Marks : 50

External Assessment : 25

Internal Assessment : 25

Objectives : The objective of the course is to expose the students to the real life applications of Statistical Tools.

Syllabus : There shall be at least fifteen computing exercises covering the applications of Statistics based on the entire syllabus of Course ST 301 (Theory).

Distribution of Internal Assesment (25 Marks)

(i)	I Assessment	:	06 marks
(ii)	II Assessment	:	06 marks
(iii)	Class Test	:	08 marks
(iv)	Attendance	:	05 marks

STATISTICS

III Semester (B.A./B.Sc.)

Title : STATISTICAL INFERENCE

Paper Code : ST 301 (Theory)	**M. Marks : 100**
Duration : 3 Hours	**Theory Examination : 80**
Credit : 4 Credit	**Internal Ass. : 20**

Objectives :

The main objectives of this course is to provide knowledge to the students about the theory of estimation, obtaining estimates of unknown parameters using different methods, testing of Hypothesis, Test of significance and use of non-parametric test in the situations where parametric tests are not applicable.

Unit I

The concept of sampling distribution, standard error and its significance, sampling distribution of Chi Square, t and F with derivations, properties of these distributions and their inter relations.

Unit II

Estimation : Problem of estimation; point estimation, interval estimation, criteria for a good estimator, unbiasedness, consistency, efficiency and sufficiency with examples. Method of moments and maximum likelihood and application of these method for obtaining estimates of parameters of binomial, Poisson and normal distributions, properties of M.L.E's (without proof), merits and demerits of these methods.

Unit III

Testing of Hypothesis : Statistical hypothesis, Null and alternative hypothesis, simple and composite hyothesis, two types of error, critical region, power of test, level of significance. Best Critical Region, NP Lemma, its applications to find most powerful in case of binomial. Poisson and normal distributions.

Unit IV

Small sample tests based on t, F and Chi-square distribution and test based on normal distribution, confidence interval for single mean, difference of means and variance (only for normal case) confidence interval for single mean, difference of means and variance (only for normal case). Test of signficance for large samples for attributes and variable, proportions and means, single sample, two samples (both paired and independent).

Unit V

Non-parametric tests : Concept of Non-parametric tests, advantages of Non-parametric tests over parametric tests. Sign test for single sample and two sample problems (for paired and independent samples), Wilcoxon-signed rank test, Mann-Whitney U-test, run test. Median test and test for independence based on Spearman's rank correlation.

Note for Paper Setting :

The question paper will contain three Sections. Section A will contain compulsory ten very short answer type questions of 1 mark each. Section B will contain 7 short answer type questions of 5 marks each at least one question from each unit and the student has to attempt any five questions. Section C will contain 10 long answer type questions, two from each unit, of 9 marks each and the student has to attempt five questions selecting one from each unit.

Internal Assessment (Total Marks : 20)

20 marks for theory paper in a subject reserved for internal assessment shall be distributed as under :

(i) Class Test :	10 marks
(ii) Two Written Assignments/Project Reports :	10 marks (05 marks each)

Books Recommended

1. Goon, Gupta and Dass Gupta : An outline of statistical inference Vol-II
2. H.C. Saxena; Statistical inference.
3. Gibbons, J.D. : Non-parametric statistical inference.
4. Kendall and Struart: The advanced theory of statistics Vol-II

5. Connor W.J. : Practical Non-parametric Inference
6. Hogg. V. and Craig A.T. : Introduction of Mathematical Statistics.
7. Mood and Graybill : An introduction to theory of Statistics.
8. Srivastava and Srivastava : Statistical Inference : Testing of Hypothesis

Unit 1 **Lesson 1**

CONCEPT OF SAMPLING DISTRIBUTION AND SAMPLING DISTRIBUTION OF CHI-SQUARE

Structure:

1.1 Introduction

1.2 Objectives

1.3 Concept of Sampling Distribution

1.4 Chi-Square distribution

1.5 Derivation of Chi-square (χ^2) distribution

1.6 Moment generating function of Chi-square (χ^2) distribution

1.7 Limiting form of Chi-square (χ^2) distribution

1.8 Mode and Skewness of χ^2-Distribution

1.9 Additive Property of χ^2-variates.

1.10 Applications of Chi-square distribution

1.11 Relation between F and χ^2 distribution

1.12 Self assessment question.

1.1 Introduction

We know how samples can be taken from populations and can use sample data to calculate statistics such as the mean and the standard deviation. If we apply what we have learned and take several samples from a population, the statistics we would compute for each sample need not be the same and most probably would vary from sample to sample.

Suppose our samples each consist of eight 20-year-old men from a city with a population of 100,000. By computing the mean height and standard deviation of that height for each of these samples, we would quickly see that the mean of each sample and the standard deviation of each sample would be different.

A probability distribution of all the possible means of the samples is a distribution of the sample means. Statisticians call this a sampling distribution of the mean.

Describing Sampling Distributions

Any probability distribution (and, therefore, any sampling distribution) can be partially described by its mean and standard deviation.

In the above example, the sampling distribution of the mean can be partially described by its mean and standard deviation.

Understanding of sampling distributions allows statisticians to take samples that are both meaningful and cost effective due to the fact that large samples are very expensive to gather, decision makers should always aim for the smallest sample that gives reliable results.

In describing distributions statisticians have their own shorthand and when they use the term **standard error** to describe a distribution they are referring to the distribution standard deviation Instead of saying "the Standard deviation of the distribution of Sample means" they say "the standard error of the mean." which indicates how spread out (dispersed) the means of the samples are.

Chi-square test is one of the most commonly used tests of significance. The chi-square test is applicable to test the hypotheses of the variance of a normal population, goodness of fit of the theoretical distribution to observed frequency distribution, in a one way classification having k-categories. It is also applied for the test of independence of attributes, when the frequencies are presented in a two-way classification called the contingency table. It is also a frequently used test in genetics, where one tests whether the observed frequencies in different crosses agree with the expected frequencies or not.

1.2 Objectives

Understanding of sampling distributions will enable the students to have basic knowledge about the behavior of sampling distributions so that samples that are both meaningful and cost effective can be taken, due to the fact that large samples are very expensive to gather, decision makers should always aim for the smallest sample that gives reliable results.

The knowledge of Chi-square test will acquaint the learners to test the hypotheses of the variance of a normal population, goodness of fit of the theoretical distribution to observed frequency distribution, in a one way classification having k-categories. It is also applied for the test of independence of attributes, when the frequencies are presented in a two-way classification called the contingency table. It is also a frequently used test in genetics, where one tests whether the observed frequencies in different crosses agree with the expected frequencies or not. In short the main objective of this lesson is to

- To introduce the Chi Square distribution and learn how to use them in statistical inferences
- To recognize situations requiring the use of Chi-square test
- To use Chi square test to check whether a particular collection of data is well described by a specified distribution

- To see whether two classifications of same data are independent of each other
- To use Chi square distribution for confidence intervals and testing hypotheses about a single population variance

1.3 Concept of Sampling Distribution

Distribution relating to an estimate of a specific population parameter is called **a sampling distribution** of that estimate. Suppose, for example, that we wish to estimate the mean family income of a particular district in a given year on the basis of a sample of say of 200 families. Assume that we use mean of the sample to estimate the population (family income of the district) mean. We can draw an infinite number of samples from the district and calculate the value of the sample mean from each sample. These values can now be arranged in the form of a (frequency) distribution which would be called a **sampling distribution** of sample mean. Note that although the population of all families in the district is a finite one, the number of samples that we can draw from this population is infinite as long as we allow each family to be included in any sample. Such sampling is called sampling with replacement. We would know all about the possible behavior of our guesses by studying the resulting sampling distribution. Had we used some other estimator, e.g., mode or median in place of mean, the resulting distribution would have been called sampling distribution of mode or median. As such we can obtain sampling distribution of any **estimator or test statistic**.

We did not refer to size of the- sample while understanding the concept of sampling distribution, it is quite obvious that samples of different sizes give different types of information about the population from which they are drawn. To avoid the effects that are due to the change in size of the samples, a sampling distribution always refers to samples of the same size. The effects of' changing the sample size are then studied by comparing the different sampling distributions built with different size of samples.

Moreover, it can also be seen that sampling distribution of sample mean, in fact, is a probability distribution; because the income of the family as well as the mean income of the sample (drawn from the district), both are random variables.

But how does sampling distribution help to obtain a good or reliable guess?

Suppose we have obtained the sampling distribution of sample mean for the above example of family incomes. In case the mean of the sampling distribution turns out to be the value which is equal to the true value of the parameter, then the mean is said to be a good guess (or **a good estimate**) for the population parameter. To generalize, we say that an estimator is said to be a good estimator if the mean of the sampling distribution of that estimator is found to be equal to the true value of the parameter. An estimator would be a perfect estimator if its sampling distribution is concentrated entirely in one point and the point is also the true value of the parameter.

But in practice perfect estimators are very rare and can be obtained only if there is no variation in the population so that every sample drawn from the population gives rise to same mean value which also happens to be the true value of the parameter. Naturally, therefore, we have to be satisfied by less than a 'perfect guess'; but again one can ask—to what limit? Statisticians provide this limit by stating some properties of an estimator that are commonly considered desirable for an estimator to be called a good estimator. The desirable statistical properties fall into two categories: small sample (or finite-sample) properties and large sample (or asymptotic) properties. Underlying both these sets of properties is the notion that an estimator has a sampling distribution.

Usually parameters are unknown and statistics are used as estimates of parameters. The probability distribution of a statistic is called its '**sampling distribution'**

Remark;-The value of a statistic varies from sample to sample; but the parameter remains a constant *However, since the parameter is constant it has neither a sampling distribution nor a standard error.*

1.4 Chi-Square distribution

So far, we have been discussing the distribution of mean obtained from all possible samples, or a large number of samples drawn from a normal population, distributed with mean μ and variance $\frac{\sigma^2}{n}$.

Now we are interested in knowing the distribution, of sample variances s^2 of these samples. Consider a random sample of size n. Let the observations of this sample be denoted by $x_1, x_2, ..., x_n$.

We know that the variance,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{for i = 1, 2,. . . n.}$$

A quantity $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$, which is a pure number, is defined as χ_k^2 .

The distribution of the random variable χ_k^2 which was first discovered by Helmert in 1876 and later independently given by Karl Pearson in 1900 when Karl Pearson used it for frequency data classified into k-mutually exclusive categories.

Another way to understand chi-square is: if X_1, X_2, . . .Xn are n independent normal variates with mean zero and variance unity, the sum of squares of these variates is distributed as chi-square with n degrees of freedom.

More precisely, the square of a standard normal variate is known as a chi-square variate with 1 degree of freedom (df.).

Thus if $X \sim N(\mu, \sigma^2) \quad then\, Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$ and

$Z^2 = \left(\dfrac{X-\mu}{\sigma}\right)^2$ is a chi-square variate with 1 d.f.

In general if Xi, (i = 1, 2, ..., n) are n independent normal variates with means μ_i and variances σ_i^2 (i = 1, 2, ..., n), then

$\chi^2 = \sum_{i=1}^{n}\left(\dfrac{X_i-\mu_i}{\sigma_i}\right)^2$ is a chi-square variate with n d.f.

1.5 Derivation of Chi-square (χ^2) distribution

If Xi, (i = 1, 2, ..., n) are n independent normal variates with means μ_i and variances aσ_i^2 (i = 1, 2, ..., n), then we want the distribution of

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n} u_i^{\,2}$$

where $u_i = \dfrac{X_i - \mu_i}{\sigma_i} \sim N(0,1)$

Since Xi's are independent, u_i s are also independent. So that

$$M_{\chi^2}(t) = M_{u_i^2}(t) = \prod_{i=1}^{n} M_{u_i^2}(t) = [M_{u_i^2}(t)]^n \qquad \text{................(1)}$$

$[\because u_i\text{'}s\ \ are\ \ i.i.d\ \ N(0,1)]$

$$M_{u_i^2}(t) = E[\exp(t_{u_i}^{\,2})] = \int_{-\infty}^{\infty} \exp(t_{u_i}^{\,2})\, f(x_i) dx_i$$

$$= \int_{-\infty}^{\infty} \exp(t_{u_i}^2) \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x_i - \mu)^2 / 2\sigma^2\} dx_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(t_{u_i}^2) \exp-(u_i)^2 / 2\} du_i \qquad \left[u_i = \frac{X_i - \mu_i}{\sigma_i} \right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ -\left(\frac{1-2t}{2} \right) u_i^2 \right\} du_i$$

$$= \frac{1}{\sqrt{2\pi}} . \frac{\sqrt{\pi}}{\left(\frac{1-2t}{2} \right)^{1/2}} = (1-2t)^{-1/2}$$

As $\left[\int_{-\infty}^{\infty} e^{-a^2x^2} dx = \frac{\sqrt{\pi}}{a} \right]$

$\therefore \quad M_{\chi^2}(t) = (1-2t)^{-n/2}$ (Using eq. 1)

which is the m.g.f of a Gamma variate with parameters $\frac{1}{2}$ and $\frac{1}{2}n$

Hence, by uniqueness theorem of m.g.f's,

$\chi^2 = \sum_{i=1}^{n} \left(\frac{X - \mu}{\sigma} \right)^2$ is a Gamma variate with parameters $\frac{1}{2}$ and $\frac{1}{2}n$

$$dP(\chi^2) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} . e^{-\chi^2/2} \left(\chi^2\right)^{(n/2)-1} d\left(\chi^2\right)$$

$$= \frac{1}{2^{n/2}\Gamma(n/2)} e^{-\chi^2/2} (\chi^2)^{(n/2)-1} d(\chi^2) \qquad 0 \le \chi^2 < \infty$$

which is the required p.d.f of Chi-square distribution with n degrees of freedom.

- If a r.v. X has a chi-square distribution with n d.f., we write $X \sim \chi^2_{(n)}$ and its p.d.f is

$$f(x)=\frac{1}{2^{n/2}\Gamma(n/2)}e^{-x/2}x^{(n/2)-1} \qquad ;0 \le x < \infty$$

- Normal distribution is a particular case of χ^2-distribution when n = 1, since for n= 1,

$$p(\chi^2))=\frac{1}{\sqrt{2}\,\Gamma(1/2)}e^{-\chi^2/2}(\chi^2)^{(1/2)-1}d\chi^2 \qquad 0 \le x < \infty$$

$$p(\chi^2))=\frac{1}{\sqrt{2\pi}}\exp(\chi^2/2)d\chi^2 \qquad ;-\infty \le x < \infty$$

1.6 Moment generating function of Chi-square distribution

Let X~$\chi^2(n)$

then $M_X(t)=E[e^{tx}]=\int_{-\infty}^{\infty}e^{tx}f(x)dx$

$$=\frac{1}{2^{n/2}\Gamma(n/2)}\int_0^\infty e^{tx}.e^{-x/2}x^{(n/2)-1}dx$$

$$=\frac{1}{2^{n/2}\Gamma(n/2)}\int_0^\infty \exp\left[-\left(\frac{1-2t}{2}\right)x\right].x^{(n/2)-1}dx$$

$$=\frac{1}{2^{n/2}\Gamma(n/2)}\frac{\Gamma(n/2)}{[(1-2t)/2]^{n/2}} \qquad \text{\{By using Gamma integral\}}$$

$$=(1-2t)^{-n/2}, \quad |2t|<1$$

which is the required m.g.f. of a χ^2-variate with n d.f

1.7 Limiting form of Chi-square (χ^2) distribution for large degrees of freedom

Let X~$\chi^2(n)$ then m.g.f. of a χ^2-variate with n degrees of freedom is given by

$$=(1-2t)^{-n/2}, \quad |2t|<1$$

The m.g.f. of standard normal variate Z is

$$M_{(X-\mu)/\sigma}(t)=e^{-\mu t/\sigma}.M_{(X)}(t/\sigma)$$

$$\Rightarrow \qquad M_Z(t)=e^{-\mu t/\sigma}.(1-2t/\sigma)^{-n/2}=e^{-nt/\sqrt{2n}}.\left(1-2t/\sqrt{2n}\right)^{-n/2}$$

Since for Chi-square (χ^2) distribution mean $\mu=n$ and variance $\sigma^2=2n$

$$\therefore \qquad K_Z(t)=\log M_Z(t)=-t\sqrt{\frac{n}{2}}.-\frac{n}{2}\log\left(1-t\sqrt{\frac{2}{n}}\right)$$

$$=-t\sqrt{\frac{n}{2}}.+\frac{n}{2}\left(t\sqrt{\frac{2}{n}}+\frac{t^2}{2}\frac{2}{n}+\frac{t^3}{3}\left(\frac{2}{n}\right)^{3/2}+..........\right)$$

$$=-t\sqrt{\frac{n}{2}}.+t\sqrt{\frac{n}{2}}+\frac{t^2}{2}+O(n^{-1/2})=\frac{t^2}{2}+O(n^{-1/2})$$

Where $O(n^{-1/2})$ are the terms containing $n^{1/2}$ and higher powers of n in the denominator

Now $\qquad \lim\limits_{n\to\infty} K_Z(t)=\frac{t^2}{2} \quad \Rightarrow \quad M_Z(t)=e^{t^2/2} \;\; as \;\; n\to\infty$

which is the m.g.f. of a standard normal variate. Hence, by uniqueness theorem of m.g.f. Z is asymptotically normal. In other words, standard χ^2 variate tends to standard normal variate as $n\to\infty$.

Thus, χ^2 distribution tends to normal distribution for large d.f.

1.8 Mode and Skewness of χ^2-Distribution

Let X ~ $\chi^2{}_{(n).}$ so that

$$f(x)=\frac{1}{2^{n/2}\Gamma(n/2)}e^{-x/2}x^{(n/2)-1} \qquad 0 \le x < \infty$$

……..(1)

Mode of the distribution is the solution of $f'(x)=0$ and $f''(x)<0$ f"(x) Logarithmic differentiation w.r.to x in (1) gives:

$$\frac{f''(x)}{f'(x)}=0\text{-}1/2+\left(\frac{n}{2-1.}\right)\frac{1}{2}=\left(\frac{n-2-x}{2x}\right)$$

Since $f(x)\neq 0$, $f''(x)=0$ $\Rightarrow x=n-2$

It can be easily seen that at the point, x = (n - 2), $f''(x)<0$

Hence mode of the chi-square distribution with n d.f. is **(n -2).**

We can write $Skewness=\dfrac{Mean-Mode}{S.D}$

$$=\frac{n-(n-2)}{\sqrt{2n}}=\sqrt{\frac{2}{n}}$$

Since Pearson's coefficient of skewness is greater than zero for n≥ 1, the χ^2 distribution is **positively skewed**.

1.9 Additive Property of χ^2-variates.

The sum of independent chi-square variates is also a χ^2-variate. More precisely, if X_i, (i = 1, 2, ..., k) are independent χ^2-variates with n_i, d.f respectively, then the sum $\sum_{i=1}^{k}X_i$, is also a chi-square variate with $\sum_{i=1}^{k}n_i$ d.f

Proof:- We have $M_X(t) = (1-2t)^{-n_i/2} \quad i = 1,2,......,k$

The m.g.f of the sum $\sum_{i=1}^{k} X_i$, is given by

$$M_{\sum X_i}(t) = M_{X_1}(t) M_{X_2}(t)......M_{X_k}(t)$$

$[\because X_i\text{'s are independent}]$

$$= (1-2t)^{-n_1/2}(1-2t)^{-n_2/2}..........(1-2t)^{-n_{k_1}/2} = (1-2t)^{-(n_1+n_2+.........+n_k)/2}$$

which is the m.g.f of a χ^2-variate with (n_1 + n_2 + ... + n_k) d.f. Hence by uniqueness theorem of m.g.f 's, $\sum_{i=1}^{k} X_i$ is a χ^2-variate with $\sum_{i=1}^{k} n_i$ d.f

Note; Converse of above theorem is also true,

1.10 APPLICATIONS OF χ^2-DISTRIBUTION

- χ^2 **test for Inferences About a Population Variance:** Suppose we want to test if a random sample $x_1, x_2.....x_n$,has been drawn from a normal population with a specified variance $\sigma^2 = \sigma_0^2$ (say). Under the null hypothesis that the population variance is $\sigma^2 = \sigma_0^2$, the statistic

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})^2}{\sigma_0^2}\right] = \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)}{n}\right] = \frac{ns^2}{\sigma_0^2}$$

follows chi-square distribution with (n -1) d.f.

By comparing the calculated value with the tabulated value of χ^2 for (n - 1) d.f at certain level of significance (usually 5%), we may retain or reject the null hypothesis.

If the sample size n is large (>30), then we can use Fisher's approximation and apply Normal Test.

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

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

- χ^2 **test for Goodness of Fit Test**. This test is used for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson and is known as "Chi-square test of goodness of fit". It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

If f_i (i =1, 2, ..., n) is a set of observed (experimental) frequencies and e_i (i = 1, 2,n) is the corresponding set of expected (theoretical or hypothetical) frequencies, then Karl Pearson's chi-square, given by

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(f_i - e_i)^2}{e_i}\right], \quad \left(\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} e_i\right)$$

The above defined statistic

follows chi-square distribution with (n - 1) d.f.

- χ^2 **Test of Independence of Attributes**: Let us consider two attributes A divided into r classes A_1, A_2, ..., A_r and B divided into s classes B_1, B_2, ..., B_s. Such a classification in which attributes are divided into more than two classes is known as manifold classification. The various cell frequencies can be expressed in the table known as r x s manifold contingency table where (A_i) is the number of persons possessing the attribute A, (i = 1, 2, ..., r), (B_j) is the number of persons possessing the attribute B_j (j = 1, 2, ..., s) and (A_iB_j) is the number of persons possessing both the attributes A_i and B_j, (i = 1, 2, ..., r;j = 1, 2, ..., s).

Here the problem is to test if the two attributes A and B under consideration independent or not.

Under the null hypothesis that the attributes are independent, the theoretical frequencies are calculated by using

$$e_{ij} = \frac{ith\ row\ total \times jth\ column\ \ total}{sample\ size}$$

the test statistic in this case is given by

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(f_{ij} - e_{ij})^2}{e_{ij}} \right],$$

Where e_{ij} is the expected frequency in column i and row j

f_{ij} = observed frequency for contingency table category in column i and row j which is distributed as a χ^2-variate with (r - 1) (s -1) degrees of freedom.

1.11 Relation between F and χ^2 distribution

In F (n_1, n_2) distribution if we let $n_2 \to \infty$, then F follows χ^2-distribution with n_1 d.f.

Proof. We have $f(F) = \dfrac{(n_1/n_2)^{n_1/2} F^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)}$.

$$\frac{\Gamma(n_1+n_2)/2}{\left(1+\frac{n_1}{n_2}F\right)^{(n_1+n_2)/2}}, \qquad 0 < F < \infty$$

As the limit $n_2 \to \infty$ we get

$$\frac{\Gamma(n_1+n_2)/2}{\left(1+\frac{n_1}{n_2}F\right)^{(n_1+n_1)/2}} \rightarrow \frac{(n_2/2)^{n_1/2}}{n_2^{\ n_1/2}} = \frac{1}{2^{n_1/2}} \qquad \ldots\ldots\ldots\ldots\ldots(1)$$

$$\left[\because \frac{\Gamma(n+k)}{\Gamma(n)} \rightarrow n^k \quad as\; n \rightarrow \infty\right]$$

Also

$$\underset{n_2\rightarrow\infty}{Lim}\left(1+\frac{n_1}{n_2}F\right)^{(n_1+n_1)/2} = \underset{n_2\rightarrow\infty}{Lim}\left(\left(1+\frac{n_1}{n_2}F\right)^{n_2}\right)^{1/2} \times \underset{n_2\rightarrow\infty}{Lim}\left(1+\frac{n_1}{n_2}F\right)^{n_1/2}$$

$$= \exp(n_1F/2) = \exp(\chi^2/2) \qquad (\because n_1F = \chi^2)$$

$$\ldots\ldots\ldots\ldots(2)$$

Hence in the limit, on using (1) and(2) the p.d.f of $\chi^2 = n_1F$ becomes

$$dP(\chi^2) = \frac{(n_1/2)^{n_1/2}e^{-\chi^2/2}}{\Gamma(n_1/2)}.\left(\frac{\chi^2}{n_1}\right)^{(n_1/2)-1} d\left(\frac{\chi^2}{n_1}\right)$$

$$= \frac{1}{2^{n_1/2}\Gamma(n_1/2)}e^{-\chi^2/2}(\chi^2)^{(n_1/2)-1}d(\chi^2)$$

Which is p.d.f of χ^2 with n_1 degrees of freedom

1.12 Self assessment question

1. Explain why we call Chi-square distribution as sampling distributions?

2. Write the parameters of the Chi-square distribution:

3. In what situation Chi-square distribution tend to normal distribution derive the condition for the same

4.By using m.g.f of Chi-square distribution find mean , variance, μ_3, μ_4 skewness and kurtosis

[Hint: Mean = n, Variance = 2n, μ_3 =8n, μ_4 = 48n + 12n^2

$\beta_1 = \frac{8}{n}$, $\beta_2 = \frac{12}{m} + 3$]

5. State the assumptions underlying Chi-Square test when applied as the test of significance for testing of null hypotheses.

6. A random sample is drawn from a normal population. The data give sample size and sample variance only. What statistic would you use to test the hypothesis that the population variance has a particular value ? Give reasons.

7. State applications of χ^2 distribution.

Unit 1 **Lesson 2**

DISTRIBUTION AND ITS PROPERTIES

Structure:

2.1 Introduction

2.2 Objectives

2.3 Concept of t distribution and its derivation

2.4 Constants of t-distribution

2.5 Limiting Form of t-distribution

2.6 Graph of t-distribution

2.7 Application of t distribution

2.8 Exercises

2.9 Self assessment question

2.1 Introduction

This distribution was discovered by W.S. Gosset in 1908. The statistician Gosset is better known by the pseudonym '**student**' and hence t-distribution is called student's t-distribution. He derived the distribution to find an exact test of a mean by making use of estimated standard deviation, based on a random sample of size n. R.A. Fisher in 1925 published that t-distribution can also be applied to the test of regression coefficient and other practical problems.

2.2 Objectives

Understanding the sampling distributions enable the learners to have basic knowledge about the behaviour of sampling distributions so that meaningful and cost effective samples in order to apply these samples in test of significance. In fact, decision makers should always aim for the smallest sample that gives reliable results.

The knowledge of t distribution and its properties test will give the learners the basic idea to test the hypotheses about single mean, two means, difference of two means, to test the significance of the observed sample correlation etc. This Lesson will also give us information about its inter-relations with the other distributions etc

- To introduce the t distribution and learn how to use them in statistical inferences
- To recognize situations requiring the use of t test
- To use t test to the hypotheses about single mean, two means, difference of two means, to test the significance of the observed sample correlation etc.

2.3 Concept of t distribution and its derivation

While deriving and defining t distribution we make use of the following assumptions

Assumption for Student's t-test. The following assumptions are made in the Student's t-test

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

(ii) The sample observations are independent, i.e., the sample is random.

(iii) The population standard deviation σ is unknown.

Student's t distribution. Suppose $x_{1,}x_2,...x_n$ be a random sample of size n drawn from a normal population with a specified mean, say μ and variance σ^2. Then

The Student's t- statistic is given by

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

where $\overline{x} = \frac{1}{n}\sum_{i-1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i-1}^{n}(x_i - \overline{x})^2$ is an unbiased estimate of population variance σ^2

The above defined statistic follows student's t distribution with $\upsilon = (n-1)$ d.f with p.d.f given by

$$f(t) = \frac{1}{\sqrt{\nu} B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}} \qquad ; -\infty < t < \infty \qquad \ldots\ldots(1)$$

Remarks about t distribution

- If we take $\nu = 1$ in the above expression (1) we get:

$$f(t) = \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \cdot \frac{1}{\left(1+t^2\right)} = \frac{1}{\pi} \cdot \frac{1}{\left(1+t^2\right)} \qquad ; -\infty < t < \infty \quad as \quad \Gamma(1/2) = \sqrt{\pi}$$

which is the p.d.f. of standard Cauchy distribution. Hence, **when ν = 1, Student's t distribution reduces to Cauchy distribution.**

Derivation of Student's t-distribution. The Student's t- statistic is given by

$t = \frac{\overline{x} - \mu}{S/\sqrt{n}}$ This expression can be rewritten as

$$t^2 = \frac{n(\overline{x} - \mu)^2}{S^2} = \frac{n(\overline{x} - \mu)^2}{ns^2 / n - 1}$$

$$\Rightarrow \qquad \frac{t^2}{(n-1)} = \frac{(\overline{x}-\mu)^2}{\sigma^2/n} \cdot \frac{1}{ns^2/\sigma^2} = \frac{(\overline{x}-\mu)^2/(\sigma^2/n)}{ns^2/\sigma^2}$$

Since x_i, (i = 1, 2, ..., n) is a random sample from the normal population with mean μ and variance σ^2 so that

$$\overline{x} \sim N(\mu, \sigma^2/n) \qquad \Rightarrow \qquad \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Hence $\frac{(\overline{x}-\mu)^2}{\sigma^2/n}$ being the square of a standard normal variate is a chi-square variate with 1 d.f. Also $\frac{ns^2}{\sigma^2}$ is a chi-square variate with (n-1) degree of freedom

Further since $\overline{x}$ and s^2 are independently distributed $\frac{t^2}{n-1}$ being the ratio of two independent χ^2-variates with 1 and (n -1) d.f. respectively, is a variate $\beta\left(\frac{1}{2}, \frac{n-1}{2}\right)$ and its distribution is given by

$$dF(t) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}} d\left(t^2/\nu\right) \qquad ; 0 \le t^2 < \infty$$

$$= \frac{1}{\sqrt{\nu}B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}} dt \qquad ; -\infty < t < \infty$$

This is the required probability density function of Student's t-distribution with $\upsilon = (n-1)$ d.f.

Remark-Factor 2 disappears since me integral from $-\infty \; to \; \infty$ must be unity.

To decide about the acceptance or rejection of null hypothesis we now compare the calculated value of $|t|$ with the tabulated value at certain level of significance α. If calculated $|t|$>tabulated t, null hypothesis is rejected and if calculated $|t|$< tab. t, H_0 may be accepted at the level of significance adopted for (n-1) degree of freedom.

Importance of Student's t-distribution in Statistics. W.S.Gosset, who wrote under pseudonym (pen-name) of Student defined his t in a slightly different way, viz., t=$(\overline{x}-\mu)^2$)/s and investigated its sampling distribution, , Prof. R.A. Fisher, later on defined his own '*t*' and gave a rigorous proof for its sampling distribution in 1926. *The salient feature of't' is that both the statistic and its sampling distribution are functionally independent of* σ*, the population standard deviation.*

The discovery of't' is regarded as a landmark in the history of statistical inference. Before Student gave his 't', it was customary to replace σ^2 in $Z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ by its unbiased estimate to give $t = \frac{\overline{x}-\mu}{S/\sqrt{n}}$ and then normal test was applied even for small samples.

It has been found that although the distribution of t is asymptotically normal for large samples it is far from normality for small samples.

The Student's t ushered in an era of exact sample distribution (and tests) and since its discovery many important contributions have been made towards the development and extension of small (exact) sample theory.

Confidence or Fiducial Limits for μ .

If $t_{0.05}$ is the tabulated value of t for $\upsilon = (n-1)$ d.f. at 5% level of significance, i.e.,

$$(P|t| > t_{0.05}) = 0.05 \;\Rightarrow\; (P|t| \le t_{0.05}) = 0.95$$

then 95% confidence limits for μ are given by:

$$|t| \le t_{0.05} \qquad i.e., \left|\frac{\bar{x}-\mu}{S/\sqrt{n}}\right| \le t_{0.05} \Rightarrow \qquad \bar{x} - t_{0.05}\, S/\sqrt{n} \le \mu \le \bar{x} - t_{0.05}\, S/\sqrt{n}$$

Thus, 95% confidence limits for μ are $\bar{x} \pm t_{0.05}\, S/\sqrt{n}$

Similarly, 99% confidence limits for μ are $\bar{x} \pm t_{0.01}\, S/\sqrt{n}$

where $t_{0.01}$ is the tabulated value of t for v = (n-1) d.f at 1% level of significance.

Fisher's 't' (Definition). It is the ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom. If ξ is a N(0, 1) and χ^2 is an independent chi-square variate with n d.f., then Fisher's t given by

$$t = \frac{\xi}{\sqrt{\chi^2/n}}$$

and it follows Student's 't' distribution with n degrees of freedom and its p.d.f is given by

$$= \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}} \qquad ; -\infty < t < \infty$$

Which is the probability density function of Student's t-distribution with n d.f

<u>Hence, Student's 't' may be regarded as a particular case of Fisher's 't'</u>

Remark-Since

$$\bar{x} \sim N(\mu, \sigma^2/n) \qquad \xi = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

And

$$\chi^2 = \frac{ns^2}{\sigma^2} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ \sigma^2 \qquad \ldots\ldots\ldots\ldots(2)$$

distributed independently as chi-square variate with (n-1) d.f. Hence Fisher's t is given by

$$t = \frac{\xi}{\sqrt{\chi^2/(n-1)}} = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \cdot \frac{\sigma}{\sqrt{(\bar{x} - \mu)/(n-1)}} = \frac{\sqrt{n}(\bar{x} - \mu)}{s}$$

$$= \frac{(\bar{x} - \mu)}{s/\sqrt{n}} \qquad \ldots\ldots\ldots(3)$$

And it follows Student's t-distribution with (n -1) d.f.Now, (3) is same as Student's 't' *.Hence Student's 't' is a particular case of Fisher's 't'*

2.4 Constants of t-distribution

Since f(t) is symmetrical about the line t=0, all the moments of odd order about origin vanish, i.e.,

$\mu'_{2r+1}(about\ origin) = 0, \qquad r = 0,1,2......$

In particular, $\mu_1^{/}$ (about origin) = 0 = Mean

Hence central moments coincide with moments about origin.

$\therefore \qquad \mu_{2r+1} = 0, \qquad r = 0,1,2......$

The moments of even order are given by

$$\mu_{2r} = \mu'_{2r}(about\,origin) = \int_{-\infty}^{\infty} t^{2r} f(t)dt = 2\int_0^{\infty} t^{2r} f(t)dt$$

$$= \frac{2}{\sqrt{n}\, B\left(\frac{1}{2},\frac{n}{2}\right)} \int_0^{\infty} \frac{t^{2r}}{\left(1+\frac{t^2}{n}\right)^{(n+1)\!/\!2}} dt$$

This integral is absolutely convergent if 2r < n.

Put $1+\frac{t^2}{n} = 1/y \qquad \Rightarrow t^2 = \frac{n(1-y)}{y} \qquad \Rightarrow 2tdt = -\frac{n}{y^2}dy$

When t = 0,y =1 and when t=∞ , y = 0 .Therefore

$$\mu_{2r} = \frac{2}{\sqrt{n}B\left(\frac{1}{2},\frac{n}{2}\right)} \int_1^0 \frac{t^{2r}}{(1/y)^{\frac{n+1}{2}}} \frac{-n}{2ty^2} dy = \frac{n}{\sqrt{n}B\left(\frac{1}{2},\frac{n}{2}\right)} \int_0^1 \left(t^2\right)^{(2r-1)/2} y^{[(n+1)/2]-2} dy$$

$$= \frac{n}{\sqrt{n}B\left(\frac{1}{2},\frac{n}{2}\right)} \int_0^1 \left[n\left(\frac{1-y}{y}\right)\right]^{r-1/2} y^{[(n+1)/2]-2} dy$$ {Using the value of t^2 as defined above}

$$= \frac{n^r}{B\left(\frac{1}{2},\frac{n}{2}\right)} \int_0^1 y^{[n/2]-r-1}(1-y)^{r-1/2} dy$$

$$= \frac{n^r}{B\left(\frac{1}{2},\frac{n}{2}\right)} B\left(\frac{n}{2}-r, r+\frac{1}{2}\right)$$

$$= n^r \frac{\Gamma[n/2-r]\Gamma[r+1/2]}{\Gamma 1/2 \ \Gamma n/2}$$

$$= n^r \frac{(r-1/2)(r-3/2)......3/2.1/2 \ \Gamma(n/2-r)}{\Gamma 1/2 \ [n/2-1][n/2-2]....[n/2-r]\Gamma[n/2-r]}$$

$$= n^r \frac{(2r-1)(2r-3)......3.2.1}{(n-2)(n-4)....(n-2r)} \qquad ; \frac{n}{2} > r$$

I n particular

$$\mu_2 = n.\frac{1}{(n-2)} = \frac{n}{(n-2)} \quad ; n > r$$

$$\mu_4 = n^2.\frac{3.1}{(n-2)(n-4)} = \frac{3n^2}{(n-2)(n-4)} \qquad ; n > 4$$

Hence $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \quad and$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \left(\frac{3n^2}{(n-2)(n-4)}\right) \div \left(\frac{n}{(n-2)}\right)^2 = 3\left(\frac{(n-2}{(n-4)}\right)$$

Note: (i) Moment generating function of this distribution does not exist

(ii).If the random variables X$_1$ and X$_2$ are independent and follow chi-square distribution with n d.f., then $\sqrt{n}(X_1 - X_2)/2\sqrt{X_1 X_2}$ distributed as Student's t with n d.f., independently of X$_1$+ X2.

2.5 Limiting Form of t-distribution:. As $n \to \infty$, the p.d.f. of t-distribution with n d.f viz.,

$$f(t) = \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \to \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} t^2\right) \qquad -\infty < t < \infty$$

Proof $$\lim_{n\to\infty} \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{\Gamma[n+1/2]}{\Gamma(n/2)\Gamma(1/2)}$$

$$= \frac{1}{\sqrt{n}} \frac{1}{\sqrt{\pi}} \left(\frac{n}{2}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2\pi}}$$

Since $\Gamma(1/2) = \sqrt{\pi}$ and $\lim_{n\to\infty} \frac{\Gamma(n+p)}{\Gamma n} = n^p$

So that

$$\therefore \quad \lim_{n\to\infty} f(t) = \lim_{n\to\infty} \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} . \lim_{n\to\infty} \left[\left(1 + \frac{t^2}{n}\right)^n\right]^{-1/2} \lim_{n\to\infty} \left(1 + \frac{t^2}{n}\right)^{\frac{-1}{2}}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} t^2\right), \qquad -\infty < t < \infty$$

Hence for large d.f. t-distribution tends to standard normal distribution.

2.6 Graph of t-distribution. The p.d.f. of t-distribution with n d.f is:

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

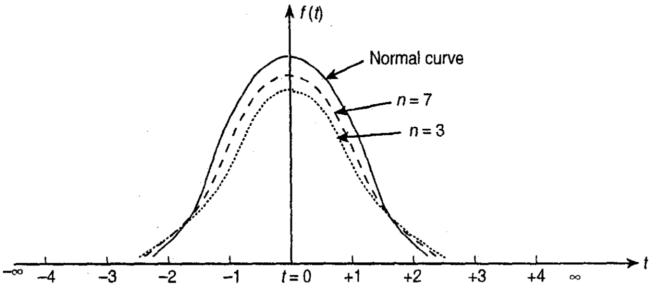
Where $c = \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)}$

Since f(-t)=f(t), the probability curve is symmetrical about the line t = 0. As t increases, f(t) decreases rapidly and tends to zero as t $\to \infty$, so that t-axis is an asymptote to the curve. We have shown that

$$\mu_2 = \frac{n}{n-2} \qquad n > 2 \;; \qquad \beta_2 = \frac{3(n-2)}{(n-4)}, \quad n > 4$$

Hence for n> 2, μ_2>1 i.e., the variance of t-distribution is greater than that of standard normal distribution and for n > 4, β_2> 3 and thus t-distribution is more flat on the top than the normal curve. In fact, for small n, we have

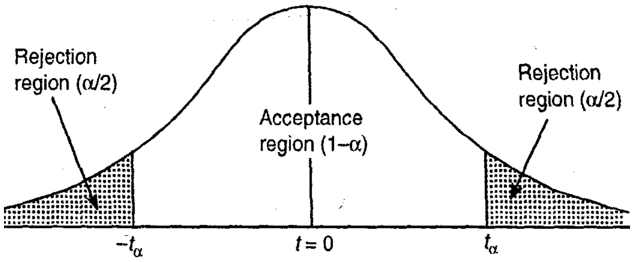
$$P[|t| \geq t_0] \geq P[|Z| \geq t_0], \qquad Z \sim N(0,1)$$



i.e., the tails of the t-distribution have a greater probability (area) than the tails of standard normal distribution.

Critical Values of t. The critical (or significant) values of t at level of significance α and d.f ν for two-tailed test are given by the equation

$$P[|t| > t_\nu(\alpha)] = \alpha \qquad \Rightarrow \qquad P[|t| \leq t_\nu(\alpha)] = 1-\alpha$$



Since t-distribution is symmetric about t = 0,so we have

$P[t > t_v(\alpha)] + P[t < -t_v(\alpha)] = \alpha \qquad \Rightarrow \qquad 2P[t > t_v(\alpha)] = \alpha$

$\Rightarrow \qquad P[t > t_v(\alpha)] = \alpha / 2$

Therefore $P[t > t_v(2\alpha)] = \alpha$

$t_v(2\alpha)$ (from the Tables) gives the significant value of t for a single-tail test [Right-tail or Left-tail-since the distribution is symmetrical], at level of significance α and v d.f. Hence the significant values of t at level of significance ' α 'for a single-tailed test can be obtained from those of two-tailed test by looking the values at level of significance 2α.

For example,

$t_8(0.05)$ for single-tail test = t_8 (0.10) for two-tail test = 1.86

2.7 APPLICATIONS OF t-DISTRIBUTION

The t-distribution has a wide number of applications in Statistics, some of which are enumerated below.

(i) To test if the sample mean ($\bar{x}$) differs significantly from the hypothetical value μ of the population mean

(ii) To test the significance of the difference between two sample means.

(iii) To test the significance of an observed sample correlation coefficient and sample regression coefficient.

(iv) To test the significance of observed partial correlation coefficient.

2.8 EXERCISES

EXERCISE NO :-1 Show that for t-distribution with n d.f., mean deviation about mean is given by

$\sqrt{n}\,\Gamma[(n-1)/2]/\sqrt{\pi}\,\Gamma n/2$

Solution. We know that E(t) = 0.

$M.D.(about\,mean) = \int_{-\infty}^{\infty} |\,t\,|\, f(t)dt$

$$= \frac{1}{\sqrt{n}\,B\left(\frac{1}{2},\frac{n}{2}\right)} \int_{-\infty}^{\infty} \frac{|\,t\,|}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}} dt$$

$$= \frac{2}{\sqrt{n}\,B\left(\frac{1}{2},\frac{n}{2}\right)} \int_{0}^{\infty} \frac{t}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}} dt$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2},\frac{n}{2}\right)} \int_{0}^{\infty} \frac{dy}{(1+y)^{(n+1)/2}}$$ by substituting $\left(\frac{t^2}{n} = y\right)$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2},\frac{n}{2}\right)} \int_{0}^{\infty} \frac{y^{1-1}}{(1+y)^{(n-1)/2 --+1}} dy = \frac{\sqrt{n}}{B\left(\frac{1}{2},\frac{n}{2}\right)} B\left(\frac{n-1}{2},1\right)$$

$$= \frac{\sqrt{n}\ \Gamma[(n-1/2)]}{\sqrt{\pi\ \Gamma(n/2)}}$$

EXERCISE No:-2 select the correct answer

Student's t-test is applicable when (a) a sample size is large, (b) a sample size is less

than five, (c) a sample size is less than thirty but greater than five.

EXERCISE No:-3 Check whether the following statement is correct:

(a) t-value lies between $-\infty$ and $+\infty$.

EXERCISE No 4.. Find the values for the following with the help of tables

(a) t_{15} when $\alpha = 0.05$ for two tailed test

(b) t_{12} when $\alpha = 0.02$ for single tailed test

(c) t_{22} when $\alpha = 0.01$ for two tailed test

(d) t_{10} when $\alpha = 0.05$ for single tailed test

(e) t_{15} when $\alpha = 0.01$ for single tailed test

2.9 Self assessment questions

1. What is Student's t distribution? When is it used to construct a confidence interval estimate of the population mean?
2. Explain the importance of t distribution as distribution as sampling distributions?
3. Describe the constants of student-t distribution

3. In what situation t distribution tends to normal distribution derive this result mathematically.

4. State the assumptions underlying Student’s t-test when applied to both single and two-sample problems.

5. Define the student’s t-test. What kind of hypotheses can be tested by the t-test.

6. Obtain formulae for 95% confidence limits of the mean of a normal population, when the mean is (i) known, (ii) unknown.

7. Obtain the formulae for 95% C.I for mean of normal population when the mean is (i) Known (ii) Unknown

DISTRIBUTION AND ITS PROPERTIES

Structure:

3.1 Introduction

3.2 Objectives

3.3 Concept of F distribution and its derivation

3.4 Constants of F-distribution

3.5 Mode and Points of Inflexion of F-distribution

3.6 Applications of F-distribution

3.7 Relation between t and F distributions

3.8 Shape of f distribution

3.9 Exercises

3.10 Self assessment question

3.1 INTRODUCTION

This distribution was discovered by G.W.Snedecor and named in the honour of the Distinguish mathematical statistician Sir R.A Fisher. It may be recalled that the t statistic is used for testing whether two population means are equal. Whenever we are required to test for the case of more than two means, this can be tested by comparing the sample variances using F distribution by the use of analysis of variance technique which consist of “separation of variation due to a group of causes from the variation due to other groups”.

F ratio is basically ratio of between column variance and between column variance, having found F ratio we can interpret it First, examine the denominator., which is based on the variance within the samples. The denominator is a good estimator of σ^2 (the population variance) whether the null hypothesis is true or not. What about the numerator? If the null hypothesis is true, then the numerator, or the variation among the sample means, is also a good estimate of σ^2 (the population variance). As a result, the denominator and numerator should be about equal if the null hypothesis is true. The nearer the F ratio comes to 1, then the more we are inclined to accept the null hypothesis Conversely, as the F ratio becomes larger, we will be more inclined to reject the null hypothesis and accept the alternative (that a difference does exist in the effects of the three training methods).

In short ,when populations are not the same, the between-column variance (which was derived from the variance among the sample means) tends to be larger than the within-column variance (which was derived from the variances within the samples), and the value of F tends to be large. This leads us to reject the null hypothesis.

Summing up, F- distribution is a very popular and useful distribution because of its utility in testing of hypothesis about the equality of several population means, two population variances and several regression coefficients in multiple regression etc. As a matter of fact, F-test is the backbone of analysis of variance.

In fact this sampling distribution is widely used in different ways while testing different null hypotheses about a variety of population parameters.

3.2 OBJECTIVES

The objectives of this lesson is

- To introduce the F distribution and learn how to use them in statistical inferences
- To recognize situations requiring the comparison of more than two means or proportions
- To compare more than two population means using analysis of variance
- To use the F distribution to test hypotheses about two population variances

3.3 CONCEPT OF F DISTRIBUTION AND ITS DERIVATION

F-distribution: If X and Y are chi-square variates with ν_1 and ν_2 degrees of freedom respectively, then F-statistic is defined by

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

Hence, F is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's F-distribution with (ν_1, ν_2) d.f. denoted by F ~ F (ν_1, ν_2) with probability function given by:

$$f(F) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{\frac{\nu_1}{2}-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{\frac{\nu_1}{2}+\frac{\nu_2}{2}}} \qquad 0 \le F < \infty$$

Derivation of Snedecor's F-distribution: Since X and Y are independent chi-square variates with ν_1 and ν_2 d.f. respectively, their joint probability density function is given by:

$$f(x,y)=\left\{\frac{1}{2^{\frac{\nu_1}{2}}\Gamma(\nu_1/2)}\exp(-x/2)x^{\frac{\nu_1}{2}-1}\right\}\times\left\{\frac{1}{2^{\frac{\nu_2}{2}}\Gamma(\nu_2/2)}\exp(-y/2)y^{\frac{\nu_2}{2}-1}\right\}$$

$$=\frac{1}{2^{\frac{(\nu_1+\nu_2)}{2}}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\exp\{-(x+y)/2)\}x^{\frac{\nu_1}{2}-1}y^{\frac{\nu_2}{2}-1},\qquad 0\le(x,y)<\infty$$

Let us transform the variables as given below

$F=\dfrac{x/\nu_1}{y/\nu_2}$ and u = y so that $0\le F<\infty$, o<u<∞

$$\therefore\qquad x=\frac{\nu_1}{\nu_2}Fu\qquad and\quad y=u$$

Jacobian of transformation is given by

$$J=\frac{\partial(x,y)}{(F,u)}=\begin{vmatrix}\frac{\nu_1}{\nu_2}u & 0\\ \frac{\nu_1}{\nu_2}F & 1\end{vmatrix}=\frac{\nu_1}{\nu_2}u$$

So that joint p.d.f of the transformed variables is

$$g(F,u)=\frac{1}{2^{\frac{(\nu_1+\nu_2)}{2}}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\}\times\left(\frac{\nu_1}{\nu_2}Fu\right)^{\frac{\nu_1}{2}-1}u^{\frac{\nu_2}{2}-1}|\ J\ |$$

$$g(F,u)=\frac{(\nu_1/\nu_2)}{2^{\frac{(\nu_1+\nu_2)}{2}}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\}\times u^{(\nu_1+\nu_2/2)-1}(F)^{\frac{\nu_1}{2}-1},$$

$0\le F<\infty$, o<u<∞

Integrating w.r. to u over the range 0 to ∞, the p.d.f. of F becomes

$$g_1(F)=\frac{(\nu_1/\nu_2)^{(\nu_1/2)}(F)^{(\nu_1/2)-1}}{2^{\frac{(\nu_1+\nu_2)}{2}}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\}\times u^{((\nu_1+\nu_2)/2)-1}du\right]$$

$$=\frac{(\nu_1/\nu_2)^{(\nu_1/2)}(F)^{(\nu_1/2)-1}}{2^{\frac{(\nu_1+\nu_2)}{2}}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\times\frac{\Gamma(\nu_1+\nu_2/2)}{\left[\frac{1}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right]^{(\nu_1+\nu_2)/2}}$$

$$\therefore \qquad g_1(F)=\frac{(\nu_1/\nu_2)^{(\nu_1/2)}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\times\frac{(F)^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, \qquad 0\le F<\infty$$

which is the required probability function of F-distribution with (ν_1,ν_2) d.f

Alternative Proof of F-distribution: If X and Y are chi-square variates with ν_1 and ν_2 degrees of freedom respectively, then F-statistic is defined by

$F=\frac{X/\nu_1}{Y/\nu_2}$ so that $\frac{\nu_1}{\nu_2}F=\frac{X}{Y}$ being the ratio of two independent chi-square variates with

ν_1 and ν_2 degrees of freedom respectively is a $\beta_2\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)$ variate. Hence the probability function of F is given by

$$dp(F)=\frac{1}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\times\frac{\left(\frac{\nu_1}{\nu_2}F\right)^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}d\left(\frac{\nu_1}{\nu_2}F\right)$$

so that $$f(F)=\frac{\left(\frac{\nu_1}{\nu_2}\right)^{(\nu_1/2)-1}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\times\frac{(F)^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, \qquad 0\le F<\infty$$

3.4 Constants of F-distribution:

$$\mu_r'\,(about\ \ origin\ ) = E[F^r] = \int_0^\infty F^r\, f(F)dF$$

$$= \frac{(\nu_1/\nu_2)^{(\nu_1/2)-1}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\int_0^\infty F^r\,\frac{(F)^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}dF$$

Put $\frac{\nu_1}{\nu_2}F = y$ *so that* $dF = \frac{\nu_2}{\nu_1}Fu$ o<u<∞

$$\therefore \qquad \mu_r' = \frac{(\nu_1/\nu_2)^{(\nu_1/2)}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\int_0^\infty \frac{\left(\frac{\nu_1}{\nu_2}y\right)^{r+(\nu_1/2)-1}}{(1+y)^{(\nu_1+\nu_2)/2}}\left(\frac{\nu_1}{\nu_2}\right)dy$$

$$= \frac{\left(\frac{\nu_1}{\nu_2}\right)^r}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\int_0^\infty \frac{y^{r+(\nu_1/2)-1}}{(1+y)^{(\nu_1/2)+r+[(\nu_2/2)-r]}}dy$$

$$= \frac{\left(\frac{\nu_1}{\nu_2}\right)^r}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}.B\left(r+\frac{\nu_1}{2},\frac{\nu_2}{2}-r\right) \qquad ,\nu_2 > 2r$$

In particular for r =1, we have

$$\mu_1' = \left(\frac{\nu_1}{\nu_2}\right)\frac{B\left(1+\frac{\nu_1}{2},\frac{\nu_2}{2}-1\right)}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}$$

$$= \left(\frac{\nu_1}{\nu_2}\right)\frac{\Gamma[1+(\nu_1/2)]\Gamma[(\nu_2/2)-1]/\Gamma[\nu_1/2,\nu_2/2]}{\Gamma[(\nu_1/2)]\Gamma[(\nu_2/2)]/\Gamma[2]}$$

$$=\left(\frac{\nu_1}{\nu_2}\right)\frac{\Gamma[1+(\nu_1/2)]\Gamma[(\nu_2/2)-1]}{\Gamma\left[(\nu_1/2)\right]\Gamma\left[(\nu_2/2)\right]}, \nu_2 > 2$$

Since $B(\mu,\nu)=\dfrac{\Gamma\mu\Gamma\nu}{\Gamma(\mu+\nu)}$

$$=\left(\frac{\nu_1}{\nu_2}\right)\frac{\left[(\nu_1/2)\right]\Gamma\left[(\nu_1/2)\right]\;\Gamma\left[(\nu_2/2)-1\right]}{\Gamma\left[(\nu_1/2)\right]\left[(\nu_2/2)-1\right]\Gamma\left[(\nu_2/2)-1\right]}=\frac{\nu_2}{\nu_2-2}\qquad,\nu_2>2$$

As $\Gamma r=(r-1)\Gamma(r-1)$

$$\mu_2^{'}=\left(\frac{\nu_1}{\nu_2}\right)^2\frac{\Gamma\left[(\nu_1/2)+2\right]\;\Gamma\left[(\nu_2/2)-2\right]}{\Gamma\left[(\nu_1/2)\right]\Gamma\left[(\nu_2/2)\right]}$$

$$=\left(\frac{\nu_1}{\nu_2}\right)^2\frac{\Gamma\left[(\nu_1/2)+1\right]\;\Gamma\left[(\nu_1/2)\right]}{\left[(\nu_2/2)-1\right]\Gamma\left[(\nu_2/2)-2\right]}=\frac{\nu_2^2(\nu_1+2)}{\nu_1(\nu_2-2)(\nu_2-4)},\quad \nu_2>4$$

$$\therefore\qquad \mu_2=\mu_2^{'}-\mu_1^{'2}=\frac{\nu_2^2(\nu_1+2)}{\nu_1(\nu_2-2)(\nu_2-4)}-\frac{\nu_2^2}{(\nu_2-2)^2}=\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}\nu_2>4$$

Similarly, on putting r =3 and 4 in $\mu_r^{'}$ we get $\mu_3^{'}$ and $\mu_4^{'}$ respectively

3.5 MODE AND POINTS OF INFLEXION OF F-DISTRIBUTION.

We have

$$f(F)=\frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}\cdot\frac{F^{\frac{\nu_1}{2}-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{\frac{\nu_1}{2}+\frac{\nu_2}{2}}}\qquad 0\le F<\infty$$

oiuu

Taking log both sides we get

$$\log\ f(F) = C + \{(\nu_1/2)-1\}\log F - \left(\frac{\nu_1+\nu_2}{2}\right)\log\left(1+\frac{\nu_1}{\nu_2}F\right)$$

C is a constant independent of F.

$$\frac{\partial}{\partial F}[\log(F)] = \{(\nu_1/2)-1\}\frac{1}{F} - \left(\frac{\nu_1+\nu_2}{2}\right)\frac{1}{\left(1+\frac{\nu_1}{\nu_2}F\right)}\cdot\frac{\nu_1}{\nu_2}$$

$$f'(F) = \frac{\partial}{\partial F}f(F) = 0 \quad \Rightarrow \quad \frac{\nu_1-2}{2F} - \frac{\nu_1(\nu_1+\nu_2)}{2(\nu_2+\nu_1F)} = 0$$

Solving for F we get

$$F = \frac{\nu_2(\nu_1-2)}{\nu_1(\nu_2+2)}$$

It can be easily verified that at this point f ''(F) <0.

Hence $\qquad Mode = \dfrac{\nu_2(\nu_1-2)}{\nu_1(\nu_2+2)}$

Since F> 0, mode exists if and only if $\nu_1 > 2$.

Mode = $\left(\dfrac{\nu_2}{\nu_2+2}\right)\left(\dfrac{\nu_1-2}{\nu_1}\right)$

Hence mode of F-distribution is always less than unity.

Hence Karl Pearson’s coefficient of skewness for F distribution is given by

$$\frac{Mean - \text{mod}e}{\sigma} > 0$$

since mean> 1 and mode < 1. hence F-distribution is highly positively skewed.

3.6 APPLICATIONS OF F-DISTRIBUTION

F- distribution is a very popular and useful distribution because of its utility in testing of hypothesis about the equality of several population means, two population variances and several regression coefficients in multiple regression etc. As a matter of fact, F-test is the backbone of analysis of variance.

In fact this sampling distribution is widely used in different ways while testing different null hypotheses about a variety of population parameters.

F-test for Equality of Two Population Variances. Suppose we want to

test (i) whether two independent samples x_i, (i = 1, 2 n_1) and y_j, (I = 1, 2 n_2) have been drawn from the normal populations with the same variance σ^2 (say), or (ii) whether the two independent estimates of the population variance are homogeneous or not.

Under the null hypothesis (H_0) that (i) $\sigma_x^2 = \sigma_y^2 = \sigma^2$, i.e., the population variances

are equal, or (ii) Two independent estimates of the population variance are homogeneous, the

statistic F is given by

$$F = \frac{S_x^2}{S_y^2}$$

Where $s_x^2 = \frac{1}{n_1 - 1}\sum_{i-1}^{n_1}(x_i - \overline{x})^2$ and $s_y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_j - \overline{y})^2$

are unbiased estimates of the common population variance σ^2 obtained from two independent samples and it follows Snedecor's F-distribution with (v_1, v_2) d,f. where $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$.

By comparing the calculated value of F obtained by using above formula for the two given samples, with the tabulated value of F for (n_1,n_2) d.f. at certain leve1 of significance (5% or 1%), H_0 is either rejected or accepted.

- **F-test for Testing the Significance of an Observed Multiple** Correlation Coefficient**:**. If R is the observed multiple correlation coefficient of a variate with k other variates in a random sample of size n from a (k+1) variate population, then Prof. R.A. Fisher proved that under the null hypothesis (H_0) that the multiple correlation coefficient in the population is zero, the statistic:

$$F = \frac{R^2}{1-R^2} . \frac{n-k-1}{k}$$

follows F distribution with (k,n-k-1) d.f

- **F-test for Testing the Significance of an Observed Sample:** Correlation Ratio η_{YX} . Under the null hypothesis that population correlation— ratio is zero, the test statistic is

$$F = \frac{\eta^2}{1-\eta^2} . \frac{N-h}{h-1} \quad \sim \quad F(h-1, N-h)$$

where N s the size of the sample (from a bi-variate normal population) arranged in h arrays

- **F-test for Testing the Linearity of Regression**: For a sample of size N arranged in h arrays, from a bi-variate normal population, the test statistic for testing the hypothesis of linearity of regression is

$$F = \frac{\eta^2 - r^2}{1-\eta^2} \cdot \frac{N-h}{h-2} \qquad \sim \qquad F(h-2, N-h) \quad .$$

- **F-test for Equality of Several Means:** This test is carried out by the technique of Analysis of Variance, which plays a very important and fundamental role in Design of Experiments in Agricultural Statistics.

3.7 RELATION BETWEEN t AND F DISTRIBUTIONS

In F-distribution with (ν_1, ν_2) d.f. , take $\nu_1 = 1$, $\nu_2 = \nu$ and t^2 = F, i.e.,*dF = 2tdt.*

Thus the probability differential of F transforms to

$$dG(t) = \frac{(1/\nu)^{1/2}}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{(t^2)^{\frac{1}{2}-1}}{\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}} 2tdt \qquad , \qquad 0 \le t^2 < \infty$$

$$= \frac{(1/\nu)^{1/2}}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}} dt \qquad -\infty < t < \infty$$

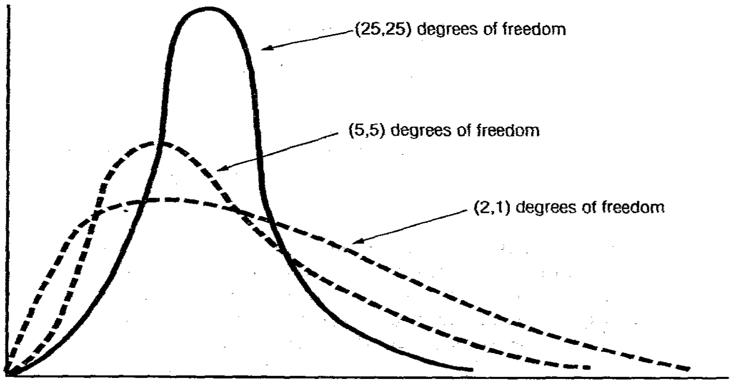
the factor 2 disappearing since the total probability in the range $(-\infty, +\infty)$ is unity. This is the probability function of Student's t-distribution with ν d.f. Hence we have the following relation between t and F distributions.

If a statistic t follows Student's t distribution with n d.f., then t^2 follows Snedecor's F-distribution with (1, n) d.f. Symbolically,

$$if \qquad t \sim t_{(n)} \qquad then \quad t^2 \sim F_{(1,n)}$$

3.8 SHAPE OF F DISTRIBUTION

As we can see in the below given figure, the F distribution has a single mode. The specific shape of an F distribution depends on the number of degrees of freedom in both the numerator and the denominator of the F ratio. But, in general, the F distribution is skewed to the right and tends to become more symmetrical as the numbers of degrees of freedom in the numerator and denominator increase.



Here we see that the probability p(F) increases steadily at first until it reaches (corresponding to the modal value which is less than 1) and then slowly so as to become tangential at $F = \infty$, i.e., F-axis is an asymptote right tail.

3.9 EXERCISES

1. Establish relationship between t and F distribution

2. Write ‘Yes’ if the statements given below are correct, otherwise write ‘No’

(a) Degrees of freedom take care of the sample size in a decision problem about a hypothesis.

(b) Randomized test also involves some statistic.

(c) Each statistic has some distribution.

(d) Standard deviation of an estimate and standard error are the same.

(e) t-value lies between 0 and ∞ .

(f) Z-value lies between 0 and 1

(g) Variance of a sample can be any value between — ∞ and $+\infty$

3. Find the values for the following with the help of tables

(a) $F_{(7,11)}$ when $\alpha = 0.05$

(b) $F_{(10,12)}$ when $\alpha = 0.02$

(c) $F_{(5,8)}$ when level of significance is 5%

4. Mention important uses of F distribution.

5. Prove that if X has the F-distribution with (m, n) if. and Y has F-distribution with (n, m) d.f., then for every a > 0,

$$P(X \le a) + P\left[Y \le \frac{1}{a}\right] = 1$$

6. If F(n_1, n_2) represent an F-variate with n_1 and n_2 degrees of freedom, prove that F(n_2, n_1) is distributed as 1/F (n_1, n_2) variate. Deduce that.

$$P[F(n_1, n_2) \ge D] = P\left[F(n_2, n_1) \le \frac{1}{D}\right]$$

Or

Show that probability points of $F(n_2, n_1)$ can be obtained from those of $F(n_1, n_2)$

7. Derive the distribution of $F = S_1^2 / S_2^2$, where S_1^2 and S_2^2 are two independent unbiased estimates of the common population variance σ^2, defined by

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

Where $S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2$ and $S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(x_{2j} - \bar{x})^2$

8. If X_1, $X_{2,}$ X_3……… X_m,X_{m+1} are independent normal variates with zero mean and standard deviation σ, obtain the distribution of

$$\sum_{i=1}^{m} X_i^2 \Big/ \sum_{i=m+1}^{m+n} X_i^2$$

9. Check whether the moment generating function of F distribution exists or not

10. Why larger of the two variances is taken as numerator while computing F statistic ?

11 State the assumptions underlying Snedecor's F-test when applied to both single and two-sample problems.

12 Obtain formulae for 95% confidence limits of the variance of a normal population, when the mean is (i) known, (ii) unknown.

13 Show that the probability curve of the distribution of F is positively skewed.

14 If X has an F distribution with n_1 and n_2 d.f., what will be the distribution of 1/X and how this result can be used ?

15 If X is t-distributed, show that X^2 is F-distributed.

Unit 2 **Lesson 4**

THEORY OF ESTIMATION

Structure:

4.1 Introduction

4.2 Objectives

4.3 Concept of Statistic and Parameter

4.4 Theory of estimation

4.5 Criterion for good estimator

4.6 Unbiasedness

4.7 Exercises

4.8 Self assessment questions

4.1 INTRODUCTION:

Whenever we take a sample, we do so with an idea of learning. something about the population from which the sample is drawn. In statistical terminology, this learning is termed as statistical inference which is of two kinds; estimator and hypothesis testing.

Everyone makes estimates. When you are ready to cross a street, you estimate the speed any car that is approaching, the distance between you and that car, and your own. Having made these quick estimates, you decide whether to wait, walk, or run. All managers must make quick estimates too. The outcome of these estimates can effect their organizations as seriously as the outcome of your decision as to whether to cross the street. University department heads make estimates of next year's enrollment in statistics, Credit managers estimate whether a purchaser will eventually pay his bills. . All, these people make estimates without worry about whether they are scientific but with the hope that the estimates bear a reasonable resemblance to the outcome. Here we can make two types of estimates about a population: **a point estimate** and an **interval estimate**. A point estimate is a single number that is used to estimate an unknown population parameter. **An interval estimate** is a range of values used to estimate a population parameter. It indicates the error in two ways: by the extent of its range and by the probability of the true population parameter lying within that range. In this whole process sampling and theory of probability plays a vital role.

The object of sampling is to study the features of the population on the basis of sample observations. A carefully selection sample is expected to reveal these features, and hence we shall infer about the population from a statistical analysis of the sample. This process is known as **Statistical Inference**.

There are two types of problems. Firstly, we may have no information at all about some characteristics of the population, especially the values of the parameters involved in the distribution, and it is required to obtain estimates of these parameters. This is the problem of **estimation.** Secondly, some information or hypothetical values of the parameters may be available, and it is required to test how far the hypothesis is tenable in the light of the information provided by the sample. This is the problem of **Test of Hypothesis or Test of Significance.**

4.2 Objectives

On careful reading of this lesson learner will be able

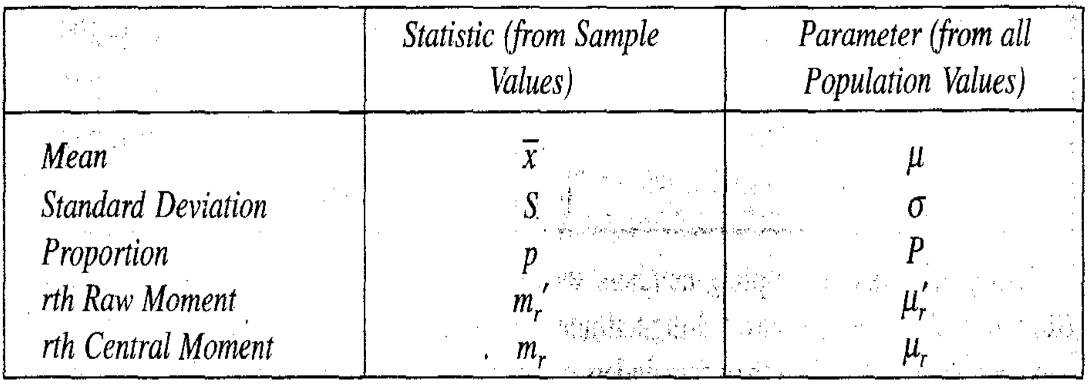
- To have the basic knowledge of theory of estimation and
- To learn how to estimate certain characteristics of a population from samples

- To learn the strengths and shortcomings of point estimates and interval estimates
- To calculate how accurate our estimates really are
- To calculate the sample size required for any desired level of precision in estimation

4.3 Concept of Statistic and Parameter, estimate and estimator

Solution: Any statistical measure calculated on the basis of sample observations is called a **Statistic;** e.g., sample mean, sample standard deviation., the proportion of defectives observed in the sample, etc. Any statistical measure based on all units in the population is called a **Parameter**; e.g., population mean, population standard deviation, proportion of defectives in the whole lot, etc. The value of a statistic varies from sample to sample; but the parameter remains a constant. Usually parameters are unknown and statistics are used as estimates of parameters. The probability distribution of a statistic is called its '**sampling distribution'** and the standard deviation in the sampling distribution is called '**standard error'** of the statistic. *However, since the parameter is constant it has neither a sampling distribution nor a standard error.*

The following notations will be used to distinguish between statistic and parameter:



	Statistic (from Sample Values)	*Parameter (from all Population Values)*
Mean	$\bar{x}$	μ
Standard Deviation	S	σ
Proportion	p	P
rth Raw Moment	m_r'	μ_r'
rth Central Moment	m_r	μ_r

Any sample statistic that is used to estimate a population parameter is called an estimator; that is, an estimator is a sample statistic used to estimate a population parameter. The sample mean $\bar{x}$ can be an estimator of the population mean μ, and the sample proportion p can be used as an estimator of the population proportion P. We can also use the sample range as an estimator of the population range.

When we have observed a specific numerical value of our estimator, we call that value an **estimate**. In other words, an estimate is a specific observed value of a statistic. We form an estimate by taking a sample and computing the value taken by our estimator in that sample. Suppose that we calculate the mean odometer reading (mileage) from a sample of used taxis and find it to be 98,000 miles. If we use this specific value to estimate the mileage for a whole fleet of used taxis, the value 98,000 miles would be an estimate populations, population parameters, estimators, and estimates.

4.4 THEORY OF ESTIMATION

Whenever we take a sample, with the aim of having idea about the population from which the sample is drawn. In statistical terminology, this learning is termed as statistical inference which is of two kinds; estimator and hypothesis testing. Both types of statistical inference utilise the information provided by the sample, for drawing some conclusions about the parameters of the population; yet each type of inference uses this information in different ways. The information by the sample is given by sample.

Suppose we have a random sample x_1, x_2, ...x_n on a variable x, whose distribution in the population involves an unknown parameterθ. It is required to find an estimate of θ on the basis of sample values. The theory of estimation is divided into two parts: point estimation and interval estimation. The theory of estimation is divided into two parts: point estimation and interval estimation.

The aim of point estimation is obtain a single value which is the best guess of the parameter interest. In interval estimation the object is to obtain interval within which the true value of the parameter may -be said to lie with some given level of probability which expresses the confidence we have that the value lies within the stipulated range.

(i) Point Estimation, and

(ii) **Interval Estimation**.

In point estimation the estimated value is given by a single quantity, which is a function of sample observations (i.e. statistic). This function is called the **estimator** of the parameter, and the value of the estimator in a particular sample is called an ‘**estimate**’.

In short point estimate is a single number that is used to estimate an unknown population parameter.

For example, a department head would make a point estimate if she said, "Our current data indicate that this course will have 350 students in the next class."

A point estimate is often insufficient, because it is either right or wrong. If you are told only that his point estimate of enrollment is wrong, you do not know how wrong it is, and you cannot be certain of the estimate's reliability. If you learn that it is off by only 10 students, you would accept 350 students as a good estimate of future enrollment. But if the estimate is off by 90 students, you would reject it as an estimate of future enrollment. Therefore, a point estimate is much more useful if it is accompanied by an estimate of the error that might be involved.

An interval estimate is a range of values used to estimate a population parameter. It indicates the error in two ways: by the extent of its range and by the probability of the true population parameter lying within that range. In this case, the department head would say something like, "I estimate that the true enrollment in this course in the fall will be between 320 and 370 and that it is very likely that the exact enrollment will fall within this interval." he has a better idea of the reliability of her estimate. If the course is taught in sections of about 100 students each, and if he had tentatively scheduled five sections, then on the basis of his estimate, he can now cancel one of those sections and offer an elective instead.

Summing up we can say that in interval estimation, an interval within which the parameter is expected to lie is given by using two quantities based on sample values. This is known as **Confidence interval**, and the two quantities which are used to specify the interval, are known as **Confidence Limits**.

4.5 POINT ESTIMATION—CRITERIA FOR GOOD ESTIMATORS

The theory of estimation is divided into two parts: point estimation and interval estimation. The aim of point estimation is obtain a single value which is the best guess of the parameter of interest. In interval estimation the object is to obtain interval within which the true value of the parameter may -be said to lie with some given level of probability which expresses the confidence we have that the value lies within the stipulated range.

There are various methods with which we may obtain point estimation or point estimates of the parameters of the phenomena under study. There is naturally a problem of choosing the one which gives us the best estimate. Also, how are we to decide whether any estimate is the best or whether it is good or better than another obtained by a different method? That is; we need to devise a criterion to call an estimator a best one. We, therefore, have to do two things: (a) to specify various properties of an estimator that go to make it a best estimator and, (b) to devise different methods that could give rise to estimators that possess at least some of these desirable properties.

Assume some random variable X whose distribution is characterized by a specific parameter, θ, which we want to estimate. Thus the parent population consists of all possible values of X and θ is one of the parametric characteristics of this population. An estimator of θ is denoted by $\hat{\theta}$ and since it is obtained by substituting the sample observations of X into a formula, we write

$$\hat{\theta} = \hat{\theta}(x_1, x_2, ...x_n)$$

which is read as "$\hat{\theta}$ is a function of X_1, X_2 ,…….,X_n

Since the accuracy of an estimator, in general, increases with the number of observations in the sample data, the desirable properties of the estimators are divided into two groups depending upon the size of sample.

Finite sample or small sample properties refer to properties of the sampling distribution of an estimator based on any fixed sample size. On the other hand asymptotic or large sample properties are the properties of the sampling distribution of the estimator which is obtained from a sample whose **size approaches infinity**

Many functions of sample observations may be proposed as estimators of the same parameter. For example, either the mean or median or mode of the sample values may be used to estimate the parameter μ of the Normal distribution with p.d.f.

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma}$$

Naturally we have to choose one among these estimators on the basis of certain criteria. The desirable properties or the main criteria for a good estimator obtained from small samples according to R.A. Fisher are

(i) Unbiasedness,

(ii) Consistency,

(iii) Efficiency,

(iv) Sufficiency.

4.6 Unbiasedness:

A statistic t is said to be an Unbiased Estimator of a parameter θ, if the expected value of t is.

E(t)= θ

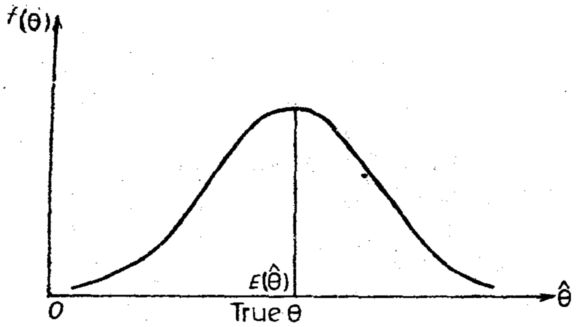
Otherwise, the estimator is said to be 'biased'. The bias of an estimator is defined as the difference between its expected value and the true value of the parameter. Mathematically, the bias of a statistic in estimating θ is given as

Bias= E(t) — θ

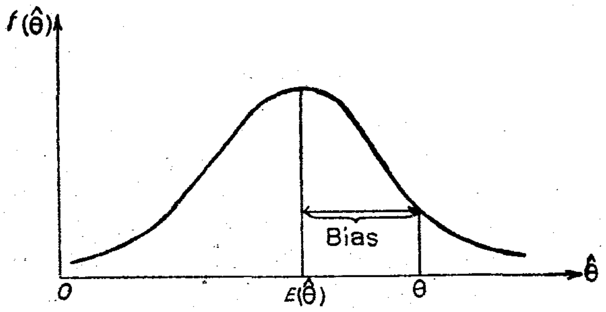
If E(t) — $\theta >0$ t is said to be positively biased.

E(t) — $\theta < 0$ t is said to be negatively biased

When the bias is positive, that is, when the mean value of the distribution is larger than its parameter, then the estimator is said to be upward biased. Conversely, when the bias is negative, the estimator is biased downwards.



θ̂ is an unbiased estimator of θ



θ̂ is a biased estimator of θ

Since the distribution is assumed to be a symmetric one, the mean is shown at the centre of the distribution, and it is equal to the true value of the parameter in Figure on the left hand side; but is not equal to the true value of the parameter in Figure on the left hand side

Remark: A concept related to bias is sampling error

Sampling error = $\hat{\theta} - \theta$

That is, sampling error is simply the difference between the value of estimator and the true value of the parameter to be estimated

Illustration of unbiased estimator

Let us consider the number set P = {2,4,6}.If we consider P as a population, then

Population Mean = μ = (2+4+6)/3 = **4.**

Variance = σ^2 = $[(2-4)^2 + (4-4)^2 + (6-4)^2]/3$ = 8/3 = 2.666667.

Standard deviation = sqrt(σ^2) = sqrt(8/3) = 1.632993.

If P is a sample, then

Sample mean = $\bar{x}$ = (2+4+6)/2 = 4.

Unbiased estimate of variance of sample mean is

s^2 = [(2-4)2 + (4-4)2 + (6-4)2]/2 = 8/2 = 4.

Sample standard deviation = $\sqrt{s^2} = \sqrt{4} = 2$

{ The formula for s^2 involves dividing by n-1. In this case, n=3. Hence n-1 = 2.}

Now, let's consider P to be a population and draw all possible samples of size 2 chosen from P, with replacement. There would be 3x3 = 9 samples.

Sample	$\bar{x}$ For sample	s^2 for sample	s for sample	S^2 for sample	S for sample
2,2	2	0	0	0	0
2,4	3	2	1.414214	1	1
2,6	4	8	2.828427	4	2

4,2	3	2	1.414214	1	1
4,4	4	0	0	0	0
4,6	5	2	1.414214	1	1
6,2	4	8	2.828427	4	2
6,4	5	2	1.414214	1	1
6,6	6	0	0	0	0
Column Means	4	2.666667	1.257079	1.333333	0.888889

To summarize, we have listed all samples of size 2 (with replacement) from a population P of size 3. We have calculated statistics for each sample of size 2. Here is an important definition:

A statistic used to estimate a population parameter is unbiased if the mean of the sampling distribution of the statistic is equal to the true value of the parameter being estimated.

The mean of the sample means (4) is equal to μ, the mean of the population P. **This illustrates that a sample mean $\overline{X}$ is an unbiased statistic**. It is sometimes stated that $\overline{X}$ is an unbiased estimator for the population parameter μ .

The mean of the sample values of s^2 (2.666667) is equal to σ^2 , the variance of the population P. This illustrates that the sample variance s^2 is an unbiased statistic. It is sometimes stated that s^2 is an unbiased estimator for the population variance σ^2.

Here we see that the sample statistic s is not an unbiased statistic. That is, the mean of the s column in the table (1.257079) is not equal to the population parameter σ = 1.632993.

Also, if we use the S^2 formula for samples, the resulting statistics are not unbiased estimates for a population parameter. Note that the means for the last two columns in the table are not equal to population parameters.

In summary, the sample statistics $\overline{x}$ and s^2 are unbiased estimators for the population mean μ and population variance σ^2, respectively.

4.7 EXERCISES

Exercise:1 If x_1, x_2, ...x_n is a random sample from an infinite population with variance σ^2, and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, show that $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is a biased estimator of σ^2 but the bias becomes negligible for large n. Give an unbiased estimator of σ^2 here.

Solution:- Let μ and σ^2 be the mean and variance of the population. Then $E(x_i) = \mu$. And Var (x_i) = $E(x_i - \mu)^2 = \sigma^2$ for each i = 1, 2, ... n. The variance of the sample is

$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ We have to show that $E(S^2) \neq \sigma^2$

Now, $$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{\sum_{i=1}^{n} x_i^2}{n - \overline{x}^2} = \frac{\sum_{i=1}^{n} y_i^2}{n - \overline{y}^2}$$ where $y_i = x_i - \mu$

($\because$ S.D is unaffected by change of origin)

$$=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-(\overline{x}-\mu)^2$$

$$\because \quad E[S^2]=\frac{1}{n}\left[\sum_{i=1}^{n}E(x_i-\mu)^2\right]-E(\overline{x}-\mu) = \frac{\sum_{i=1}^{n}\sigma^2}{n}-\text{var}(\overline{x})$$

$$=\sigma^2-\frac{\sigma^2}{n}=\frac{n-1}{n}\sigma^2\neq\sigma^2$$

This show that $S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$ is a biased estimator of σ^2

Thus for large n bias will be negligibly small if we write

$$s^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

We see that

$s^2=\frac{n}{n-1}S^2$ so that $E[s^2]=\frac{n}{n-1}E[S^2]$

$$=\frac{n}{n-1}\frac{n-1}{n}\sigma^2=\sigma^2$$

This shows that s^2 is unbiased estimator of σ^2

Note : the distinction between S^2 and s^2 in which only the denominators are different.S^2 is the variance of the sample observations, but s^2 is the ‘unbiased estimator’ of the variance (σ^2) in the population.

Exercise:2 Show that the sample mean based on a simple random sample with replacement (**srswr**) is an unbiased estimator of the population mean.

Solution:- Suppose we have a sample x_1, x_2, ...x_n obtained by using simple random sampling with replacement obtained from a finite population of size N i.e$X_1, X_2, \ldots\ldots\ldots.., X_N$

We have to show that $E[x] = \mu$

In SRSWR any of the population members X_1, X_2, ... Xn may appear at the i-th drawing, i.e. x is a random variable with the following probability distribution:

x_i	X_1	X_2	………	X_N	Total
Prob	1/N	1/N		1/N	1

Therfore $E[x_i] = \left(\frac{1}{N}\right)X_1 + \left(\frac{1}{N}\right)X_2 + \ldots\ldots\ldots\ldots + \left(\frac{1}{N}\right)X_N$

$= (X_1+X_2+\ldots\ldots\ldots..,+X_N)/N = \mu$

Hence $E(\overline{x}) = E\left[\frac{1}{n}(x_1 + x_2 + ..x_n)\right] =$

$\frac{1}{n}[E(x_1) + E(x_2) + \ldots.. + E(x_n)]$

$= (\_\ \mu + \mu + \ldots. + \mu) / n = n\mu / n = \mu$

This shows that $\overline{x}$ is an unbiased estimator of μ

[**Note**: This result holds in all cases of random sampling, irrespective of whether the sample is drawn ‘with replacement’ or ‘without replacement’ from a finite population or from an infinite population.]

Exercise:3 If $x_1, x_2, ...x_n$ is a random sample from N (μ ,1) show that t $= \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ is an unbiased estimate of $\mu^2 + 1$

Sol: We are given $E[x_i] = \mu$ and $v(x_i) = 1$; $i = 1,2,...n$

Now $E[x_i^2] = v(xi) + E[x_i]^2 = 1 + \mu^2$

$E[t] = \frac{1}{n}E[x_i^2] = \frac{1}{n}\sum_{i=1}^{n}[1+\mu^2] = [1+\mu^2]$ hence t $= \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ is an unbiased estimate of $\mu^2 + 1$

Exercise:4 Show that $\frac{\sum_{i=1}^{n} x_i(\sum x_i - 1)}{n(n-1)}$ is an unbiased estimate of θ^2 for the sample values $x_1, x_2, ...x_n$ drawn on X which takes the values 0 and 1 with respective probabilities θ and (1-θ)

Sol: Since $x_1, x_2, ...x_n$ is a random sample from Bernoullian population so that

T = $\sum_{i=1}^{n} x_i \sim B(n,p)$ so that E[t] = nθ and var (T)=nθ(1-θ)

$$E\left[\frac{\sum_{i=1}^{n} x_i(\sum x_i - 1)}{n(n-1)}\right] = \left[\frac{T(T-1)}{n(n-1)}\right] = \frac{1}{n(n-1)}E[T^2 - T] = \frac{1}{n(n-1)}[E(T^2) - E(T)]$$

$$= \frac{1}{n(n-1)}[V(T) + E(T)^2 - E(T)]$$

$$= \frac{1}{n(n-1)}[n\theta(1-\theta) + n^2\theta^2 - n\theta] = \frac{n(n-1)\theta^2}{n(n-1)} = \theta^2$$

Hence proved

4.8 SELF ASSESSMENT QUESTIONS

1. Show that the sample mean ($\overline{x}$)is an unbiased estimator of the population mean(μ)

$$E[\overline{x}] = \mu$$

2. Prove that the sample variance S^2 is a biased estimator of the population variance σ^2 because

Hint: $\left\{ E[S^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \right\}$

3. An unbiased estimator of the population variance σ^2 is given by

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

4. Point out the distinction between s^2 and S^2

5. Let X be distributed in Poisson form with parameter θ show that only unbiased estimate of $\exp\{-(k+1)\theta\}$, k>0 is $T(x)=[-k]^x$ so that T(x)> if x is even and T(x)<0 if x is odd**.**

THEORY OF ESTIMATION

Structure:

5.1 Introduction

5.2 Objectives

5.3 Concept of Consistency, efficiency

5.4 Sufficiency

5.5 Exercises

5.6 Self assessment questions

5.1 INTRODUCTION

A given sample statistic is not always the best estimator of its analogous population parameter. If we consider a symmetrically distributed population in which the values of the median and the mean coincide. In this case, the sample mean would be an unbiased estimator of population median. Also, the sample mean would be a consistent estimator of the population median because, as the sample size increases, the value of the sample mean would tend to come very close to the population median. And the sample mean would be a more efficient estimator of the population median than the sample median itself because in large samples, the sample mean has a smaller standard error than the sample median. At the same time, the sample median in a symmetrically distributed population would be an unbiased and consistent estimator of the population mean but not the most efficient estimator because in large samples, its standard error is larger than that of the sample mean.

In short the main objective of present lesson is get a better idea to decide among a class of unbiased estimator which one is best, consistency and efficiency (in addition to sufficiency) are the criterion on the basis of which question under consideration can be addressed.

5.2 OBJECTIVES

After careful reading of this lesson learner will be able

- To have the basic knowledge about the concepts of consistency , sufficiency and efficiency
- To learn how to decide about certain characteristics of a population from samples
- To calculate how accurate our estimates really are
- To have an understanding about the criterion of good estimator
- To decide about the best estimator

5.3 CONCEPT OF CONSISTENCY, EFFICIENCY

Consistency:

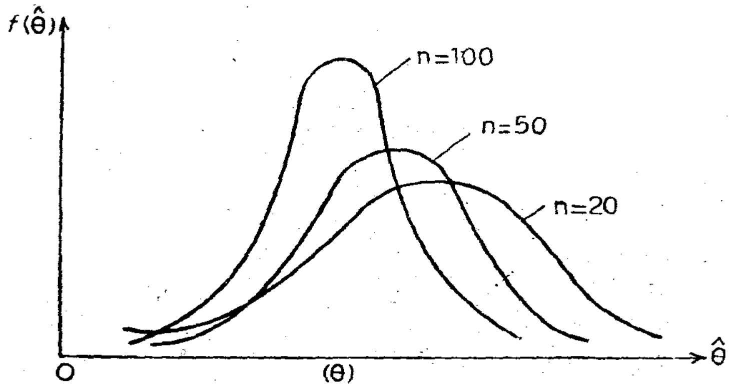
A statistic is a consistent estimator of a population parameter if as the sample size increases; it becomes almost certain that the value of the statistic comes very close to the value of the population parameter. If an estimator is consistent, it becomes more reliable with large samples. Thus, if some one is wondering whether to increase the sample size to get more information about a population parameter, find out first whether your statistic is a consistent estimator. If it is not, one will waste time and money by taking larger samples. A desirable property of a good estimator is that its accuracy should increase when the sample size becomes larger. That is, the estimator is expected to come closer to the parameter as the size of the sample increases.

A statistic t computed from a sample of n observations is said to be a Consistent Estimator of a parameter θ, if it converges in probability to 0 as n tends to infinity. This means that the larger the sample size (n), the less is the chance that the difference between t, and θ will exceed any fixed value. Given any arbitrary small positive quantity $\in$,

$$\underset{n\to\infty}{\mathrm{Lt}}\ P\{|t_n - \theta| > \in\} = 0$$

$$\textit{If}\quad E[t_n] \to \theta \ \textit{ and } \quad Var[t_n] \to 0 \ \textit{as}\, n \to \infty \, \textit{then}\, t_n \ \textit{will be a consistent estomator of } \theta$$

To see whether an estimator is consistent, we should therefore examine its bias and variance as sample size is increased. If both bias and variance decrease as n becomes larger, and at the limit (as $n \to \infty$) both become zero, then estimator is assumed to possess the property of consistency. This is illustrated in Figure which is given below, which shows that as the sample size increases from 20 to 100 observations both bias of and its variance decrease.



Since sum of squared bias and variance is equal to the MSE, the disappearance of the bias and variance as $n \to \infty$ is equivalent to the disappearance of the MSE, so that we can also say;

A statistic t is said to be a consistent estimator of a parameter θ, if $\underset{n \to \infty}{\mathrm{Lt}}\ \mathrm{MSE}[t] = 0$

For example, in sampling from a Normal population N (μ, σ^2). both the sample mean and the sample median are consistent estimators

Efficiency and Minimum Variance: Unbiasedness is a desirable property but not particularly important by itself. It is because this property tells us nothing about the dispersion of the distribution of the estimator. An estimator which is unbiased, but one which has a large variance, will frequently lead to estimates that are quite different from true value of the parameter. On the other hand an estimator which has a very small variance but is biased, is equally (and even more) less useful. In the light of this argument it seems desirable to examine the variance of the distribution of the estimator also.

This criterion based upon the variances of the sampling distributions of the estimators which enables us to choose between the estimators with the comm. on property of consistency usually known as **efficiency.** Of two consistent estimators for the same parameter, the statistic with the smaller sampling variance is said to be "more efficient". Thus if t and t' are both consistent estimators of θ, and

$$Var(t) < Var(t')$$

then t is 'more efficient' than t' in estimating θ,

If a consistent estimator exists whose sampling variance is less than that of any other consistent estimator, it is said to be **"most efficient**"; and it provides a standard for the measurement of 'efficiency' of a statistic. If V_0 be the variance of the most efficient estimator and V be the variance of any other estimator, then the **efficiency** of the estimator is defined as

$$\text{Efficiency} = \frac{V_o}{V}$$

Obviously, the measure of efficiency cannot exceed 1.

In sampling from a Normal population N (μ, σ^2), both the sample mean and the sample median are consistent estimators of t, but

$$\text{Var}(\overline{x}) = \frac{\sigma^2}{n,} \qquad \text{Var(Median)} = \frac{\pi\sigma^2}{2n}$$

Since Var ($\overline{x}$) is smaller than Var(median), mean is more efficient than median in estimating the parameter μ. It can be shown that the sample mean is the most efficient estimator. Hence

$$\text{Efficiency of median} = \frac{\dfrac{\sigma^2}{n,}}{\dfrac{\pi\sigma^2}{2n}} = \frac{2}{\pi} = 0.64 \text{ approx.}$$

Asymptotic efficiency

t is an asymptotically efficient estimator of θ, if

(a) t is consistent, and

(b) t has a smaller asymptotic variance as compared to any other consistent estimator.

The establishment of the first condition does not pose any difficulty. To establish whether consistent estimator satisfies the second condition is more difficult. It is because the variance of any consistent estimator goes to zero as $n \to \infty$

So in the present situation when we are comparing consistent estimators, we choose the one whose variance goes faster to zero (as $n \to \infty$) and call it asymptotically more efficient.

For example consider two estimators t and t* whose distributions have the following mean and variance;

$$\text{Mean} \quad : \quad E[t] = \left(\frac{n-1}{n}\right)\theta; \qquad E[t^*] = \left(\frac{n+1}{n}\right)\theta$$

$$\text{Variance: } Var(t) = \frac{\sigma^2}{n^2}; \qquad Var\,(t^*) = \frac{\sigma^2}{n}$$

Both estimators are asymptotically unbiased and consistent; since their bias and variance become zero as $n \to \infty$ and we can prove that

$$\underset{n\to\infty}{Limit}\, E[t] = \theta; \qquad \underset{n\to\infty}{Limit}\, E[t^*] = \theta;$$

$$\underset{n\to\infty}{Limit}\, Var[t] = 0; \qquad \underset{n\to\infty}{Limit}\, Var[t^*] = 0;$$

However, the variance of t goes faster to zero as $n \to \infty$. Thus t is asymptotically more efficient than the alternative consistent estimator t*.

Minimum Variance Unbiased Estimator (MVUE):If a statistic t =t(x_1, x_2, ...x_n) based on the sample of the size n is such that

(i) t is unbiased for θ for all $\theta \in \Theta$

(ii) It has smallest variance among the class of all the unbiased estimator of θ

Then t is called as minimum variance unbiased estimate of θ .

More precisely t is MVUE of θ if

$E_\theta(t) = \theta$ for all $\theta \in \Theta$ and

$Var_\theta(t) \leq Var_\theta(t')$ for all $\theta \in \Theta$ where t' is any other unbiased estimate of θ

5.4 SUFFICIENCY

An estimator is sufficient if it makes so much use of the information in the sample that no other estimator could extract from the sample additional information about the population parameter being estimated.

A statistic is said to be a 'sufficient estimator' of a parameterθ, if it contains all information in the sample about θ If a statistic t exists such that the joint distribution of the sample is expressible as the product of two factors, one of which is the sampling distribution of t and contains θ, but the other factor is independent of θ, then t will be a sufficient estimator of θ

Thus if $x_1, x_2, \ldots\ldots\ldots x_n$ is a random sample from a population whose p.m.f or p.d.f is

$f(x.\theta)$ an d t is sufficient statistic for the estimation ofθ,we can write

$$f(x_1.\theta),\ f(x_2.\theta),\ f(x_3.\theta)\ldots\ldots\ldots\ \ldots\ldots,\ f(x_n.\theta)$$
$$=\ g(t.\theta),\ h(x_1, x_2.x_3\ldots\ldots\ldots\ ,x_n)$$

Where $g(t.\theta),$ is the sampling distribution of t and contains only θ, but $h(x_1, x_2.x_3\ldots\ldots\ldots, x_n)$is independent of$\theta$, since parameter θoccurring in the joint distribution of all the sample observations can be contained the distribution of statistic t, it is said that t alone can provide all the information regarding θ, therefore sufficient for θ

Fisher-Neyman criterion: A statistic t =t(x_1, x_2, ...x_n) is sufficient for parameter if and only if the likelihood function(joint p.d.f of the sample) can be expressed as θ

$$L = \prod_{i-1}^{n} f(x_i, \theta) = g(t, \theta)\ k(x_1, x_2, ... x_n)$$

Where $g(t, \theta)$ is the p.d.f of the statistic t and $k(x_1, x_2, ... x_n)$ is the function of sample observations only, independent of θ

5.5 Exercises

Exercise:-1 Examine the desirable properties (Unbiasedness, consistency, sufficiency and asymptotic properties) in case of the following three estimators which have been proposed to estimate true mean (μ) from a random sample of observations on X_1, X_2,………X_n (It is assumed that parent population is normally distributed.)

(i) $$(\overline{x}) = \frac{\sum X_i}{n}$$

(ii) $$\hat{\mu} = \frac{\sum X_i}{n+1}$$

(iii) $$\mu^* = \frac{X_1}{2} + \frac{\sum_{i=2}^{n} X_i}{2n}$$

Solution:- Unbiasedness:

(i) $$E(\overline{x}) = E\left[\frac{\sum X_i}{n}\right] = \frac{1}{n}\sum E[X_i] = \mu$$

Hence $\overline{x}$ is unbiased estimator of μ

(ii) $$E[\hat{\mu}] = \frac{1}{n+1}\sum E[X_i] = \left(\frac{1}{n+1}\right)E[X_i] = \left(\frac{1}{n+1}\right)\mu$$

Hence $\hat{\mu}$ is a biased estimator of μ

(iii) $$E[\mu^*] = E\left[\frac{X_1}{2} + \frac{\sum_{i=2}^{n} X_i}{2n}\right] == \frac{1}{2}E[X_1] + \frac{1}{2n}\sum_{i=2}^{n} E[X_i]$$

$$= \frac{1}{2}\mu + \left(\frac{n-1}{2n}\right)\mu = \left(\frac{2n-1}{2n}\right)\mu \neq \mu$$

Hence μ^* is a biased estimator of μ

Efficiency

Only $\overline{X}$ is to be examined for this property ($\because$ other two estimators are biased)

Var $(\overline{X}) = \frac{\sigma^2}{n}$ and it can be shown that $\frac{\sigma^2}{n}$ is the minimum variance amongst the unbiased estimators of μ. Thus $\overline{X}$' is an efficient estimator μ

Asymptotic Properties:

$$\underset{n\to\infty}{Limit}\ [\overline{X}] = \underset{n\to\infty}{Limit}\ [\mu] = \mu$$

Hence $\overline{X}$ is **Asymptotically** unbiased estimator of μ

(ii) $$\underset{n\to\infty}{\text{Limit}}\, E[\hat{\mu}] = \underset{n\to\infty}{\text{Limit}}\left(\frac{n}{n+1}\right)\mu = \mu$$

Hence $\hat{\mu}$ is **Asymptotically** unbiased estimator of μ

(iii) $$\underset{n\to\infty}{\text{Limit}}\, E[\mu^*] = \underset{n\to\infty}{\text{Limit}}\left(\frac{2n-1}{2n}\right)\mu = \mu$$

Hence μ^* is **Asymptotically** unbiased estimator of μ

Consistency:

(i) $$\text{Var}\,(\overline{X}) = \frac{\sigma^2}{n} \therefore \qquad \underset{n\to\infty}{\text{Limit}}\frac{\sigma^2}{n} = 0$$

$\overline{X}$ **is consistent estimators.**

(ii) $$\text{Var}[\hat{\mu}] = \text{Var}\left(\frac{1}{n+1}\right)\sum_i X_i = \left(\frac{1}{n+1}\right)^2 \sum_i \text{Var}(X_i) = \left(\frac{1}{n+1}\right)^2 n\sigma^2$$

$$\therefore \qquad \underset{n\to\infty}{\text{Limit}}\frac{n}{(n+1)^2}\sigma^2 = 0$$

Hence $\hat{\mu}$ **is consistent estimators.**

(iii) $$Var[\mu^*] = Var\left[\frac{X_1}{2} + \frac{\sum_{i=2}^{n} X_i}{2n}\right] = \frac{1}{4}Var(X_1) + \left(\frac{1}{2n}\right)^2 \sum_{i=2}^{n} Var(X_i)$$

$$= \left(\frac{n^2+n}{4n^2}\right)^2 \sigma^2$$

$$\therefore \qquad \underset{n\to\infty}{\text{Limit}}\left(\frac{n^2+n}{4n^2}\right)^2 \sigma^2 = \frac{\sigma^2}{4} \neq 0$$

Hence μ^* **is not a consistent estimators of** μ**.**

Asymptotic efficiency:

Only $\overline{X}$ and $\hat{\mu}$ satisfy the condition of consistency and thus needs to be examined for this property. $\overline{X}$ is efficient even in case of small samples, hence it is asymptotically efficient as well.

$$\text{Var}[\hat{\mu}] = \left(\frac{1}{n+1}\right)^2 n\sigma^2 = \left(\frac{n}{n+1}\right)^2 \frac{\sigma^2}{n}$$

In large samples $\left(\frac{n}{n+1}\right)$ will be close to infinity; as such asymptotic variance of $\hat{\mu} = \frac{\sigma^2}{n}$; which is same as that of $\overline{X}$.It follows, therefore, that $\hat{\mu}$ is **also asymptotically efficient.**

Exercise:-2 If x_1, x_2, ...x_5 is a random sample of size 5 from normal population with mean μ. Consider the following estimate of estimate of μ

(i) $t_1 = \frac{x_1 + x_2 + ... + x_5}{5}$ (ii) $t_2 = \frac{x_1 + x_2}{2} + x_3$ (iii) $t_3 = \frac{2x_1 + x_2 + \lambda\, x_3}{3}$

Where λ is such that t_3 is unbiased estimate of μ. Find λ, are t_1 and t_2 unbiased, state giving reasons which is the best among t_1, t_2 and t_3.

Sol**:** Since sample is from normal population with mean. So that $E(x_i) = \mu$ nd $v(x_i) = \sigma^2$ and cov(x_i, x_j)=0 ; i= 1,2…n Now

(i) E(t_1)= $\frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}\sum_{i=1}^{5} \mu = \mu$ It means that t_1 is unbiased estimate of μ

(ii) E(t_2)= $\frac{1}{2}E(x_1 + x_2) + Ex_3) = \frac{1}{2}(\mu + \mu) + \mu = 2\mu$

It means that t_2 is biased estimate of μ

(iii) E(t$_3$)= $\mu \Rightarrow \frac{1}{3}E(2x_1+x_2+\lambda x_3)=\mu$ or $E(2x_1+x_2+\lambda x_3)=3\mu$ or $2\mu+\mu+\lambda\mu=3\mu \Rightarrow \lambda=0$

Now Var(t$_1$)=$\frac{V(x_1)+V(x_2)+...+V(x_5)}{25}=\frac{1}{5}\sigma^2$

Var(t$_2$)= $\frac{V(x_1)+V(x_2)}{4}+V(x_3)=\frac{3}{2}\sigma^2$

Var(t$_2$)= $\frac{4V(x_1)+V(x_2)}{9}+V(x_3)=\frac{5}{9}\sigma^2$

Since the variance of t$_1$ is minimum so t$_1$ is the best estimate of μ

Exercise-3 Let x$_1$, x$_2$, ...x$_n$ be a random sample from N (μ,σ^2)find the sufficient statistic for μ and σ^2

Sol Let us write θ=(μ,σ^2) the

$$L=\prod_{i-1}^{n} f(x_i,\theta)=\left(\frac{1}{\sigma\sqrt{2\Pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)=$$

$$\left(\frac{1}{\sigma\sqrt{2\Pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i)^2-2\mu\sum_{i=1}^{n}x_i+n\mu^2\right]\right)=g_\theta[t(x)]h(x)$$

Where $g_\theta[t(x)]=\left(\frac{1}{\sigma\sqrt{2\Pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}t_2x-2\mu\, t_1(x)+n\mu^2\right]\right)$

And t(x) ={t$_1$(x), t$_2$(x}=$\left[\sum_{i=1}^{n}(x_i),\ \sum_{i=1}^{n}(x_i)^2\right]$ and h(x)=1

Thus t$_1$(x)= $\sum_{i=1}^{n}(x_i)$, is sufficient for μ and t$_2$(x}= $\sum_{i=1}^{n}(x_i)^2$ is sufficient for σ^2

5.6 SELF ASSESSMENT QUESTIONS

Question:-1. When would you say that estimate of a parameter is good? In particular, discuss the requirements of consistency and Unbiasedness of an estimate. Give an example to show that a consistent estimate need not be unbiased.

Question:-2. Discuss the terms (i) estimate, (ii) consistent estimate, (iii) unbiased estimate, of a parameter and show that sample mean is both consistent and unbiased estimate of the population mean.

Question:-3 (b) If X_1, X_2, X_3, ..., Xn,. are the sample means based on samples of sizes n_l, n_2, n_3 ..., n_r. respectively, an unbiased estimator

$$t = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + ... + n_r\overline{x}_r}{k}$$

Has been defined to estimate μ .Find the value of k.

Question:-4 We are given that

$$f(x, \mu, \sigma^2) = \frac{1}{\sigma 2} \exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]; \qquad \mu \le x < \infty, \qquad -\infty < \mu < \infty,$$

$$\text{and } 0 < \sigma < \infty$$

Obtain

(i) an unbiased estimate of μ when σ is known

(ii) an unbiased estimate of σ when μ is known

Question:-5

(i) Does the consistency of an estimator imply that its variance approaches zero as the sample size increases without limit?

(ii) Why is asymptotic efficiency defined only for consistent estimators?

Question:-6 Examine whether following statement are true or false.

(i) Consistent estimators re asymptotically unbiased.

(ii) Bias and error are the two statistical terms which refer to the same characteristic of an estimator.

(iii) Mean Square Error is the difference of two quantities:

variance and square of bias.

(iv) Sample variance is unbiased estimator of the population variance.

Question:-7 Discuss whether Unbiasedness or efficiency is the more desirable property of an estimator to be used to estimate the annual exports of each product of a developing country, when:

(a) Suppose you wish to establish a long run average annual growth rate for total exports;

(b) Suppose you wish to establish import controls for a given year based on amount of foreign exchange available from exports average annual growth rate for total exports

Unit 2 **Lesson 6**

METHODS OF ESTIMATION

Structure:

6.1 Introduction

6.2 Objectives

6.3 Method of Maximum Likelihood

6.4 Exercises based on Method of Maximum Likelihood

6.5 Method of Moments

6.6 Exercises based on Method of moments

6.7 Self assessment questions

6.1 INTRODUCTION

Whenever we take a sample, we do so with an idea of learning. something about the population from which the sample is drawn. In statistical terminology, this learning is termed as statistical inference which is of two kinds; estimation and hypothesis testing. Both types of statistical inference utilise the information provided by the sample, for drawing some conclusions about the parameters of the population; yet each type of inference uses this information in different ways.

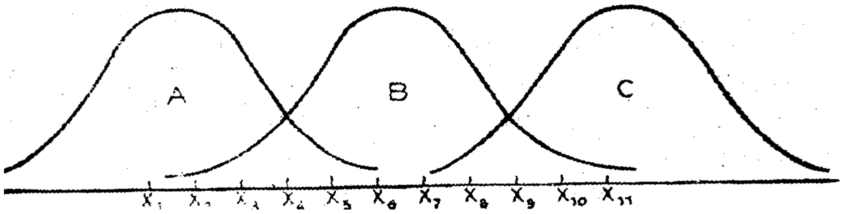
So far we have been discussing the requirements of a good estimator There are various methods of estimation which lead us to estimators that possess different properties. These estimators are known by the names that indicate the nature of the technique used in deriving the formula. The method of moments, least squares method and the maximum likelihood method; all three methods lead to estimators which are known by the names of these techniques.

6.2 OBJECTIVES

The main aim of this lesson is to enable the learners to obtain the estimators which possess the requirements of a of a good estimator by making use of the different estimation techniques such as method of moments, least squares method and the maximum likelihood method

6.3 METHOD OF MAXIMUM LIKELIHOOD

The most important procedure of estimation is the method of maximum likelihood. The basic principle underlying this technique of estimation is that different populations generate different samples, and that any given sample is more likely to have come from some population than from others. Assume that we obtain a sample of n—observations of whose parent population is normal. In fact our sample might have been generated by many different normal populations. But suppose the mean of our observed sample is 10. Now, we ask ourselves; to which population does our sample most likely belong? In general, as we have said, any normal population could be its parent population and the one which has mean equal to 10 (or near about 10) is likely to generate samples with mean equal to 10. As shown in the below given figure.



If X_1, X_2, X_3, ..., X_{11} depict 11 specific sample observations. These observations could have come from any of the normal populations A, B or C. The probability of obtaining our sample from A, or C appears to be very small, but the probability of getting the same sample from population B is very high. As such we say that the particular sample is more likely to have come from population B than from populations A or C. Here we did not refer to the variance of the different populations, and as we know, every population is characterised by its mean and variance. A sample with large variance is more likely to be obtained from a population with large variance than from a population with a small variance. In other words we ought to consider combinations of specific mean and variance of the population in relation to combinations of specific mean and variance of the (observed) sample.

With this background let us now define the maximum likelihood estimator in a formal way.

This method was initially formulated by C.F. Gauss but as a general method of estimation was introduced by Prof. R.A. Fisher.

Let x_1, x_2, ... x_n, be a random sample from a population with p.m.f. (for discrete case) or p.d.f. (for continuous case)f(x, θ), where θ is the parameter. Then the joint distribution of the sample observations viz.

$$L = f(x_1.\theta),\ f(x_2.\theta),\ f(x_3.\theta).........\ .....,\ f(x_n.\theta) = \prod_{i-1}^{n} f(x_i, \theta)$$

is called the Likelihood Function of the sample.

The Method of Maximum Likelihood consists in choosing as an estimator of θ that statistic, which when substituted for θ, maximizes the likelihood function L. Such a statistic is called a maximum likelihood estimator (m.l.e.) denoted by θ_0

Since log L is maximum when L is maximum, in practice the m.l.e. of θ is obtained by maximizing log L. This is achieved by differentiating log L partially with respect to θ, and using the two relations

$$\left[\frac{\partial}{\partial\theta}\log \mathrm{L}\right]_{\theta=\theta_0} = 0, \quad \left[\frac{\partial^2}{\partial\theta^2}\log \mathrm{L}\right]_{\theta=\theta_0} < 0 \qquad ......(1)$$

Here L>0 and Log L are non decreasing function of L. Eq (1) can be rewritten by

$$\frac{1}{\mathrm{L}}\frac{\partial \mathrm{L}}{\partial\theta} = 0 \qquad \Rightarrow \frac{1}{\mathrm{L}}\frac{\partial\,\mathrm{LogL}}{\partial\theta} = 0 \qquad ....(2)$$

Here (2) is termed as likelihood equation for estimating θ

Properties of maximum likelihood estimator (m.l.e.)

We make the following assumptions, known as the Regularity Conditions:

(i) The first and second order denvatives, viz., $\frac{\partial}{\partial\theta}\log \mathrm{L}$ and $\frac{\partial^2}{\partial\theta^2}\log \mathrm{L}$ exist and are continuous functions of θ in a range R (including the true value θ_0 of the parameter) for almost all x. For every θ in R

$$\left[\frac{\partial}{\partial\theta}\log \mathrm{L}\right] < \mathrm{F}_1(\mathrm{x}), \qquad \left[\frac{\partial^2}{\partial\theta^2}\log L\right] < F_2(x)$$

where F$_1$(x) and F$_2$(x) are integrable functions over $(-\infty, \infty)$

(ii) The third order derivative $\frac{\partial^3 \log \mathrm{L}}{\partial\theta^3}$ exists such that

$$\left[\frac{\partial^3}{\partial\theta^3}.\log \mathrm{L}\right] < \mathrm{M}(\mathrm{x})$$

where E[M(x)] <K, a positive quantity.

(iii) For every θ in R,

$$\mathrm{E}\left[-\frac{\partial^2}{\partial\theta^2}\log \mathrm{L}\right] = \int_{-\infty}^{\infty}\left(-\frac{\partial^2}{\partial\theta^2}\log \mathrm{L}\right)\mathrm{L}\,dx = \mathrm{I}(\theta)$$

is finite and non-zero.

(iv) The range of integration is independent of θ. But if the range of integration

depends on θ,then $f(x, \theta)$ vanishes at the extremes depending on θ.This assumption is to make the differentiation under the integral sign valid.

Under the above assumptions M.L.E. possesses a number of important properties,

(1) "With probability approaching unity as $n \to \infty$, the likelihood equation $\frac{\partial}{\partial\theta}\log \mathrm{L} = 0$, has a solution which converges in probability to the true value θ_0"

In other words ML.E. 's are consistent. The m.1.e. is consistent, most efficient, and also sufficient, provided a sufficient estimator sexists.

(2) Any consistent solution of the likelihood equation provides a maximum of the likelihood with probability tending to unity as the sample size (n) tends to infinity.

(3) (Asymptotic Normality of MLE's). A consistent solution of the likelihood equation is asymptotically normally distributed about the true value θ_0.

Thus, $\hat{\theta}$ is asymptotically $\mathrm{N}\left(\theta_0, \frac{\mathrm{I}}{\mathrm{I}(\theta_0)}\right)$, as $n \to \infty$.

4 The m.l.e. is invariant under functional transformations. This means that if t is an m.l.e. of θ, and g(θ) is a function of θ, then g(t) is the m.l.e. of g(θ).

5. If M.L.E. exists, it is the most efficient in the class of such estimators.

Remark: M.L.E's are always consistent estimators but need not be unbiased. For example in sampling from N (μ,) population,

6.4 EXERCISES BASED ON METHOD OF MAXIMUM LIKELIHOOD

Exercise: On the basis of a random sample find the maximum likelihood estimator of the parameter λ of a Poisson distribution.

Solution:- The Poisson distribution with parameter m has p.m.f as given below

$$f(m,x) = \frac{e^{-m} m^{x}}{x!} \qquad (x = 0,1,2 \ldots\ldots \infty)$$

The likelihood function of the sample observations is

$$L = f(x_1.m), f(x_2.m), f(x_3.m) \ldots\ldots\ldots \; \ldots.., \; f(x_n.m)$$

And

$$\log L = \log f(x_1.m) + \log f(x_2.m) + \log f(x_3.m) \ldots\ldots\ldots \; .,+ \log f(x_n.m)$$

$$\sum_{i=1}^{n} \log f(x_i, m) = \sum[-m + x_i(\log m) - \log x_i!] = -nm + \log(m)\Sigma x_i - \Sigma \log(x_i!)$$

Taking partial derivative of log L with respect to the parameter m,

$$\left[\frac{\partial}{\partial m} \log L\right] = -n + \frac{\Sigma x_i}{m} = -n + \frac{n\overline{x}}{m}$$

Now replacing m by m_0 and equating the result to zero,

$$\left[\frac{\partial}{\partial m} \log L\right]_{m=m_0} = -n + \frac{n\overline{x}}{m_0}$$

Solving we get $\mathrm{m}_0 = \overline{\mathrm{x}}$, again

$$\left[\frac{\partial^2}{\partial \mathrm{m}^2}\log \mathrm{L}\right]_{\mathrm{m}=\mathrm{m}_0} = -\frac{\mathrm{n}\overline{\mathrm{x}}}{\mathrm{m}_0^2} = -\frac{\mathrm{n}\overline{\mathrm{x}}}{\overline{\mathrm{x}}^2} = -\frac{\mathrm{n}}{\overline{\mathrm{x}}} \qquad \text{which is negative}$$

This shows that log L is maximum at m = m$_0$ = $\overline{x}$ .That is the m.l.e.of m is m$_0$= $\overline{x}$, the sample mean:

Exercise: Find the maximum likelihood estimator of the variance $\sigma^2$of a Normal population N(u, σ^2), when the parameter μ is known. Show that this estimator is unbiased.

Solution:- The p.d.f of Normal distribution is

$$\mathrm{f}(\mathrm{x},\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}[\mathrm{x}-\mu]^2\right) \qquad ;(-\infty < \mathrm{x} < \infty)$$

And its likelihood function is

$$\mathrm{L} = \prod_{\mathrm{i}-1}^{\mathrm{n}} \mathrm{f}(\mathrm{x}_\mathrm{i},\mu,\sigma^2) = \left(\frac{1}{\sigma\sqrt{2\Pi}}\right)^{\mathrm{n}} \exp\left(-\frac{1}{2\sigma^2}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i}-\mu)^2\right)$$

The logarithm of likelihood function L is

$$\log L = \sum_{i=1}^{n}\log(x_i,\mu,\sigma^2) = \Sigma\left[-\log\sigma - \frac{1}{2}\log(2\pi) - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) - \frac{\Sigma(x_i-\mu)^2}{2\sigma^2}$$

Differentiating partially with respect to σ^2

$$\frac{\partial \log \mathrm{L}}{\partial(\sigma^2)} = -\frac{\mathrm{n}}{2\sigma^2} + \frac{\Sigma(\mathrm{x}_\mathrm{i}-\mu)^2}{2(\sigma^2)^2}$$

The m.l.e. σ^2 of is obtained by solving

$$-\frac{n}{2\sigma^2}+\frac{\Sigma(x_i-\mu)^2}{2\sigma_0^4}=0$$

$\therefore$ $$\sigma_0^2=\frac{\Sigma(x_i-\mu)^2}{n}$$

It can be shown that $$\left[\frac{\partial^2}{\partial\sigma^2}\log L\right]_{\sigma^2=\sigma^2{}_0}=-\frac{n}{2\sigma_0^4}$$

which is negative. Thus the maximum likelihood estimator of σ^2 is

$$\sigma_0^2=\frac{\Sigma(x_i-\mu)^2}{n}, \qquad \text{(known)}$$

Again, since x_1, x_2, ...x_n is a random sample and μ is the population mean we have

$E(x_i-\mu)^2=\sigma^2$ therefore,

$$E(\sigma_0^2)=\frac{\Sigma E(x_i-\mu)^2}{n}=\frac{\Sigma\sigma^2}{n}=\sigma^2$$

Thus $\sigma^2{}_0$ is unbiased estimator of σ^2

Exercise: Find the m.l.e. of the parameters μ and σ^2 in random samples from a N(μ, σ^2) population, when both the parameters are unknown.

Solution $$f(x,\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}[x-\mu]^2\right) \qquad ;(-\infty<x<\infty)$$

$$\log L=-\frac{n}{2}\log\sigma^2-n\log\sqrt{(2\pi)}-\frac{\Sigma(x_i-\mu)^2}{2\sigma^2}$$

$$\therefore \qquad \left[\frac{\partial \log L}{\partial \mu}\right]_{\mu=\mu_0} = -\frac{1}{2\sigma^2}\Sigma 2(x_i-\mu_0)(-1)=0$$

This gives $\Sigma(x_i-\mu_0)=0;$ i.e., $\mu_0=\overline{x}$ the sample mean. He m.l.e of the parameter μ is the

sample mean $\overline{x}$ (This estimator is unbiased)

Proceeding as in the above example

$\sigma_0^2 = \dfrac{\Sigma(x_i-\mu)^2}{n},$ Since parameter μ is unknown it is replaced by its m.l.e and we use $\mu=\overline{x}$ to get

$\sigma_0^2 = \dfrac{\Sigma(x_i-\overline{x})^2}{n}, = S^2$ Which is the sample variance

Exercise: If n_1 trials conducted are of Bernoullian type following binomial distribution, find the maximum likelihood estimate of p.

Solution. We know that probability function of binomial distribution is

$$f(n_1,x_i)=\binom{n_1}{x_i}p^{x_i}(1-p)^{n_1-x_i} \qquad \text{for i=1,2.........n}$$

The likelihood function,

$$L(x/p)=\prod_{i=1}^{n}\binom{n_1}{x_i}p^{x_i}(1-p)^{n_1-x_i}$$

Taking logarithm of both sides,

$$\log L=\sum_{i=1}^{n}\log\binom{n_1}{x_i}+\sum_{i=1}^{n}x_i\log p+\sum_{i=1}^{n}(n_1-x_i)\log(1-p)$$

Differentiating partially w.r.t p and equating to zero.

$$\left[\frac{\partial \log L}{\partial p}\right] = 0 + \sum_{i=1}^{n} x_i \frac{l}{\hat{p}} + \frac{nn_1 - \sum_{i=1}^{n} x_i}{(1-\hat{p})} = 0 \qquad \text{i.e., } nn_1 = \Sigma x_i$$

or $p = \frac{\Sigma x_i}{nn_1} = \frac{\overline{x}}{n_1}$

It is the trivial to show that $\frac{\overline{x}}{n_1}$ is the maximum likelihood estimate of p.

6.5 Method of Moments

This is the oldest estimation method in statistics. The underlying principle in this method is that the sample moments reflect the population characteristics in the sense that the expected values of the sample moments are equal to the population moments.

It was first put forward by Karl Pearson in 1894. The method of moments consists of equating the sample moments to the corresponding moments of distribution, which are the functions of the unknown parameters. Here, we equate as many sample moments as there are unknown parameters. Solving these equations simultaneously we get the estimates of the moments of the population in terms of sample variates.

Here we equate the moments of the population with the corresponding moments of the sample, i.e. setting

$$\mu_r' = m_r'$$

Where μ_r' = E(x^r) and $m_r' = \frac{1}{n}\sum_{i=1}^{n} x_i^r$ Also μ_1'=E(x)=μ

These relations when solved for the parameters give the estimates by the method of moments. This method is applicable only when the population moments exist. The method is generally applied for fitting theoretical distributions to observed data.

Properties of the estimates obtained by the method of moments.

(i) Under fairly general conditions, the estimates obtained by the method of moments will have asymptotically normal distribution for large n.

(ii) The mean of the distribution of estimate will differ from the true value of the parameter by

a quantity of order 1/n

(iii) The variance of the distribution of estimate will be of the type c^2/n.

(iv) In general, the deviation estimators obtained by the method of moments are less efficient than the maximum likelihood estimators. In particular cases, they are equivalent.

6.6 EXERCISES BASED ON METHOD OF MOMENTS

Exercise: Estimate the parameter np of the binomial distribution by the method of moments (when n is known**).**

Solution:-If X~B(n,p) then its p.m.f is given by

$$f(n,x)=\binom{n}{x}p^x(1-p)^{n-x} \qquad \textbf{x =0,1,2.......n}$$

And $\mu_1^{'}=E(x)=np$ Also $m_1^{'}=\overline{x}$

Setting $\mu_1^{'}=m_1^{'}$ we have $np=\overline{x}$ Thus $p=\dfrac{\overline{x}}{n}$

i.e. the estimated value of p is given by the sample mean divided by the parameter n (known).

Exercise: Find the estimates of μ and σ in the Normal population N(μ , σ^2) by the method of moments.

Ans. Let X_1, X_2, ..., X n be a random sample from a normal population N(μ, σ^2)

We know that $\mu_1^{'} = m_1^{'} = \mu = \overline{x}$

and $\mu_2 = \mu_2^{'} - (\mu_1^{'})^2 = \frac{1}{n}\Sigma X_i{}^2 - \overline{X}^2 = \frac{1}{n}\left(\Sigma X_i{}^2 - n\overline{X}^2\right) = \frac{1}{n}\sum\left(X_i - \overline{X}^2\right)^2$

for i =0,1,2…….n

Therefore, $\overline{x}$ is an unbiased estimator of μ whereas sample variance $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is not an unbiased estimator for σ^2.

Exercise: Find the estimate of the parameter λ the Poisson distribution $\frac{e^{-\lambda}\lambda^x}{x!}$ by the method of moments.

Solution: Let X_1, X_2, ..., X n be a random sample from a normal population a Poisson distribution P (x; λ). We know in case of Poisson distribution, its mean and variance are equal. The mean,

$$\mu_1^{'} = m_1^{'} = E[X] = \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = \lambda\sum_{x=1}^{\infty} e^{-\lambda}\frac{\lambda^{x-1}}{(x-1)!} \qquad (x = 0,1,2.......\infty)$$

$$= \lambda = \overline{x}$$

Thus, the estimate of the parameter λ by the method of moments is the sample mean $\overline{x}$.

6.7 SELF ASSESSMENT QUESTIONS

1. Why do the decision makers often measure the sample rather than the entire population. What is the disadvantage?

2. Explain the shortcoming that occurs n the point estimation but not in an interval estimation. What measure is included with an interval estimation that compensate for this ?

3. What is an estimator ? How does an estimate differ from an estimator?

4. List and briefly describe the criteria of a good estimator.

5. Describe the M.L method of estimation and discuss five of its optimal properties.

6. Describe the method of moments for estimation .What are the properties of the estimator obtained by the method of moments ?

7. What two basic tools are used in making statistical inferences?

8. Why do decision makers often measure samples rather than entire populations? What is the disadvantage?

9. Explain a shortcoming that occurs in a point estimate but not in an interval estimate. What measure is included with an interval estimate to compensate for this?

10. What is an estimator? How does an estimate differ from an estimator?

11. List and describe briefly the criteria of a good estimator.

12. What role does consistency play in determining sample size?

13. State and explain the principle of maximum likelihood for estimation of population parameter.

14. Describe the M.L. method of estimation and discuss five of its optimal properties.

15. Compute the likelihood function for a random sample of size n for the each of the following populations.

(i) Normal (θ, σ^2) (ii) Binomial (n,p)

(iii) Poisson (t) (iv) Uniform (a,b)

16. Describe the method of moments for estimating the parameters. What are the properties of the estimates obtained by this method?

17. Let X_1, X_2 ……….Xn be a random sample from the p.d.f.

$$f(x,\theta) = \theta e^{-\theta x} \qquad 0 < x < \infty,\ \theta > 0$$
$$= 0, \quad elsewhere$$

Estimate θ using the method of moments.

18. Explain the methods of estimation-method of moments and maximum likelihood. Do these lead to the same estimates in respect of the standard deviation of a normal population? Examine the properties of the estimates from the point of view of consistency and Unbiasedness.

19. For the distribution with probability function:

$$f(x,\theta) = \frac{e^{-\theta}\theta^x}{x!(1-e^{-\theta})}; \qquad x = 1,2,3.......$$

Obtain the estimate of θ by the method of moments.

Unit 2 **Lesson NO.7**

CONFIDENCE INTERVAL AND CONFIDENCE INTERVAL

Structure:

7.1 Introduction

7.2 Objectives

7.3 Definition

7.4 Meaning and interpretation

7.5 Desirable properties

7.6 Method of derivation

7.7 Exercises topic

7.8 Summary

7.9 Self assessment Questions

7.1 INTRODUCTION

Interval estimation can be contrasted with point estimate. . A point estimate is a single number that is used to estimate an unknown population parameter. **An interval estimate** is a range of values used to estimate a population parameter. Confidence interval are commonly reported in tables or graphs along with point estimates of the same parameter, to show the reliability of the estimates.

7.2 OBJECTIVES

In statistics, a confidence interval (C.I)is a particular type of interval estimation of a population parameter and is used to indicate the reliability of an estimate. It is an observed interval (i.e. it is calculated from the observations), in principle different from sample to sample, that frequently includes the parameter of interest, if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the confidence level or confidence coefficient.

A confidence interval with a particular confidence level is intended to give the assurance that, if the statistical model is correct, then taken over all the data that might have been obtained, the procedure for constructing the interval would deliver a confidence interval that included the true value of the parameter the proportion of the time set by the confidence level. More specifically, the meaning of the term "confidence level" is that, if confidence intervals are constructed across many separate data analyses of repeated (and possibly different) experiments, the Proportion of such intervals that contain the true value of the parameter will approximately match the confidence level; this is guaranteed by the reasoning underlying the construction of confidence intervals.

A confidence interval does not predict that the true value of the parameter has a particular probability of being in the confidence interval given the data actually obtained. An interval to have such property is called a credible interval, can be estimated by using Bayesian method; but such methods bring with them their own distinct strengths and Weaknesses.

7.3 DEFINITION

Let X be a random sample from, probability distribution with parameter θ , which is the quantity to be estimated , and let φ represents the quantity not of immediate interest. A confidence interval for the parameter θ, with confidence level γ , is an interval with random endpoints (u(x), v(x)) determined by the pair of statistics u(x) and v(x), with the property

$$\gamma = P_{\theta,\varphi}(u(x) < \theta < v(x))$$

The quantities φ in which there is no immediate interest are called nuisance parameter, as Statistical theory Still needs to find some way to deal with them.The number γ, With typical values close to but not greater than 1 is Sometimes given in the form $1-\alpha$ (Or as a Percentage 100% ($1-\alpha$), where α , a small nonnegative number close to 0.

Here $P_r \theta, \varphi$ is used to indicate the probability when the random variable X has the distribution characterized by (θ, φ). An important part of this specification is that the random interval (U,V) covers the unknown value θ with a high probability no matter what the true value of θ actually is.

Note that here $P_r \theta, \varphi$ p need not refer to an explicitly given parameterized family of distributions, although it often does. Just as the random variable X notionally corresponds to other possible realizations of x from the same population or from the same version of reality, the parameters (θ, φ) indicate that we need to consider other versions of reality in which the distribution of X might have different characteristics.

In a specific situation, when x is the outcome of the sample X, the interval (u(x), v(x)) is also referred to as a confidence interval for θ. Note that it is no longer possible to say that the (observed) interval (u(x), v(X)) has probability γ to contain the parameterθ. This observed interval is just one realization of all possible intervals for which the probability statement holds.

CONFIDENCE INTERVAL AND CONFIDENCE LIMITS: Let us consider a random sample x_i, (i = 1, 2, ..., n) of n observations from a population involving a single unknown parameterθ, (say). With probability function

f(x,θ) and let us suppose that this distribution is continuous. Let

$t = t(x_1, x_2, ..., x_n)$

be a function of the sample values be an estimate of the population parameter θ, with the sampling distribution given by

g(t,θ).

After obtaining the value of the statistic t from a given sample, the problem is, "*Can we make some reasonable probability statements about the unknown parameter θ in the population, from which the sample has been drawn*?" This question is very well answered by the technique of Confidence interval due to Neyman

We choose once for all some small value of α (5% or 1%) and then determine two constants say, c_1 and c_2 such that

$$P(c_1 < \theta < c_2 I\ t) = 1 - \alpha$$

The quantities c_1 and c_2, so determined, are known as the confidence limits or fiducial limits and the interval $[c_1, c_2]$ within which the unknown value of the population parameter is expected to lie, is called the confidence interval and $(1-\alpha)$ is called the confidence coefficient.

E.g., if we take a = 0.05 we shall get 95% confidence limits.

7.4 MEANING AND INTERPRETATION

The confidence interval can be expressed in terms of samples (or repeated samples): "Were this procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter 90% of the time." Note that this need not be repeated sampling from the same population, just repeated sampling.

The explanation of a confidence interval can amount to something like: The confidence interval represents values for the population parameter for which "the difference between the parameter and the observed estimate is not statistically significant at the 10% level". In fact, this relates to one particular way in which a confidence interval may be constructed.

The probability associated with a confidence interval may also be considered from a pre-experiment point of view, in the same context in which arguments for the random allocation of treatments to study items are made. Here the experimenter sets out the way in which they intend to calculate a confidence interval and know, before they do the actual experiment, that the interval they will end up calculating has a certain chance of covering the true but unknown value. This is Very Similar to the "repeated sample" interpretation above, except that it avoids relying on considering hypothetical repeats of a sampling procedure that may not be repeatable in any meaningful sense.

In each of the above, the following applies: If the true value of the parameter lies outside the 90% confidence interval once it has been calculated, then an event has occurred which had a probability of 10% (or less) of happening by chance.

The probability associated with a confidence interval may also be coinsidered from a pre-experiment point of view, in the same context in which arguments for the random allocation of treatments to study items are made. Here the experimenter sets out the way in which they intend to calculate a confidence interval and know, before they do the actual experiment, that the interval they will end up calculating has a certain chance of covering the true but unknown value. This is Very Similar to the "repeated sample" interpretation above, except that it avoids relying on considering hypothetical repeats of a sampling procedure that may not be repeatable in any meaningful sense.

In each of the above, the following applies: If the true value of the parameter lies outside the 90% confidence interval once it has been calculated, then an event has occurred which had a probability of 10% (or less) of happening by chance.

7.4.1 MEANING OF THE TERM "CONFIDENCE"

There is a difference in meaning between the common usage of the word "confidence" and its statistical usage, which is often confusing to the layman, and this is one of the critiques of confidence intervals, namely that in application by non-statisticians, the term "confidence" is misleading.

In common usage, a claim to 95% confidence in something is normally taken as indicating virtual certainty. In statistics, a claim to 95% confidence simply means that the researcher has seen something occur that happens only one time in 20 or less. If one were to roll two dice and get double six (which happens 1/36th of time , about 3%) a few would claim this as proof that the dice were fixed, although statistically although statistically speaking one could have 97% confidence that they were. Similarly, the finding of a statistical link at 95% confidence is not proof, nor even very good evidence, that there is any real connection between the things linked.

When a study involves multiple statistical tests, people tend to assume that the confidence associated with individual tests is the confidence one should have in the results of the study itself. In fact, the results of all the statistical tests conducted during a study must be judged as a whole in determining what confidence one may place in the positive links it produces. For example, say a study is conducted which involves 40 statistical tests at 95% confidence, and which produces 3 positive results. Each test has a 5% chance of producing a false positive, so such a study will produce 3 false positives about two times in three. Thus the confidence one can have that any of the study's positive conclusions are correct is only about 32%, well below the 95% the researchers have set as their standard of acceptance

7.5 DESIRABLE PROPERTIES

When applying standard statistical procedures, there will often be standard ways of constructing confidence intervals. These will have been devised so as to meet certain desirable properties, which will hold given that the assumptions on which the procedures rely are true. These desirable properties may be described as: validity, optimality and invariance. Of these "validity" is most important, followed closely by "optimality". "Invariance" may be considered as a property of the method of derivation of a confidence interval rather than of the rule for constructing the interval. In non-standard applications, the same desirable properties would be sought.

Validity: This means that the nominal coverage probability (confidence level) of the confidence interval should hold, either exactly or to a good approximation.

Optimality: This means that the rule for constructing the confidence interval should make as much use of the information in the data-set as possible. Recall that one could throw away half of a data set and still be able to derive a valid confidence interval. One way of assessing optimality is by the length of the interval, so that a rule for constructing a confidence interval is judged better than another if it leads to intervals whose lengths are typically shorter.

Invariance: In many applications the quantity being estimated might not be tightly defined as such. For example, a Survey might result in an estimate of the median income in a population, but it might equally be considered as providing an estimate of the logarithm of the median income, given that this is a common scale for presenting graphical results. It would be desirable that the method used for constructing a confidence interval for the median income would give equivalent result when applies to constructing a confidence interval of logarithm of the median income; specifically values at the ends of the latter interval would be the logarithms of the values at the ends of the former interval.

7.6 METHODS OF DERIVATION

For non-standard applications, there are several routes that might be taken to derive a rule for the construction of confidence intervals. Established rules for Standard procedures might be justified or explained via several of these routes. For Typically a rule for constructing confidence intervals is closely tied to a particular that Way of finding a point estimate of the quantity being considered.

7.6.1 STATISTICS

This is closely related to the method of moments for estimation. A simple example arises where the quantity to be estimated is the mean, in which case a natural estimate is the sample mean. The usual arguments indicate that the sample a variance can be used to estimate the variance of the sample mean. A new confidence interval for the true mean can be constructed centered on the sample mean with a width which is a multiple of the square root of the sample variance.

7.6.2 LIKELIHOOD THEORY

Where estimates are constructed using the maximum likelihood principle, the theory for this provides two ways of constructing confidence intervals or confidence regions for the estimates.

7.6.3 ESTIMATING EQUATIONS

The estimation approach here can be considered as both a generalization of the method of moments and a generalization of the maximum likelihood approach. There are corresponding generalizations of the results of maximum likelihood theory that allow confidence intervals to be constructed based on estimates derived from estimation equation.

Via significance testing

If significance tests are available for general values of a parameter, then confidence intervals/regions can be constructed by including in the 100p% confidence region all those points for which the significance test of the null hypothesis that the true value is the given value is not rejected at a significance level of

7.7.1 STATISTICAL HYPOTHESIS TESTING

Confidence intervals are closely related to statistical significance testing. For example, if for some estimated parameter θ one wants to test the null hypothesis that $\theta = 0$ against the alternative that $\theta \neq 0$, then this test can be performed by determining whether the confidence interval for θ contains 0.

More generally, given the availability of a hypothesis testing procedure that can test the null hypotheses $\theta = \theta_0$ against the alternative that $\theta \neq \theta_0$ for any value of θ. Then a confidence interval confidence level $1-\gamma$ with can be defined as containing any number θ_0 for which the corresponding null hypothesis is not rejected at significance level α '

In consequence, if the estimates of two parameters (for example, the mean values of a variable in two independent groups of objects) have confidence intervals at a given value γ that does not overlap, then the difference between the two values is significant at the corresponding value of α . However, this test is too conservative. If two confidence intervals overlap, the difference between the two means still may be significantly different.

7.7.2 CONFIDENCE REGION

Confidence region generalize the confidence interval concept to deal with multiple quantities. Such regions can indicate not only the extent of likely sampling error but can also reveal whether (for example) it is the case that if the estimate for one quantity is unreliable then the other is also likely to be unreliable.

In applied practice, confidence intervals are typically stated at 95% confidence level. However, when presented graphically, confidence intervals can be shown at several levels, for example 50%, 95% and 99%.

7.7.3 INTERVALS FOR RANDOM OUTCOMES

Confidence intervals can be defined for random quantities as well as for fixed quantities as in the above. For this, consider an additional single-valued random variable Y which may or may not be statistically dependent on X. Then the rule for constructing the interval (u(x), v(x)) provides a confidence interval for the as-yet-to-be observed value y of Y if

$$\gamma = P_{\theta,\varphi}(u(x) < \theta < v(x))$$

Here $P_{\theta,\varphi}$ is used to indicate the probability over the joint distribution of the random variables (X, Y) when this is characterised by parameters (θ,φ).

Approximate confidence intervals

For non-standard applications it is sometimes not possible to fine rules for constructing confidence intervals that have exactly the required properties. But practically useful intervals çạn still be found. The probability $c(\theta,\varphi)$ for a random interval is defined by

$$\Pr_{\theta,\varphi}(u(x) < \theta < v(x)) = c(\theta,\phi)$$

And rule for constructing the interval may be accepted as providing a confidence interval if

$$c(\theta,\phi) \cong 1-\alpha \qquad for\ all\,(\theta,\phi)$$

to an acceptable level of approximation.

EXAMPLE:

Find 100(1-α)% confidence for (i) θ (ii) σ^2 in normal population with p.d.f

$$f(x,\theta,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad -\infty < x < \infty$$

Sol: Let us consider a random sample x_i, (i = 1, 2, ..., n) of n observations from density function $f(x,\theta,\sigma^2)$ and suppose $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ then statistic

$$t = \frac{\overline{x}-\theta}{s/\sqrt{n}} \sim t_{(n-1)}$$

Hence 100(1-α)% confidence for θ are given by

$$P\left[|t| \le t_\alpha\right] = 1-\alpha \quad \text{or} \quad P\left[\left|\frac{\overline{x}-\theta}{s/\sqrt{n}}\right| \le t_\alpha\right] = 1-\alpha$$

or $P\left[\ |\overline{x}-\theta| \le s/\sqrt{n}\, t_\alpha\ \right] = 1-\alpha$

$$\therefore \qquad P\left[\ \overline{x} - t_\alpha\, s/\sqrt{n} \le \theta \le \overline{x} + t_\alpha\, s/\sqrt{n}\right] = 1-\alpha$$

where t_α is the tabulated value of t for (n-1) degrees of freedom at α level of significance. Hence required level of significance is

$$\left[\ \overline{x} - t_\alpha\, s/\sqrt{n}\ ,\ \ \overline{x} + t_\alpha\, s/\sqrt{n}\right]$$

(ii) let θ is unknown = μ (say) then

$$\frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}=\frac{ns^2}{\sigma^2}\sim\chi_n{}^2$$

If we define chi-square as such

$$P\left[\chi^2>\chi_\alpha^2\right]=\int_{\chi_\alpha^2}^{\alpha}p(\chi^2)d\chi^2=\alpha$$

Hence reqd. confidence interval is given by

$$P\left[\chi^2_{(1-\alpha/2)}\le\chi^2\le\chi^2_{\alpha/2}\right]=1-\alpha$$

$$\Rightarrow P\left[\chi^2_{(1-\alpha/2)}\le\frac{ns^2}{\sigma^2}\le\chi^2_{\alpha/2}\right]=1-\alpha \qquad \text{...............(1)}$$

Now $\frac{ns^2}{\sigma^2}\le\chi^2_{\alpha/2}\Rightarrow\frac{ns^2}{\chi^2_{\alpha/2}}\le\sigma^2$ and $\chi^2_{(1-\alpha/2)}\le\frac{ns^2}{\sigma^2}\Rightarrow\sigma^2\le\frac{ns^2}{\chi^2_{(1-\alpha/2)}}$

Hence from (1)

$$\Rightarrow P\left[\frac{ns^2}{\chi^2_{\alpha/2}}\le\sigma^2\le\frac{ns^2}{\chi^2_{(1-\alpha/2)}}\right]=1-\alpha$$

Which is the required confidence interval

7.8 Summary

Confidence intervals is to define a $100(1-\alpha)\%$ confidence interval of all these values of θ for which a test of the hypothesis $\theta=\theta_0$ is not rejected at a significance level of $100\alpha\%$. Such an approach may not always be available Since it presupposes the practical availability of an appropriate significance test. Naturally, any assumptions required for the significance test would carry over to the confidence intervals.

It may be convenient to make the general correspondence that parameter values within a confidence interval are equivalent to those values that would not be rejected by an hypothesis test, but this would be dangerous. In many instances the confidence intervals that are quoted are only approximately valid, perhaps derived from "plus or minus twice the standard error", and the implications of this for the supposedly corresponding hypothesis tests are usually unknown.

7.9 SELF ASSESSMENT QUESTIONS

1. What do you understand by confidence and explain its desirable properties.

2. Obtain $100(1-\alpha)\%$ confidence interval for parameter u in the random sample from normal population:

   $df(x) = ue^{-ux} \quad x > 0, \; u > 0$

3. Obtain $100(1-\alpha)\%$ confidence interval for unknown parameter p of a binomial distribution when the parameter n is known in the random sample from normal population

Unit 3 **Lesson 8**

TESTING OF HYPOTHESES

Structure:

8.1 Introduction

8.2 Objectives

8.3 Concepts Basic to the Hypothesis-Testing Procedure

8.4 Test of Significance

8.5 Critical region

8.6 One-and two-tailed Tests

8.7 Size (Level of significance) and Power of a Test

8.8 Degrees of freedom

8.9 P-Values

8.10 Self Assessment Questions

8.1 INTRODUCTION

Hypotheses testing begin with an assumption; called hypotheses that we make about a population parameter. Then we collect sample data, produce sample statistics, and use this information to decide how likely it is that our hypothesized population parameter is correct. Say that we assume a certain value for a population mean. To test the validity of our assumption, we gather sample data and determine the difference between the hypothesized value and the actual value of the sample mean. Then we judge whether the difference is significant. The smaller the difference, the greater the likelihood that our hypothesized value for the mean is correct. The larger the difference, the smaller the likelihood.

Unfortunately, the difference between the hypothesized population parameter and the actual statistic is more often neither so large that we automatically reject our hypothesis nor so small that we just as quickly accept it. So in hypothesis testing, as in most significant real-life decisions, clear-cut solutions are the exception, not the rule.

8.2 OBJECTIVES

Objectives of this lesson is to enable the learners

1. To learn how to use samples to decide whether a population possesses a particular characteristic
2. To understand the basis of testing procedure
4. To learn when to use one- tailed tests and when to use two-tailed tests
5. To hypothesis and its types hypotheses
6. To understand the concept of critical region and P-values

8.3 CONCEPTS BASIC TO THE HYPOTHESIS-TESTING PROCEDURE

Hypothesis testing begins by making an assumption about the population parameter. Then we gather sample data and determine the sample statistic. To test the validity of our hypothesis the difference between the hypothesized value and the actual value of the sample statistic will be determined. If the difference between the hypothesized population parameter and the actual value is large then we automatically reject our hypothesis. If it is small, we accept it.

The theory of testing of Hypothesis was initiated by J. Neyman and E.S. Pearson. In Neyman Pearson Theory we use statistical methods to arrive at a decision in certain situations where there is lack of certainty on the basis of the sample where size is fixed in advance while in Wald sequential theory the sample size is not fixed in advance but regarded as a random variable.

TYPES OF HYPOTHESIS

In attempting to arrive at decision about the population on the basis of sample information, it is necessary to make assumptions or guesses about the population parameters involved. Such an assumption is called a statistical hypothesis which may or may not be true. The procedure which enables us to decide on the basis of a sample, whether a hypothesis is true or not, is called Test of Hypothesis or Test of Significance. There are two hypotheses:

- Null Hypothesis
- Alternative Hypothesis.

NULL HYPOTHESIS

In tests of hypothesis, we always begin with an assumption, the null hypothesis. The null hypothesis asserts that there is no (significant) difference between the statistic and the population parameter and whatever observed difference is there, it is merely due to chance (fluctuations in sampling from the same population). Null hypothesis is usually denoted by the symbol H_0.

A hypothesis which is to be actually tested for acceptance or rejection is termed as null hypothesis. As the name suggests it is always taken as hypothesis of no difference. The decision maker should adopt a null or neutral attitude regarding the outcome of the test. It is denoted by H_0.In the words of prof. R.A Fisher

"Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true"

In hypothesis testing, a statistician or decision-maker should not be motivated by prospects of profit or loss resulting from the acceptance or rejection of the hypothesis.

Much, therefore, depends upon how the hypothesis is framed. Hence the best course is to adopt the thesis of no difference. If we want to test the significance of difference between a statistic and a parameter or between two sample statistics, and then we set up null hypothesis H_0 **that the difference is not significant.** This means that the difference is just due to the fluctuations of sampling. E.g $H_0 : \mu = \mu_0$ where μ_0 is some specified value of μ

ALTERNATIVE HYPOTHESIS

Any hypothesis which contradicts the null hypothesis H_0 is called an Alternative Hypothesis and is denoted by the symbol H_1.We can say that it is a statement about the population parameter or parameters, which gives an alternative to the null hypothesis (H_0), within the range of pertinent values of the parameter, i.e., if H_0 is accepted, what hypothesis is to be rejected and vice versa.

It is desirable to state what is called an alternative hypothesis in respect of every statistical hypothesis being tested because the acceptance or rejection of null hypothesis is meaningful only when it is being tested against a rival hypothesis which should rather be explicitly mentioned. If null hypothesis is $H_0 : \mu = \mu_0$ where μ_0 is some specified value of μ then alternative could be

(1) $H_1 : \mu \neq \mu_0$ i.e,. $\mu < \mu_0$ or $\mu > \mu_0$ (Two tailed alternative)

(2) $H_1 : \mu < \mu_0$ left tail (one tailed alternative)

(3) $H_1 : \mu > \mu_0$ right tail (one tailed alternative)

The Alternative Hypothesis in (1) is known as a two-tailed alternative and in (2) and (3) is known as left-tailed and right-tailed alternatives respectively. The corresponding tests of hypotheses are called two-tailed (or two-sided), right-tailed (one-sided) and left-tailed (one-sided) tests respectively.

STATISTICAL HYPOTHESIS (Simple and Composite)

A statistical hypothesis is a statement, an idea or an assertion about a population or equivalently about probability distribution characterizing a population which we want to verify on the basis of information available from the sample.

'A hypothesis is an assertion or conjecture about the parameter(s) of population distribution(s)"

If statistical hypothesis specifies the population completely then it is termed as simple. If statistical hypothesis does not specifies the population completely then it is termed as Composite.

Example:-1 let for a random sample $x_1, x_2, ...x_n$ from a normal population with mean μ and variance σ^2 the hypothesis

$H_0 : \mu = \mu_0$, $\sigma^2 = \sigma_0^2$ is simple hypothesis as in $N(\mu, \sigma^2)$ $\Theta = [\mu = \mu_0 \ , 0 > \sigma^2 > \infty]$

Where as (i) $\mu = \mu_0$

(ii) $\sigma^2 = \sigma_0^2$

(iii) $\mu < \mu_0, \quad \sigma^2 = \sigma_0^2$

(iv) $\mu > \mu_0, \quad \sigma^2 = \sigma_0^2$ are composite hypothesis

A hypothesis which doesn't completely specify the 'r' parameters of the population is termed as composite hypothesis with 'r' degrees of freedom. A hypothesis may be simple or composite depending upon the alternative hypothesis.

Example:-2. For instance, we consider a normal population $N(\mu,\sigma^2)$, where σ^2 is known and we want to test the hypothesis, $H_0 : \mu = 25$ against H_1: μ =30. From these hypotheses we know that μ can take either of the two values, 25 or 30, In this case, H_0 and H_1 are both simple. But generally $H_1 : \mu \neq 25$ is composite, i.e. of the form, $H_1 : \mu < 25$ or $H_1 : \mu > 25$. Likewise, simple and composite hypothesis for any other parameter(s) can be stated.

8.4 TEST OF SIGNIFICANCE

A research worker or an experimenter has always some fixed ideas about certain population parameter(s) based on prior experiments, surveys or experience. Sometimes these ideas might have been fixed in the mind. There is a need to ascertain whether these ideas or claims are correct or not by collecting information in the form of data. In this way, we come across two types of problems, first is to draw inferences about the population on the basis of sample data and the other is to decide whether our sample observations have come from a postulated population or not.

By hypothesis we mean to give postulated or stipulated value(s) of a parameter. Also, instead of giving values, some relationship between parameters is postulated in the case of two or more populations. On the basis of observational data, a test is performed to decide whether the postulated hypothesis be accepted or not. This involves certain amount of risk. This amount of risk is termed as a level of significance. When the hypothesis is accepted, we consider it a nonsignificant result and if the reverse situation occurs, **it is called a significant result**.

STATISTICAL TEST

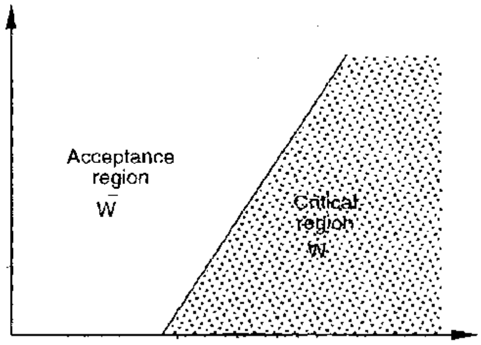
A test is defined as, "A statistical test is a procedure governed by certain rules, which leads to take a decision about the hypothesis, for its acceptance or rejection on the basis of sample values."

USES OF STATISTICAL TESTS

Statistical tests of hypotheses play an important role in industry, biological sciences, behavioral sciences and economics, etc. The use of tests has been made clear through a number of practical problems.

1. A feed manufacturer announces that his feed contains forty per cent protein. Now to make sure whether his claim is correct or not, one has to take a random sample of the product and by chemical analysis, find the protein percentages in the samples. From these observed values, he would decide about the manufacturer's claim for his product. This is done by performing a test of significance.
2. Psychologists are often interested in knowing whether the level of IQ of a group of school boys is up to a certain standard or not. In this case, some boys are selected and an intelligence test is conducted. The scores obtained by them pass through a statistical test and a decision is made whether their IQ is up to the standard or not.

8.5 CRITICAL REGION (C.R.)



A statistic is used to test the hypothesis H_0. The test statistic follows some known distribution. In a test, the area under the probability density curve is divided into two regions, viz., the region of acceptance and the region of rejection. The region of rejection is the region in which H_0 is rejected. It means that if the value of test statistics lies in this region, H_0 (null hypothesis) will be rejected.

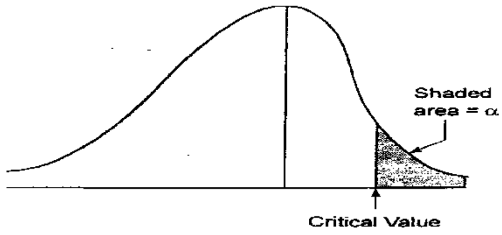
The region of rejection is called a **critical region**. Moreover, the area of the critical region is equal to the level of significance α. The critical region is always on the tail of the distribution curve. It may be on both the tails or on one tail, depending upon the alternative hypothesis.

In short the value of the standard statistic beyond which we reject the null hypothesis; the boundary between the acceptance and rejection regions.

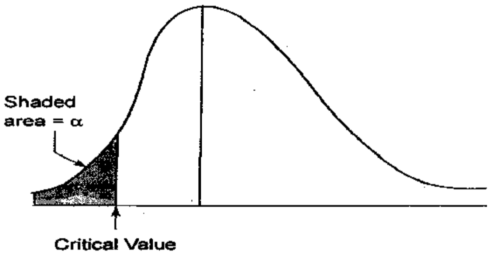
8.6 ONE AND TWO-TAILED TESTS

If the alternative hypothesis, H_1 is of the type $\mu > \mu_0$ or $\mu < \mu_0$ etc., the critical region lies on only one tail of the probability density curve. In this situation the test is called **one-tailed test**.

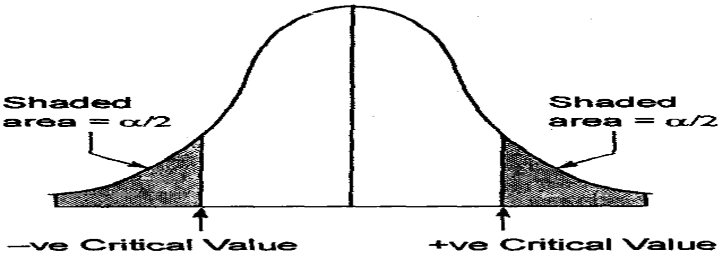
If H_1 is of the type $H_1 : \mu > \mu_0$ the critical region is towards the right tail as shown below



On contrary to this, if the alternative hypothesis, H_1 is of the type $\mu < \mu_0$ the critical region lies on only one tail (left tail) of the probability density curve. In this situation ($H_1 : \mu > \mu_0$) the critical region is towards the left tail



If the test is two-tailed, i.e., it is of the type $H_1 : \mu \neq \mu_0$ then the test is called two-tailed test and in such a case the critical region lies in both the right and left tails of the sampling distribution of the test statistic, with total area equal to the level of significance as shown in diagram.



If the alternative hypothesis is of the type $H_1 : \mu \neq \mu_0$ i.e,. $\mu < \mu_0$ or $\mu > \mu_0$ the critical region lies at the both the tails , in this situation test is called two tailed test and an area equal to $\alpha/2$ lies at the both the tails.

8.7 SIZE (LEVEL OF SIGNIFICANCE) AND POWER OF A TEST

The main purpose of hypothesis testing is not to question the computed value of the sample statistic, but to make judgment about the difference between the sample statistic and a hypothesized population parameter. After stating the Null and Alternative Hypotheses, we have to decide what criterion to be used for deciding whether to accept or reject the null hypothesis.

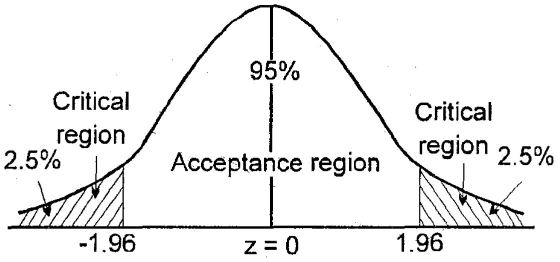
In testing a given hypothesis the minimum probability with which we would be willing to risk a type one error is called as level of significance. The size of a test is the probability of rejecting the null hypothesis when it is true, and is usually denoted by α. The level of, significance and size are synonymous in a practical sense. Therefore.

$$P[\text{rejet}\, H_0 / H_0] = \alpha \qquad ...(1)$$

For example when we choose 5% level of significance in a test procedure, there are about 5 cases in 100 that we would reject the hypothesis when it should be accepted, that is, we are about 95% confident that we have made the right decision. Similarly, if we choose 1% level of significance in testing a hypothesis, then there is only 1 case in 100 that we would reject the hypothesis when it should be accepted.

Suppose, that under a given hypothesis the sampling distribution of a statistic θ is approximately a normal distribution with mean $E(\theta)$ and standard deviation (Standard Error) σ_θ.

Then $z = \dfrac{\text{Observed value - Expected value}}{\text{Standard Error of } \theta}$ is called the standardized normal variable or z-score, and its distribution is the standardized normal distribution with mean 0 and standard deviation 1, the graph of which is shown below.



From the above figure, we see that if the test statistic z of a sample statistic θ lies between —1.96 and 1.96, then we are 95% confident that the hypothesis is true.

But if for a simple random sample we find that the test statistic (or z-score) z lies outside the range —1.96 to 1.96, i.e. if z > 1.96, we would say that such an event could happen with probability of only 0.05 (total shaded area in the above figure if the given hypothesis were true). In this case, we say that z-score differed significantly from the value expected under the hypothesis and hence, the hypothesis is to be rejected at 5% (or 0.05) level of significance. Here the total shaded area 0.05 in the above figure represents the probability of being wrong in rejecting the hypothesis. Thus if z > 1.96, we say that the hypothesis is rejected at a 5% level of significance.

Remark: - The set of z scores outside the range —1.96 and 1.96, constitutes the **critical region or region of rejection** of the hypothesis or the region of significance. Thus critical region is the area under the sampling distribution in which the test statistic value has to fall for the null hypothesis to be rejected.

Thus choosing a certain level of probability with which we would be willing to risk error of type-I, is called level of significance.

POWER OF A TEST :-The power of a test is defined as the probability of rejecting the null hypothesis when it is actually false, i.e., when H_1 is true. It is ability of a statistical test to detect the alternative hypothesis when it is true .

Power= $P[\text{rejet}\, H_0 / H_1] = 1 - P[\text{accept}\, H_0 / H_1]$

$$= 1 - \text{Prob. of type} - \text{II error} = 1 - \beta$$

where β is the probability of type II error.

Among a class of tests, the best test is the one which has the maximum power for the same size i.e., same level of significance α.

8.8 DEGREES OF FREEDOM

In a test of hypothesis, a sample is drawn from the population of which the parameter is under test. The size of the sample varies since it depends either on the experimenter or on the resources available; moreover, the test statistic involves the estimated value of the parameter which depends on the number of observations. Hence, the sample size plays an important role in testing of hypothesis and is taken care of by degrees of freedom.

Summing up we can say that the number of values in a sample we can specify freely, once we know something about that sample is known as degrees of freedom

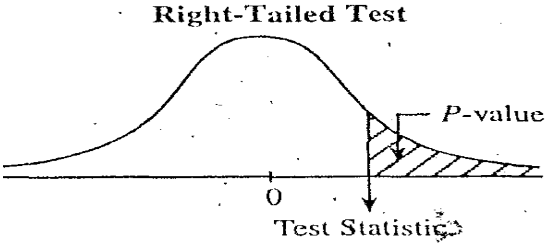
Definition: Degree Of Freedom is the number of independent observations in a set.

8.9 P-VALUES

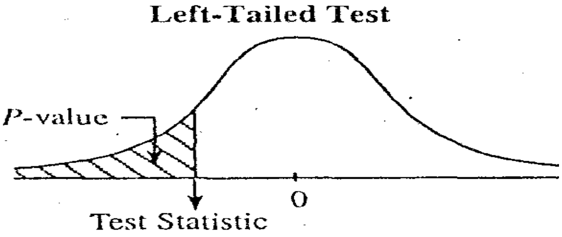
It may be defined as the smallest level of α at which H_0 is rejected. In this situation , it is not inferred whether H_0 is accepted or rejected at level 0.05, 0.01 or any other value, but the statistician only give the smallest level of α at which H_0 is rejected.

This facilitates the individual to decide for himself as to how much significant the data are. This approach avoids the imposition of fixed level of significance.

1. For the right tailed test, the P—value is the area to the right of the computed value of the test statistic under H_0



2. For the left tailed test, the P-value is the area to the left of the computed value of the test statistic under Ho

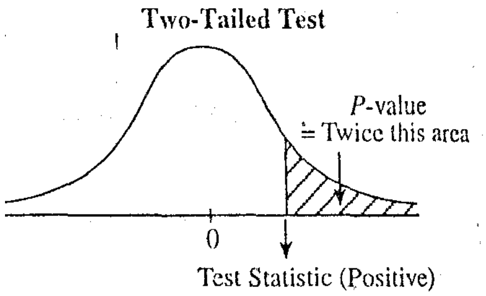


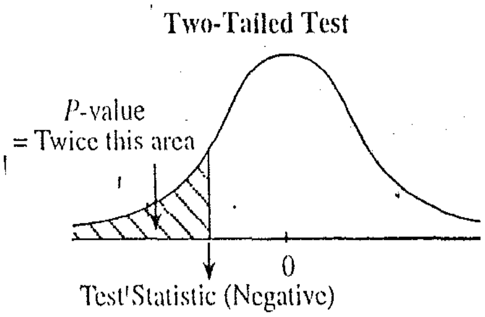
3. For the two-tailed test, P-value is

(a) Twice the area to the left of the computed value of test statistic under H_0, if it is negative or,

(b) Twice the area to the right of the computed value of test statistic under Ho, if it is Positive

The P-value for two-tailed test is twice the area on either tail (left or right) of the computed value of test statistic under Ho





8.10 SELF ASSESSMENT QUESTIONS

Question No:-1 What is a critical region and on what basis, are we able to know about the position of critical region(s)?

Question No:-2 Why are the degrees of freedom so important in taking a decision about the rejection or acceptance of a hypothesis?

Question No:3 Define the following terms:

(a) one tailed and two tailed test.

(b) Test of significance.

(c) Degrees of freedom.

(d) Level of significance.

(e) Composite hypothesis.

Question No:4 Write short notes on:

(a) Randomized test.

(b) One-tailed test.

(c) Critical region.

(d) Statistic.

(e) P-value concept.

Question No:4 What is a critical region and on what basis, are we able to know about the position of critical region(s)?

Unit 3 **Lesson 9**

TESTING OF HYPOTHESES

Structure:

9.1 Introduction

9.2 Objectives

9.3 Types of Error

9.4 Procedure for testing the hypothesis

9.5 Illustration

9.6 Sampling from attributes

9.7 Self Assessment Questions

9.1 INTRODUCTION

We observed that there are essentially two kinds of statistical inferences, estimation and hypothesis testing. Both are concerned with learning something about an unknown aspect of a population on the basis of sample information. We have so far discussed the problems relating to estimation; presently our concern shall be the problem of testing hypotheses. A hypothesis is a theoretical proposition that is capable of empirical verification or disproof. It may be viewed as an explanation of some event or events, and which may be true or false explanation. Three forms of hypotheses are generally described in statistics; maintained, simple and composite. Those assumptions that are not exposed to any test are called the maintained hypotheses; while the remaining are called testable hypotheses. The A hypothesis is a theoretical proposition that is capable of empirical verification or disproof. It may be viewed as an explanation of some event or events, and which may be true or false explanation.

Three forms of hypotheses are generally described in statistics; maintained, simple and composite. Those assumptions that are not exposed to any test are called the maintained hypotheses; while the remaining are called testable hypothesis.

The procedure, by which we are able to reject our null hypothesis, is called criterion of test. In other words criterion of test refers to setting up of the boundary between critical and acceptance regions which is determined by many considerations; such as, the prior information concerning the distribution of the test-statistic, by the specification of the alternative hypothesis and so on.

The test criterion, however, may not always give us correct conclusions. In making any decision we are liable to commit one of the two types of error in this lesson

9.2 OBJECTIVES

Objectives of this lesson is to enable the learners

1. To learn how to use samples to decide whether a population possesses a particular characteristic
2. To determine how unlikely it is that an observed sample could have come from a hypothesized population and further, how to check the validity of our assertion about the population
3. To understand the two types of errors possible when testing hypotheses
4. To learn when to use one- tailed tests and when to use two-tailed tests
5. To learn the five-step process for testing hypotheses
6. To understand how and when to use the normal distribution for testing hypotheses about population means and proportions

9.3 TYPES OF ERROR

We make decision about the H_0 (Null hypothesis) on the basis of the information supplied by the observed sample observations. The conclusion drawn on the basis of a particular sample observation may not always be true in the in respect of the population. There are four possible situations which may arise if a statistical hypothesis is tested.

True State	Decision from the sample	
	Reject H_0	Accept H_0
H_0 true	Incorrect decision (Type-I Error)	Correct decision
H_0 false	Correct decision	Incorrect decision (Type-II Error)

If a statistical hypothesis is tested, as shown in the above table, we may get the following four possible cases:

a. The null hypothesis is true and it is accepted;

b. The null hypothesis is false and it is rejected;

c. The null hypothesis is true, but it is rejected;

d. The null hypothesis is false, but it is accepted.

Clearly, the last two cases lead to errors which are called errors of sampling. The error made in (c) is called Type I Error. The error committed in (d) is called Type II Error. In either case a wrong decision is taken.

Thus we can say that, Error of rejecting H_0 when it is true is called as the type-I error and the error of accepting H_0 when it is false is called as type-II error. The probabilities of type-I and Type-II errors are denoted respectively by α and β .Thus

α= Possibility of type-I error= Probability of rejecting H_0 when H_0 is true.

β =Possibility of type-II error= Probability of accepting H_0 when H_0 is false.

Symbolically

$P[x \in \omega / H_0] = \alpha$, where $x = x_1, x_2, ...x_n$

$\Rightarrow \int_{\omega} L_0 dx = \alpha$

where L_0 is the likelihood function of sample observations under H_0 and $\int dx$ represents n-fold integral $\int \int ....\int dx_1 dx_2 .....dx_n$ .If we find $x \in \omega$ we reject H_0 and if we find $x \in \overline{\omega}$ we accept H_0. Where ω and $\overline{\omega}$ are two disjoint and exhaustive subsets of the set S(The set of all possible outcomes of the variable x).

Again

$P[x \in \overline{\omega} / H_0] = \beta$ or $\int_{\omega} L_1 dx = \beta$

where L_1 is the likelihood function observations under H_1

Since we have

$$\int_{\omega} L_1 dx + \int_{\bar{\omega}} L_1 dx = 1 \quad \Rightarrow \int_{\omega} L_1 dx = 1 - \int_{\bar{\omega}} L_1 dx = 1 - \beta$$

or $P[x \in \omega / H_1] = 1 - \beta$ .... $(*)$

Note:

1. α, the probability of the type –I error is known as the level of significance. It is also called as the size of the test.

2. $1-\beta$ as defined in $(*)$ is called as the power function of the test for testing H_0 against the alternative H_1. The value of the power function at a particular point is called as the power of the test at that point.

3. An ideal test would be one which properly keeps under control both the type of errors, unfortunately for fixed sample size n, α and β are so related (like producers and consumers in sampling inspection plan) that the reduction in one results in an increase in the other. Consequently simultaneous minimizing of both the errors is not possible. Since error of type-I seems to be more serious. The usual practice is to control α at predetermined low level and subject to this restriction, choose a test which minimizes β or maximize the power function $1-\beta$, generally we choose α=0.05 or .01.

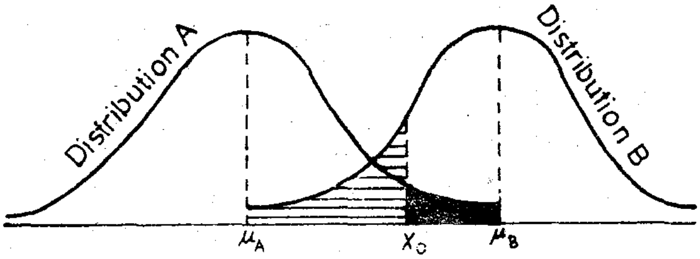
The general idea behind the two types of errors. can also be illustrated by an example of testing null hypothesis against simple alternative hypothesis.. Assume that we obtain a sample of n-observations. We are to examine whether this sample belongs to a normal population A with mean X or population B with mean = Y (We are not considering the variances of the populations sample for the time being.)

So we have: $H_0 : \mu = \mu_X$

H_1: $\mu = \mu_Y$

The level of significance may be chosen a priori as, say, 5 per cent. Since the alternative hypothesis is $\mu = \mu_Y$ that and assuming that $\mu_X > \mu_Y =$ only high values of observed sample mean $\overline{X}$ (which is test-statistic in the present case) relative to μ_X would constitute evidence against H_0.

The two distributions A and B are compared diagrammatically in below given figure where in we show the probabilities of two types of error involved in hypothesis testing.



Error type I is committed whenever $\overline{X}$ falls to the right of the boundary point X_0 (assuming that H_0 is true) and its probability is given by the chosen level of significance (i.e., 5 per cent) and corresponds to the blackened area. The error type II occurs whenever we do not reject H_0 when it is in fact false. This happens whenever $\overline{X}$ falls to the left of X_0 (assuming that H_0 is not true). The probability of making this error is given by the striped area in above figure. As could be seen, the decrease in the probability of one type of error can be brought about only at the cost of increase in the probability of another type of error. We can decrease the probability of error type I by shifting the boundary point X_0 farther to the right. But by doing so we would obviously increase the striped area which represents the probabilities of error type II. Then, the question arises, how to decrease the probabilities of both types of error simultaneously? The only way in which we can reduce the probabilities of both kinds of error at the same time is by increasing the size of sample.

9.4 PROCEDURE FOR TESTING THE HYPOTHESIS

The first step in hypothesis testing is that of formulation of the null hypothesis and its alternative. The next step consists of devising a criterion of test that would enable us to decide whether the null hypothesis is to be rejected or not. For this purpose the whole set of values of the population is divided into two regions:

The acceptance region and rejection regions. The acceptance region includes the values of the population which have a high probability of being observed, and the rejection region or critical region includes those values which are highly unlikely to be observed. The test is then performed with reference to test-statistic. The empirical tests that are used for testing the hypothesis are called tests of significance. If the value of the test-statistic falls in the critical region, the null hypothesis is rejected; while if the value of test-statistic falls in the acceptance region, the null hypothesis is not rejected.

The various steps involved in testing of a statistical hypothesis are as under.

1. ***Null Hypothesis***: we set up the Null Hypothesis H_o.

2. ***Alternative Hypothesis.*** Next we set up the alternative hypothesis H_1. This will enable us to decide whether we have to use a single-tailed (right or left) test or two-tailed test.

3. ***Level of Significance***. Appropriate level of significance α is chosen depending on the reliability of the estimates and permissible risk. This is to be decided before sample is drawn,

4. ***Test Statistic*: we c**ompute the test statistic:

$$Z = \frac{t - E(t)}{S.E(t)} \qquad \text{under} \qquad H_0$$

5. ***Conclusion***. We compare the computed value of Z in step 4 with the significant value (tabulated value) Z_α at the given level of significance, 'α'.

If $|Z| < Z_\alpha$ i.e., if the calculated value of Z (in modulus value) is less than Z_α, we say it is not significant. By this we mean that the difference t- E(t) is just due to fluctuations of sampling and the sample data do not provide us sufficient evidence against the null hypothesis which may, therefore, be accepted.

If $|Z| > Z_\alpha$ i.e., if the ***computed*** value of test statistic is greater than the critical or significant value, then we say that it is significant and the null hypothesis is rejected at level of significance α, i.e., with confidence coefficient (1-α)

Let us examine each step separately in a detailed manner as described below.

Step 1: The object of statistical inference is to derive conclusions about the population parameter, from the sample statistics. Certain rules are to be followed to measure the Level of uncertainty and decide whether to accept or reject our conclusions. To do this, the best way is to compare the sample estimate with the true value of the population parameter. When true value of population parameter is unknown, some assumption about the value of the true population parameter is made. This is then formulation of null hypothesis. There could be a very large number of hypothetical values which may be compatible with our sample estimate. To avoid such problem, it has become customary to make the hypothesis that the true population parameter is equal to zero.

Step 2: In making any decision, one is liable to commit one of the following types of errors:

Error type I: Rejects the null hypothesis, when it is actually true.

Error type II: Accepts the null hypothesis, when it is actually wrong.

One would like to minimise type I and type H errors. But unfortunately, for any given sample size, it is not possible to minimise both the errors simultaneously. The classical approach to this problem is to assume that a type I error is likely to be more serious in practice than a type-II error. Therefore, one should try to keep the probability of committing a type I error at a fairly low level, such as 0.01 or 0.05, and then try to minimise the type 11 error as much as possible. The probability of type I error is called the level of significance. Choosing a certain level of significance would mean specifying the probability of committing a type I error.

Step 3: The critical region includes only those values that correspond to the level of significance. But the critical region may be chosen at

(i) the right end

(ii) the left end

(iii) half at each end of the distribution of the variable.

In the first and second cases, it involves one-tail test and in the third case it involves a two-tail test. The decision on, which of the two to choose' would depend on the form in which the alternative hypothesis is expressed.

(1) $H_1 : \mu \neq \mu_0$ i.e,. $\mu < \mu_0$ or $\mu > \mu_0$ (Two tailed alternative)

(2) $H_1 : \mu > \mu_0$ right tail (one tailed alternative)

(3) $H_1 : \mu < \mu_0$ left tail (one tailed alternative)

The location of the critical region would depend on the direction at which the inequality sign points One has to choose the right tail as the critical region if the inequality sign is greater than; the left hand tail as the critical region if the inequality sign is less than, and a two-tail critical region when the inequality sign is not equal to.

Step 4: The choice among the various tests of significance depends on two things

(a) Size of sample, and (b) information on population variance

(i) If the variance of parent population is known, Z-test is appropriate (irrespective of the normality of the population and the sample size).

(ii) If the variance of the parent population is not known but the size of sample is large (it is greater than 30 observations), Z-test is still appropriate because the estimate of the population variance from a large sample is a satisfactory estimate of the true population variance.

(iii) If the variance is not known and also the size of sample is small (less than 30 observations), t-test is appropriate provided that the parent population is normal. And so on

Step 5: Once the decision has been taken about the particular test of significance, the test-statistic has to be computed from the observed sample observations to conduct the required test.

Step 6: The final step of the hypothesis testing is to compare the computed value of the test-statistic with that of tabulated theoretical value of this statistic.

9.5 ILLUSTRATION

Suppose that we are given the following information

$n = 36$, $s = 6$, $\overline{X} = 499$, $\alpha = 10\%$. And let us suppose that we have to test the hypothesis that population mean is equal to 500 against the alternative that it is less than 500 i.e

1. $H_0 : \mu = 500$ and $H_1 : \mu < 500$

This is a left-tail test with a α=10%.

2. $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = 6/6 = 1$ X ~N(500,1)

3. Critical region is z <-1.28

The z-score = 499 — 500 = —l

Since —1 <—1.28, H_0 is accepted.

What is the probability that we are wrong?

β = P(The Null Hypothesis is false but sample statistic falls in the acceptance region)

The acceptance region for the standard normal curve is $Z \geq$ —1.28.

Therefore, the acceptance region for the distribution of $\overline{X}$ is

$$Z = \frac{\overline{X} - 500}{\sigma_{\overline{X}}} \geq —1.28$$

$$\overline{X} \geq -1.28\sigma_{\overline{X}} + 500 = 500 - 1.28 \times 1 = 498.72$$

Therefore,

$$\beta = P(\overline{X} \geq 498.72 | \mu \neq 500)$$

We cannot compute the probability unless μ is known. Suppose

$\mu = 499.5$

Then $\beta = P(\overline{X} \geq 498.72 | \mu = 499.5)$

$$= P[(\overline{X} - \mu)/\sigma_{\overline{X}} \geq (498.72)/\sigma_{\overline{X}}] | \mu = 499.5, \sigma_{\overline{X}} = 1)$$

$$= P[Z \geq (498.72 - 499.5)/1]$$

$$= P[Z \geq -0.78] = 0.5 + 0.2823 = 0.7823$$

$(1-\beta)$ indicates how powerful the test is. A high $(1-\beta)$ (that is, close to 1) implies that the test is doing exactly what it should be doing: Rejecting H_0 when it is false. And a low $(1-\beta)$ indicates poor performance.

9.6 SAMPLING FROM ATTRIBUTES

Let us consider a sample from a population which is divided into two mutually exclusive and collectively exhaustive classes-one class possessing a particular attribute A (say), and the other class not possessing that attribute, The presence of an attribute in sampled unit may be termed as success and its absence as failure. In this case a sample of n observations is identified with that of a series of n independent Bernoulli trials with constant probability P of success for each trial. Then the probability of x successes in n trials, as given by the binomial probability distribution is: $p(x) = {}^{n}_{x}c\, p^{x} q^{n-x}$; x = 0, 1, 2... ….n.

as well. The only difference is that now, since we're dealing with a proportion, the binomial distribution is the correct sampling distribution to use. We know that as long as n is large enough to make both np and nq at least 5, we can use the normal distribution to approximate the binomial. If that is the case, we proceed exactly as we did with interval estimates of the mean.

9.7 SELF ASSESSMENT QUESTIONS

Question No:-1 Throw light on the need of the testing of hypothesis.

Question No:-2 Discuss a hypothesis, What types of hypotheses do you know? Discuss each of them.

Question No:-3 Discuss two types of errors in the testing of hypotheses. What is their role in testing?

Question No:-4 What do you understand by a large sample test?

Question No:-5 Why are the degrees of freedom so important in taking a decision about the rejection or acceptance of a hypothesis?

Question No:- 6. Define the following terms:

(a) Type II error.

(b) Power of a test.

(c) Degrees of freedom.

(d) Level of significance.

(e) Composite hypothesis.

Question No:-7 what is the role of an alternative hypothesis in hypotheses testing?

Question No:-8. Explain the basic principle of interval estimation as invented by J.Neyman.

Question No:-9 Write ‘Yes’ if the statements given below are correct, otherwise write ‘No’

(a) Degrees of freedom take care of the sample size in a decision problem about a hypothesis.

(b) Randomized test also involves some statistic.

(c) Each statistic has some distribution.

(d) Critical region is always on one tail only.

(e) Standard deviation of an estimate and standard error are the same.

(I) Interval estimate is better than point estimate.

Question No:-10 Write whether the following statements are correct:

(a) Z-value lies between 0 and ∞ .

(c) Variance of a sample can be any value between $-\infty$ and ∞.

(d) When $\alpha = 0$, we have to accept H_0.

(e) When $\alpha=1$, we have to reject H0.

Unit 3 **Lesson 10**

TESTS OF SIGNIFICANCE OF DIFFERENCE OF PROPORTION

Structure:

10.1 Introduction

10.2 Objectives

10.3 Test of significance for single proportion

10.4 Examples based test of significance for single proportion

10.5 Test of significance of difference of proportions

10.6 Examples based test of significance for difference of proportion

10.7 Self Assessment Questions

10.1 INTRODUCTION

Let us consider a sample from a population which is divided into two mutually exclusive and collectively exhaustive classes-one class possessing a particular attribute A (say), and the other class not possessing that attribute, The presence of an attribute in sampled unit may be termed as success and its absence as failure. In this case a sample of n observations is identified with that of a series of n independent Bernoulli trials with constant probability P of success for each trial. Since we are dealing with a proportion, the binomial distribution is the correct sampling distribution to use. We know that as long as n is large enough to make both np and nq at least 5, we can use the normal distribution to approximate the binomial. If that is the case, we proceed exactly as we did with interval estimates of the mean. So while dealing with the testing the significance of proportions we make use of the binomial distribution further we can use the normal distribution to approximate the binomial for large samples.

10.2 OBJECTIVES

Objectives of this lesson is to enable the learners

1. To learn how to use samples to decide whether a population possesses a particular characteristic
2. To determine how unlikely it is that an observed sample could have come from a hypothesized population and further, how to check the validity of our assertion about the population proportion
3. To understand the use test of significance when testing the significance for single proportion
4. To test of significance when testing the significance difference of two proportions in case of large population .
5. In general, to understand how and when to use the normal distribution for testing hypotheses about population means and proportions

10.3 TEST OF SIGNIFICANCE FOR SINGLE PROPORTION

Let us consider a sample from a population which is divided into two mutually exclusive and collectively exhaustive classes-one class possessing a particular attribute A (say), and the other class not possessing that attribute, The presence of an attribute in sampled unit may be termed as success and its absence as failure. In this case a sample of n observations is identified with that of a series of n independent Bernoulli trials with constant probability P of success for each trial the binomial distribution is the correct sampling distribution to use.

Then the probability of x successes in n trials, as given by the binomial probability distribution is: $p(x) = {}^{n}_{x}c\, p^{x} q^{n-x}$; x = 0, 1, 2... ….n.

Further we can use the normal distribution to approximate the binomial. If that is the case, we proceed exactly as we did with interval estimates of the mean.

If X is the number of individuals (units) possessing the given attribute in n independent trials with constant probability P of success for each trial, then

p= observed sample proportion= x/n

E (X) = nP and V (X)= nPQ,

where Q = 1- P. is the probability of failure. For large samples, the binomial distribution tends to normal distribution.

Hence for large n, X~ N (nP, nPQ), , the standard normal variate corresponding to the statistic p is

$$Z = \frac{p - E[p]}{S.E(p)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0,1)$$

If we have a sampling from finite population of size N S.E of p is given by

$$\sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}}$$

Probable limits for observed proportion of success is given by

$$E(p) \pm 3\ S.E(p) = P \pm 3\sqrt{PQ/n}$$

if P is not known than we take p (sample proportion) as the estimate of P and probable limits for observed proportion of success are

$$p \pm 3\sqrt{pq/n}$$

Rejection rule for $H_0 : P = P_0$

Suppose that we are taking 5% level of significance then for testing significance at 5% level, the rules are as follows:

(i) If the alternative hypothesis is that the population proportion P is 'different'

from P_0,i.e $P \neq P_0$ reject H_0 when the value of z lies outside the range-1 .96 to 1.96.

H_1: $P \neq P_0$; Critical Region $|Z| \geq 1.96$

(ii) If the alternative hypothesis is that the population proportion P is 'greater' than P_0, reject H_0 when the value of Z is greater than 1.645.

H_1:($P > P_0$); Critical Region Z> 1.645

(iii) If the alternative hypothesis is that the population proportion P is 'less' than P_0, reject H_0 when the value of z is less than - 1.645.

H_1:($P < P_0$); Critical Region Z- 1.645

Otherwise, do not reject the null hypothesis H_0. Similarly for testing at 1% level and 10% the rejection rule are given below in a tabular manner

Level of significance α	10%	5%	1%
Critical region for $P \neq P_0$	$\|Z\|$> 1.64	$\|Z\|$> 1.96	$\|Z\|$>2.58
Critical region for $P < P_0$	Z<-1.28	z <-1.64	z<-2.33
critica1 region for $P > P_0$	Z>1.28	Z > 1.64	Z> 2.33

For large samples (n > 30), the sampling distributions of many statistics are approximately normal distribution. In such cases, we can use the results of the table given above to formulate decision rules.

10.4 Examples based test of significance for single proportion

Example::In order to check that what proportion of the employees prefer to provide their own retirement benefits in lieu of a company-sponsored plan. a simple random sample of 75 employee was taken and we find that 0.4 of them are interested in providing their own retirement plans. Management wants to find an interval about which they can be 99 percent confident that it contains the true population proportion.

Sol: In usual notations we have

n = 75 $\rightarrow$ Sample size, p= 0.4 $\rightarrow$ Sample proportion in favor

q= 0.6 $\rightarrow$ Sample proportion not in favor

Now the standard error of sample proportions is estimated by

$$\hat{\sigma}_p = \sqrt{\hat{p}\hat{q}/n} = \sqrt{\frac{(0.4)(0.6)}{75}} = \sqrt{0.0032} = 0.057$$

$\rightarrow$Estimated standard error of proportion

A 99 percent confidence level would include 49.5 percent of the area on either side of the mean in the sampling distribution. From the table we see that 0.495 of the area under the normal curve is located between the mean and a point 2.58 standard errors from the mean. Thus, 99 percent of the area is contained between plus and minus 2.58 standard errors from the mean. Our confidence limits then become

$$p \pm 2.58\hat{\sigma}_p = 0.4 \pm 2.58 \times 0.057 = 0.547, 0.253$$

Thus, we estimate from our sample of 75 employees that with 99 percent confidence we believe that the proportion of the total population of employees who wish to establish their own retirement plans lies between 0.253 and 0.547.

Example: In a sample of 1,000 people from a particular area, 540 prefer diet A and the rest prefer diet B. Can we assume that both rice and wheat are equally popular in this State at1% level of significance?

Solution.

In the usual notations, we are given : n = 1,000 X = Who prefer diet A = 540

p = Sample proportion of those Who prefer diet A = x/n =540/1000 = 0.54

Null Hypothesis, H_0: Both diet A and diet B are equally popular in the area so that

P =Population proportion of diet A = 0•5 ⇒ Q =1- P = 0.5.

Alternative Hypothesis, H_1 : $P \neq 0.5$ (two-tailed alternative)

Under H0, the test statistic is

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

$$Z = \frac{0.54 - 0.50}{\sqrt{0.50 \times 0.50/1000}} = \frac{0.04}{0.0138} = 2.532$$

Conclusion. The significant or critical value of Z at 1% level of significance for two- tailed test is 2.58, Since computed Z = 2.532 is less than 2.58, it is not significant at 1% level of significance. Hence the null hypothesis is accepted and we may conclude that diet A and diet B are equally popular in that area.

Example: : A die is thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the die cannot be regarded as an unbiased one and find the limits between which the probability of a throw of 3 or 4 lies.

Solution. If the coming of 3 or 4 is called a success, then in usual notations n 9,000; X = Number of successes = 3,240

Under the null hypothesis (H_0) that the die is an unbiased one, we get

P = Probability of success = Probability of getting a 3 or 4=1/6+1/6 =1/3

Alternative hypothesis, $H_1 : p \neq 1/3$, (i.e., die is biased).

We have $Z = \frac{X - nP}{\sqrt{nPQ}} \sim N(0,1)$

since n is large

$$Z = \frac{3240 - 9000 \times (1/3)}{\sqrt{9000 \times (1/3) \times (2/3)}} = \frac{240}{\sqrt{2000}} = 5.36 \sim N(0,1)$$

Since $|Z| > 3$, H_0 is rejected and we conclude that the die is almost certainly biased.

Since die is not unbiased, $P \neq 1/3$. The probable limits for 'P' are given by:

$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = p \pm \sqrt{\hat{p}\hat{q}/n}$ where

$\hat{P} = p = \frac{3240}{9000 = 0.36}$ and $\hat{Q} = 1 - p = 1 - 0.36 = 0.64$

Probable limits for population proportion of successes may be taken as:

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0.36 \pm \sqrt{\frac{0.36 \times 0.64}{9000}}$$

$$= 0.36 \pm \frac{0.6 \times 0.8}{30\sqrt{10}} = 0.345, 0.375$$

Hence the probability of getting 3 or 4 almost certainly lies between 0.345 and

0.375.

Example: A random sample of 500 oranges was taken from a large consignment and 65 were found to be bad. Show that the S.E. of the proportion of bad ones in a sample of this size is 0.015 and deduce that the percentage of bad oranges in the consignment almost certainly lies between 8.5 and 17.5.

Solution.

Here we are given: n = 500 X = Number of bad pineapples in the sample =65

p =Proportion of bad pineapples in the sample =65/500 = 0.13 q =1-p = 0.87

Since p, the proportion of bad pineapples in the consignment is not known, we may take

$$\hat{P} = p = 0.13,\ \hat{Q} = q = 0.87.$$

$$\text{S.E of proportion} = \sqrt{\hat{P}\hat{Q}/n} = \sqrt{0.13 \times 0.87/500} = 0.015$$

Thus, the limits for the proportion of bad pineapples in the consignment are:

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0.130 \pm 3 \times 0.015 = 0.30 \pm 0.045 = (0.085, 0.175)$$

Hence the percentage of bad oranges in the consignment lies almost certainly between 85 and 17.5.

Example: Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more, at 5% level. (Use Large Sample Test.)

Solution. In the usual notations, we are given. n = 20, X = Number of persons who survived after attack by a disease = 18 p = Proportion of persons survived in the sample = 0.90

Null Hypothesis, H0 : P= 0.85, i.e., the proportion of persons survived after attack by a disease in the lot is 85%.

Alternative Hypothesis, H_1 : P> 0.85 (Right-tailed alternative).

Under H_0,

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

$$Z = \frac{0.90 - 0.85}{\sqrt{0.85 \times 0.15/20}} = \frac{0.05}{0.079} = 0.633$$

Conclusion. Since the alternative hypothesis is one-sided (right-tailed), we shall apply right-tailed test for testing significance of. Z. The significant value of Z at 5% level of significance for right-tailed test is + 1.645. Since computed value of Z= 0.633 is less than 1.645, it is not significant and we may accept the null hypothesis at 5% level of significance.

10.5 TEST OF SIGNIFICANCE OF DIFFERENCE OF PROPORTIONS

Suppose we have to compare two large populations say A and B with respect to the prevalence of a certain attribute among their members. Let x_1, x_2 be the number of persons possessing the given attribute in large random samples of sizes n_1 and n_2 from the two populations respectively.

Then sample proportions are given by

p_1= observed proportion of success in a sample from population A= X_1/n_1

P_2 = observed proportion of success in a sample from population B= X_2/n_2.

If P_1 and P_2 are population proportions, then

$E(p_1) = P_1$. $E(p_2) = P_2$

$$V(p_1) = \sqrt{\frac{P_1Q_1}{n_1}} \quad \text{and} \quad V(p_2) = \sqrt{\frac{P_2Q_2}{n_2}}$$

Since for large samples, p_1 and p_2 are independently and asymptotically normally distributed, (p_1 -p_2) is also normally distributed. Then the standard variable corresponding to the difference (p_1-p_2) is given by:

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0,1)$$

Under the null hypothesis, **H_0 : P_1=P_2**. i.e., there is no significant difference between the sample proportions, we have

$$E(p_1-p_2)=E(p_1)-E(p_2)=P_1-P_2=0$$

$$\text{Also} \quad V(p_1-p_2)=V(p_1)+V(p_2)$$

the covariance term Cov(p_1, p_2) vanishes, since sample proportions are independent.

$$\Rightarrow \qquad V(p_1-p_2)=\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}=PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

$$\left[\because \text{under} H_0:P_1=P_2=P \text{ say,and} Q_1=Q_2=Q\right]$$

Hence, under H_0 P_1 = P_2, the test statistic for the difference of proportions becomes

$$Z=\frac{(p_1-p_2)}{\sqrt{PQ(1/n_1+1/n_2)}}\sim N(0,1)$$

In general, common population proportion P under H_0 is not known. Under H_0: P_1 =P_2 = P (say), an unbiased estimate of the population proportion P based on both the samples is

$$\hat{P}=\frac{n_1p_1+n_2p_2}{n_1+n_2}=\frac{X_1+X_2}{n_1+n_2}$$

Thus by using $\hat{P}$ in above stated test statistic we test for null hypothesis

10.6EXAMPLES BASED ON TEST OF SIGNIFICANCE OF DIFFERENCE OF TWO PROPORTIONS

Example: Random samples of 400 men and 600 women were asked whether they would like to have a cinema hall in their locality. 200 men and 325 women were in favor of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal are same against that they are not, at 1% level.

Sol. Null Hypothesis H_0:P_1 = P_2 = P (say), i.e., there is no significant difference between the opinions of men and women as far as proposal of flyover is concerned.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

We are given: n_1 = 400, X_1 = Number of men favoring the proposal = 200, n_2= 600, X_2 = Number of women favoring the proposal = 325

p_1 =Proportion of men favoring the proposal in the sample $= {x_1}/{n_1}$ = 220/400=0.5

p_2= Proportion of women favoring the proposal in the sample $= x_2/n_2$

=325/600 = 0.541

Since samples are large, the test statistic under the Null Hypothesis,H_0 is:

$$Z = \frac{(p_1 - p_2)}{\sqrt{PQ(1/n_1 + 1/n_2)}} \sim N(0,1)$$

Under H_0: P_1 =P_2 = P (say), an unbiased estimate of the population proportion P based on both the samples is

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{200 + 325}{400 + 600} = 0.525$$

$$\Rightarrow \hat{Q} = 1 - \hat{P} = 1 - 0.525 = 0.475$$

$$\therefore \qquad Z = \frac{-0.041}{\sqrt{0.001039}} = \frac{-0.041}{0.0323} = -1.269$$

Conclusion. Since $|z|$ =1.269 which is less than 2.58, it is not significant at 1% level of significance. Hence H_0 may be accepted at 5% level of significance and we may conclude that men and women do not differ significantly as regards proposal of flyover is concerned

Example: In a large city A, 20 per cent of a random sample of 900 school children had defective eye-sight. In other large city B, 15 per cent of random sample of 1,600 children had the same defect. Is this difference between the two proportions significant? Obtain 95% confidence limits for the difference in the population pro portions.

Sol Sol In usual notations we have

$p_2 = 0.15 \quad \Rightarrow q_2 = 0.85$ and $p_1 = 0.20 \quad \Rightarrow q_1 = 0.80$

Under H_0: $P_1 = P_2$ the test statistic for large samples is

$$Z = \frac{(p_1 - p_2)}{\sqrt{PQ(1/n_1 + 1/n_2)}} = 3.21 \sim N(0,1)$$

Where under H_0: $P_1 = P_2 = P$ (say), an unbiased estimate of the population proportion P based on both the samples is

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} \qquad \Rightarrow \hat{Q} = 1 - \hat{P}$$

Conclusion. Since the calculated value of Z is greater than 1.96, it is significant at 5% level. We, therefore, reject the null hypothesis H_0 and conclude that the difference between the two proportions is significant.

The 95% confidence limits for the difference P_1 -P_2 are

$(p_1 - p_2) \pm 1.96 \, \mathrm{S.E \; of} \, (p_1 - p_2):$

Where $\quad S.E\,of\,(p_1-p_2)=\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}\approx\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$

$$=\sqrt{\frac{0.20\times0.80}{900}+\frac{0.15\times0.85}{1600}}=0.016$$

Hence 95% confidence limits for the difference P_1 -P_2 are

$$(0.20-0.15)\pm1.96(0.016)=0.05\pm0.031=0.019,0.081$$

Where 0.019 is the upper confidence limit and 0.018 is the lower confidence limit.

Example: A company has the head office at Delhi and a branch at Mumbai.The H.R director wanted to know if the workers at the two places would like the introduction of a new scheme of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Delhi, 62% favoured the new plan. At Mumbai out of a sample of 400. 42% were against the new plan. Is there any significant difference between their attitude towards the new plan at 1% level?

SOL: Under H_0: there is no significant difference between their attitude towards the new plan the, test statistic for large samples is:

$$Z=\frac{(p_1-p_2)}{S.E(p_1-p_2)}=\frac{(p_1-p_2)}{\sqrt{\hat{P}\hat{Q}(1/n_1+1/n_2)}}\sim N(0,1)$$

$$\hat{P}=\frac{n_1p_1+n_2p_2}{n_1+n_2}=\frac{500\times0.62+400\times0.59}{500+400}=0.607\quad\Rightarrow\hat{Q}=1-\hat{P}=1-0.607=0.393$$

$$\therefore\qquad Z=\frac{0.62-0.59}{\sqrt{0.607*0.393\left(\frac{1}{500}+\frac{1}{400}\right)}}=\frac{0.03}{\sqrt{0.00107}}=0.917\sim N(0,1)$$

Conclusion. Since the calculated value of Z is 0.917 which is less than 2.58, it is insignificant at 1% level. We, therefore, accept the null hypothesis H_0 and conclude that the there is no significant difference between the attitude of employees posted at Delhi and Mumbai as for as the introduction of new scheme is concerned.

10.7 SELF ASSESSMENT QUESTIONS

Question No:-1 Describe briefly the test of significance of difference of single proportion.

Question No:-2 Write down the steps involved in testing the significance of difference of two proportions

Question No:-3 Obtain 95% confidence limits in following cases

(a) For single proportion.(b) For two proportions

Question No:-4 when a sample of 70 retail executives was surveyed regarding the poor performance of the retail industry 66 percent believed that decreased sales were due to unseasonably warm temperatures, resulting in consumers' delaying purchase of cold-weather items.

(a) Estimate the standard error of the proportion of retail executives who blame warm weather for low sales.

(b) Find the upper and lower confidence limits for this proportion, given a 95 percent confidence level.

Question No:-5 A noted social psychologist, surveyed 150 top executives and found that 42 percent of them were unable to add fractions correctly.

(a) Estimate the standard error of the proportion.

(b) Construct a 99 percent confidence interval for the true proportion of top executives who cannot correctly add fractions.

Question No:-6 In a random sample of 200 men taken from area A, 90 were found to be consuming alcohol. In another sample of 300 men taken from area B, 100 were found to be consuming alcohol. Do the two areas differ significantly in respect of the proportion of men who consume alcohol?

Question No:-7 In a random sample of 500 men from a particular district of Maharashtra., 500 are found to be smokers. In one of 5,000 men from another district, 650 are smokers. Do the data indicate that the two districts are significantly different with respect to the prevalence of smoking among men?

Question No:-8 A factory is producing 40,000 pairs of shoes daily. From a sample of 400 pairs, 3% were found to be of sub-standard quality. Estimate the number of pairs that can be reasonably expected to be spoiled in the daily production and assign limits at 99% level of confidence.

Question No:-9 A manufacturer claimed that at least 98% of the steel pipes which he supplied to a factory conformed to specifications. An examination of a sample of 500 pieces of pipes revealed that 30 were defective. Test this claim at a significance level of (1) 0.05, (ii) 0•01.

Question No:-10 A random sample of size 1100 selected from a large bulk of mass produced machine parts contains 7% defectives. What information can be inferred about the percentage of defective in the bulk?

TESTS OF SIGNIFICANCE FOR MEANS

Structure:

11.1 Introduction

11.2 Objectives

11.3 Test of significance for single mean

11.4 Examples based test of significance for single mean

11.5 Test of significance of difference of two means

11.6 Examples based test of significance for difference of two means

11.7 Self Assessment Questions

11.1 INTRODUCTION

For large samples (n > 30), the sampling distributions of many statistics are approximately normal distribution. The test of hypothesis about a population mean or two population means, by the t-test, is applicable under the circumstances that population variance(s) is/are not known and the sample(s) is/are of small size. In cases where the population variance(s) is/are known, we use Z-test (normal test). Moreover, when the sample size is large, sample variance approaches population variance and is deemed to be almost equal to population variance. In this way, the population variance is known even if we have sample data and hence the normal test is applicable. The distribution of Z is always normal with a mean zero and a variance 1. The value of Z can be read from the table for the area under the normal curve,

11.2 OBJECTIVES

Objectives of this lesson is to enable the learners

1. To learn how to use samples to decide whether a population possesses a particular characteristic.
2. To determine how unlikely it is that an observed sample could have come from a hypothesized population and further, how to check the validity of our assertion about the population mean
3. To understand the use test of significance when testing the significance for single proportion
4. To learn how to use the test of significance of the significance difference of two means in case of large population .
5. In general, to understand how and when to use the normal distribution for testing hypotheses about population means

11.3 TEST OF SIGNIFICANCE FOR SINGLE MEAN

TEST OF SIGNIFICANCE FOR SINGLE MEAN. We know that if x_i (i= 1, 2, ..., n) is a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean is distributed normally with mean μ and variance $\frac{\sigma^2}{n}$ i.e.,

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

However, this result holds, i.e., $\overline{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, even in random sampling from non-normal population provided the sample size n is large [Central Limit Theorem]. Thus for large samples, the standard normal variate corresponding $\overline{x}$ to is:

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \qquad \text{…………………(1)}$$

Under the null hypothesis H$_0$, that the sample has been drawn from a population with mean μ. and variance σ^2, i.e., there is no significant difference between the sample mean ($\overline{x}$) and population mean (μ), the test statistic (for large samples), is:

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \text{……………………(2)}$$

If the population s.d. σ is unknown then we use its estimate provided by the

sample variance given by $\hat{\sigma}^2 = s^2$ or $\hat{\sigma} = s$ (for large samples).Then from(2)

$$Z = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim N(0,1)$$

Confidence limits for μ: 95% confidence interval for μ is given by:

$$|Z| \le 1.96, \qquad \text{i.e.,} \qquad \left|\frac{\overline{x} - \mu}{\sigma/\sqrt{n}}\right| \le 1.96,$$

$$\Rightarrow \overline{x} - 1.96(\sigma/\sqrt{n}) \le \mu \le \overline{x} + 1.96(\sigma/\sqrt{n})$$

and $\overline{x} \pm 1.96 \,\sigma/\sqrt{n}$ are known as 95% confidence limits for μ. Similarly, 99% confidence limits for μ. are

$$\overline{x} - 2.58(\sigma/\sqrt{n}) \le \mu \le \overline{x} + 2.58(\sigma/\sqrt{n})$$

If the population s.d. σ is unknown then we use its estimate provided by the sample variance given by $\hat{\sigma}^2 = s^2$ or $\hat{\sigma} = s$ (for large samples).

However, in sampling from a finite population of size N, the corresponding 95% and 99% confidence limits for μ are respectively

$$\overline{x} \pm 1.96 \,\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}} \quad \text{and} \quad \overline{x} \pm 2.58 \,\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}}$$

For testing at 1% level and 10% the rejection rule are given below in a tabular manner

Level of significance α	10%	5%	1%
Critical region for $\mu \ne \mu_0$	$\lvert Z \rvert$> 1.64	$\lvert Z \rvert$> 1.96	$\lvert Z \rvert$>2.58
Critical region for $\mu < \mu_0$	Z<-1.28	z <-1.64	z<-2.33
critica1 region for $\mu > \mu_0$	Z>1.28	Z > 1.64	Z> 2.33

11.4 EXAMPLES BASED TEST OF SIGNIFICANCE FOR SINGLE MEAN

Example :A cinema hall has a cool drinks fountain supplying Orange and Colas. When the machine is turned on, it fills a 550 ml cup with 500 ml of the required drink.

The manager has two problems on hand.

i. The clients have been complaining that the machine supplies less than 500 ml.

ii. The two colas are supplied by two different manufacturers, each pressurizing him to drop the other supplier. Should he drop one?

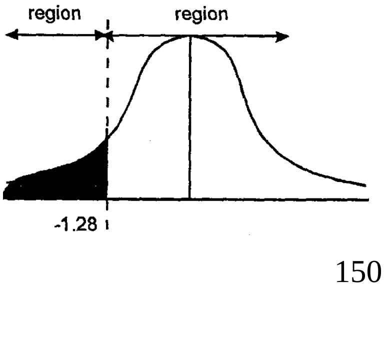
On a particular day, he took a survey of 36 clients and $\overline{X}$ comes out to be499 ml, specifications of the machine gave a s.d of 1 ml, Suppose that manager wants to minimize the customer complaints, Here we can set the hypothesis in three ways

Case-1

$H_0 : \mu = 500 \quad H_I : \mu < 500$ and the test statistic under H_o is

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ where } \frac{\sigma}{\sqrt{n}} = \frac{1}{6} = 0.17 \text{ so that}$$

$$Z = \frac{499 - 500}{0.17} = -6$$ This is a left-tailed test with level of significance 10%.

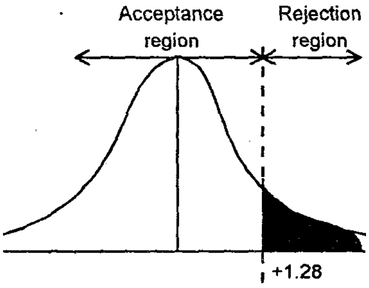


Critical Region: Z <-1.28 Since -6 <-1.28, we reject H_0.The machine was not set up properly.

Case-2 Suppose the manager ignores customer's complaints and instead wants to control the volume. That is, on an average, he does not want an excess outflow. We may set up the test as follows**.** $H_0 : \mu = 500 \quad H_l : \mu > 500$ and the test statistic is

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \text{ where } \frac{\sigma}{\sqrt{n}} = \frac{1}{6} = 0.17 \text{ so that}$$

$$Z = \frac{499 - 500}{0.17} = -6 \text{ This is a right-tailed test with } \alpha = 10\%.$$



Critical Region : z > +1.28, Since, this being a right tailed test the acceptance region is given by Z < 1.28 and therefore we accept H_0.

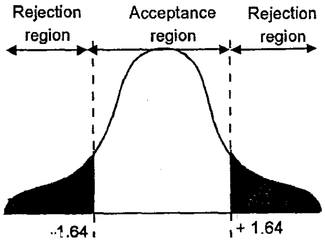
Case 3: Suppose, we combine Case 1 and Case 2. That is the manager intends to minimize customer complaints and does not want excess outflow. We may set up the test as follows:

$$H_0 : \mu = 500 \quad H_l : \mu \neq 500$$

The test statistic is

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \quad \text{where} \quad \frac{\sigma}{\sqrt{n}} = \frac{1}{6} = 0.17 \text{ so that}$$

$Z = \frac{499-500}{0.17} = -6$ This is a two-tailed test with α =10%.



Critical Region: Z >1.64, that is Z<-1.64 and Z > 1.64. Since -6 lies in one of the rejection regions, we reject H_0.

Example: A sample of 900 members has a mean 3.4 cms. and s.d. 2.61 cms. Is the sample from a large population of mean 3.25 cms. and s.d. 2.61 cms. ?If the population is normal and its mean is unknown, find the 95% and 98% fiducial limits of true mean.

Solution. Null Hypothesis, (H_0): The sample has been drawn from the population with mean μ = 3.25 cms. and S.D. $= \sigma$ =2.6l cms.

Alternative Hypothesis, H1: $\mu \neq 3.25$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is: $Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$ (Since n is large.)

Here, we are given: $\overline{x}$ = 3.4 cms., n = 900 μ = 3.25 cms. and σ = 2.61 cms.

$$Z = \frac{3.45 - 3.25}{2.61 / \sqrt{900}} = \frac{0.15 \times 30}{2.61} = 1.73$$

Since $|Z| <$ 1•96, we conclude that the data don’t provide us any evidence against the null hypothesis (H_0) which may, therefore, be accepted at 5% level of significance.

95% fiducial limits for the population mean μ are:

$$\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right) = 3.40 \pm 1.96\left(\frac{2.61}{\sqrt{900}}\right) = 3.40 \pm 0.1705$$

i,e., 3.5705 and 3.2295

Example: A sample of size 400 was drawn and mean was found to be 99. Test whether this sample could have came from a normal population with mean100 and standard deviation 8 at 5% level of significance

Sol: Here, we are given: $\bar{x}$ = 99, n = 400 μ = 100 and σ = 8

Null Hypothesis, (H_0): The sample has been drawn from the normal population with mean μ = 100 and S.D. = σ =8

Alternative Hypothesis, H1: $\mu \neq 100$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is: $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

(Since n is large.)

$$Z = \frac{99-100}{8/\sqrt{400}} = -\frac{5}{2} = -2.5 \quad \Rightarrow |Z| = 2.5$$

Since $|Z| >$ 1•96, i.e the calculated value of Z is greater that its critical value ($|Z| > 1.96$) we reject the null hypothesis and conclude the sample has not been drawn from the normal population with mean 100 and S.D. = 8.

11.5 Test of Significance for Difference of Means

Let us consider two independent large samples of sizes n$_1$ and n$_2$ from two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively. Let $\bar{x}_1$ and $\bar{x}_2$ be the corresponding sample means Then, since sample sizes are large,

$$\bar{x}_1 \sim N(\mu_1, \sigma_1^2 / n_1) \qquad \text{and} \qquad \bar{x}_2 \sim N(\mu_2, \sigma_2^2 / n_2)$$

Also $\bar{x}_1$ - $\bar{x}_2$, being the difference of two independent normal variates is also a normal variate with mean μ_1- μ_2 and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ The value of Z (S.N.V.) corresponding to($\bar{x}_1$ - $\bar{x}_2$) is given by:

$$Z = \frac{(\bar{x}_1 - \bar{x}_1) - E(\bar{x}_2 - \bar{x}_2)}{S.E(\bar{x}_1 - \bar{x}_2)} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Under the null hypothesis, H$_0$: $\mu_1 = \mu_2$ i.e., there is no significant difference between the sample means,

Thus under H$_0$: $\mu_1 = \mu_2$, the test statistic becomes (for large samples),

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, i.e., if the samples have been drawn from the populations with common S.D. a, then under H$_0$: $\mu_1 = \mu_2$

$$Z = \frac{(\overline{x}_1 - \overline{x}_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Remark: 1. If σ is not known, then its estimate based on the sample variances is used. For large samples, the following estimate of σ^2 is used

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$

2. If $\sigma_1^2 \neq \sigma_2^2$ and σ_1 and σ_2 are not known, then they are estimated from sample values, for large samples we use

$$\hat{\sigma} = S_1^2 \approx s_1^2 \quad \text{and} \quad \hat{\sigma}_2^2 = S_2^2 \approx s_2^2$$

$$Z = \frac{(\overline{x}_1 - \overline{x}_1)}{\sigma\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

11.6 EXAMPLES BASED TEST OF SIGNIFICANCE FOR DIFFERENCE OF TWO MEANS

Example: The means of two single large samples of 1,000 and 2,000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2.5 inches? (Test at 5% level of significance.)

Sol. In usual notations, we are given: n1=1,000, n_2 = 2,000, $\overline{x}_1$ = 67.5 inches, $\overline{x}_2$= 68.0 inches.

Null hypothesis, $H_0\ \mu_1 = \mu_2$ and σ = 25 inches, i.e., the samples have been drawn from the same population of standard deviation 2.5 inches.

Alternative hypothesis $H_1 : \mu \neq \mu_0$

Under H_0 the test statistic is

$$Z = \frac{(\overline{x}_1 - \overline{x}_1)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1) \qquad 0r\ Z = \frac{67.5 - 68.0}{2.5\sqrt{\frac{1}{1000} + \frac{1}{2000}}} = -5.1$$

Conclusion. Since $|Z|$ >3,, the value is highly significant and we reject the null hypothesis and conclude that samples are certainly not from the same population with standard deviation 2.5.

EXAMPLE : in a survey of buying habits, 400 women shoppers are chosen at random in super market 'A' located in a certain section of the city. Their average weekly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For 400 women shoppers chosen at random in super market 'B' in another section of the city, the average weekly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test at 1% level of significance whether the average weekly food expenditure of the two populations of shoppers are equal.

Solution, in the usual notations, we are given that

$$n_1 = 400 \qquad \overline{x}_1 = 250 \qquad s_1 = 40$$
$$n_2 = 400 \qquad \overline{x}_2 = 220 \qquad s_2 = 55$$

Null hypothesis, $H_0\ \mu_1 = \mu_2$ i.e., the average weekly food expenditures of the two populations of shoppers are equal.

Alternative hypothesis $H_1 : \mu \neq \mu_0$ (Two-tailed)

Test Statistic. Since samples are large, under H_0, the test statistic is:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Since σ_1^2 and σ_2^2, the population standard deviations are not known, we can take for large samples $\sigma_1^2 = s_1^2$ and $\sigma_2^2 = s_2^2$ and then Z is given by:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{250 - 220}{\sqrt{\frac{(40)^2}{400} + \frac{(55)^2}{400}}} = 8.82 \sim N(0,1)$$

Conclusion. Since $|Z|$ is much greater than 2.58, the null hypothesis ($\mu_1 = \mu_2$) is rejected at 1% level of significance and we conclude that the average weekly expenditures of two populations of shoppers in markets A and B differ significantly.

11.7 SELF ASSESSMENT QUESTIONS

Question No:-1 Under what circumstances can the normal distribution be used to find confidence limits of the populations mean?

Question No:-2 On the basis of a random sample from a normal population with a known variance 2, obtain 99% confidence limits for the population mean u. What will be the confidence limits, if the variance is unknown?

Question No:-3 The manufacturer of television tubes knows from past experience that the average life of a tube is 2,000 hours with a standard deviation of 200 hours. A sample of 100 tubes has an average life of 1,950 hours. Test at the 0.05 level of significance if this sample came from a normal population of mean 2,000 hours.

State your null and alternative hypotheses and indicate clearly whether a one-tail or a two- tail test is used and why? Is the result of the test significant?

Question No:-4 A sample of 100 items, drawn from a universe with mean value 64 and S.D. 3, has a mean value 63.5. Is the difference in the means significant? What will be your inference, if the sample had 200 items?

Question No:-5 A sample of 400 individuals is found to have a mean height of 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height of 67•39 inches and standard deviation 1.30 inches?

Question No:-6 A random sample of 400 is taken from a large number of coins. The mean weight of the coins in the sample is 28.57 gms and the s.d. is 1.25 gms. What are the limits which have a 95% chance of including the mean weight of all the coins?

Question No:-7 : A random sample of 110 days shows an average daily sale of Rs 60 with a s.d. of Rs 10 in a particular shop. Assuming a normal distribution, construct a 95% confidence interval for the expected sale per day.

Question No:-8 Throw light on the need of the testing of hypothesis.

Question No:-9. Discuss a hypothesis. What types of hypotheses do you know ? Discuss each of them.

Question No:-10.The yield of two strains of a crop was found to be as given below:

Strain 1	15.1, 20, 15, 38.7, 9, 12.3, 17, 36.5, 36
Strain 2	13.8, 19, 12, 9.04, 7.6, 19, 29, 34.1, 18.08, 19.2, 16

Test whether the mean yields of the two strains in general are equal. Perform the test at $\alpha = 0.05$.

NEYMAN-PEARSON LEMMA

Structure:

12.1 Introduction and objectives

12.2 Derivation of N P Lemma

12.3 Examples based on N P Lemma

12.4 Self assessment Questions

12.1 INTRODUCTION AND OBJECTIVES

In statistics, the Neyman–Pearson lemma, named after Jerzy Neyman and Egon Pearson. The N-P lemma tells us that the best test for a simple hypothesis is a likelihood ratio test. While performing test of significance for simple null versus simple alternative it provides most powerful test at the level of significance. In practice, the likelihood ratio is often used directly to construct tests. However it can also be used to suggest particular test-statistics that might be of interest or to suggest simplified tests for this, one considers algebraic manipulation of the ratio to see if there are key statistics in it related to the size of the ratio (i.e. whether a large statistic corresponds to a small ratio or to a large one).

12.2 DERIVATION OF N P LEMMA

Let ω be a critical region of the size α and k>0 be a constant such that

$$\omega = \left\{ x \in S : \frac{f(x,\theta_1)}{f(x,\theta_0)} > k \right\}$$

$$\Rightarrow \omega = \left\{ x \in S : \frac{L_1}{L_0} > k \right\} \qquad \text{..................(i)}$$

and $\overline{\omega} = \left\{ x \in S : \frac{L_1}{L_0} \leq k \right\}$ …..………………….(ii)

Where L_1 and L_0 are the likelihood functions of the sample observations $x = (x_1, x_2, ...x_n)$ under H_1 and H_0 respectively and ω is the most powerful critical region of the test hypothesis $H_0 : \theta = \theta_0$ v/s $H_1 : \theta = \theta_1$

Proof: We are given

$P[x \in \omega / H_0] = \alpha$ ……………(iii)

Power of the region

$P[x \in \omega / H_1] = \int_{\omega} L_1 dx = 1 - \beta$ ……………(iv)

(say)

Now in order to establish the lemma, we have to prove that there exists no other critical region of size less than equal to α which is more powerful than ω

Let ω_1 be another critical region of the size $\alpha_1 \leq \alpha$ and its power is $1 - \beta_1$

So that

$P[x \in \omega_1 / H_0] = \alpha_1 = \int_{\omega_1} L_0 dx$

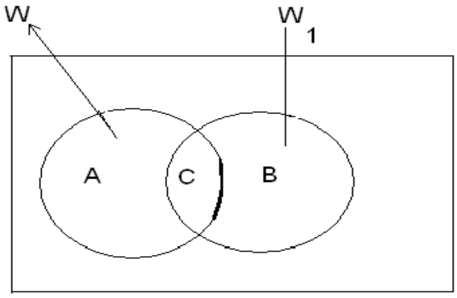
…………………(v)

$P[x \in \omega_1 / H_1] = 1 - \beta_1 = \int_{\omega_1} L_1 dx$ ……………….(vi)

Now we have to prove that

$1 - \beta \geq 1 - \beta_1$

Let $\omega = A \cup C$ and $\omega_1 = B \cup C$ (C may be empty i.e ω and ω_1 may be disjoint) as shown in the figure



If $\alpha_1 \leq \alpha$ then we have

$$\int_{\omega_1} L_0 dx \leq \int_{\omega} L_0 dx$$

$$\Rightarrow \int_{B\cup C} L_0 dx \leq \int_{A\cup C} L_0 dx \Rightarrow \int_{B} L_0 dx \leq \int_{A} L_0 dx$$

or $\int_{A} L_0 dx \geq \int_{B} L_0 dx$ ……………..(vi)

Since $A \subset \omega$, (i) $\Rightarrow$ $\int_{A} L_1 dx > k\int_{A} L_0 dx \geq k\int_{B} L_0 dx$

…………….(vii)

[using (vi)]

Now (ii) also implies that $\int_{\overline{\omega}} L_1 dx \leq k\int_{\overline{\omega}} L_0 dx$

This result also holds for any subset of $\overline{\omega}$ say $\overline{\omega} \cap \omega_1 = B$

Hence $\int_{B} L_1 dx \leq k\int_{B} L_0 dx \leq \int_{A} L_1 dx$ [Using (vii)]

Now adding $\int_{C} L_1 dx$ to both sides we get

$\int_{\omega_1} L_1 dx \leq \int_{\omega} L_1 dx$ $\Rightarrow 1-\beta \geq 1-\beta_1$

Hence the lemma.

12.3 EXAMPLES BASED ON NEYMAN PEARSON LEMMA

Example: If $X \sim N(\mu,4)$ to test $H_0 : \mu = -1$ against $H_1 : \mu = 1$ based on the sample of size 10 from the population whose critical region $x_1 + 2x_2 + 3x_3 + ... + 10x_{10} \geq 0$. What is the size α and power $(1-\beta)$ of the test.

Solution: critical region $\omega = x_1 + 2x_2 + 3x_3 + ... + 10x_{10} \geq 0$

Let $u = x_1 + 2x_2 + 3x_3 + ... + 10x_{10}$ since x_i's are i.i.d. $N(\mu,4)$

then $u \sim (55\mu, 385\sigma^2) \Rightarrow u \sim N(55\mu, 385 \times 4) = N(55\mu, 1540)$

The size α of the critical region is given by

$\alpha = P(x \in \omega / H_0) \quad = P(u \geq 0 / H_0)$

Now under Ho $\mu = -1$

$\because \quad u \sim (-55, 1540)$ and we have $Z = \frac{u - E[u]}{\sigma_u} = \frac{u+55}{\sqrt{1540}}$

under H_0 when $u = 0$

$$Z = {}^{55}\!/_{39.2428} = 1.4015$$

$\because \quad \alpha = P(Z \geq 1.4015) = 0.5 - P(Z \leq 1.4015)$=0.5-0.4192=0.1808 (from normal probability tables)

Now power of the test is

$(1-\beta) = P(x \in \omega / H_1) = P(u \geq 0 / H_1)$

under $H_1 : \mu = 1$, $u \sim (55{,}1540)$

$\Rightarrow \qquad Z = -55 \big/ 39.2428 = -1.4015$ and

$$(1-\beta) = P(Z \geq -1.4015) = 0.5 - P(-1.4015 \leq Z \leq 0) + 0.5$$

$$= 0.4192 + 0.5 = 0.9192$$

Exercise: If X has a p.d.f. of the form $f(x,\theta) = \frac{1}{\theta} e^{-\frac{x}{2}}$ $, 0 < x < \infty , \theta > 0 = 0$ otherwise

To test $H_0 : \theta = 2$ against $H_1 : \theta = 1$ use a random sample of size 2 and define critical region $\omega = \{x_1, x_2, x_1 + x_2 \geq 9.5\}$. Find α and β.

Soln: $\omega = \{x_1, x_2, x_1 + x_2 \geq 9.5\}$

Now $\alpha = P(x \in \omega / H_0) = P\{x_1 + x_2 \geq 9.5 / H_0\}$ ………………..(1)

In sampling from exponential distribution

$$\frac{2}{\theta}\sum_{i=1}^{n} x_i \sim \chi^2_{(2n)} \Rightarrow u = \frac{2}{\theta}(x_1 + x_2) \sim \chi^2_4$$

$$\because \quad \alpha = P\left[\frac{2}{\theta}(x_1 + x_2) \geq \frac{2}{\theta} \times 9.5 / H_0\right] = P[\chi^2_4 \geq 9.5]$$

$$\alpha = 0.5$$

Power of the test is given by

$$(1-\beta) = P(x \in \omega / H_1) = P[(x_1 + x_2) \geq 9.5 / H_1]$$

$$= P\left[\frac{2}{\theta}(x_1 + x_2) \geq \frac{2}{\theta} \times 9.5 / H_1\right] = P[\chi^2_4 \geq 19]$$

Example: Use Neyman-Pearson lemma to obtain the region for testing $\theta=\theta_0$ against $\theta=\theta_1>\theta_0$ and $\theta=\theta_1>\theta_0$, in the case of normal population $N(\theta,\sigma^2)$ where σ^2 is known. Hence find the power of the test.

Soln. $$L=\prod_{i=1}^{n} f(x_i,\theta)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}$$

Using N-P lemma best critical region is given by

$$\frac{L_1}{L_0}=\frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2\right\}}\geq k$$

$$\Rightarrow \exp\left[\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_1)^2-\sum_{i=1}^{n}(x_i-\theta_0)^2\right\}\right]\geq k$$

$$\Rightarrow -\frac{n}{2\sigma^2}(\theta_1^2-\theta_0^2)+\frac{1}{\sigma^2}(\theta_1-\theta_0)\sum_{i=1}^{n}x_i\geq \log k$$

$$\Rightarrow \bar{x}(\theta_1-\theta_0)\geq\frac{\sigma^2}{n}\log k+\frac{(\theta_1^2-\theta_0^2)}{2}$$

Case (i) if $\theta_1>\theta_0$ then BCR is determined by the relation (right tailed test)

$$\bar{x}>\frac{\sigma^2}{n}.\frac{\log k}{(\theta_1-\theta_0)}+\frac{(\theta_1+\theta_0)}{2} \qquad \Rightarrow \bar{x}>\lambda_1(\text{say})$$

$\because$ BCR is $\omega=\{x:\bar{x}>\lambda_1\}$ ………………….(i)

Case (ii) If $\theta_1<\theta_0$ then BCR is determined by the relation (left handed test)

$$\bar{x} < \frac{\sigma^2}{n} \cdot \frac{\log k}{(\theta_1 - \theta_0)} + \frac{(\theta_1 + \theta_0)}{2} \qquad = \lambda_2 \text{ (say)}$$

Hence BCR is $\omega = \{x : \bar{x} \le \lambda_2\}$ ……………………..(ii)

The constants λ_1 and λ_2 are so chosen as to make the probability of each of the relation (i) and (ii) equal to when H_0 is true.

Now sampling distribution of $\bar{x}$ when H_i is true is $(\theta_i, \frac{\sigma^2}{n})$, $(i = 0,1)$. Therefore the constants λ_1 and λ_2 are determined from the relations:

$$P(\bar{x} > \lambda_1 / H_0) = \alpha \text{ and } P(\bar{x} < \lambda_2 / H_0) = \alpha$$

$$P(\bar{x} > \lambda_1 / H_0) = P\left[Z > \frac{\lambda_1 - \theta_0}{\sigma/\sqrt{n}} \right] = \alpha \qquad ; Z \sim N(0,1)$$

$$\Rightarrow \frac{\lambda_1 - \theta_0}{\sigma/\sqrt{n}} = Z_\alpha \qquad \Rightarrow \lambda_1 = \theta_0 + \frac{\sigma}{\sqrt{n}} Z_\alpha$$

……………………(iii)

Where z_α ids the upper α -point of the standard normal variate given by

$$P[\,Z < z_\alpha\,] = \alpha$$

Also $P(\bar{x} > \lambda_2 / H_0) = \alpha \quad \Rightarrow P(\bar{x} \ge \lambda_2 / H_0) = 1 - \alpha \quad \Rightarrow$

$$P\left[Z \ge \frac{\lambda_2 - \theta_0}{\sigma/\sqrt{n}} \right] = 1 - \alpha$$

$$\Rightarrow \frac{\lambda_2 - \theta_0}{\sigma/\sqrt{n}} = Z_{1-\alpha} \qquad \Rightarrow \lambda_2 = \theta_0 + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha}$$ …………………(iv)

Power of the test

$$1-\beta = P(\bar{x} \in \omega / H_1) = P(\bar{x} \geq \lambda_1 / H_1) = P[Z \geq \frac{\lambda_1 - \theta_1}{\sigma/\sqrt{n}}]$$

$$= P[Z \geq \theta_0 + \frac{\sigma}{\sqrt{n}} Z_\alpha - \theta_1]$$

$$= P\left[Z \geq Z_\alpha - \frac{\theta_1 - \theta_0}{\sigma/\sqrt{n}} \right] \therefore \qquad \theta_1 > \theta_0$$

$=1-P[Z \leq \lambda_3] = 1-\phi[\lambda_3]$ Similarly in the case of (ii)

$$1-\beta \quad = P(\bar{x} \geq \lambda_2 / H_1) = P[Z < \frac{\lambda_2 - \theta_1}{\sigma/\sqrt{n}}] = P[Z < \theta_0 + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha} - \theta_0] \text{ using (iv)}$$

$$= P\left[Z < Z_{1-\alpha} + \frac{\theta_0 - \theta_1}{\sigma/\sqrt{n}} \right] = 1-\phi[\lambda_3]$$

Eq. (i) and (ii) provide BCR for testing $H_0 : \theta = \theta_0$ v/s $H_1 : \theta = \theta_1$ provided $\theta_1 > \theta_0$ in the first case and $\theta_1 < \theta_0$ in the second case.

Thus BCR for testing $H_0 : \theta = \theta_0$ v/s $H_1 : \theta = \theta_1 + c \;\; ; c > 0$ will not serve as BCR for testing $H_0 : \theta = \theta_0$ v/s $H_1 : \theta = \theta_1 - c \;\; c > 0$.

Hence in this problem no UMP test exists for testing simple hypothesis $H_0 : \theta = \theta_0$ v/s $H_1 : \theta \neq \theta_0$

Example::Show that for a normal population with mean zero and variance σ^2 the BCR for testing $H_0 : \sigma = \sigma_0$ v/s $H_1 : \sigma = \sigma_1$ is of the form

$$\sum_{i=1}^{n} x_i^2 \le a\alpha \quad \text{for } \sigma_0 > \sigma_1 \quad \text{and} \quad \sum_{i=1}^{n} x_i^2 \le b\alpha \quad \text{for } \sigma_0 < \sigma_1$$

show that the power of the BCR when $\sigma_0 > \sigma_1$ is $F\left[\frac{\sigma_0^2}{\sigma_1^2}\ , \chi^2_{\alpha,n}\right]$ where $\chi^2_{\alpha,n}$ is the lower $100\alpha\%$ and $F(.)$ is the density function of χ^2 wi8th n-degrees of freedom.

Soln: According to N P Lemma BCR is given by $\frac{L_1}{L_0} > k$ or $\frac{L_0}{L_1} \le \frac{1}{k} = A\alpha$ (say)

$$\text{Or} \quad \frac{L_0}{L_1} = \left[\frac{\sigma_1}{\sigma_0}\right]^n \exp\left[\left\{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right\}. -\frac{1}{2}\sum x_i^2\right] \le A\alpha$$

$$= \left[\frac{\sigma_1}{\sigma_0}\right]^n \exp\left[-\frac{1}{2}\left\{\frac{\sigma_1^2 - \sigma_0^2}{\sigma_0^2\sigma_1^2}\right\}\sum x_i^2\right] \le A\alpha$$

$$\text{or} = n\log\left[\frac{\sigma_1}{\sigma_0}\right] - \frac{1}{2}\left[\left\{\frac{\sigma_1^2 - \sigma_0^2}{\sigma_0^2\sigma_1^2}\right\}\sum x_i^2\right] \le A\alpha$$

$$\left\{\frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2}\right\}\sum x_i^2 \le \text{Log}A\alpha - n\log\left[\frac{\sigma_1}{\sigma_0}\right] \qquad \text{..................}(*)$$

When $\sigma_1 < \sigma_0$ BCR from (*)

$$\sum x_i^2 \le \left[\text{Log}A\alpha - n\log\left[\frac{\sigma_1}{\sigma_0}\right]\right]\frac{2\sigma_0^2\sigma_1^2}{\sigma_0^2 - \sigma_1^2} = a\alpha \ \text{(say)}$$

i.e., $\omega = [x : \sum x_i^2 \le a\alpha]$;for $\sigma_1 < \sigma_0$

When $\sigma_1 > \sigma_0$ BCR from (*)

$$\sum x_i^2 \geq \left[\text{Log}A\alpha - n\log\left[\frac{\sigma_1}{\sigma_0}\right]\right]\frac{2\sigma_0^2\sigma_1^2}{\sigma_0^2 - \sigma_1^2} = b\alpha \text{ (say)}$$

i.e., $\omega = [x : \sum x_i^2 \geq b\alpha]$;for $\sigma_1 > \sigma_0$

The constants $a\alpha$ and $b\alpha$ are so chosen so that size of critical region is α

$P[\sum x_i^2 \leq a\alpha] = \alpha$ under H_0

$$\text{or } P\left[\frac{\sum x_i^2}{\sigma_0^2} \leq \frac{a\alpha}{\sigma_0^2}\right] = \alpha \qquad \text{or} P\left[\chi_n^2 \leq \frac{a\alpha}{\sigma_0^2}\right] = \alpha$$

$$\Rightarrow \left[\frac{a\alpha}{\sigma_0^2}\right] = \chi^2_{\frac{\alpha}{n}} \qquad \Rightarrow a\alpha = \sigma_0^2\chi^2_{\frac{\alpha}{n}}$$

where $\chi^2_{\frac{\alpha}{n}}$ is the lower $100\alpha\%$ of chi-square distribution with n degrees of freedom given by

$$P\left[\chi^2 \leq \chi^2_{\frac{a}{n}}\right] = \alpha$$

Hence BCR $H_0 : \sigma = \sigma_0$ v/s $H_1 : \sigma = \sigma_1$ $(< \sigma_0)$ is $\omega = [x : \sum x_i^2 \leq \sigma_0^2\chi^2_{\frac{\alpha}{n}}]$

By definition power of the test is

$$1 - \beta = P(x \geq \omega / H_1) = P\left[\sum x_i^2 \leq a\alpha / H_0\right]$$

$$= P\left[\frac{\sum x_i^2}{\sigma_0^2} \leq \frac{a\alpha}{\sigma_0^2}\right] = P\left[\frac{\sum x_i^2}{\sigma_0^2} \leq \chi^2_{\alpha,n} / H_0\right]$$

$$= P\left[\frac{\sum x_i^2}{\sigma_1^2} \leq \frac{\sigma_0^2}{\sigma_1^2}\chi_{\alpha,n}^2 / H_1\right]$$

12.4 SELF ASSESSMENT QUESTIONS

Question No 1 Explain the concept of the most powerful tests and discuss how the Neyman-Pearson lemma enables us to obtain the most powerful crsitical region for testing a simple hypothesis against a simple alternative.

Question No 2 State and prove Neyman-Pearson Fundamental Lemma for testing a simple hypothesis against a simple alternative.

Question No 3 State Neyman-Pearson Lemma. Prove that if ω an MP region for testing

H_0: $\theta = \theta_0 = 0$ against H_1 : $\theta = \theta_1$

then it is necessarily unbiased. Also prove that the same holds good if. ω is an

UMP region.

Unit 4 **Lesson 13**

SMALL SAMPLE TESTS

Structure:

13.1 Introduction

13.2 Objectives

13.3 Concept of small sample tests

13.4 t-Test of significance for Single Mean

13.5 Assumption for Student's t-test

13.6 Examples based on t-Test of significance for Single Mean

13.7 Self Assessment Questions

13.1 INTRODUCTION

In a test of hypothesis, a sample is drawn from the population of which the parameter is under test. The size of sample varies since it depends either on experimenter or resources available, moreover, the test statistic involves the estimated value of the parameter which depends upon the number of observations, So the sample size play a very important role in testing of hypothesis.

When the sample size is small , such hypothesis testing can be achieved by using t-test, discovered by W.S. Gosset in 1908. He derived the distribution to find an exact test of a mean by making use of estimated standard deviation, based on a random sample of size n. R.A. Fisher in 1925 published that t-distribution can also be applied to the test of regression coefficient and other practical problems. In the present lesson we will learn to use the t test for the equality of single mean also the related confidence interval for population mean .

13.2 OBJECTIVES

The main objectives of this lesson are

1. To learn how to use samples to decide whether a population possesses a particular characteristic
2. To determine how unlikely it is that an observed sample could have come from a hypothesized population.
3. To understand how and when to use t distribution for testing hypotheses about population mean.
4. To learn when to use one- tailed tests and when to use two-tailed tests while testing the hypothesis for equality of single mean.
5. To understand the basic Concept of small sample tests

13.3 CONCEPT OF SMALL SAMPLE TESTS

In a test of hypothesis, a sample is drawn from the population of which the parameter is under test. The size of sample varies since it depends either on experimenter or resources available, moreover, the test statistic involves the estimated value of the parameter which depends upon the number of observations, So the sample size play a very important role in testing of hypothesis. For large samples (n>30) almost all the sampling distributions can be approximated to the normal, probability curve. However for small samples such hypothesis testing can be achieved by using t-test, F-test Chi-square test, Fisher's Z transformations etc.

13.4 t-TEST FOR SINGLE MEAN

Suppose we want to test

(i) If a random sample x_i (i = 1, 2…. n) of size n has been drawn from a normal population with a specified mean, say μ_0 or

(ii) If the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

Under the null hypothesis, H_0:

(i) The sample has been drawn from the population with mean μ_0 or

(ii) There is no significant difference between the sample mean and the population mean μ_0

The statistic $t = \frac{\overline{x} - \mu_0}{S/\sqrt{n}}$ follows Student's t-distribution with (n-1) d.f

where $\overline{x} = \frac{1}{n}\sum_{i-1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i-1}^{n}(x_i - \overline{x})^2$ is an unbiased estimate of σ^2

To decide about the acceptance or rejection of null hypothesis we now compare the calculated value of $|t|$ with the tabulated value at certain level of significance α. If calculated $|t|$>tabulated t, null hypothesis is rejected and if calculated $|t|$< tab. t, H_0 may be accepted at the level of significance adopted for (n-1) degree of freedom.

In many situations we may have limited data about the population so that we are required to estimate the confidence interval for the population mean μ with the help of a small sample.

In such cases we can use the estimating methods outlined below provided the population is normal.

Let us assume that the population is normal. Then we have to see whether the standard deviation of the population is known. If it is, then we can proceed as in the case of large samples provided we use c in computing the confidence interval,

The confidence interval will be $\left[\overline{x} \pm \sigma Z\right]$

If the population standard deviation is not known. In such a case we use the t distribution instead of the normal distribution.

The 95% confidence limits for population mean is given by

$\overline{x} \pm t_{0.05} S/\sqrt{n}$ where $\overline{x} + t_{0.05} S/\sqrt{n}$ is the upper confidence limit

And $\overline{x} - t_{0.05} S/\sqrt{n}$ is the lower confidence limit

Similarly 99% confidence limits for population mean is given by $\overline{x} \pm t_{0.01} S/\sqrt{n}$

Where degrees of freedom is (n -1). (This is because we have lost one degree of freedom by estimating σ using the n sample values.)

13.5 Assumption for Student's t-test.

The following assumptions are made in the Student's t-test

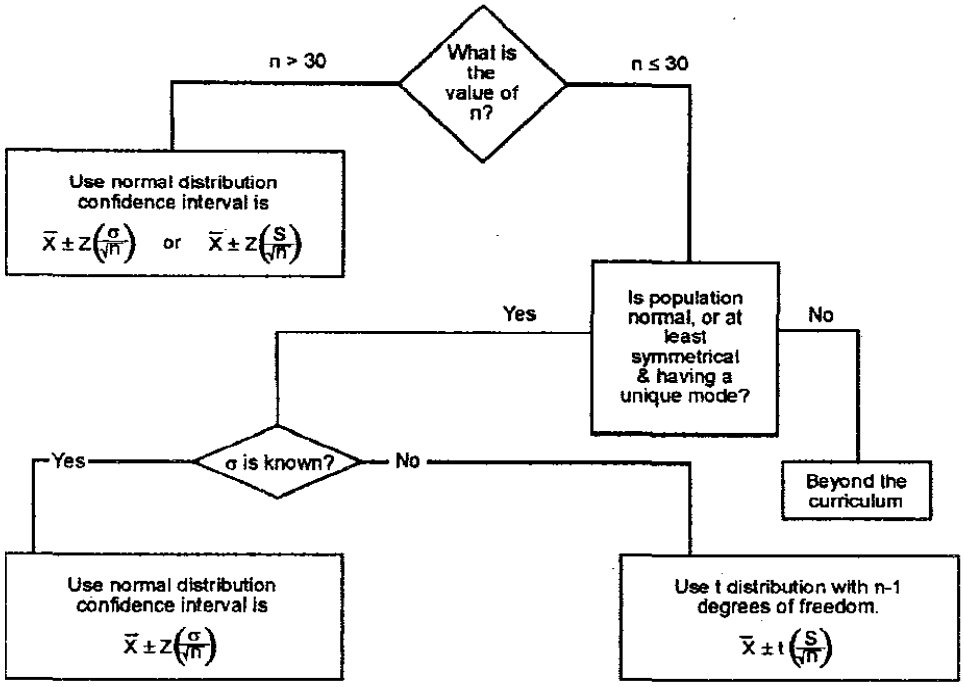
(i) The parent population from which the sample is drawn is normal.

(ii) The sample observations are independent, i.e., the sample is random.

(iii) The population standard deviation σ is unknown.

FOR CONFIDENCE INTERVAL

We can now prepare a flow chart for estimating a confidence interval for L, the population parameter


What is the value of n?
n > 30
n ≤ 30
Use normal distribution confidence interval is
$\bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$ or $\bar{X} \pm Z\left(\frac{S}{\sqrt{n}}\right)$
Is population normal, or at least symmetrical & having a unique mode?
Yes
No
Beyond the curriculum
σ is known?
Yes
No
Use normal distribution confidence interval is
$\bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$
Use t distribution with n-1 degrees of freedom.
$\bar{X} \pm t\left(\frac{S}{\sqrt{n}}\right)$

13.6 EXAMPLES BASED ON T-TEST FOR SINGLE MEAN

EXAMPLE:-A random sample of size 11 is selected from a symmetrical population with a unique mode. The sample mean and standard deviation are 200 and 30 respectively. Find the 90% confidence interval in which the population mean μ will lie.

Here, $\bar{X} = 200$ s = 30 n = 11

Degrees of freedom = n- 1 = 11-1 = 10

If we refer to the t table we see that for 10 degrees of freedom, the area in both tails combined is 0.10 or 10%, when t = 1.812.

Hence, area under the curve between $\bar{x} - t(s/\sqrt{n})$ and $\bar{x} + t(s/\sqrt{n})$ is 90% when t = 1.812.

Hence, we are 90% confident that the population mean lies in the interval 183.61 to 216.39.

Example :-A cinema hall has a cool drinks fountain supplying Orange and Colas. When the machine is turned on, it fills a 550 ml cup with 500 ml of the required drink.

The manager has two problems on hand.

i. The clients have been complaining that the machine supplies less than 500 ml.

ii. The two colas are supplied by two different manufacturers, each pressurizing him to drop the other supplier. Should he drop one?

A random sample of size 16 is taken with $\overline{x}$ = 499 ml. and n = 16, VQC) is unknown but the sample variance 1.96 (ml)^2.

Case I

Suppose the manager wants to minimize customer complaints. We may set up the test as follows:

$H_o : \mu = 500 \quad v/s \quad H_1 : \mu < 500$

Under the null hypothesis our test statistic is

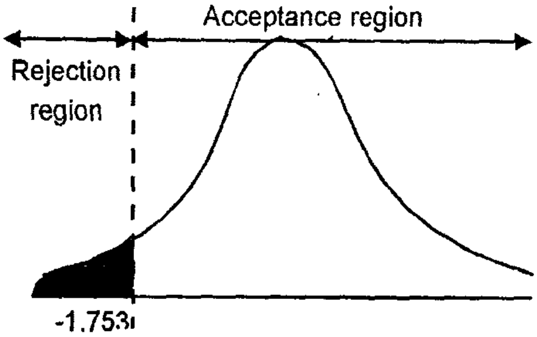
$$t = \frac{\overline{x} - \mu_0}{S/\sqrt{n}} = \frac{\overline{x} - 500}{1.4/\sqrt{16}} = \frac{\overline{x} - 500}{0.35} = \frac{4.99 - 500}{0.35}$$
$$= -2.857$$

follows Student's t-distribution with (n-1)=15 d.f

where $\overline{x} = \frac{1}{n}\sum_{i-1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i-1}^{n}(x_i - \overline{x})^2$ is an unbiased estimate of σ^2

This is a left tailed test. But the critical point is being obtained from the t tables.

Since n = 16, we need to look at the t_{15} distribution. If $\alpha = 5\%$, the critical point is -1.753. (Since tables indicate that P(—l .753 <t_{15} < 1.753) =90%)



Critical region t_{15} <—1.753

Since the observed value —2.857 <—1.753.

Therefore H0 is rejected.

Case 2

Suppose the manager ignores customer's complaints and instead wants to control the volume. That is, on an average, he does not want an excess outflow. We may set up the test as follows.

$$H_0 : \mu = 500 \quad v/s \quad H_1 : \mu > 500$$

Under the null hypothesis our test statistic is

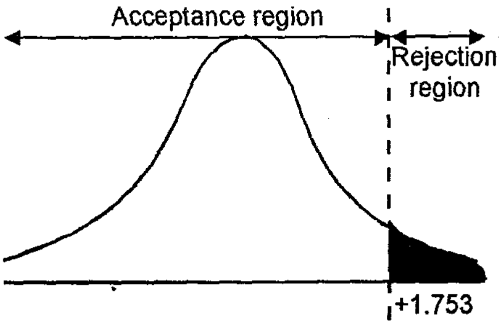
$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{\bar{x} - 500}{1.4/\sqrt{16}}$$

$$= \frac{\bar{x} - 500}{0.35} = \frac{4.99 - 500}{0.35} = -2.857$$

follows Student's t-distribution with (n-1)=15 d.f

This is a right tailed test. The tabulate value of t obtained from the t tables for 15 degrees of freedom at $\alpha = 5\%$ is 1.753 .

So critival region is $t_{15} > 1.753$.



Since —2.857 < 1.753, we accept H_0

Case 3

Suppose, we combine case 1 and case 2. That is, the manager intends to minimize customer complaints and does not want an excess outflow, we may set up the test as follows.

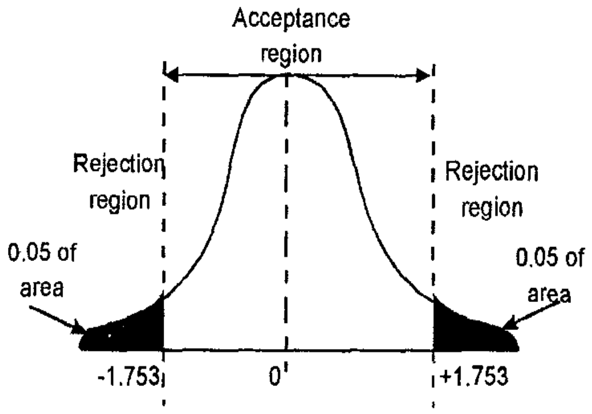
$$H_o : \mu = 500 \quad v/s \quad H_1 : \mu \neq 500$$

Under the null hypothesis our test statistic is

$$t = \frac{\overline{x} - \mu_0}{S/\sqrt{n}} = \frac{\overline{x} - 500}{1.4/\sqrt{16}} = \frac{\overline{x} - 500}{0.35} = \frac{4.99 - 500}{0.35}$$
$$= -2.857$$

follows Student’s t-distribution with (n-1)=15 d.f

This is a two tailed test. The tabulate value of t obtained from the t tables for 15 degrees of freedom at $\alpha = 5\%$ is 1.753 .



Critical region | t | > 1.753

Since | – 2.857 |>1.753 or we can say —2.857 lies in one (left) of the rejection regions, we reject H_0.

Remark:-If we use the t distribution for a one-tailed test, we need to determine the area located in only one tail. So to find the appropriate t value for a one-tailed test at a significance level of say 0.05 with 15 degrees of freedom, we would look under 0.10 column opposite the 15 degrees of freedom row. This is true because the 0.10 column represents 0.10 of the area under the curve contained in both the tails combined, and so it also represents 0.05 of the area under the curve contained in each of the tails separately.

Example: A machinist is making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a standard deviation of O.040 inch. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.

Solution. Here we are given: μ=0.700, $\bar{X}$=0.742 s=0.040 and n=10

Null Hypothesis, H_0 : $\mu = 0.700$, i.e., the product is conforming to specifications.

Alternative Hypothesis, H_1 : $\mu \neq 0.700$

Test Statistic. Under H0, the test statistic is:

$$t = \frac{\overline{x} - \mu_0}{S/\sqrt{n}} = \frac{\overline{x} - \mu_0}{s/\sqrt{n-1}}$$ follows Student's t-dist.^{n} with (n-1) d.f

$$t = \frac{(0.742 - 0.700) - \mu_0}{0.040/\sqrt{9}} = 3.15$$

How to proceed further: Here the test statistic 't' follows Student's t-distribution with 10-1 = 9 d.f. We compare the calculated value with the tabulated value of t for 9 d.f. and at certain level of significance, say 5%. Let this tabulated value be denoted by t_0.

(1) If calculated 't', viz., 3.15 > t_0, we say that the value of t is significant. This implies that $\overline{x}$ differs significantly from μ and H_0 is rejected at this level of significance and we conclude that the product is not meeting the specifications.

(ii) If calculated $t < t_0$, we say that the value of t is not significant, i.e., there is no significant difference between $\overline{x}$ and t_0. In other words, the deviation ($\overline{x}$ - μ) is just due to fluctuations of sampling and null hypothesis H_0 may be retained at 5% level of significance, i.e., we may take the product conforming to specifications.

Ex: The mean weekly sales of soap bars in departmental stores as 146.3 bars per store. After advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful?

Solution. We are given: n= 22, $\overline{x}$ = 153.7, s =17.2.

Null Hypothesis. The advertising campaign is not successful, i.e., $H_0\, \mu$ = 146.3

Alternative Hypothesis, H_1: μ > 146.3 (Right-tail).

Test Statistic. Under H_0, the test statistic is : $t = \frac{\overline{x} - \mu_0}{\sqrt{S^2/n}}$ follows Student's t-dist.^{n} with (22-1) d.f

Or $t = \frac{153.7 - 146.3}{\sqrt{(17.2)^2/21}} = 9.03$

Conclusion. Tabulated value of t for 21 d.f at 5% level of significance for single tailed test is 1.72. Since calculated value is much greater than the tabulated value, it is highly significant. Hence we reject the null hypothesis and conclude that the advertising campaign was definitely successful in promoting sales.

Examples:A random sample of 10 boys had the following I.Q.'s : 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. 100 ? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.

Solution. Null hypothesis, H_0 : The data are consistent with the assumption of a mean I.Q. of 100 in the population, i.e., μ = 100.

Alternative hypothesis, H_1: $\mu \neq$ 100

Under H0, the test statistic is

$$t = \frac{\overline{x} - \mu_0}{\sqrt{S^2/n}} \sim t_{(n-1)}$$

Where $\overline{x}$ and s^2 are to be computed from the sample values of I.Q.'s.

CALCULATIONS FOR SAMPLE MEAN AND S.D

x	x-x	$\sum_{i=1}^{n}(x_i-\bar{x})^2$
70	-22.7	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
Total 972		1833.60

Here n=10,

$$\bar{x}=\frac{972}{10}=97.2\ and \qquad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1833.60}{9}=203.73$$

Or $$|\,t\,|=\frac{|97.2-100|}{\sqrt{203.73/10}}=\frac{2.8}{\sqrt{20.37}}=\frac{2.8}{4.514}=0.62$$

Tabulated $t_{0.05}$ for (10-1), i.e., 9 d.f. for two-tailed test is 2.262.

Conclusion. Since calculated t is less than tabulated $t_{0.05}$ for 9 d.f., H_0 may be accepted at 5% level of significance and we may conclude that the data are consistent with the assumption of mean I.Q. of 100 in the population.

The 95% confidence limits within which the mean I.Q. values of samples of 10 boys will lie are given by

$$\overline{x} \pm t_{0.05} S/\sqrt{n} = 97.2 \pm 2.262 \times 4.514 = 97.2 \pm 10.21 = 107.41 \text{ and } 86.99$$

Hence the required 95% confidence interval is [86.99, 107.41].

Ex: The heights of 10 males of a given locality are found to be 70, 67, 62, 68,61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches ? Test at 5% significance level assuming that for 9 degrees of freedom P (t> l.83) =0.O5.

Solution. Null Hypothesis, H_0: μ = 64 inches

Alternative Hypothesis, H_1: μ > 64 inches

X	70	67	62	68	61	68	70	64,	64	66	Total 660
$\left(X-\overline{X}\right)$	4	1	-4	2	-5	2	4	-2	-2	0	
$\left(X-\overline{X}\right)^2$	16	1	16	4	25	4	16	4	4	0	90

Here

$$\overline{x} = \frac{\Sigma X}{N} = \frac{660}{10} = 66 \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i-1}^{n}(x_i - \overline{x})^2 = \frac{90}{9} = 10$$

Under H_0, the test statistic is: $t = \dfrac{\overline{x}-\mu}{\sqrt{S^2/n}} = \dfrac{66-64}{\sqrt{10/10}} = 2$

which follows Student's t-distribution with 10-1 = 9 d.f. Tabulated value of t for 9 d.f at 5% level of significance for single (right) tail-test is 1.833. (This is the value $t_{0.10}$ for 9 d.f in the two-tailed tables)

Conclusion. Since calculated value of t is greater than the tabulated value, it is significant. Hence H_0 is rejected at 5% level of significance and we conclude that the average height is greater than 60 inches.

EXAMPLE: A random sample of 16 values front a normal population showed a mean of 41 •5 inches and the sum of squares of deviations front this mean equal to 135 square inches. Show that the assumption of a mean of 43.5 inches for the population is not reasonable. Obtain 95 per cent and 99 per cent fiducial limits for the same.

Solution;-We are given

n=16,

$\overline{x} = 41.5$ inches and $\sum_{i-1}^{n}(x_i - \overline{x})^2 = 135$ square inches

$$|t| = \frac{|41.5-43.5|}{3/4} = \frac{8}{3} == 2.667$$

Conclusion:- Since calculated $|t|$ is greater than 2.131, null hypothesis is rejected

99% fiducial limits for population mean μ is given $\overline{x} \pm t_{0.05}\dfrac{S}{\sqrt{n}}$

$$\overline{x} \pm t_{0.05} \frac{S}{\sqrt{n}}$$

$$= 41.5 \pm 2.131 \times \frac{3}{4} = 41.5 \pm 1.598 \qquad \Rightarrow \qquad 39.902 < \mu < 43.098$$

Similarly 99% fiducial limits for population mean is given by

$$\overline{x} \pm t_{0.01} S / \sqrt{n}$$

$$= 41.5 \pm 2.947 \times 3/4 = 43.71 \text{ and } 39.29 \qquad 39.29 < \mu < 43.098$$

13.7 Self assessment questions

1. A random sample of size 12 taken from a population of size 64 with s.d of 3 inches. Check the assumption that population mean height is 65 , also set up the probable limits for mean height of the population.

2. A random sample of size 200 is taken from a large number of coin. The mean weight of the sample of coins is 25.50 gms and s.d of 1.21 gms. Construct 95% CI for mean weight of coin in the population.

3. A random sample of 10 days shows an average daily sale of Rs.50 with a s.d of Rs.10 is taken from a large number of coin. The mean weight of the sample of coins is 25.50 gms and s.d of 1.21 gms. Construct 95% CI for mean weight of coin in the population.

4. (a) Under what circumstances can normal distribution be used to find confidence limit of population mean .

   (b) When we use to construct confidence interval estimate of population mean .

5. A random sample of size 10 taken from a normal population has mean 40 with s.d of 12 inches. Check the assumption that population mean height is 45 , also set up the 95% probable limits for mean height of the population.

   (Given $t_{0.05} = 3.25$ *for* $9 d.f$)

Unit 4 **Lesson 14**

SMALL SAMPLE TESTS

14.1 Introduction

14.2 Objectives

14.3 t-Test for Difference of Means

14.4 Paired t-test for Difference of Means

14.5 Examples based on t-Test for Difference of Means and Paired t-test

14.6 t-test for testing the significance of observed sample correlation coefficient

14.7 Examples based on t-test for testing the significance of observed sample correlation coefficient

14.8 Self Assessment Questions

14.1 INTRODUCTION

When the sample sizes are small, there are two technical changes in our procedure for testing the differences between means. The first involves the way we compute the estimated standard error of the difference between the two sample means. Here we base our small-sample tests on the t distribution, rather than the normal distribution.

In the present lesson we have demonstrated how to use samples from two populations to test hypotheses about how the populations are related , how hypothesis tests for differences between population means take different forms when samples are large or small. Further after the careful study of this lesson will enable the learner to distinguish between independent and dependent samples when comparing two means and to learn how to reduce a hypothesis test for the difference of means from dependent .

14.2 Objectives

1. To learn how to use samples from two populations to test hypotheses about how the populations are related
2. To learn how hypothesis tests for differences between population means take

   different forms; when the samples are of small size
3. To distinguish between independent and dependent samples when comparing two means
4. To learn how to reduce a hypothesis test for the difference of means from dependent samples to a test about a single mean
5. To understand how probability values can be used in testing hypotheses

14.3 t-TEST FOR DIFFERENCE OF MEANS.

Suppose we want to test if two independent samples x_i, (i = 1,2 …n_1) and Y_j' (j= 1, 2, ..., n_2) of sizes n_1 and n_2 have been drawn from two normal populations with means μ_x and μ_y respectively.

Under the null hypothesis (H_0) that the samples have been drawn from the normal populations with means μ_x and μ_y i.e., $H_o: \mu_x = \mu_y$ or $H_o: \mu_x - \mu_y = 0$ and under the assumption that the population variance are equal, i.e., $\sigma_x^2 = \sigma_y^2 = \sigma^2$ (say) but unknown, the statistic

$$t = \frac{(\bar{x}-\bar{y})-(\mu_x\mu_y)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{(n_1+n_2\text{-}2)} \text{ at } \alpha \text{ level of significance}$$

Under the null hypothesis H$_o$: $\mu_x = \mu_y$ the test statistic is given by

$$t = \frac{(\bar{x}-\bar{y}),}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

where $\bar{x} = \frac{1}{n}\sum_{i-1}^{n} x_i$

and $s^2 = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2 + \sum_{j=1}^{n_2}(y_j-\bar{y})^2\right]$

is an unbiased estimate of σ^2 based on both the samples

By comparing the calculated value of $|\,t\,|$ with its tabulated value for (n_1+n_2-2) degrees of freedom at α level of significance (usually 5% or 1%) we either reject or retain the null hypothesis

14.4 PAIRED T-TEST FOR DIFFERENCE OF MEANS

This test is applicable only when the two samples are not independent and the observations are taken into pairs. Let us now consider the case when

(i) the sample sizes are equal, i.e., $n_1 = n_2 = n$ (say), and

(ii) the two samples are not independent but the sample observations are paired together, i.e., the pair of observations (x_i, y_i), (i = 1, 2, ..., n) corresponds to the same (ith) sample unit.

The problem is to test if the sample means differ significantly or not.

For example, suppose we want to test the efficacy of a particular drug, say, for inducing sleep. Let x_i and y_i ,(i = 1, 2, ..., n) be the readings, in hours of sleep, on the ith individual, before and after the drug is given respectively. Here the observations are paired and dependent so we apply the paired t-test. Here we consider the increments, di= x_i- y_i , (i = 1, 2, ..., n)

Under the null hypothesis, H_0 that increments are due to fluctuations of sampling, i.e., the drug is not responsible for these increments, the statistic

$$t = \frac{\bar{d}}{\sqrt{S^2/n}}$$

where where $\bar{d} = \frac{1}{n}\sum_{i-1}^{n} d_i$ and $S^2 = \frac{1}{n-1}\sum_{i-1}^{n}(d_i - \bar{d})^2$ is an unbiased estimate of σ^2 follows Student's t-distribution with (n-1) d.f.

Depending upon whether t_{cal} is less than or greater than tabulated value of at (n-1) degrees of freedom at α level of significance we may accept or reject the null hypothesis.

14.5 EXAMPLES BASED ON T-TEST FOR DIFFERENCE OF MEANS AND PAIRED T-TEST

EXAMPLE. Below are given the gain in weights (in kgs.) of pigs fed on two diets A and B.

Gain in weight

Diet A : 25, 32, 30, 34, 24, 14, 32, 24, 30, 32, 35, 25

Diet B : 44, 34, 22, 10, 47, 32, 40, 30, 32, 35, 22, 35, 29, 22

Test, if the two diets differ significantly as regards their effect on increase in weight.

Solution. **Null hypothesis**, H_0 $\mu_x = \mu_y$ i.e., there is no significant difference between the mean increase in weight due to diets A and B.

Alternative hypothesis, H_1 $\mu_x \neq \mu_y$ (two-tailed).

Diet A			Diet B		
x	(x-$\overline{x}$)	$(x - \overline{x})^2$	Y	(y-$\overline{y}$)	$(y - \overline{y})^2$
25	-3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	-8	64
34	6	36	10	-20	400
24	-4	16	47	17	289
14	-14	196	31	1	1
32	4	16	40	10	100
24	-4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	-12	14
25	-3	9	21	-9	81
			35	5	25
			29	-1	1
			22	-8	64
$\Sigma x = 336$ $(x - \overline{x})^2 = 380$			$\Sigma y = 450$ $(y - \overline{y})^2 = 1410$		

Here n_1=12, n_2=15 n=16, $\bar{x}=\frac{336}{12}=28$ and $\bar{y}=\frac{450}{15}=30$

$$S^2=\frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2+\sum_{j=1}^{n_2}(y_j-\bar{y})^2\right]=71.6$$

Under null hypothesis (Ho): $t=\frac{(\bar{x}-\bar{y}),}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim t_{(n_1+n_2-2)}$

So that $t=\frac{28.30}{\sqrt{71.6\left(\frac{1}{12}+\frac{1}{15}\right)}}=\frac{-2}{\sqrt{10.74}}=-0.609$

Tabulated value of $t_{0.05}$ for (12+15-2)=25 degrees of freedom is 2.06

Conclusion. Since calculated $|t|$ is less than tabulated t, H_0 may be accepted at 5% level of significance and we may conclude that the two diets do not differ significantly as regards their effect on increase in weight.

EXAMPLE: Samples of to types of electric light bulbs were tested for length of life and fol1owing data were obtained:

	Type I	Type II
Sample No	$n_1=8$	$n_2=7$
Sample means	$\bar{x}_1=1234$	$\bar{x}_2=1036$
Sample S.D.'s	$s_1=36$	$s_2=40$

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life?

Solution. Null Hypothesis, H_0 $\mu_1 = \mu_2$ i.e., the two types I and II of electric bulbs are identical.

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

Test Statistic. Under H_0, the test statistic is

$$t = \frac{(\overline{x}_1 - \overline{x}_2),}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

Where $s^2 = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_{1i} - \overline{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \overline{x}_2)^2\right]$

$$= \frac{1}{n_1+n_2-2}\left[n_1 s_1^2 + n_2 s_2^2\right]$$

$$= \frac{1}{8+7-2}\left[8(36)^2 + 7(40)^2\right] = 1659.08 \text{ and}$$

$$t = \frac{1234 - 1036}{\sqrt{1659\left(\frac{1}{8} + \frac{1}{7}\right)}} = 9.39$$

Tabulated value of t for 13 df. at 5% level of significance for right (single)-tailed test is 1.77.

Conclusion. Since calculated It' is much greater than tabulated value of t', it is highly significant and H_0 is rejected. Hence the two types of electric bulbs differ significantly.

EXAMPLE: To test the claim that the resistance of electric wire can be reduced by at least 0.05 ohm by alloying, 25 values obtained for each alloyed wire and standard wire produced the following results

	Mean	Standard deviation
Alloyed wire	0.083 ohm	0.003 ohm
Standard wire	0.136 ohm	0.002 ohm

Test at 5% level whether or not the claim is substantiated.

Solution. Null Hypothesis, H$_0$ $\mu_x - \mu_y \geq 0.05$ i.e., the claim is sustained.

Alternative Hypothesis, H$_1$:$\mu_x - \mu_y < 0.05$, left tailed test.

Under H$_0$, the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1, n_2 - 2)} \quad \text{at } \alpha \text{ level of significance}$$

Where
$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{25(0.003)^2 + 25(0.002)^2}{25 + 25 - 2}$$
$$= \frac{0.000225 + 0.0001}{48} = 0.0000067$$

So that
$$t = \frac{(0.083 - 0.136) - 0.05}{\sqrt{0.0000067\left(\frac{1}{25} + \frac{1}{25}\right)}} = \frac{-0.103}{0.00071} = -145.07$$

The (critical) tabulated value of t for 48 d.f., at 5% level of significance for left tailed test is - 1•645.

Conclusion: Learner are advised to write the conclusion themselves.

Example: In a certain experiment to compare two types of animal foods A and B, the following results of increase in weights were observed in animals:

Animal number		1	2	3	4	5	6	7	8	Total
Increase in	Food A	49	53	51	52	47	50	52	53	407
weight	Food B	52	55	52	53	50	54	54	53	423

(i) Assuming that the two samples of animals are independent, can we conclude that food B is better than food A ?

(ii) Also examine the case when the same set of eight animals were used in both the foods.

Solution. Null Hypothesis, H_0 : If the increase in weights due to foods A and B are denoted by X and Y respectively, then $H_0: \mu_x = \mu_y$ i.e., there is no significant difference in increase in weights due to diets A and B.

Alternative Hypothesis, H_1: $\mu_x < \mu_y$ (Left-tailed).

(i) If the two samples of animals be assumed to be independent, then we will apply t-test for difference of means to test H_0. $\mu_x = \mu_y$ and test statistic is

$$t = \frac{(\overline{x} - \overline{y}),}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

where

$$s^2 = \frac{1}{n_1 + n_2 - 1}\left[\sum_{i=1}^{n_1}(x_i - \overline{x})^2 + \sum_{j=1}^{n_2}(y_j - \overline{y})^2\right] = 3.41$$

and $$t = \frac{50.875 - 52.875}{\sqrt{3.41\left(\frac{1}{8} + \frac{1}{8}\right)}} = -2.17$$

And t =Tabulated $t_{0.05}$ for (8 +8 -2) = 14 d.f for one-tail test is 1•76.

Conclusion. The critical region for the left-tail test is t <-1.76. Since calculated t is less than-—1.76, H_0 is rejected at 5% level of significance. Hence we conclude that the foods A and B differ significantly as regards their effect on increase in weight.

(ii) If the same set of animals is used in both the cases, then the readings X and Y are not independent but they are paired together and we apply the paired t-test for testing H_0.

$$t = \frac{\bar{d}}{\sqrt{S^2 / n}} \qquad \text{Where} \qquad \bar{d} = \frac{1}{n}\sum_{i-1}^{n} d_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i-1}^{n}(d_i - \bar{d})^2$$

X	49	53	51	52	47	50	52	53	Total
Y	52	55	52	53	50	54	54	53	
d=X-Y	-3	-2	-1	-1	-3	-4	-2	0	-16
d^2	9	4	1	1	9	16	4	0	44

$$\bar{d} = \frac{1}{n}\sum_{i-1}^{n} d_i = \frac{-16}{8} = -2 \qquad \text{and } S^2 = 1.714$$

$$|\, t \,| = \frac{|\,\bar{d}\,|}{\sqrt{S^2 / n}} = \frac{2}{\sqrt{1.7143/8}} = \frac{2}{04629} = 4.32$$

Tabulated $t_{0.s05}$ for (8- 1) = 7 d.f. for one-tail test is 1.90.

Conclusion. the observed value of 't' is significant at 5% level of significance and we conclude that food B is superior to food A.

14.6 T-TEST TOR TESTING THE SIGNIFICANCE OF AN OBSERVED SAMPLE CORRELATION COEFFICIENT

If r is the observed correlation coefficient in a sample of n pairs of observations from a bi-variate normal population, then Prof. Fisher proved that under the null hypothesis, H_0: ρ =0, i.e., population correlation coefficient is zero, the statistic

$$t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)}$$

follows Student's t-distribution with (n- 2) d.f.

If the value of t comes out to be significant, we reject H_0 at the level of significance adopted and conclude that $\rho \neq 0$, i.e., 'r' is significant of correlation in the population. If t comes out to be non-significant, then H_0 may be accepted and we conclude that variables may be regarded as uncorrelated in the population.

14.7 EXAMPLES BASED ON T-TEST FOR TESTING THE SIGNIFICANCE OF OBSERVED SAMPLE CORRELATION COEFFICIENT

EXAMPLE: (a) A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of 0.6. Is this significant of correlation in the population?

(b) Find the least value of r in a sample of 18 pairs of observations from a bi-variate normal population, significant at 5% level of significance.

Solution. (a) We set up the null hypothesis, H$_0$: ρ =0, i.e., the observed sample correlation coefficient is not significant of any correlation in the population, under null hypothesis our test statistic is

$t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)}$ follows Student's t-distribution with (n- 2) d.f.

$$t = \frac{0.6}{\sqrt{(1-0.36)}}\sqrt{(27-2)} = \frac{3}{\sqrt{0.64}} = 3.75$$

Tabulated $t_{0.05}$ for (27-2) = 25 d.f. is 2.06.

Conclusion. Since calculated t is much greater than the tabulated t, it is significant and hence H$_0$ is discredited at 5% level of significance. Thus we conclude that the variables are correlated in the population.

(b) Here n = 18. From the tables t0.05 for (18-2) = 16 d.f. is 2.12.Under null hypothesis, H$_0$: ρ =0

$t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)}$ follows Student's t-distribution with (n- 2) d.f.

In order that the calculated value of t is significant at 5% level of significance, we should have

$$\left|\frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)}\right| > t_{0.05} \qquad \Rightarrow \left|\frac{r}{\sqrt{(1-r^2)}}\sqrt{16}\right| > 2.12$$

$$\Rightarrow \qquad 16r^2 > (2.12)^2(1-r^2) \qquad \text{or} \qquad 20.493r^2 > 4.493$$

$$\text{or} \qquad r^2 > \frac{4.493}{20.493} = 0.2192$$

Hence $|\ r\ | > 0.4682$

EXAMPLE: A coefficient of correlation of 0.2 is derived from a random sample of 625 pairs of observations. (1) Is this value of r significant ? (ii) What are the 95% and 99% confidence limits to the correlation coefficient in the population ?

Sol: under null hypothesis, H_0: ρ =0, the test statistic is

$$t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)} = \frac{0.2}{\sqrt{(1-0.04)}}\sqrt{(625-2)} = 5.09$$

Since d.f 625- 2 = 623, the significant values of t are same as in the case of normal distribution, viz., $t_{0.05}$=1.96 and $t_{0.01}$ = 2.58. Since calculated t is much greater. Write result accordingly

95% confidence limits to the correlation coefficient of population are

$$r \pm 1.96\, S.E(r) = r \pm 1.96 \frac{(1-r^2)}{\sqrt{n}}$$

$$0.2 \pm 1.96 \frac{0.96}{\sqrt{625}} = 0.2 \pm 0.075 = (0.125, 0.275)$$

and 99% confidence limits to the correlation coefficient in the population are

$$0.2 \pm 2.58 \times \frac{0.96}{\sqrt{625}} = 0.2 \pm 0.099 = (0.101, 0.299)$$

14.8 SELF ASSESSMENT QUESTIONS

Question No:-1Two independent random samples of sizes 8 and 6 are drawn from normal population with unknown means , μ_1 and μ_2 and variances

32 and 30 respectively. If the sample means are 68.1 and 60.4 respectively, Teat for the equality of population means, find 95% confidence limits for the difference of population means.

Question No:-2 Two independent random sample of size 8 and 6 are drawn from two normal populations whose means and variances are unknown. If the samples have means 28.3 and 20.8, and standard deviations 6 and 5 respectively, find 95% confidence limits for the difference of population means. State the necessary assumption. (Value oft for 12 d.f. is $t_{0.25}$ = 2.18).

Question No:-3 Nine computer-components dealers in major metropolitan areas were asked for their prices on two similar color inkjet printers. The results of this survey are given below. At α = 0.05, is it reasonable to assert that, on average, the Apson printer is less expensive than the HP printer?

Dealer	1	2	3	4	5	6	7	8	9
Apson price	250	319	285	260	305	295	289	309	275
HP price	270	325	269	275	289	285	295	325	300

Question No:-4 Sherri Welch is a quality control engineer with the windshield wiper manufacturing division of Emsco, Inc. Emsco is currently considering two new synthetic rubbers for its wiper blades, and Sherri was charged with seeing whether blades made with the two new compounds wear equally well. She equipped 12 cars belonging to other Emsco employees with one blade made of each of the two compounds. On cars 1 to 6, the right blade was made of compound A and the left blade was made of compound B; on cars 7 to 12, compound A was used for the left blade. The cars were driven under normal operating conditions until the blades no longer did a satisfactory job of clearing the windshield of rain. The data below give the usable life (in days) of the blades.

At α = 0.05, do the two compounds wear equally well?

Car	1	2	3	4	5	6	7	8	9	10	11	12
Left blade	162	323	220	274	165	271	233	156	238	211	241	154
Right blade	183	347	247	269	189	257	224	178	263	199	263	148

Question No:-5 Ten soldiers visit a rifle range for two consecutive weeks. For the first week their scores are: 67, 24, 57, 55, 63, 54, 56, 68, 33, 43 and during the second week they score in the same order— 70, 38, 58, 58, 56, 67, 68, 72, 42, 38

Examine if there is any significant difference in their performance.

Unit 4 **Lesson 15**

SMALL SAMPLE TESTS

Structure:

15.1 Introduction

15.2 Objectives

15.3 F-test for Equality of Two Population Variances

15.4 Shape of F-distribution and tabulated values

15.5 Illustration

15.6 Examples

15.7 Self Assessment Questions

15.1 INTRODUCTION

F- Distribution is a very popular and useful distribution because of its utility in testing of hypothesis about the equality of several population means, two population variances and several regression coefficients in multiple regression etc. As a matter of fact, F-test is the backbone of analysis of variance.

This distribution was discovered by G.W.Snedecor and named in the honor of the Distinguish mathematical statistician Sir R.A Fisher. It may be recalled that the t statistic is used for testing whether two population means are equal. Whenever we are required to test for the case of more than two means, this can be tested by comparing the sample variances using F distribution by the use of analysis of variance technique which consist of “separation of variation due to a group of causes from the variation due to other groups". F ration is basically ratio of between column variance and within column variance, having found F ratio we can interpret it First, examine the denominator., which is based on the variance within the samples. The denominator is a good estimator of σ^2 (the population variance) whether the null hypothesis is true or not. What about the numerator? If the null hypothesis is true, then the numerator, or the variation among the sample means, is also a good estimate of σ^2 (the population variance). As a result, the denominator and numerator should be about equal if the null hypothesis is true. The nearer the F ratio comes to 1, then the more we are inclined to accept the null hypothesis Conversely, as the F ratio becomes larger, we will be more inclined to reject the null hypothesis and accept the alternative (that a difference does exist in the effects of the three training methods).

In short ,when populations are not the same, the between-column variance (which was derived from the variance among the sample means) tends to be larger than the within-column variance (which was derived from the variances within the samples), and the value of F tends to be large. This leads us to reject the null hypothesis.

Summing up, F- distribution is a very popular and useful distribution because of its utility in testing of hypothesis about the equality of several population means, two population variances and several regression coefficients in multiple regressions etc.

In fact this sampling distribution is widely used in different ways while testing different null hypotheses about a variety of population parameters.

15.2 Objectives

The objectives of this lesson is

- To introduce the F distribution and learn how to use them in statistical inferences
- To recognize situations requiring the comparison of more than two means or proportions
- To compare more than two population means using analysis of variance
- To use the F distribution to test hypotheses about equality of two population variances

15.3 F-test for Equality of Two Population Variances:

F- distribution is a very popular and useful distribution because of its utility in testing of hypothesis about the equality of several population means, two population variances and several regression coefficients in multiple regression. In the present section we will use this test statistic t for Equality of Two Population Variances

Suppose we want to Test

(i) whether two independent samples x_i, (i = 1, 2 n_1) and y_j, (I = 1, 2 n_2) have been drawn from the normal populations with the same variance σ^2 (say), or

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

Under the null hypothesis (H_0) that (i) $\sigma_x^2 = \sigma_y^2 = \sigma^2$,

i.e., the population variances are equal, or

(iii) Two independent estimates of the population variance are homogeneous, the statistic F is given by

$$F = \frac{S_x^2}{S_y^2}$$

Where $$S_x^2 = \frac{1}{n_1 - 1}\sum_{i-1}^{n_1}(x_i - \overline{x})^2$$

and $$S_y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_j - \overline{y})^2$$

are unbiased estimates of the common population variance σ^2 obtained from two independent samples and it follows Snedecor's F-distribution with (ν_1, ν_2) d,f. where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

By comparing the calculated value of F obtained by using above formula for the two given samples, with the tabulated value of F for (n_1,n_2) d.f. at certain leve1 of significance (5% or 1%), H_0 is either rejected or accepted.

Proof:-

$$F = \frac{S_x^2}{S_y^2}$$

$$= \left[\frac{n_1}{n_1 - 1}s_x^2\right] \Big/ \left[\frac{n_2}{n_2 - 1}s_y^2\right] = \left[\frac{n_1 s_x^2}{\sigma_x^2}.\frac{1}{n_1 - 1}\right] \Big/ \left[\frac{n_2 s_y^2}{\sigma_y^2}.\frac{1}{n_2 - 1}\right]$$

As under null hypothesis $\sigma_x^2 = \sigma_y^2 = \sigma^2$

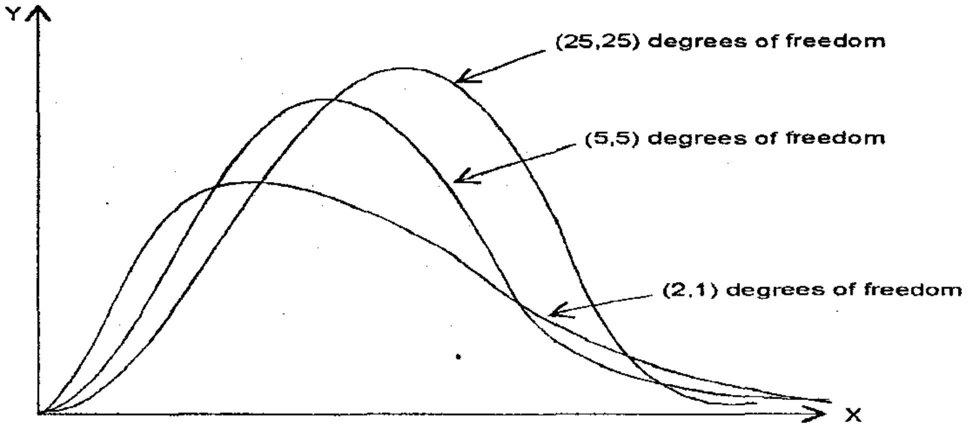
Since $\frac{n_1 s_x^2}{\sigma_x^2}$ and $\frac{n_2 s_y^2}{\sigma_y^2}$ are independent chi-square variates with (n_1- 1) and (n_2-1) d.f. respectively, follows Snedecor's F-distribution with (n_1-1, n_2-1) d.f.

As a norm larger of two variances is taken along numerator and the degrees of freedom corresponding to it is denoted by υ_1

If the calculated value of F is greater than its tabulated value for (n_1- 1, n_2- 1) degrees of freedom at α level, of significance we reject the null hypothesis otherwise we may retain it.

15.4 SHAPE OF F-DISTRIBUTION AND TABULATED VALUES

Similar to Chi Square distribution, F distribution is also a family of distributions. As the number of degrees of freedom varies, so is the shape of the distribution. For small numbers of degrees of freedom the curve is skewed extremely to the right and as the number of degrees of freedom increases the distribution tends to become symmetrical. The degree of skewness for some of the degrees of freedom is shown below.



In F distribution, we have two degrees of freedom as compared to one in Chi Square distribution. The number of degrees of freedom is expressed as (ν_1, ν_2). Where ν_1 referring to number of degrees of freedom for the numerator and ν_2 representing the same for the denominator. From the F tables, the value of the F statistic is obtained at the point of intersection of ν_1 and ν_2 at the corresponding level of significance. For the numerator we have to move column wise whereas for the denominator, we move row wise. For this distribution also the tables are given for the significance levels most often used.

15.5 ILLUSTRATIONS

Illustration:-A Quality Control Engineer at Zen Automobiles wants to check the variability in the number of defects in the cars coming from two assembly lines A and B. When he collected data it was as shown below.

Number of Defects

	Assembly_Line A	Assembly Line B
Mean	10	11
Variance	9	25
Sample Size	20	16

Can he conclude that the assembly line B has more variability than line A? Test the hypothesis at significance level of 5%.

Solution:-We set-up the hypothesis such that we do not have to test the hypothesis at the lower tail of the distribution. The null and the alternative hypothesis will be as follows:

$H_0 : \sigma_1^2 = \sigma_2^2$ (Null Hypothesis: The number of defects has same variability)

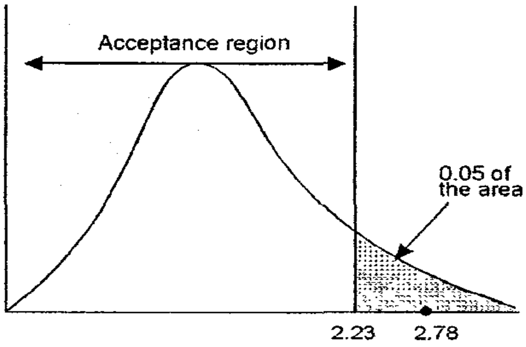
$H_0 : \sigma_1^2 > \sigma_2^2$ (Alternative Hypothesis: The number of defects from the assembly line B is more than that from line A)

Significance level = 5%.

Under null hypothesis We now calculate the F statistic

$$F = \frac{s_1^2}{s_2^2} = \frac{25}{9} = 2.78$$

From the tables, at a significance level of 0.05, 5 degrees of freedom in the numerator and 19 degrees of freedom in the denominator, the value of the F statistic is 2.23. This is represented in the figure below.



Since, the calculated value falls outside the acceptance region we reject the null hypothesis.

Illustration(two-tailed test): Two populations which are believed to have same variance were taken. However, on examination of the samples it was found that sample A (sample size 16) had a variance of 3.75 and for sample B (sample size 10) had a variance of 5.38. Formulate an appropriate hypothesis and test it at a significance level of 10% and state your conclusion.

Solution: We are given that

n_1 = 16 and s_1^2 =3.75

n_1 =10 and s_2^2 = 5.38

The hypothesis will be

$H_0 : \sigma_1^2 = \sigma_2^2$ (Null Hypothesis: Populations have the same variance)

$H_1 : \sigma_1^2 \neq \sigma_2^2$ (Alternative Hypothesis: Populations do not have the same variance)

Under null hypothesis our test statistic is

$$F = \frac{s_1^2}{s_2^2} = \frac{3.75}{5.38} = 0.697$$

The number of degrees of freedom in the numerator is 16 - 1 = 15 and in the denominator it is 10 - 1 = 9. Since we require both the limits, the limit F(l5, 9, 0.05) is directly obtained from the tables. Its value is 3.01. Now how do we get the value for the limit F(15, 9, 0.95), as at this level the values are not given in the tables. Here also we take the inverse of $\frac{s_1^2}{s_2^2}$, . The inverse of $\frac{s_1^2}{s_2^2}$ will be $\frac{s_2^2}{s_1^2}$.

We know that $F(n, d, \alpha) = \frac{1}{F(n, d, 1 - \alpha)}$

where,

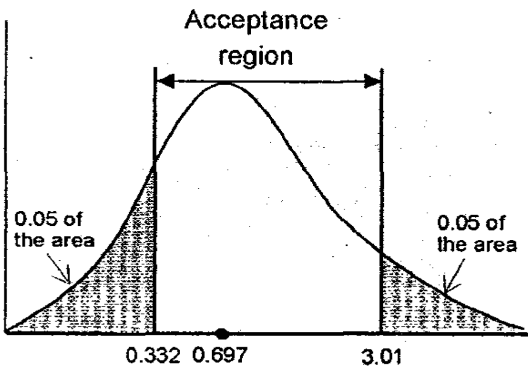
n is the degree of freedom in the numerator

d is the degree of freedom in the denominator

α is the significance level.

The value of $\frac{s_2^2}{s_1^2}$ will be 0.332.

Both these values give our limits as shown in the figure below.



Conclusion: Since the value of the calculated statistic falls in the acceptance region, we conclude that the samples belong to two populations which have the same variance.

14.6 EXAMPLES

EXAMPLE: Below given are the two random samples of sizes 12 and 15 respectively with values as given below

Sample-A 25, 32, 30, 34, 24, 14, 32, 24, 30, 31 35, 25,

Sample- B 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Check the hypothesis that the samples came from the same normal populations with identical variances

Solution Let us set the null hypothesis

$H_0 : \sigma_1^2 = \sigma_2^2$ (Null Hypothesis: Populations have the same variance)

$H_1 : \sigma_1^2 \neq \sigma_2^2$ (Alternative Hypothesis: Populations do not have the same variance)

CALCULAION TABLE

Sample-A			Sample- B		
x	(x- $\overline{x}$)	$(x-\overline{x})^2$	Y	(y- $\overline{y}$)	$(y-\overline{y})^2$
25	-3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	-8	64
34	6	36	10	-20	400
24	-4	16	47	17	289
14	-14	196	31	1	1
32	4	16	40	10	100
24	-4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	-12	14
25	-3	9	21	-9	81
			35	5	25
			29	-1	1
			22	-8	64
$\Sigma x = 336$		$(x-\overline{x})^2 = 380$	$\Sigma y = 450$ $(y-\overline{y})^2 = 1410$		

Under the null hypothesis our test statistic is

$$F = \frac{S_x^2}{S_y^2}$$

It follows Snedecor's F-distribution with $(\nu_1,\ \nu_2)$ d,f. where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

Where $S_x^2 = \frac{1}{n_1 - 1}\sum_{i-1}^{n_1}(x_i - \overline{x})^2$ and $S_y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_j - \overline{y})^2$

Here $S_x^2 = \frac{1}{n_1 - 1}\sum_{i-1}^{n_1}(x_i - \overline{x})^2 = \frac{1}{11}\times 380 = 34.54$

and $S_y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_j - \overline{y})^2 = \frac{1}{14}\times 1410 = 100.71$

$F = \frac{S_y^2}{S_x^2} = \frac{100.71}{34.54} = 2.9157$ follows Snedecor's F-distribution with $(\nu_2, \nu_1\ )$

The tabulated value of F(14,11) at 5% level of significance is 2.72 which is less that the calculated 2.9147 value so we reject the null hypothesis and conclude that Populations do not have the same variance

EXAMPLE: In one sample of 8 observations, the sum of the squares of deviations of the sample values from the sample mean was 84.4 and in the other sample of 10 observations it was 102.6. Test whether this difference is significant at 5 per cent level, given that the 5 per cent point of F for $n_1 = 7$ and $n_2 = 9$ degrees of freedom is 3.29.

Solution:-Here

$n_1 = 8$ $n_2 = 10$ and $\Sigma(y - \overline{y})^2 = 102.6$ and $\Sigma(x - \overline{x})^2 = 84.4$

$$S_x^2 = \frac{1}{n_1 - 1}\sum_{i-1}^{n_1}(x_i - \overline{x})^2 = \frac{84.4}{7} = 12.057$$

and $$S_y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_j - \overline{y})^2 = \frac{102.6}{9} = 11.4$$

Let us make the following assumption

$\sigma_x^2 = \sigma_y^2 = \sigma^2$ i.e., the estimates of σ^2 given by the samples are homogeneous.

Then the test statistic under null hypothesis is

$$F = \frac{S_x^2}{S_y^2} = \frac{12.057}{11.4} = 1.057$$

Tabulated $F_{0.05}$ for (7, 9) d.f is 3.29Since calculated F $<F_{0.05}$, H_0 may be accepted at 5% level of significance.

EXAMPLE:. Two random samples gave the following results

Sample No	1	2
Sample Size	10	12
Sample mean	15	14
Sum of squares of deviations from mean	90	108

Test whether the samples come from the same normal population at 5% level of significance.

$$\left[Given: F_{0.05}(9,11) = 2.90 \qquad F_{0.05}(11,9) = 3.10\, and \;\; t_{0.05}(20) = 2.086, \quad t_{0.05}(22) = 2.07,\right]$$

Solution. A normal population has two parameters, viz., mean μ. and variance σ^2. To test if two independent samples have been drawn from the same normal population, we have to test (i) the equality of population means, and (ii) the equality of population variances. $H_0 : \mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$

Null Hypothesis: The two samples have been drawn from the same normal population, Equality of means will be tested by applying t-test and equality of variances will be tested by applying F-test. Since t-test assumes $\sigma_1^2 = \sigma_2^2$, we shall first apply F-test and then t-test. In usual notations,

we are given:

$n_1 = 10 \;\; n_2 = 12 \qquad \bar{x}_1 = 15 \;\; \bar{x}_2 = 14 \; and \quad \Sigma(x_1 - \bar{x}_1)^2 = 90$
$\Sigma(x_2 - \bar{x}_2)^2 = 108$

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i-1}^{n_1}(x_1 - \bar{x}_1)^2 = \frac{90}{9} = 10$$

and $$S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(x_2 - \bar{x}_2)^2 = \frac{108}{11} = 9.82$$

$$F = \frac{S_1^2}{S_2^2} = \frac{10}{9.82} = 1.018$$ follows F distribution with $(n_1 - 1,$ $n_2 - 1)$ degrees of freedom

Tabulated $F_{0.05}$ (9,11) = 2.90. Since calculated F is less than tabulated F, it is not significant. Hence null hypothesis of equality of population variances may be accepted.

t-test: under the null hypothesis our test statistic is

$$t = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1, n_2 - 2)} = t_{20} \quad \text{at } \alpha \text{ level of significance}$$

Where $s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(x_1 - \overline{x}_1)^2 + \sum_{j=1}^{n_2}(x_2 - \overline{x}_2)^2\right]$

$$t = \frac{15 - 14}{\sqrt{9.9\left(\frac{1}{10} + \frac{1}{12}\right)}} = \frac{1}{\sqrt{9.9\left(\frac{11}{60}\right)}} = \frac{1}{\sqrt{1.815}} = 0.742$$

Tabulated $t_{0.05}$ for 20 d.f. = 2.086.Since the calculated value of t is less than its tabulated value se we accept the null hypothesis regarding the equality of population means

Since both the hypotheses are accepted, we may regard that the given samples have been drawn from the same normal population.

15.7 SELF ASSESSMENT QUESTIONS

QUESTION No 1: A random sample of 16 values from a normal population has a mean of 41.5 inches and sum of squares of deviations from the mean is equal to 135 inches. Another sample of 20 values from art unknown population has a mean of 43.0 inches and sum of squares of deviations from their mean is equal to 171 inches. Show that the two samples may be regarded as coming from the same normal population.

QUESTION No 2: The household net income from property and entrepreneurship in France

And Germany. follows:

China : 15.0, 8.0, 3.8, 6.4, 27.4, 19.0, 35.3, 13.6

Japan: 18.8, 23.1, 10.3, 8.0, 18.0, 10.2, 15.2, 19.0, 20.2

Test the equality of variances of household net income in China and Japan

QUESTION No 3 Following data give the distribution of women ever married by age.

Test whether the data have come from a normal population.

Age group	No. of women
15—19	3
19—23	43
23—27	62
27—31	38
31—35	24
35—39	14
39—43	11
43—47	5
47—51	2

QUESTION No 4 Two random samples taken from two normal populations are as follows. Estimate the variances of populations and test that the two populations have equal variances.

Sample I: 20, 16, 26, 27, 23, 22, 18, 24, 25, 19

Sample II:17, 23, 32, 25, 22, 24, 28, 18, 31, 33, 20, 27

QUESTION No 5 Given the following information about two samples from two normal populations, $n_1 = 10$, $s_1 = 1.97$, $n_2 = 8$ and $s_2 = 3.21$.

Can it be concluded that both the samples have come from populations having the same variability.

Unit 4 **Lesson 16**

SMALL SAMPLE TESTS

Structure:

16.1 Introduction

16.2 Objectives

16.3 Precautions about Using the Chi-Square Test

16.4 χ^2 test for Inferences about a Population Variance

16.5 Illustrations

16.6 Examples based on χ^2 test for Inferences about a Population Variance

16.7 χ^2 test for goodness of fit

16.8 Examples based on χ^2 test for Goodness of Fit Test.

16.9 χ^2 Test of Independence of Attributes

16.10 Examples based on χ^2 Test of Independence of Attributes

16.11 Self Assessment Questions

16.1 INTRODUCTION

We know how samples can be taken from populations and can use sample data to calculate statistics such as the mean and the standard deviation. If we apply what we have learned and take several samples from a population, the statistics we would compute for each sample need not be the same and most probably would vary from sample to sample.

Chi-square test is one of the most commonly used tests of significance. The chi-square test is applicable to test the hypotheses of the variance of a normal population, goodness of fit of the theoretical distribution to observed frequency distribution, in a one way classification having k-categories. It is also applied for the test of independence of attributes, when the frequencies are presented in a two-way classification called the contingency table. It is also a frequently used test in genetics, where one tests whether the observed frequencies in different crosses agree with the expected frequencies or not.

16.2 OBJECTIVES

Understanding of sampling distributions will enable the students to have basic knowledge about the behavior of sampling distributions so that samples that are both meaningful and cost effective can be taken, due to the fact that large samples are very expensive to gather, decision makers should always aim for the smallest sample that gives reliable results.

The knowledge of Chi-square test will acquaint the learners to test the hypotheses of the variance of a normal population, goodness of fit of the theoretical distribution to observed frequency distribution, in a one way classification having k-categories. It is also applied for the test of independence of attributes, when the frequencies are presented in a two-way classification called the contingency table. It is also a frequently used test in genetics, where one tests whether the observed frequencies in different crosses agree with the expected frequencies or not. In short the main objective of this lesson is to

- To introduce the Chi Square distribution and learn how to use them in statistical inferences
- To recognize situations requiring the use of Chi-square test
- To use Chi square test to check whether a particular collection of data is well described by a specified distribution
- To see whether two classifications of same data are independent of each other

- To use Chi square distribution for confidence intervals and testing hypotheses about a single population variance

16.3 PRECAUTIONS ABOUT USING THE CHI-SQUARE TEST

To use a chi-square hypothesis test, we must have a sample size large enough to guarantee the similarity between the theoretically correct distribution and our sampling distribution of the chi-square statistic. When the expected frequencies are too small, the value of χ^2 will be overestimated and will result in too many rejections of the null hypothesis. To avoid making incorrect inferences from χ^2 hypothesis tests, follow the general rule that an expected frequency of less than 5 in one cell of a contingency table is too small to use.When the table contains more than one cell with an expected frequency of less than 5, we can combine these in order to get an expected frequency of 5 or more. But in doing this, we reduce the number of categories of data and will gain less information from the contingency table.

This rule will enable us to use the chi-square hypothesis test properly, but unfortunately, each test can only reflect (and not improve) the quality of the data we feed into it. So far, we have rejected the null hypothesis if the difference between the observed and expected frequencies—that is, the computed chi-square value—is too large. In the case of the job-review preferences, we would reject the null hypothesis at a 0.10 level of significance if our chi-square value was 6.251 or more. But if the chi-square value was zero, we should be careful to question whether absolutely no difference exists between observed and expected frequencies. If we have strong feelings that some difference ought to exist, we should examine either the way the data were collected or the manner in which measurements were taken, or both, to be certain that existing differences were not obscured or missed in collecting sample data.

In the 1 860s, experiments with the characteristics of peas led the monk Gregor Mendel to propose the existence of genes. Mendel's experimental results were astoundingly close

16.4 χ^2 TEST FOR INFERENCES ABOUT A POPULATION VARIANCE

Suppose we want to test if a random sample $x_1, x_2 \ldots\ldots x_n$ has been drawn from a normal population with a specified variance $\sigma^2 = \sigma_0^2$ (say).

Under the null hypothesis that the population variance is $\sigma^2 = \sigma_0^2$, the statistic

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(x_i - \overline{x})^2}{\sigma_0^2}\right] = \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)}{n}\right] = \frac{ns^2}{\sigma_0^2}$$

follows chi-square distribution with (n -1) d.f.

By comparing the calculated value with the tabulated value of χ^2 for (n -1) d.f at certain level of significance (usually 5%), we may retain or reject the null hypothesis.

If the sample size n is large (>30), then we can use Fisher's approximation and apply Normal Test.

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

16.5 Illustration

A psychologist after a survey of children with age below 5 years old regarding the variability in their attention span finds that $\sigma = 8$ minutes. To convince herself that the attention span of six years old should be different from that of five years old, she conducts another survey of 20 children and finds that the sample variance as s^2=28 minutes. What would be null and the alternative hypothesis? At a significance level of $\alpha = 5\%$, what is the probable conclusion she would reach.

Solution:- We are given that n = 20 and s^2 =28.

We would set up the hypothesis as follows:

H_0: σ^2 = 64 (Null Hypothesis: the population variance is equal to 64)

H1: $\sigma^2 \neq$ 64 (Alternate Hypothesis: the population variance is not equal to 64)

Significance level =5%.

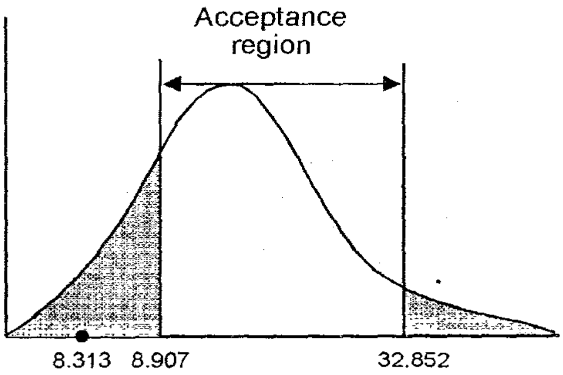
We observe that this is a two-tailed test and therefore we ought to look at both the limits.

The value of the χ^2 statistic is given by

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(20-1)(28)}{64} = 8.313$$

At 19 degrees of freedom and a significance level of 5%, the values of χ^2 where 0.025 of the area will lie at both the tails is 8.907 and 32.852 respectively. Since the calculated value of x2 does not fall in the acceptance region as shown in the figure below, we reject the null hypothesis.

Therefore the conclusion she would reach is that the attention span of six years old varies from that of the five years old.



16.6 EXAMPLES BASED ON χ^2 TEST FOR INFERENCES ABOUT A POPULATION VARIANCE

EXAMPLE: It is believed that the precision (as measured by the variance) of an instrument is no more than 0.16. Write down the null and alternative hypothesis for testing this belief Carry out the test at 1% level given ii measurements of the same subject on the instrument: 2.5, 2.3,24, 2.3, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, 2.5

Solution. Null Hypothesis, H_0: $\sigma^2 = 0.16$ Alternative Hypothesis, H_1: $\sigma^2 > 0.16$

Under the null hypothesis, H_0: $\sigma^2 = 0.16$, the test statistic is:

$$\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sigma_0^2} = \frac{0.1891}{0.16} = 1.182$$

which follows χ^2-distribution with d.f. n- 1= (11- 1) =10.

Since the calculated value of χ^2 is less than the tabulated value 23.2 of χ^2 for 10 df at 1% level of significance, it is not significant. Hence H_0 may be accepted and we conclude that the data are consistent with the hypothesis that the precision of the instrument is 0.16.

EXAMPLE: Test the hypothesis that $\sigma = 10$, given that s = 15 for a random sample of size 50 from a normal population.

Solution. Null Hypothesis, H_0 : $\sigma = 10$.

$$\chi^2 = \frac{ns^2}{\sigma_0^2} == \frac{50 \times 225}{100} = 112.5$$

If the sample size n is large (>30), then we can use Fisher's approximation and apply Normal Test.

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

$$\therefore \qquad Z = \sqrt{225} - (\sqrt{99}) = 15 - 9.95 = 5.05 \sim N(0,1)$$

Since $|Z| > 3$, it is significant at all levels of significance and hence H_0 is rejected and we conclude that $\sigma \neq 10$

EXAMPLE : An owner of a company agrees to purchase the product of a factory, if the produced items do not have variance of more than 0.5 mm^2 in their length. To make sure of the specifications, the buyer selects a sample of 18 items from his lot. The length of each item was measured to be as follows:

Length (mm)
18.57, 18.10, 18.61, 18.32, 18.33, 18.46, 18.12, 18.34, 18.57,
18.22, 18.63, 18.43, 18.37, 18.64, 18.58, 18.34, 18.43, 18.63

Solution:-On the basis of the sample data, the hypothesis

Null Hypothesis, $H_0: \sigma^2 = 0.5$ Alternative Hypothesis, $H_1: \sigma^2 > 0.5$

Under the null hypothesis, $H_0: \sigma^2 = 0.5$, the test statistic is:

$$\chi^2 = \frac{\sum_{i=1}^{18} (x_i - \overline{x})^2}{\sigma_0^2}$$

We calculate $\sum_{i=1} (x_i - \overline{x})^2 = \sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}$

For the given data, $\sum_i x_i^2 = 6112.64$ $\sum_i x_i = 331.69$

So that

$$\sum_{i=1} (x_i - \overline{x})^2 = \sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n} = 6112.64 — \frac{(33169)^2}{18}$$

$$= 6112.640 — 6112.125 = 0.515$$

Thus

$$\chi^2 = \frac{\sum_{i=1}^{18} (x_i - \overline{x})^2}{\sigma_0^2} = \frac{0.515}{0.5} = 1.03$$

For $\alpha = 0.05$, χ^2 for 17 degrees of freedom is 27.587. Since the calculated value of χ^2 is 1.03 which is not greater than 27.587, we accept the null hypothesis, $: \sigma^2 = 0.5$ at $\alpha = .05$. and we conclude that the buyer should purchase the lot.

Example : Future Technologies Ltd. manufactures high resolution telescopes. The management wants its products to have a variation of less than 2 standard deviations in resolution, while focusing on objects which are beyond 500 light years. When they tested their newly manufactured telescope for 30 times, to focus on an object 500 light years away, they found that the sample standard deviation to be 1.46. State the hypothesis and test it at a significance level of 1%. Can the management accept to sell this product?

Solutio: We are given, n=30 and $s^2 = (1.46)^2$

We set up the hypothesis as follows:

Null Hypothesis, $H_0: \sigma^2 = 4$ (Null hypothesis: Population variance is equal to 4)

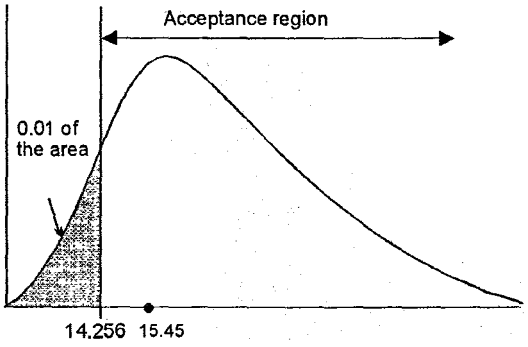
Alternative Hypothesis, $H_1: \sigma^2 < 4$ (Alternative Hypothesis: Population variance is less than four)

Level of significance $\alpha = 1\%$ We observe that this is a one-tailed test

Under the null hypothesis, $H_0: \sigma^2 = 4$, the test statistic is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(30-1)(1.46)^2}{(2)^2} = 15.45$$

Now referring to the tables, we find that at 29 degrees of freedom, the value of χ^2 that leaves an area of 0.01 in the lower tail of the curve is 14.256 (since we are testing at the lower end, this value is got at 1 -0.01 = 0.99 of the area under the right tail). This is shown in the figure as given below



Conclusion: Since the calculated value of chi square falls in the acceptance region, we accept the null hypothesis and conclude that standard deviation is equal to 2. Therefore the management will not aloe the sale of its telephone.

16.7 χ^2 test for Goodness of Fit Test.

This test is used for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson and is known as "Chi-square test of goodness of fit". It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

If f_i (i =1, 2, ..., n) is a set of observed (experimental) frequencies and e_i (i = 1, 2,n) is the corresponding set of expected (theoretical or hypothetical) frequencies, then Karl Pearson's chi-square, given by

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

follows chi-square distribution with (n - 1) d.f.

How to decide :Accept H_0 if $\chi^2 \le \chi^2_\alpha$ (n- 1) and reject H_0 if $\chi^2 > \chi^2_\alpha$ (n - 1), where χ^2 is the calculated value of chi-square and χ^2_α (n-1) is the tabulated value of chi-square for (n-1) d.f. and level of significance α .

16.8 EXAMPLES BASED ON χ^2 TEST FOR GOODNESS OF FIT TEST.

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

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

Test the hypothesis that the number of parts demanded does not depend on the day of the week. (Given : the values of χ^2 significance at 5, 6, 7, d.f. are respectively 11.07,12.59, 14.07 at the 5% level of significance.)

Solution. Here we set up the null hypothesis, H_0 that the number of parts demanded does not depend on the day of week.

Under the null hypothesis, the expected frequencies of the spare part demanded on each of the six days would be:1/6(1124+ 1125 + 1110 + 1120 + 1126 + 1115)=6720/6=1120

And our test static is

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(f_i - e_i)^2}{e_i} \right], \quad \left(\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} e_i \right)$$

CALCULATIONS FOR χ^2

Days	Observed frequency(fi)	Expected frequency(ei)	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
Monday	1124	1120	16	0.014
Tuesday	1125	1120	25	0.022
Wednesday	1110	1120	100	0.089
Thursday	1120	1120	0	0
Friday	1126	1120	36	0.032
Saturday	1115	1120	25	0.022
Total	6720			0.179

So $\chi^2 = \sum_{i=1}^{n}\left[\frac{(f_i - e_i)^2}{e_i}\right] = 0.179$

The tabulated χ^2 0.05 for 5 d.f. = 11.07.

Conclusion: Since calculated value of χ^2 is less than the tabulated value, it is not significant and the null hypothesis may accepted at 5% level of significance. Hence we conclude that the numbers of parts demanded are same over the 6-day period.

EXAMPLE: The following figures show the distribution of digits in numbers chosen at random from a telephone directory:

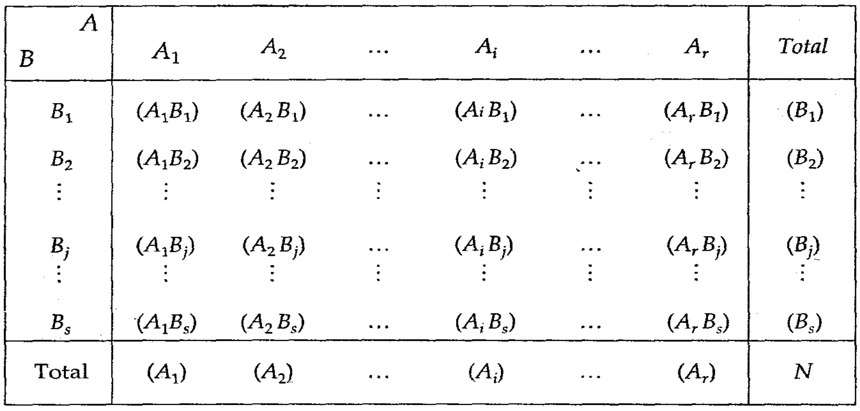
Digits :	0	1	2	3	4	5	6	7	8	9
Freq :	1026	1107	997	966	1075	933	1107	972	964	853
Total	10,000									

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

16.9 χ^2 TEST OF INDEPENDENCE OF ATTRIBUTES

Let us consider two attributes A divided into r classes A_1, A_2, ..., A_r and B divided into s classes B_1, B_2, ..., B_s. Such a classification in which attributes are divided into more than two classes is known as manifold classification. The various cell frequencies can be expressed in the following table known as r x s manifold contingency table where (A_i) is the number of persons possessing the attribute A, (i = 1, 2, ..., r), (B_j) is the number of persons possessing the attribute

B_j (j = 1, 2, ..., s) and (A_iB_j) is the number of persons possessing both the attributes A_i and B_j, (i = 1, 2, ..., r;j = 1, 2, ..., s).



A B	A_1	A_2	$\cdots$	A_i	$\cdots$	A_r	*Total*
B_1	(A_1B_1)	(A_2B_1)	$\cdots$	(A_iB_1)	$\cdots$	(A_rB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	$\cdots$	(A_iB_2)	$\cdots$	(A_rB_2)	(B_2)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_j	(A_1B_j)	(A_2B_j)	$\cdots$	(A_iB_j)	$\cdots$	(A_rB_j)	(B_j)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_s	(A_1B_s)	(A_2B_s)	$\cdots$	(A_iB_s)	$\cdots$	(A_rB_s)	(B_s)
Total	(A_1)	(A_2)	$\cdots$	(A_i)	$\cdots$	(A_r)	N

Here the problem is to test if the two attributes A and B under consideration independent or not.

Under the null hypothesis that the attributes are independent, the theoretical frequencies are calculated by using

$$e_{ij} = \frac{ith\ row\ total \times jth\ column\ \ total}{sample\ size}$$

the test statistic in this case is given by

$$\chi^2 = \sum_I \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

Where e_{ij} is the expected frequency in column i and row j

f_{ij} = observed frequency for contingency table category in column i and row j

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

16.10 EXAMPLES BASED ON χ^2 TEST OF INDEPENDENCE OF ATTRIBUTES

EXAMPLE: Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the adjoining table. Examine whether the nature of the area is related to voting preference in this election.

Area	Vote for A	Vote for B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

Sol: Under the null hypothesis that the nature of the area is independent of the voting preference in the election, we get the expected frequencies as follows

$$\text{E}(620) = \frac{1170\times1000}{2000,} = 585 \quad \text{and} \quad \text{E}(380) = \frac{830\times1000}{2000} = 415$$

$$\text{E}(550) = \frac{1170\times1000}{2000,} = 585 \quad \text{and} \quad \text{E}(380) = \frac{830\times1000}{2000} = 415$$

$$\chi^2 = \sum_i \left[\frac{(f_i - e_i)^2}{e_i}\right],$$

$$= \frac{(620-585)^2}{585} + \frac{(380-415)^2}{415} + \frac{(550-585)^2}{585} + \frac{(450-415)^2}{415} = 10.0891$$

Tabulated χ^2 0.05 for (2-1) (2-1) =1 d.f. is 3.841. Since calculated χ^2 is much greater than the tabulated value, it is highly significant and null hypothesis is rejected at 5% level of significance. Thus we conclude that nature of area is related to voting preference in the election.

EXAMPLE: (2×2 CONTINGENCY TABLE). For the 2 x2 table,

a	b
c	d

Prove that chi-square test of independence gives

$$\chi^2 = \left(\frac{N[ad-bc]^2}{(a+b)(a+c)(b+d)(c+d)} \right)$$ Where

N=a+b+c+d

Solution. Under the hypothesis of independence of attributes,

$$E(a) = \frac{(a+b)(a+c)}{N}, \qquad E(b) = \frac{(a+b)(b+d)}{N}$$

$$E(c) = \frac{(a+c)(c+d)}{N} \qquad E(d) = \frac{(b+d)(c+d)}{N}$$

a	b	a+b
c	d	C+d
a+c	b+d	N

$$\chi^2 = \frac{[a-E(a)]^2}{E(a)} + \frac{[b-E(b)]^2}{E(b)} + \frac{[c-E(c)]^2}{E(c)} + \frac{[d-E(d)]^2}{E(d)}$$

………..(1)

Now

$$a-E(a)=a-\frac{(a+b)(a+c)}{N}=\frac{a(a+b+c+d)-(a^2+ac+ab+bc)}{N}$$

$$=\frac{ad-bc}{N}$$

Similarly we get

$$\mathrm{b-E(b)=-\frac{ad-bc}{N}=c-E(c);\qquad d-E(d)=\frac{ad-bc}{N}}$$

Substituting in (1), we get

$$\chi^2=\frac{[\mathrm{ad-bc}]^2}{\mathrm{N}^2}\left[\frac{1}{\mathrm{E(a)}}+\frac{1}{\mathrm{E(b)}}+\frac{1}{\mathrm{E(c)}}+\frac{1}{\mathrm{E(d)}}\right]$$

$$=\frac{[ad-bc]^2}{N}\left[\left(\frac{1}{(a+b)(a+c)}+\frac{1}{(a+b)(b+d)}\right)+\left(\frac{1}{(a+c)(c+d)}+\frac{1}{(b+d)(c+d)}\right)\right]$$

$$=\frac{[\mathrm{ad-bc}]^2}{\mathrm{N}}\left[\left(\frac{\mathrm{a+b+c+d}}{\mathrm{(a+b)(a+c)(b+d)}}\right)+\left(\frac{\mathrm{b+d+a+c}}{\mathrm{(a+c)(c+d)(b+d)}}\right)\right]$$

$$=\frac{[\mathrm{ad-bc}]^2}{\mathrm{N}}\left[\left(\frac{\mathrm{a+b+c+d}}{\mathrm{(a+b)(a+c)(b+d)}}\right)+\left(\frac{\mathrm{b+d+a+c}}{\mathrm{(a+c)(c+d)(b+d)}}\right)\right]$$

$$=[\mathrm{ad-bc}]^2\left[\left(\frac{\mathrm{c+d+a+b}}{\mathrm{(a+b)(a+c)(b+d)}}\right)+\left(\frac{\mathrm{b+d+a+c}}{\mathrm{(a+c)(c+d)(b+d)}}\right)\right]$$

$$=\frac{[\mathrm{ad-bc}]^2}{\mathrm{N}}\left[\left(\frac{\mathrm{a+b+c+d}}{\mathrm{(a+b)(a+c)(b+d)(c+d)}}\right)\right]$$

$$=\left(\frac{\mathrm{N[ad-bc]^2}}{\mathrm{(a+b)(a+c)(b+d)(c+d)}}\right)$$

Hence Proved

16.11 Self Assessment Questions

Question No 1 what are the types of observational data suitable for the chi-square test, in a contingency table?

Question No 2 **What** do you understand by the test of goodness of fit?

Question No 3 Discuss a contingency table.

Question No 4 Following table gives the data regarding the field of study in the university and their field of specialization in High School.

Specialization in High School	Field of study in the University		
Biology	Biology	Medicine	Agriculture
Physics and Maths	26	52	23
Agriculture	3	44	8
Humanities	4	1	15
	6	11	10

Check whether is there any dependency of the field of study in the university on their field of specialization in High School.

Question No 5 Following table gives the number of births according to their sex and condition at the time of birth.

Sex	Condition -	
	Normal	Abnormal
Male	19	5

Female	30	6

Test at $\alpha = 0.05$, whether the condition at birth depends on the sex of the child.

Question No 6 In a departmental examination, the candidates of both the sexes yielded results as presented here in the (2 x 2) table.

Sex	Pass	Fail
Male	42	2
Female	14	6

Can it be inferred that he result of the test is related to the sex of the candidates. Perform a suitable statistical test to arrive at the correct decision, using a 5 per cent level.8

Question No 7 Describe the use of the χ^2 test in testing of independence of attributes in a (2 X 2) contingency table

Question No 8 A private coaching school claims that 60% of the students, coached in the school, will be selected in a competition, 55 candidates sought admission in the school and only 24 candidates got selected.

Do the result of the candidates justify the claim of the school authorities at 1 per cent level of significance.

Question No 9 The following table reveals the condition of the house and the condition of the children. Using the chi-square test, find out whether the condition of house affects the condition of children

Condition of children	Condition of house		Total
	Clean	Not clean	
Very clean	76	43	119
Clean	47	17	55
Dirty	38	25	72
Total	139	107	246

Question No:-10 What are the kinds of hypotheses that can be tested by the chi-square test ?

Question No:-11 Given below are the number of accidents of airplanes that occurred on different days of a week. Find out whether the airplane accidents are uniformly distributed over the seven days of the week.

Days	Sun	Mon	Tues	Wed	Thurs	Fri	Sat	Total
No. of Accidents	16	18	10	14	13	11	16	98

Question No:-12 Answer the following in not more than three lines.

(a) Expected frequencies are obtained under which hypothesis?

(b) Why can the chi-square not be negative?

(c) Why can the value of F-statistic not be negative?

Unit 5 **Lesson 17**

NON-PARAMETRIC TESTS

Structure:

17.1 Introduction

17.2 Objectives

17.3 Concept of Non Parametric Tests

17.4 Non Parametric Methods V/S Parametric Metohds

17.5 Non Parametric Tests for Univariate Distributions

17.5.1 Test for Randomness (Run Test)

17.5.2 Examples based on Test for Randomness (Run Test)

17.5.3 The Sign Test

17.5.4 Examples based on Sign Test

17.5.5 The Wilcoxon Test

17.6 Self Assessment Questions

17.1 Introduction

In most of the Statistical tests which we have so for studied we have some of the features which are to be comply with if we have to apply these statistical test correctly, for example we make the assumption of normality of parent population from which we draw the random Samples or we may apply limit theorem for sufficiently large Samples to relax the assumption of normality etc.

A second assumption upon which most of the statistical tests rest is that meaningful sample statistic, such as mean standard deviation, can be derived from the samples and used to estimate the corresponding population parameters. But the data which is nominal in nature or ordinal do not yield the good results.

For such instances statistician have devised alternative procedures which can be used to test the data which are nominal or ordinal in nature or for which meaningful statistics cannot be calculated. A most important feature of these alternative procedures is that they do not depend upon the shape of frequency distribution. Since they do not depend upon the shape of frequency distribution they are termed as distribution free tests. Such tests do not depend upon the population parameters such as mean and variance, they are also called as non parametric tests.

17.2 OBJECTIVES

The main objectives of this lesson are.

1. To offer a different approach to many of the decision problems
2. To understand the basic concept of non parametric tests.
3. To make a comparison between parametric and nonparametric tests
4. To know how to apply these tests to univariate data in a variety of problems.

17.3 CONCEPT OF NON PARAMETRIC TESTS

In most of the statistical tests which we have so far studied we have two features common

(i) The form of the frequency function of the parent population from which the samples have been drawn is assumed to be known, and

(ii) They were concerned with testing statistical hypothesis about the parameters of this frequency function or estimating its parameters.

For example, almost all the exact (small) sample tests of significance are based on the fundamental assumption that the parent population is normal and are concerned with testing or estimating the means and variances of these populations. Such tests, which deal with the parameters of the population, are known as **Parametric Tests**.

Thus, a parametric statistical test is a test whose model specifies certain conditions about the parameters of the population from which the samples are drawn.

On the other hand, a Non-parametric (NP.) Test is a test that does not depend on the particular form of the basic frequency function from which the samples are drawn. *In other words, non-parametric test does not make any assumption regarding the form of the population.*

In short, most of the statistical tests which we have so for studied we have some of the features which are to be comply with if we have to apply these Statistical test correctly, for example we make the assumption of normality of parent population form which we draw the random samples or We may apply central limit theorem for sufficiently large Samples to relax the normality assumption etc.

Second assumption upon which most of the statistical tests rest is that meaningful sample statistic, such as mean, standard deviation, But the data Which is nominal in nature or Ordinal do not yield the good results.

For such instances statistician have devised alternative procedures which can be used to test the data which are nominal or ordinal in nature or for which meaningful statistics cannot be calculated is distribution free tests. Such tests do not depend upon the population parameters such as mean and variance they are also called as non parametric tests.

However, certain **assumptions** associated with N.P. tests are:

(i) Sample observations are independent.

(ii) The variable under study is continuous.

(iii) p.d.f. is continuous. This is postulated to determine the sampling distributions

(iv) Lower order moments exist.

Median is as good an index of central tendency as mean. We know, for symmetrical distributions, mean and median coincide. Hence, in nonparametric statistics median is taken as a measure of location parameter instead of mean.

Obviously these assumptions are fewer and much weaker than those associated with parametric tests.

In the above mentioned questions/problems, the question first is known as the problem of fit. The question second deals with the testing of randomness of the sample and third question deals with the testing of hypothesis whether a particular sample has been drawn from a specified population or not.

17.4 ADVANTAGES AND DRAWBACKS OF NON-PARAMETRIC METHODS OVER PARAMETRIC METHODS.

Advantages

1. N.P. methods are readily comprehensible, very simple and easy to apply and do not require complicated sample theory.
2. No assumption is made about the form of the frequency function of the parent population from which sampling is done.

3. No parametric technique will apply to the data which are mere classification (i.e., which are measured in nominal scale), while N.P. methods exist to deal with such data.
4. Since the socio-economic data are not, in general, normally distributed, N.P. tests have found applications in Psychometry, Sociology and Educational Statistics.

Drawbacks

1. NP tests can be used only if the measurements are nominal or ordinal. Even in that case, if a parametric test exists it is more powerful than the NP test. In other words, if all the assumptions of a statistical model are satisfied by the data and if the measurements are of required strength, then the NP. tests are wasteful of time and data.
2. So far, no NP methods exist for testing interactions in 'Analysis of Variance' model unless special assumptions about the additivity of the model are made.
3. N.P. tests are designed to test statistical hypothesis only and not for estimating the parameters

17.5 NON-PARAMETRICTESTS FOR UNIVARIATE DISTRIBUTION

The one sample tests are generally, used to answer the questions such as:

(i) Is there a significance difference between the observed and expected frequencies?

(ii) Is it reasonable to accept that the sample is a random Sample from Some known population?

(ii) Is it reasonable to believe that the sample has been drawn from a Specified population?

In the above mentioned questions/problems, the question first is known as the problem of fit which has been in Chi-square that of Goodness of fit and test of homogeneity.

The question second deals with the testing of randomness of the sample and third question deals with the testing of hypothesis whether a particular sample has been drawn from a specified population or not.

Here a random sample of size n is drawn from a population and the sample values are arranged in order of magnitude and ranked accordingly, if need be. Various tests lead us to decide whether the sample has come from a particular population. Also, we test whether the median of the population is equal to a known value or not. Such tests are classified as tests for goodness of fit like chi-square test.

17.5.1 TEST FOR RANDOMNESS

ONE SAMPLE RUN TEST FOR RANDOMNESS

One application of run test is in testing the randomness of a given set of the observations. The run test for randomness tests the **null hypothesis that a sequence of events has occurred randomly, against the alternative hypothesis that this sequence of events has not occurred randomly.**

To test this hypothesis we have the following procedure

Procedure*:* Let x_1 , x_2, ..x_n, be the set of sample observations arranged in the order in which they occur. Then for each of observations we write A if the observation x’s is above the median (Mo) and B if it is below the median, Then we determine total number of runs, where **a run is defined as a sequence of letters of one kind surrounded by sequence of letters of other kind**.

Let R$_1$ and R$_2$ are the number of runs of type I and type II respectively, thus **R=R$_1$+R$_2$** be the total number of runs in the sample. It is our test statistic. We compare this value of R with the critical value of R for given n$_1$, n$_2$ and level of significance α .If n is greater than 25, then to test the above hypothesis we use normal approximation as

$$Z = \frac{R - E[R]}{S.D(R)} \sim N(0,1)$$

Where $$E[R] = \frac{2n_1 n_2}{n_1 + n_2} + 1$$ and

$$S.D[R] = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

where, n$_1$,and n$_2$ are the number of observations of type I and type II respectively.

17.5.2 EXAMPLE BASED ON RUN TEST FOR RANDOMNESS

Example: Test the randomness of the 15 observations in the order obtained

5.42, 5.28, 5.43, 5.54, 5.44, 5.31, 5.32, 5.34, 5.46, 5.33, 5.32, 5.31, 5.47, 5.48, 5.20

Solution: First of all we arrange the given (above) observations in ascending order and then obtain median as

$$Sample\ median = \left(\frac{15+1}{2}\right)^{th} term = 8^{th}\ term = 5.34$$

Writing A and B under each observation according it as above or below the median (discard tied observations) Therefore, we. obtained the following sequence of A and B.

ABAAABBABBBAAB

thus $n_1=7$,$n_2=7$,$R_1=4$, $R_2=4$ and $R=R_1+R_2=8$ and $n=7+7=14$

To test Ho **: The sample is random**.

To test this hypothesis, we will use run test. Since both the n_1 and n_2 are less than 25, the critical value of R at 5 percent level of significance is 3 and 13(from table) Thus we do not reject the null hypothesis of randomness.

EXAMPLE: Following is a sequence of heads (H) and tails (T) in tossing of a coin 14 times.

HTTHHHTHTTHHTH

Test whether the heads and tails occur in random order. [Given: $\alpha = 0.05$, $r_L= 2$, $r_u = 12$]

Ans. For the given sequence, The sample size, n = 14, No. of heads,$n_1= 8$

No. of tails,$n_2= 6$ No. of runs of H,$r_1 =5$ No. of runs of T, $r_2= 4$

so that $r=5+4=9$

Since the observed value of r = 9 lies between the critical values 3 and 12, we accept H_0. It means that the heads and tails occur in random order or it can be said that the coin is unbiased.

17.5.3 THE SIGN TEST

The sign test is the simplest of the non-parametric tests. Its name comes from the fact that it is based on the direction (i.e. signs of + and -) of observations and not on their numerical magnitude. The sign test is applied to make the hypothesis test about preferences, single median and the median of paired differences for two dependent populations.

Let us use this test to test the hypothesis that the sample has been drawn from a population with c.d.f. $f_0(x)$ or with median M_0. That is

$$H_0 : F(x) = F_0(x) \qquad against \qquad H_1 : F(x) \neq F_0(x)$$

or equivalently

$$H_0 : M = M_0 \qquad against \qquad H_1 : M \neq M_0$$

$$H_0 : p = \frac{1}{2} \qquad against \qquad H_1 : p \neq \frac{1}{2}$$

Where p is the probability that the number of observations less than the median.

To test this hypothesis the following test procedure will be followed.

Procedure : Let x_1 , x_2, ..x_n, be a random sample drawn from a continuous population with median M. Let M_0 be the value of median under Ho. Further let D_i= X_i-M_0, i = 1, 2, 3, .n and put plus sign (+) or negative sign (-) according as Di >0 or Di<O and discard those Di's for which D_i, =0.

Let S denotes the number of plus signs, then S follows a binomial distribution with parameters n and p. where p =1/2 and n is the sum of plus and minus signs. To test the above Ho, we find the probability

$$P[S \geq s / H_0] = \sum_{X=s}^{n} \binom{n}{s} \left(\frac{1}{2}\right)^{s} \left(\frac{1}{2}\right)^{n-s}$$

If this probability is greater than the given level of significance (α), we accept Ho, otherwise reject Ho. Further if n is large, then to test Ho we use normal approximation as

$$Z = \frac{S - np}{\sqrt{npq}} \sim N(0,1)$$

17.5.4 EXAMPLE BASED ON THE SIGN TEST

Example : The average income (as measured be median) of women employees in a firm is Rs 3500 per month. A sample of 13 men chosen from the men employee in that firm. On the basis of their incomes given below, is there evidence that the average income of men exceeds that of women? Income in thousand of Rupees

4.0, *3.5, 4.6, 4.4,* 3.7, 3.4, 3.9, 4.1, 4.1, *9.9,* 3.6, 3.3 and 4.2

Sot: Let M is the median of the distribution of X, where X is the income of men in thousand

Then we have to test the hypothesis.

Ho: $M=M_0=3.5$ against H_1: $M>3.5$

To test H_o, let $D = X_i - Mo = X_i - 3.5$ for i = 1 ,2 3 …n and then put plus signs or minus signs according as D_i, >0 or D_i <0 as given below in the table

X	D = X_i –Mo	- signs	+ signs
4.0	+0.5		+
3.5	0		
4.6	1.1		+
4.4	0.9		+

3.7	0.2		+
3.4	-0.1	-	
3.9	0.4		+
4.1	0.6		+
4.1	0.6		+
9.9	6.4		+
3.6	0.1		+
3.3	-0.2	-	
4.2	0.7		+

Here S number of plus signs = 10

Thus $\quad S \sim B\left(12, \frac{1}{2}\right)$

To test Ho, we find the probability

$$P[S \geq s / H_0] = P[S \geq 10 / H_0] = \sum_{X=10}^{12} \binom{12}{s} \left(\frac{1}{2}\right)^s \left(\frac{1}{2}\right)^{n-s} = \left(\frac{1}{2}\right)^{12} \sum_{X=10}^{12} \binom{12}{s}$$

$$= \left(\frac{1}{2}\right)^{12} (66 + 12 + 1) = 0.0193$$

If we take α =0.05, then critical region is given by

$W = \{S; S \geq r_\alpha\}$

Since P = 0.0193 is less that of $\alpha = 0.05$ so we reject Ho.

That is, this provides reasonable evidence that the average income of men exceeds that of women in the firm.

17.5.5 THE WILCOXON TEST (WILCOXON RANK SUM TEST)

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed or data is on ordinal scale.

The test is named for Frank Wilcoxon (1892–1965) who, in a single paper, proposed both it and the rank-sum test for two independent samples (Wilcoxon, 1945). The test was popularized by Siegel (1956) in his influential text book on non-parametric statistics. Siegel used the symbol T . In consequence, the test is sometimes referred to as the Wilcoxon T test, and the test statistic is reported as a value of T.

The ordinary sign test takes into account only the signs of differences between each observation and the hypothesized median M_0 whereas magnitudes of these differences are ignored. If we take the assumption that the population is symmetric, the Wilcoxon test (or Wilcoxon signed-rank test) provides an alternative test of location which utilizes both the magnitudes and signs of these differences.

Let X_1, X_2... X_n be a random sample from a continuous population with c.d.f. F(x) and median M. The test of location that takes into account not only the sign of deviations $\{x_i - M_0\}$, i = 1, 2,... n but also the magnitudes of the deviations, where Mo is the median under Ho. Here we also assume that the probability density function of X, f(x), is symmetric about the median M_0 of the distribution. Here our hypothesis is

$$H_0 : M = M_0 \qquad \textit{against} \qquad H_1 : M \neq M_0$$

Now Define $D_i = X_i - M_0$, i = 1, 2….n and define the absolute difference without regard to sign i.e.

$$|D_i| = |X_i - M_0|$$

Discard the tied observations i.e. $X_i - M_0 = 0$.

Then under Ho, the D_i's are symmetrically distributed about median zero. Therefore, positive and negative differences of equal magnitude have the same probability of occurrance i.e.

$$P|D_i \leq c| = P|D_i \geq c|$$

Now we arrange the D_i's in ascending order of magnitudes and then assign ranks from1 to n.

Now we define T^+ and T^-

$$T^+ = Sum of \ \ ranks \ \ for \ \ which \ \ D_i > 0$$

$$T^- = Sum of \ \ ranks \ \ for \ \ which \ \ D_i < 0$$

Since sum of all raks is constant i.e.,

$$T^+ + T^- = \sum_{i=1}^{n} i = \sum_{i=1}^{n} \frac{n(n+1)}{2}$$

the test based on $T^+ + T^-$ and $T^+ - T^-$ will be equivalent. In practice, the smallest of $T^+ + T^-$ and $T^+ - T^-$ is used as the test statistic

Let T= Min(T^+, T^-) and t_α be such that $P[T \leq t_\alpha] = \alpha$, then the critical regions for different types of alternatives will be as given below

Appropriate Alternative	Critical Region
H$_1$:M>Mo	$T^- \leq t_\alpha$
H$_1$:M<Mo	$T^+ \leq t_\alpha$
H$_1$:M$\neq$Mo	$T^+ \leq t_{\frac{\alpha}{2}}$ or $T^- \leq t_{\frac{\alpha}{2}}$

The tables of the left-hand critical values are given by Wilcoxon.

For large n (n >25), the distribution of standardized T may be taken to be N (0, 1). Under H$_0$, the distribution of

$$Z = \frac{T - E(T^+)}{\sqrt{V(T^+)}} \sim N(0,1)$$

Where E[T$^+$] = $\frac{n(n+1)}{4}$ and $V(T^+) = \frac{n(n+1)(2n+1)}{24}$

17.6 SELF ASSESSMENT QUESTIONS

1. Explain the non parametric methods how they are different from the parametric methods?
2. Derive the sign test stating clearly the assumptions made for it.
3. Explain the main difference between non parametric methods and parametric methods.
4. Explain the median test, how it is applied.
5. Give the advantages of non parametric methods over the parametric methods.
6. What are runs, how they are helpful in non parametric inferences?

Unit 5 **Lesson 18**

NON-PARAMETRIC TESTS

Structure :

18.1 Introduction

18.2 Objectives

18.3 Non Parametric Tests for bivariate distributions

18.3.1 The Sign Test for paired samples

18.3.2 Examples based on Sign Test for paired samples

18.3.3 Wilcoxon Signed Rank Test for paired Data

18.3.4 Two sample tests for unpaired data.

18.3.5 Examples based Two sample tests for Unpaired data (Wilcoxon test)

18.3.6 Median test

18.4 Self assessment Questions

18.1 INTRODUCTION

Non Parametric tests based on two samples are classified into two categories; non-parametric tests based on two paired (**dependent samples**) samples and tests based on unpaired samples (**independent samples**).

(i) When there are pairs of observations on two things being compared

(ii) For any given pair, each of the two observations is made under similar extraneous conditions.

(iii) Different pairs are compared under different conditions

Here (ii) give rise to "dependent pairs of observations so require different treatment from (i) and (ii) dependent pairs of observations

18.2 OBJECTIVES

The main objectives of this lesson is

1. To offer a different approach to many of the decision problems
2. To know how to apply these tests to bivariate data in a variety of problems.
3. To apply non parametric test for dependent Samples.
4. To apply non parametric test for independent samples

18.3 NON PARAMETRIC TESTS FOR BIVARIATE DISTRIBUTIONS

Non Parametric tests based on two samples are classified into two Categories; non-parametric tests based on two paired (**dependent samples**) Samples and tests based on unpaired samples **(independent samples).**

18.3.1 SIGN TEST FOR PAIRED SAMPLES

The single sample sign test procedure for testing of hypothesis is equally applicable to paired sample data. That is the observations in the two samples are matched pairs such as

(i) X denotes a worker’s daily output before training and Y denotes his daily output after the training.

(ii) X and Y are pre and post treatment observations when considering the effect of a single treatment.

Suppose X and Y are two random variables and we want to test the null hypothesis that the distributions of X and Y are identical. That is both the samples have been drawn from two populations with same c.d.f.s.

Let $(X_1, Y_1), (X_2, Y_2), \ldots\ldots., (X_n, Y_n)$ be paired samples of observations drawn from two populations with c.d.f,s F(x) and F(y) . Thus our problem is of testing null hypotheses

F(x) = F(y)

To test the null hypothesis H_0

Let $D_i = X_i - Y_i$, i = 1, 2, 3,…. .,n

Let M be the median of population difference D, where this population is assumed to be continuous so that

P[D=M]=0 and P[D>M] = P[D<M] = p

And under $H_0 : p = \frac{1}{2}$

Thus null hypothesis reduces to

$$P[D > M] = \frac{1}{2} \qquad and \qquad P[D < M] = \frac{1}{2}$$

If $X_i > Y_i$ *i.e.*, $D_i > 0$ put plus sign (+) and negative sign (-), if $X_i < Y_i$ *i.e.*, $D_i < 0$ and discard those Di's for which D_i, =0.

Let S denotes the number of plus signs, then S follows a binomial distribution with parameters n and p. where p =1/2 and n is the sum of plus and minus signs. To test the above Ho, we find the probability

$$P[S \geq s / H_0] = \sum_{S=s}^{n} \binom{n}{s} \left(\frac{1}{2}\right)^{s} \left(\frac{1}{2}\right)^{n-s}$$

If this probability is greater than the given level of significance (α), we accept Ho, otherwise reject Ho. Further if n is large (n >30), we use normal approximation as

$$Z = \frac{S - E[S]}{S.D(S)} \sim N(0,1)$$

$$Z = \frac{S - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \sim N(0,1)$$

18.3.2 Examples based on Sign Test for paired samples

Example : Suppose there are 16 positive and 4 minus signs in the set of the 20 paired observations (X, Y). Test the hypothesis that two samples are drawn from the same population.

Solution : Here we want to test the hypothesis.

H_0: The two populations have an identical distribution.

It is given that the number of plus signs i.e. S = 16, n = 20 and under H_0,

p = 1/2. Thus under H_0, S~ B(20,1/2)). To test H_0, we find

$$P = P(S \geq 16/H_0)$$

$$= P \sum_{S=16}^{20} \binom{20}{s} \left(\frac{1}{2}\right)^{s} \left(\frac{1}{2}\right)^{20-s}$$

$$= \left(\frac{1}{2}\right)^{20} \left[\binom{20}{16} + \binom{20}{17} + \binom{20}{18} + \binom{20}{19} + \binom{20}{20}\right]$$

$$= 0.0059$$

If take α = 0.**01**, then we reject the null hypothesis of identical distribution as P < 0.01.

18.3.3 Wilcoxon Signed Rank Test for paired Data

Let $(X_1, Y_1), (X_2, Y_2), \ldots\ldots., (X_n, Y_n)$ be paired samples of observations drawn from two populations with c.d.f,s F(x) and F(y) . Thus our problem is of testing null hypotheses

F(x) = F(y)

To test this hypothesis following procedure is used

To test the null hypothesis H_0

Let $D_i = X_i - Y_i$; i = 1, 2, 3,…. .,n and find $|Di|$ $for\ i = 1,2,\ldots., n$ discard those Di's for which D_i, =0.Now arrange $|Di|$ in ascending order of their magnitude and assign them rank from 1 to n. Next we determine T^+ and T^- ,where

$$T^+ = Sum of\ ranks\ for\ which\ D_i > 0$$

$$T^- = Sum of\ ranks\ for\ which\ D_i < 0$$

Then our test statistic is

$$T_w = \min(T^+, T^-)$$

This value of Tw can be compared with critical value of Tw, obtained from table for given value of n , α and T.If calculated value of Tw is less than or equal to critical value of Tw , we reject H_0 , otherwise we accept the null hypothesis.> f n is large we use normal approximation

$$Z = \frac{T_W - E[T_W]}{S.D(T_W)} \sim N(0,1)$$

Where $E[T_W] = \frac{n(n+1)}{4}$ and $V[T_W] = \frac{n(n+1)(2n+1)}{24}$

18.3.4 TWO SAMPLE TESTS FOR UNPAIRED DATA.

Two Sample Tests For Unpaired Data (Wilcoxon test)

Let X_1, X_2,. ..Xm be a random sample of size m from a population with c.d.f. F(x) and Y_1, Y_2,. ..,Yn be another random sample of size n from a population with c.d.f F(y). These samples are drawn independently from each of the two populations. The hypothesis of interest is that the two samples are drawn from identical populations i.e.

H_0: F(x) = F(y), for all x.

In order to test this hypothesis the following is the testing procedure.

Combine the two sample observations. Arrange the pooled observations in ascending order of magnitude and assign them ranks from 1 to m+n. Find the sum of the ranks of first sample and second sample and let W_1 and W_2 denote these sum respectively.

Let $W=W_1+W_2$

Sum of the ranks of all the m+n N observations

Wilcoxon proposed a test for accepting the one-sided location alternative if the W_1 in the combined sample is too large or two small according the alternative hypothesis. The two-sided location alternative hypothesis is accepted if the sum of ranks of first sample is either too large or two small. The Wilcoxon test statistic is

$$T_W = \sum_{i=1}^{N} iZ_i \qquad N = m + n$$

Where Zi is defined as

Zi = 1 if the observation in combined sample is from first sample

= 0 otherwise.

If N = (m+n) is large, we use normal approximation as

$$Z = \frac{T_W - E[T_W]}{\sqrt{V(T_W)}} \sim N(0,1)$$

Where $E[T_W] = \frac{m(N+1)}{2}$ and $V[T_W] = \frac{mn(N+1)}{12}$

Note For this test, the normal approximation is good even for N = 12

18.3.5 EXAMPLES BASED TWO SAMPLE TESTS FOR UNPAIRED DATA (WILCOXON TEST)

Example Consider the data given in the table below describing the lifetimes of certain types of tubes manufactured by two methods

New Method :	259	254	249	256	252	260
Old Method :	250	247	253	244	251	258

Does this indicate that the life time with new method has increased?

Solution Here we formulate the following null hypothesis.

H_0: There is no significance difference between the life times of tubes manufactured by two methods.

To test this hypothesis, we first combined the observations of both the samples and assign ranks after arranging in ascending order of their magnitude

Life time	Method		Rank	
	Old	New	Old	New
244	Old		1	
247	Old		2	
249		New		3
250	Old		4	
251	Old		5	
252		New		6
253	Old		7	
254		New		8
256		New		9
258	Old		10	
259		New		11
260		New		12
Sum			29	49

Therefore, Tw = sum of ranks of first sample =49

Since N = m+n = 6+6 = 12 is large, so, we use normal approximation.

$$Z = \frac{T_W - E[T_W]}{\sqrt{V(T_W)}} \sim N(0,1)$$

Where $E[T_W] = \frac{m(N+1)}{2} = \frac{6\times 13}{2} = 39$

and

$$V[T_W] = \frac{mn(N+1)}{12}$$

$$= \frac{6\times 6\times 13}{12} = 39$$

There fore

$$Z = \frac{T_W - E[T_W]}{\sqrt{V(T_W)}}$$

$$= \frac{49-39}{\sqrt{39}} = 1.60 \sim N(0,1)$$

If $\alpha = 0.05$ *then* $Z_\alpha = 1.645$

Since $Z < 1.645$ we accept null hypothesis.

18.3.6 MEDIAN TEST

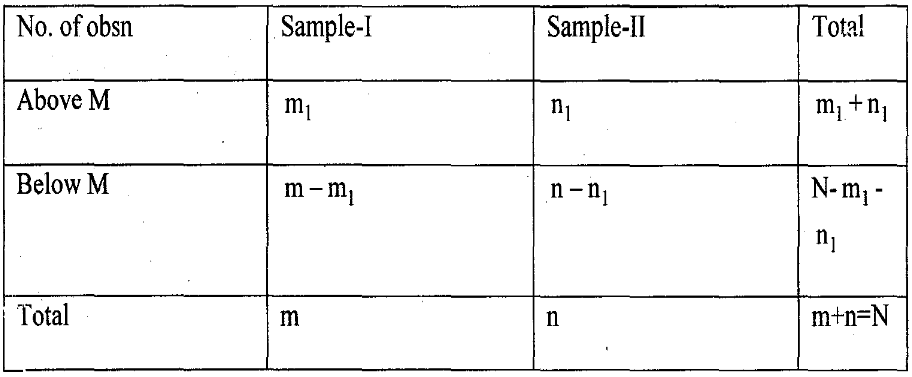
In statistics, Mood's median test is a special case of Pearson's chi-Square test. It is a nonparametric test that tests the null hypothesis that the medians of the populations from which two samples are drawn are identical. The data in each Sample are assigned to two groups, one consisting of data whose values are higher than the median Value in the two groups combined, and the other consisting of data whose values are at the median or below. A Pearson's chi-square test is then used to determine whether the observed frequencies in each group differ from expected frequencies derived from a distribution combining the two groups.

Let X_1, X_2,...Xm and Y_1, Y_2,...Yn be two independent random samples of size m and n from populations with c.d.f. F(x) and F(y) respectively. The hypothesis of interest is that

$$H_0 : F_X(x) = F_Y(y) \quad \textit{for all } x \textit{ or } H_0 : M_1 = M_2$$

where M_1 and M_2 are the median of first and second sample respectively. To test this hypothesis, the test procedure is given below

First of all we arrange the two sample observations together in increasing order and calculate the median for combined sample. Let it is M. Now we classify the sample values of both samples in the following 2x2 table



No. of obsn	Sample-I	Sample-II	Total
Above M	m_1	n_1	$m_1 + n_1$
Below M	$m - m_1$	$n - n_1$	N- m_1 - n_1
Total	m	n	m+n=N

Let m_1 and n_1 be the number of observations of first and second sample greater than the median. Now to test the hypothesis, we find the following probability

$$P = \frac{\binom{m}{m_1}\binom{n}{n_1}}{\binom{m+n}{m_1+n_1}}$$

We compare this probability P with level of significance α .If $P \geq \alpha$,we accept null hypothesis , otherwise we reject it.If frequencies in 2x2 table are large, we may use χ^2 test with 1 degree of freedom for testing null hypothesis.

$$\chi^2 = \frac{N[m_1(n-n_1)-n_1(n_1-m_1)]^2}{(m_1+n_1)(N-m_1-n_1)mn}$$

18.4 SELF ASSESSMENT QUESTIONS

1. Explain non-parametric methods how can they be used for bivariate data
2. Differentiate clearly between dependent and independent pairs of observations in reference to non-parametric tests.
3. Derive the median test stating clearly assumptions made for it.
4. Derive two sample test for unpaired data (WIlcoxon) and for paired data.
5. Explain median test how it is applied
6. Explain clearly stating assumptions if any, the sign test for paired samples.

Unit-5 **Lesson 19**

NON-PARAMETRIC TESTS

STRUCTURE

19.1 Introduction

19.2 Objectives

19.3 Mann-Whitney U test

19.3.1 Assumptions and formal statement of hypotheses in Mann-Whitney U test

19.3.2 Procedure for Mann-Whitney U test

19.3.3 Examples

19.3.4. Normal approximation of Mann Whitney U test

19.3.5 Relation of Mann Whitney U test with other tests

19.4 Test for independence based on Spearman’s rank correlation method

19.4.1 Example based on Spearman's rank correlation method

19.5 Self assessment questions

19.1 INTRODUCTION

We classify Non Parametric tests based on two samples into two categories; non-parametric tests based on two paired (dependent samples) samples and tests based on unpaired samples (independent samples)

19.2 OBJECTIVES

The main objectives of this lesson are

1. To offer a different approach to many of the decision problems

2. To know how to apply these tests to bivariate data in a variety of problems.
3. To know how to apply Mann Whitney U test.
4. To know how to apply Spearmen's rank correlation method to non-parametric problems
5. To apply non parametric test for independent samples

19.3 MANN-WHITNEY U TEST

In statistics, the Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW) or Wilcoxon rank-sum test) is a non-parametric statistical hypothesis test for assessing whether two independent samples of observations have equally large values. It is one of the most well-known non-parametric significance tests. It was proposed initially by Frank Wilcoxon in 1945, for equal sample sizes, and extended to arbitrary sample sizes and in other ways by Henry Mann and his student Donald Ransom Whitney in 1947.

19.3.1 ASSUMPTIONS AND FORMAL STATEMENT OF HYPOTHESES

Although Mann and Whitney developed the MWW test under the assumption of continuous responses with the alternative hypothesis being that one distribution is stochastically greater than the other, there are many other ways to formulate the null and alternative hypotheses such that the MWW test will give a valid test.

A Very general formulation is to assume that:

1. All the observations from both groups are independent of each other,

2. The responses are ordinal or continuous measurements(i.e. one can atleast say, of any two observations, which is the greater),

3. Under the null hypothesis the distributions of both groups are equal, so that the probability of an observation from one population (X) exceeding an observation from the second population (Y)equals the probability of an observation from Y exceeding an observation from X, that is, there is a symmetry between populations with respect to probability of random drawing of a larger observation.

4. Under the alternative hypothesis the probability of an observation from one population (X) exceeding an observation from the second population (Y) (after correcting forties) is not equal to 0.5. The alternative may also be stated in terms of a one-sided test, for example: P(X > Y) + 0,5 - P(X = Y) > 0.5.

If we add more strict assumptions than those above such that the responses are assumed continuous and the alternative is a location shift (i.e. $F_1(x) = F_2(x+\delta)$), then we can interpret a significant MWW test as showing a significant difference in medians. Under this location shift assumption, we can also interpret the MW W as assessing whether the Hodges-Lehmann estimate of the difference in central tendency between the two populations differs significantly from zero. The Hodges-Lehmann estimate for this two-sample problem is the median of all possible differences between an observation in the first sample and an observation in the second sample.

19.3.2 PROCEDURE MANN-WHITNEY U TEST

The test involves the calculation of a statistic, usually called U, whose distribution under the null hypothesis is known. In the case of small samples, the distribution is tabulated, but for sample sizes above ~20 there is a good approximation using the normal distribution. Some books tabulate statistics equivalent to U, such as the sum of ranks in one of the samples, rather than U itself

The U test is included in most modern statistical packages. It is also easily calculated by hand, especially for small samples. There are two ways of doing this.

First, arrange all the observations into a single ranked series. That is, rank all the observations without regard to which sample they are in.

For Small Samples a direct method is recommended. It is very quick, and gives an insight into the meaning of the U statistic.

1. Choose the sample for which the ranks seem to be smaller (The only reason to do this is to make computation easier). Call this "sample 1" and call the other sample "sample 2."
2. Taking each observation in sample 1, count the number of observations in sample 2 that have a smaller rank (count a half for any that are equal to it). The sum of these counts is U.

For larger samples, a formula can be used:

1. Add up the ranks for the observations which came from sample 1. The sum of ranks in Sample 2 follows by calculation, since the sum of all the ranks equals N(N + 1)/2 where N is the total number of observations.

2. U is then given by:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

where n_1 is the sample size for sample 1, and R_1 is the sum of the ranks in sample 1

Note that there is no specification as to which sample is considered sample 1. An equally valid formula for *U* is

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

The smaller value of U_1 and U_2 is the one used when consulting significance tables. The sum of the two values is given by

$$U_1 + U_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2}$$

Knowing that $R_1 + R_2 = N(N + 1)/2$ and $N = n_1 + n_2$, we find that the sum is

$$U_1 + U_2 = n_1 n_2$$

The maximum value of U is the product of sample sizes of two samples.

19.3.3 EXAMPLES

EXAMPLE:Suppose that Aesop is dissatisfied with his classic experiment in which one tortoise was found to beat one hare in a race, and decides to carry out a significance test to discover whether the results could be extended to tortoises and hares in general. He collects a sample of 6 tortoises and 6 hares, and makes them all run his race at once. The order in which they reach the finishing post (their rank order, from first to last crossing the finish line) is as follows, writing T for a tortoise and H for a hare:

T H H H H H T T T T T H

What is the value of *U*?

- Using the direct method, we take each tortoise in turn, and count the number of hares it is beaten by, getting 0, 5, 5, 5, 5, 5, which means $U = 25$. Alternatively, we could take each hare in turn, and count the number of tortoises it is beaten by. In this case, we get 1, 1, 1, 1, 1, 6. So $U = 6 + 1 + 1 + 1 + 1 + 1 = 11$. Note that the sum of these two values for *U* is 36, which is 6 × 6.

- Using the indirect method: the sum of the ranks achieved by the tortoises is 1 + 7 + 8 + 9 + 10 + 11 = 46.

Therefore $U = 46 - (6 \times 7)/2 = 46 - 21 = 25$.

the sum of the ranks achieved by the hares is 2 + 3 + 4 + 5 + 6 + 12 = 32, leading to $U = 32 - 21 = 11$.

Example:Consider another hare and tortoise race, with 19 participants of each species, in which the outcomes are as follows:

H H H H H H H H H T T T T T T T T T T H H H H H H H H H H T T T T T T T T T

The median tortoise here comes in at position 19, and thus actually beats the median hare, which comes in at position 20.

However, the value of *U* (for hares) is 100

(9 Hares beaten by (x) 0 tortoises) + (10 hares beaten by (x) 10 tortoises) = 0 + 100 = 100

Value of *U* (for tortoises) is 261

(10 tortoises beaten by 9 hares) + (9 tortoises beaten

by 19 hares) = 90 + 171 = 261

Consulting tables, or using the approximation below, shows that this *U* value gives significant evidence that hares tend to do better than tortoises ($p < 0.05$, two-tailed). Obviously this is an extreme distribution that would be spotted easily, but in a larger sample something similar could happen without it being so apparent. Notice that the problem here is not that the two distributions of ranks have different variances; they are mirror images of each other, so their variances are the same, but they have very different skewness.

19.3.4 Normal approximation of Mann Whitney U test

For large samples, U is approximately normally distributed. In that case, the standardized value

$$Z = \frac{U - m_U}{\sigma_U}$$

where m_U and σ_U are the mean and standard deviation of U, are given by

$$m_U = \frac{n_1 n_2}{2}$$

And $\sigma_U = \dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}$

Since $U_1 + U_2 = n_1\, n_2$, the mean $\frac{n_1 n_2}{2}$ used in the normal approximation is the mean of the two values of *U*. Therefore, the absolute value of the *z* statistic calculated will be same whichever value of *U* is used.

19.3.5 Relation of Mann Whitney U test with other tests

Non-parametric tests are basically used in order to overcome the underlying assumption of normality in parametric tests. Quite general assumptions regarding the population are used in these tests.

A case in point is the Mann-Whitney U-test (Also known as the Mann-Whitney-Wilcoxon (MWW)). Unlike its parametric counterpart, the t-test for two samples, this test does not assume that the difference between the samples is normally distributed, or that the variances of the two populations are equal. Thus when the validity of the assumptions of t-test are questionable, the Mann-Whitney U-Test comes into play and hence has wider applicability.

Summarizing the above discussion we can say that:The U test is useful in the same situations as the independent samples Student's t-test, and the question arises of which should be preferred.

Ordinal data

U remains the logical choice when the data are ordinal but not interval scaled, so that the spacing between adjacent values cannot be assumed to be constant.

Robustness

As it compares the sums of ranks, the Mann–Whitney test is less likely than the t-test to spuriously indicate significance because of the presence of outliers – i.e. Mann–Whitney is more robust.

19.3.6 EFFICIENCY OF MANN WHITNEY U TEST

When normality holds, MWW has an (asymptotic) efficiency of 3 / π or about 0.95 when compared to the t test. For distributions sufficiently far from normal and for sufficiently large sample sizes, the MWW can be considerably more efficient than the t.

Overall, the robustness makes the MWW more widely applicable than the t test, and for large samples from the normal distribution, the efficiency loss compared to the t test is only 5%, so one can recommend MWW as the default test for comparing interval or ordinal measurements with similar distributions.

The relation between efficiency and power in concrete situations isn't trivial though. For small sample sizes one should investigate the power of the MWW v/s t.

MWW will give very similar results to performing an ordinary parametric two-sample t test on the rankings of the data.

19.4 TEST FOR INDEPENDENCE BASED ON SPEARMAN'S RANK CORRELATION METHOD

Let (X$_1$, Y$_1$), (X2, Y$_2$).....(Xn ,Yn) be a random samples from bivariate population. We know that the coefficient of correlation is defined by

$$R=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad \text{............(1)}$$

where $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ and $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$

If the sample values X$_1$, X$_2$,...Xn and Y$_1$, Y$_2$,...Yn are each ranked from 1 to n in increasing order of their magnitude and if X$_i$'s and Y$_i$'s have continuous degrees of freedom we get a unique set of rankings and the data will reduce to n pairs of rankings .

Let us write R$_i$ =Rank (Xi), S$_i$= Rank(Yi) ; i=1,2…..n

$$\sum_{i=1}^{n}R_i=\sum_{i=1}^{n}S_i=\frac{n(n+1)}{2} \qquad \text{..................(2)}$$

So that $\bar{R}=\bar{S}=\frac{(n+1)}{2}$ ....(3)

And $\sum_{i=1}^{n}(R_i-\bar{R})^2=\sum_{i=1}^{n}(S_i-\bar{S})^2=\frac{n(n^2-1)}{12}$ substituting in eq (1) we get

$$R = \frac{12\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{n^3 - n} = \frac{12\sum_{i=1}^{n}(R_i S_i)}{n(n^2-1)} - \frac{3(n+1)}{n-1}$$

………………(4)

Let D$_i$ = R$_i$ - S$_i$ = $(R_i - \overline{R})(S_i - \overline{S})$ then we have

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n}(R_i - \overline{R})^2 + \sum_{i=1}^{n}(S_i - \overline{S})^2 - 2\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})$$

$$= \frac{n(n^2-1)}{6} - 2\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})$$

$$= \frac{n(n^2-1)}{6} - \frac{1}{6}R.n(n^2-1) = \frac{n(n^2-1)}{6}[1-R.]$$

$$\Rightarrow \qquad R = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad ……………..(5)$$

The statistic defined in (4) and (5) is called as spearman's rank correlation coefficient. From eq.(4) we see that

$$\text{E[R]} = \frac{12E\left[\sum_{i=1}^{n} R_i S_i\right]}{n(n^2-1)} - \frac{3(n+1)}{n-1}$$

$$= \qquad \frac{12n}{n(n^2-1)}E[R_i S_i] - \frac{3(n+1)}{n-1} = \qquad \frac{12}{(n^2-1)}E[R_i S_i] - \frac{3(n+1)}{n-1}$$

……………. (6)

Under H$_0$, the random variables X and Y are independent, so that the ranks R$_i$ and S$_i$ are also independent. It means that

$$E_{H_0}[R_i S_i] = E[R_i].E[S_i] = \frac{n+1}{2}.\frac{n+1}{2} = \left(\frac{n+1}{2}\right)^2 \qquad \text{so that from (6)}$$

$$E_{H0}[R]=\frac{12}{(n^2-1)}\left(\frac{n+1}{2}\right)^2-\frac{3(n+1)}{n-1}=0$$

..................(7)

Thus we should reject Ho if the tabulated value of R is large , reject H0 if $|R|=R_\alpha$

$$P_{Ho}\left[|R|>R_\alpha\right]\le\alpha$$

Critical Values of r, the critical values of r_s can be obtained by the table for critical values for the Spearman Rho rank correlation coefficient test for given sample size and significance level. If test is two tailed, we use two critical values, one negative and one. positive. For left tailed test we use negative values of r_s, and use positive value of r_s if test is right tailed test. Where is Spearman's rank correlation

19.4.1 EXAMPLE BASED ON SPEARMAN'S RANK CORRELATION METHOD

Example: The following table shows the per capita income (in thousands) and food expenditure of the family in different states.

Per capita income	11	16	18	8	6	15	10	5
Expenditures	5	7	8	3	2	8	4	2

Based on above data, we can conclude that there is no significance (linear) correlation between the per capita incomes and expenditures, use $\alpha=0.05$

Solution:Here, the null hypothesis H_0 is there is no correlation between per capita incomes and expenditure. And alternative hypothesis is H_1 is correlation between per capita incomes and expenditures.

$H_0 : \rho = 0 \quad and \quad H_1 : \rho \neq 0$ where ρ is the rank correlation coefficient.

Critical region for statistic: here, n =8 and α=0.05 for two tailed test the critical values are + 0.738. So we will reject null hypothesis if the observed values of r_s is either - 0.738 or less, or + 0.738 or above.

-0.738 or less	-0.738 to +0.738	+0.738 or above
Rejection	Non rejection	Rejection

Here, r_s = 0.869048 is higher than 0.738 so falls in rejection region, so null hypothesis reject. Then we conclude that there is correlation between the per capita income and expenditure.

19.5 SELF ASSESSMENT QUESTIONS

l. Give advantages of non parametric methods over the parametric methods

2. Explain non parametric how they can be used in case of bivariate data.

3. Elaborate Mann Whitney U test with the help of suitable example

4. Derive expression for the test Statistic for Spearman's rank correlation test for independence.

5. Develop the following nonparametric tests, stating clearly he underlying assumptions and the null hypothesis

   (a) Mann -Whitney-Wilcoxon Test

   (b) Spearman's rank correlation test for independence

6. Describe the median test when there are two independent samples. What non parametric test you would like to use When theses samples are related.

7. Discuss Mann-Whitney-Wilcoxon test for equality of two population distribution Functions

8. Critically examine the utility of Non Parametric tests

9. Highlight the advantages of Non Parametric tests in certain experimental conditions

10. Use appropriate tests to see if there is a difference between numbers of days required to collect receivable amount before and after a new collection policy

Before:	32	35	33	36	44	41	36	32	39	31
After:	36	37	34	40	40	42	36	40	42	33
Before:	47	30	34	29	41					
After:	36	37	34	40	39					