COM-1046

GAUHATI UNIVERSITY
Centre for Distance and Online Education

M.A. First Semester

(Under CBCS)

MASTER OF COMMERCE

Paper: COM 1046
BUSINESS STATISTICS

CONTENTS:

SLM Development Team:

Head, Department of Commerce, GU
Co-Ordinator, M.Com Programme, GUCDOE
Prof. S.K Mahapatra, Dept of Commerce, GU
Prof. Prashanta Sharma, Dept of Commerce, GU
Mr. Rajen Chetry, Assistant Professor, GUCDOE

Course Coordination:

Dr. Debahari Talukdar	Director, GUCDOE
Programme Coordinator	M.Com, GUCDOE
	Professor, Dept. of Commerce, G.U.
Mr. Rajen Chetry	Assistant Professor, GUCDOE

Contributors:

Dr. Pranjal Sarma Block I & Block II
Assistant Professor, LCB College
Dr Mahuya Deb Block III, Block IV & Block V
Assistant Professor, Deptt. of Commerce, GU

Content Editing:

Dr. Pranjal Sarma Block III, Block IV & Block V
Assistant Professor, LCB College
Dr Mahuya Deb Block I & Block II
Assistant Professor, Deptt. of Commerce, GU

Cover Page Design & Type Setting:

Bhaskar Jyoti Goswami	GUCDOE
Nishanta Das	GUCDOE

ISBN:
October, 2023

Published on behalf of Gauhati University Centre for Distance and Online Education by the Director, and printed at Gauhati University Press, Guwahati-781014.

Block-1

UNIT 1: SAMPLING DISTRIBUTION AND THEORY OF ESTIMATION

Unit Structure:

1.0 INTRODUCTION

In case of any statistical investigation, our interest lies in studying the various characteristics of a particular collection of objects or observations usually called the target population, simply population or universe. By definition, the collection of all the observations under study in any statistical investigation, is called population or universe for that specific study. The number of observations included in a population is termed as the size of the population or population size.

Again, a subcollection of the population is known as sample. In other words, a sample may be difined as a part of a population so selected with a view to represent the population. The number of units in a sample is called sample size and the units forming the sample are known as “Sampling Units”. Again, a detailed and complete list of all the sampling units is termed as a “Sampling Frame”. It is a must to have a updated sampling frame complete in all respects before the samples are actually drawn.

Any statistical characteristics such as mean, median, quartile, standard deviation, moments etc. of the population under study, is called a

parameter while any statistical characteristics such as above of a sample drawn from a population is called a statistic. Very often, the values of various parameters and unknown and these are estimated by the corresponding statistic. For example, sample mean $\bar{x}$ is used as a estimator of population mean , sample standard deviation & is used as an estimator of population standard deviation , etc. The difference between a statistic and the corresponding parameter is known as sampling error. The study of Sampling theory as well as theory of estimation help us to estimate the true value of the population parameters by minimizing the sampling errors.

1.1 OBJECTIVES

Having studied this chapter, you should be able to

* know the concept of sampling distribution;
* develop the understanding of the methods of estimation;
* differentiate between point estimation and interval estimation.

1.2 SAMPLING FLUCTUATIONS

The value of parameter is considered as constant. But, if we compute the value of a statistic, say mean or median or mode or s.d., etc, it is quite natural that the value of the sample statistic may vary from sample to sample as the sampling units of one sample may be different from that of another sample bisides sample sizes are same. The variation in the values of a statistic from sample to sample is termed as “Sampling Fluctuations” or “Sampling Variation”.

1.3 SAMPLING DISTRIBUTION OF A STATISTIC

If it is possible to obtain the values of a statistic (t) from all the possible samples of a fixe sample size along with the corresponding probabilities, then we can arrange the values of the statistic, which is to be treated as a random variable, in the form of a probability distribution. Such a probability distribution is known as the sampling distribution of the statistic.

Starting with a population of N units, we can draw many a sample of a fixed size n. In case of sampling with replacement, the total number of samples that can be drawn is and consequently we shall N^n different values of any statistic (t) like mean, median, S.D. etc. computed for Nn samples. Again, when sampling is done without replacement of the sampling units, the total number of samples that can be drawn is $N_{C_n=m}$ (say). We can compute any statistic (t) like mean, median, S,d. etc. for these m samples resulting in m values of the statistic. These N^n values of the statistic (t) in case of sampling with replacement and $N_{C_n=m}$ values in case of sampling without replacement may be arranged in the form of a probability distribution known as the sampling distribution of the statistic.

The sampling distribution, just like a theoretical probability distribution possess different properties. One of these is the 'Law of Large Numbers' which asserts that a positive integer n can be ditermined such that if a random sample of size n or large is drawn from a population having mean μ, the probability that the sample mean will deviate from by less than any arbitrarily small quantity can be made to be as close to 1. This implies that a fairly reliable inference can be made about an infinite population by taking only a finite sample of sufficiently large size. Another interesting result in this connection is the 'Central Limit theorem' which is discussed elaborately in the Unit 2, possible sample of size 2 with replacement = 3 =1. These are exhibited along with the corresponding sample mean in the following table :

Sl. No.	Sample	Sample Mean ( )
1	1, 1	1
2	1, 5	3
3	1, 3	2
4	5, 1	3
5	5, 5	5
6	5, 3	4
7	3, 1	2

8	3, 5	4
9	3, 3	3

This sampling distribution of the sample mean is given as follows:

	1	2	3	4	5	Total
p :	1/9	2/9	3/9	2/9	1/9	1

(ii) Without replacement :

As N = 3 and n = 2, the total number of possible samples without replacement = . Possible samples of size 2 and corresponding sample means are given below :

Serial No.	Sample	Sample Mean ($\bar{x}$)
1	1, 3	2
2	1, 5	3
3	3, 5	4

The sampling distribution of the sample mean is given as follows:

:	2	3	4	Total
p :	1/3	1/3	1/3	1

Example : Compute the standard deviation of sample mean for the last problem. Obtain the SE of sample mean and show that they are equal.

Solution : We consider the following cases :

(i) With replacement :

Let u = . The sampling distribution of u is given by

u :	1	2	3	4	5
p :	1/9	2/9	3/9	2/9	1/9

$$= \frac{1}{9}\times 1 + \frac{2}{9}\times 2 + \frac{3}{9}\times 3 + \frac{2}{9}\times 4 + \frac{1}{9}\times 5$$

$$= 3$$

$$\therefore E(u^2) = \sum p_i u_i^{\,2}$$

$$= \frac{1}{9}\times 1^2 + \frac{2}{9}\times 2^2 + \frac{3}{9}\times 3^2 + \frac{2}{9}\times 4^2 + \frac{1}{9}\times 5^2$$

$$= \frac{31}{3}$$

$$\therefore V(u) = E(u^2) - [E(u)]^2$$

$$= \frac{31}{3} - 3^2 = \frac{4}{3}$$

Hence $SE_{\bar{x}} = \frac{2}{\sqrt{3}} \longrightarrow (1)$

Again, the population mean (μ) is given by

and the population variance (σ^2) is given by

$$\sigma^2 = \frac{1}{N}\sum (x-\mu)^2 \qquad \mu = \frac{1+5+3}{3} = 3$$

$$= \frac{1}{3}\left[(1-3)^2 + (5-3)^2 + (3-3)^2\right] = \frac{8}{3}$$

$$\therefore SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{8}{3}}\times\frac{1}{\sqrt{2}} = \frac{2}{\sqrt{3}} \longrightarrow (2)$$

Thus comparing (1) and (2), we are able to verify the validity of the formula.

(ii) Without replacement :

In this case, the sampling distribution of $v = \bar{x}$ is given by

v :	2	3	4
p :	1/3	1/3	1/3

= 3

$$V(\bar{x}) = Var(v) = E(v^2) - [E(v)]^2$$

$$= \frac{1}{3} \times 2^2 + \frac{1}{3} \times 3^2 + \frac{1}{3} \times 4^2 - 3^2$$

$$= \frac{29}{3} - 9$$

$$= 2/3$$

$$\therefore SE_{\bar{x}} = \sqrt{\frac{2}{3}}$$

$\therefore$ SE for without replacement for population is given by

$$= \sqrt{\frac{8}{3}} \times \frac{1}{\sqrt{2}} \times \sqrt{\frac{3-2}{3-1}}$$

$$= \sqrt{\frac{2}{3}}$$

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad \therefore E(v) = \frac{1}{3} \times 2 + \frac{1}{3} \times 3 + \frac{1}{3} \times 4$$

and thereby, we make the same conclusion as in previous case.

Example : Construct a sampling distribution of the sample mean for the following population when random samples of size 2 are taken from it (a) with replacement and (b) without replacement. Also find the mean and standard error of the distribution in each case.

Population Unit :	1	2	3	4
Observation :	22	24	26	28

Solution :

The mean and standard diviation of population are

$$\mu = \frac{22 + 24 + 26 + 28}{4} = 25 \quad \text{and}$$

$$\sigma = \sqrt{\frac{22^2 + 24^2 + 26^2 + 28^2}{4} - 25^2}$$

$$= \sqrt{5} = 2.236 \quad \text{respectively.}$$

(a) With replacement :

When random samples of size 2 are drawn, we have $4^2 = 16$ samples, shown below :

Simple No	Sample Values	$\bar{x}$
1	22, 22	22
2	22, 24	23
3	22, 26	24
4	22, 28	25
5	24, 22	23
6	24, 24	24
7	24, 26	25
8	24, 28	26
9	26, 22	24
10	26, 24	25
11	26, 26	26
12	26, 28	27
13	28, 22	25
14	28, 24	26
15	28, 26	27
16	28, 28	28

Since all of the above samples are equally likely, therefore, the probability of each value of is . Thus, we can write the sampling distribution of $\bar{x}$ as given below :

:	22	23	24	25	26	27	28	Total
p :		$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	1

$$\therefore \mathrm{E}(\bar{x}) = 22 \times \frac{1}{16} + 23 \times \frac{2}{16} + 24 \times \frac{3}{16} + 25 \times \frac{4}{16} +$$

$$26\times\frac{3}{16}+27\times\frac{2}{16}+28\times\frac{1}{16}$$

= 25

$$V(\bar{x}) = E(\bar{x}^2) - [E(\bar{x})]^2$$

$$= \left[22^2\times\frac{1}{16}+23^2\times\frac{2}{16}+24^2\times\frac{3}{16}+25^2\times\frac{4}{16}+\right.$$

$$\left.26^2\times\frac{3}{16}+27^2\times\frac{2}{16}+28^2\times\frac{1}{16}\right]-25^2$$

$$= 627.5 - 625 = 2.5$$

Which is equal to $\frac{\sigma}{\sqrt{n}}$

(b) Without replacement :

When random samples of size 2 are drawn without replacement, we have ${}^4C_2$ samples, shown below :

Simple No	Sample Values	$\bar{x}$
1	22, 24	23
2	22, 26	24
3	22, 28	25
4	24, 26	25
5	24, 28	26
6	26, 28	27

$$S.E.(\bar{x}) = \sqrt{v(\bar{x})} = \sqrt{2.5}$$

Since all the samples are equally likely, the probability of each value

of is . Thus, we can write the sampling distribution of $\bar{x}$ as

:	23	24	25	26	27	Total
p :		$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

$$\therefore E(\bar{x}) = \frac{1}{6}[23+24+25\times2+26+27]$$

= 25

$$\therefore V(\bar{x}) = E(\bar{x}^2) - [E(x)]^2$$

$$= \left[\frac{1}{6}\times 23^2 + \frac{1}{6}\times 24^2 + \frac{2}{6}\times 25^2 + \frac{1}{6}\times 26^2 + \frac{1}{6}\times 27^2\right] - 25^2$$

$$= 626.67 - 625$$

$$= 1.67$$

Alternatively, population S.E. is given as

$$S.E.(\bar{x}) = \sqrt{\frac{N-n}{N-1}.\frac{\sigma^2}{n}} = \sqrt{\frac{4-2}{3}\times\frac{5}{2}}$$

$$= \sqrt{1.67} = 1.292$$

1.4 Summing Up

A sampling distribution is an array of sample studies relating to a population. If we select a number of independent random samples of a definite size from a given population and calculate some statistic like the mean, standard deviation etc. from each sample, we shall get an array of values of these statistics. The distribution so obtained by these values of the statistic is called the sampling distribution of that statistic. the standard deviation of the sampling distribution would be called the standard error which is abbreviated as S.E. The concept of Standard Error is very useful in testing statistical hypothesis and in the theory of estimation.

$$\therefore S.E.(\bar{x}) = \sqrt{V(\bar{x})} = \sqrt{1.67} = 1.292$$

1.5 References and Suggested Readings

1. Gupta S.C. & Kapoor V.K.; Fundamentals of Mathematical Statistics; Sultan Chand & Sons.
2. Hazarika P.; Essential Statistics for Economics and Commerce; Akansha Publishing House.
3. Rao Radhakrishna C.; Linear Statistical Inference and its Applications; Wiley Eastern Limited.

1.6 Model Questions

Objective Questions:

1. A population has N items. Samples of size n are selected without replacement. Find the number of possible samples.
2. Standard error is always non-negative. (True or False)
3. If the mean of population is ()) then the mean sampling distribution is ..... . (fill in the blank)

4. Consider a population containing N items and n are selected as a sample with replacement. find the numver of possible samples.
5. Sampling distribution describes the distribution of sample ..... . (fill in the blank)

Descriptive Questions:

1. Explain the concept of sampling distribution of a statistic.
2. A population consists of four numbers 3, 4, 2, 5. Consider all possible distinct samples of size two that can be drawn without replacement and verify that the population mean is equal to the mean of the sample means.
3. A simple random sample of size 36 is drawn from a finite population of 101 units. If the population S.D. is 12.6, find the standard error of the sample mean when the sample is drawn (i) with replacement, (ii) without replacement.
4. Consider a population of 6 units with values 1,2,3,4,5,6. Write down all possible samples of size 2 (without replacement) from this population and construct a sampling distribution of the sample mean. Also find the mean and standard error of the distribution.
5. What do you mean by 'Sampling Fluctuations'? Describe briefly.

1.8 Answer of objective questions

1. (((((())))
2. True
3. ((()))
4. N^n
5. Statistics

Block-1

UNIT 2:

Unit Structure:

2.0 INTRODUCTION

It is seen that most of the distributions like Binomial, Poisson, etc. tend to normal distribution when the size of the sample is too large. For this reason and for otherwise also the distribution of sample mean, whatever be the nature of the parent population, will approach to the normal distribution as the size of the sample increases. This fact leads to the Central limit theorem, first proved by the French mathematician Pierre-Simon Laplace in 1810.

The theorem is applicable to all the populations in practice except a few which are very much different from the normal. It should be noted that the efficiency of the theorem increases with an increase in the sample size regardless of whether the source population is normal or skewed.

2.1 OBJECTIVES

This unit is based on concept of Central Limit Theorem. After compition of this unit, one should be able to

* have the basic concept of Central Limit Theorem
* understood the applications of Central Limit Theorem
* learn the technique of finding estimation of the mean and the variance
* know the concept of standard error

2.2 CENTRAL LIMIT THEOREM

This theorem states that :

"If $\{x_1, x_2, ..., x_n\}$ is a random sample of size n from a non-normal population of size N with mean μ and standard deviation , then the sampling distribution of sample mean will approach normal distribution with mean and standard error as n becomes larger and larger."

It should be noted that as a general rule, when $n \geq 30$, then sampling distribution of $\bar{x}$ is taken to be normal for practical purposes. Moreover, the larger the sample size the bettir will be the approximation.

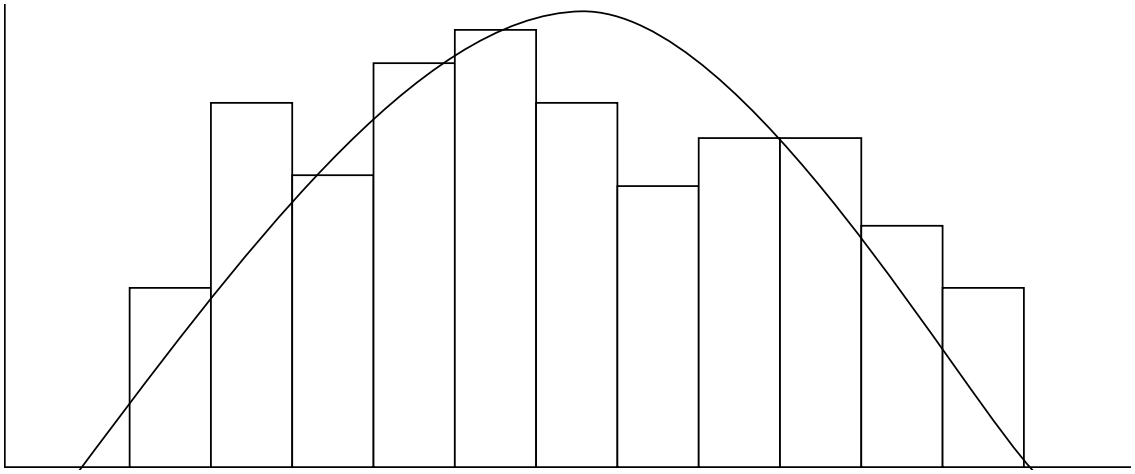
The statement of the above Central limit theorem is actually deducted from the generalized central limit theorem which is given as :

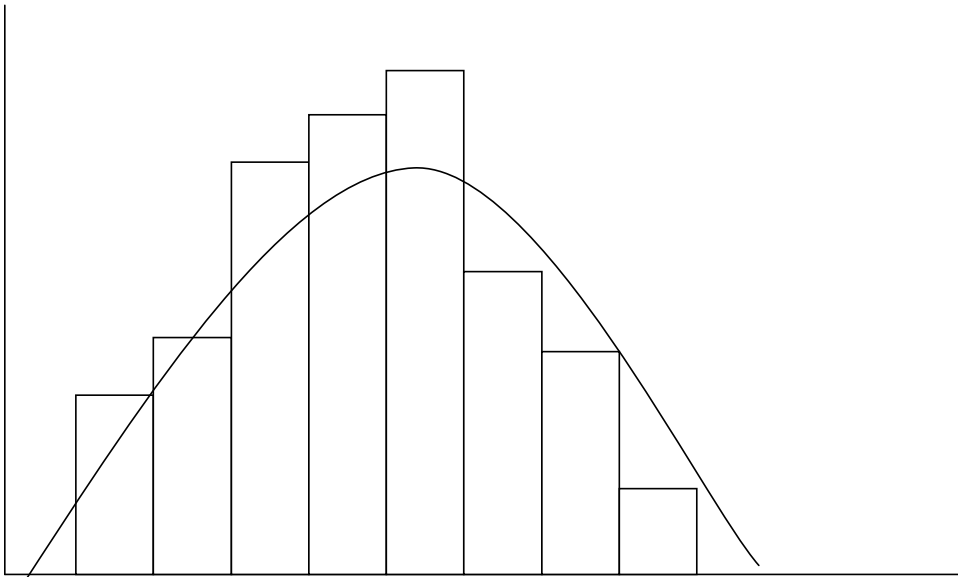
"If are independent random variables following any distribution, then under certain very general conditions, their sum $\sum x = x_1 + x_2 + ... + x_n$ is asymptotically normally distributed, i.e. $\sum x$ follows normal distribution as $n \longrightarrow \alpha$."

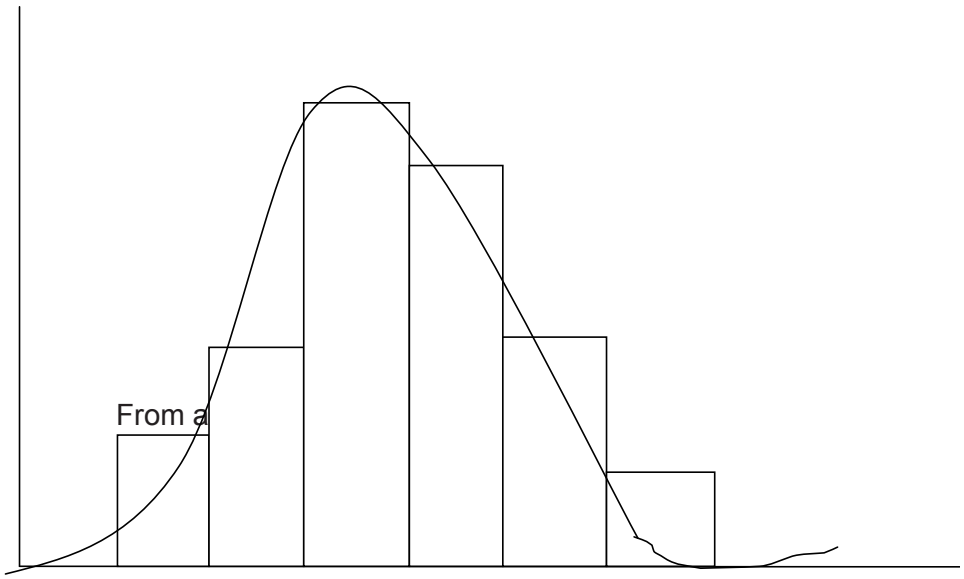
Thus the Central Limit Theorem asserts that for any statistic t, the random variable $Z = \dfrac{t - E(t)}{S.E.(t)}$ approaches the standard normal distribution of the population as n tends to infinity. This result is extensively used in Large Sample tests and in construction of confidence limits for the parameters provided the samples are relatively large.

How Does the Central Limit Theorem Works?

Basically the probability distributions are based on the concept of the Central Limit Theorem. For repeated sampling, the theorem provides us the behavour of the population parameters estimates. When sample values are plotted on a graph, the theorem gives us the shape of the distribution formed by means. As the sample sizes get larger, the distribution of the means from the repeated sample tends to normalize and forms a normal distribution. Statistically, when sample size (n) is more than or equal to 30, the Central Limit Theorem works better. But in case, even though n is less than 30, the distribution of sample means may tend to normal if the source population is normally distributed.







From a

The averages of samples have approximately follows normal distribution.

Morever, as sample size increases, the Distribution of Averages normal and the curve becomes narrow.

Example: A certain group of people receives government welfare benefit of Rs. 110/- per week with a standard deviation of Rs. 20/-. If a random sample of size 25 people is drawn, what is that probability that their mean benefit will be greater than Rs. 120/- per week?

Solution : We are given,

$$\mu = 110$$
$$\bar{x} = 120$$
$$\sigma = 20$$

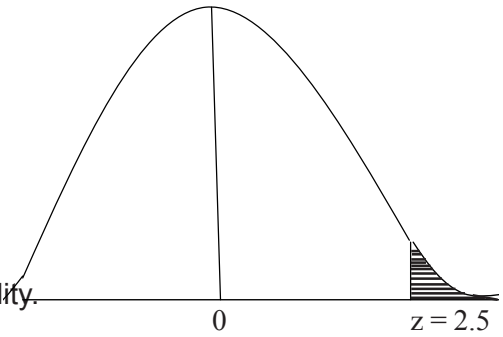
and n = 25

$$\therefore Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{120}{20/\sqrt{25}} = \frac{10}{4} = 2.5$$

We are to find,

$$P(z > 2.5)$$
$$= 0.5 - P(0 < z < 2.5)$$
$$= 0.5 - 0.4938$$
$$= 0.0062$$

whic is the reguired probability.



2.3 STANDARD ERROR OF A STATISTIC

The standard deviation of a statistic is termed as standard error. We know that, the population standard deviation describes the variation among values of members of the population, whereas the standard deviation of sampling distribution measures the variability among the values of the statistic (such as mean values, median values, etc) due to sampling errors. Thus knowledge of sampling distribution of a statistic enables us to find the probability of sampling error of the given magnitude. Consequently standard deviation of sampling distribution of a sample statistic measures sampling

error of the statistic. If t be any statistic calculated for different samples, then the standard error of the statistic t is generally denoted by S.E. (t).

The S.E.(t) measures not only the amount of chance error in the sampling process but also the accuracy disired in estimation of population parameters. Some of the common results of standard error of different statistic are given below :

1. $S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}}$ $(sample\ drawn\ with\ replacement)$,

   $S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}}\left(\frac{N-n}{N-1}\right)(sample drawn without replacement)$

2. $S.E.(p) = \sqrt{\frac{PQ}{n}}$ $(sample\ drawn\ with\ replacement)$,

   $S.E.(p) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}(sample\ drawn\ without\ replacement)$

3. $S.E.(s) = \frac{\sigma}{\sqrt{2n}}$

4. $S.E.(sample\ median) = \sqrt{\frac{\pi}{2n}}\quad \sigma = \frac{1.125332\sigma}{\sqrt{n}}$ $\qquad S.E.(r) = \frac{1-p^2}{\sqrt{n}}$

   $(\therefore \pi = 3.1416)$

5.

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

7. $S.E.(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

8. $S.E.(S_1 - S_2) = \sqrt{\frac{\sigma_1^2}{2x_1} + \frac{\sigma_2^2}{2n_2}}$

2.4 ESTIMATIONS OF THE MEAN AND THE VARIANCE OF THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

Let $x_1, x_2 ....., x_n$ be a random sample of size n from a large population $x_1, x_2 ....., x_n$ of size N whose mean is μ and variance is .

The mean of the sampling distribution of the sample mean

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Now, $E(\bar{x}) = E\left[\frac{x_1 + x_2 + ... + x_n}{n}\right]$

$$= \frac{1}{n} E\left[x_1 + x_2 ... + x_n\right]$$

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

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

Since $x_i (i = 1,2,...n)$ is a sample observation from the population $X_i (i = 1,2,...N)$, hence it can take any one of the values $x_1, x_2 ..., x_N$ each with equal probability $\frac{1}{N}$.

$$\therefore \mathrm{E}(x_i) = \frac{1}{\mathrm{N}} x_1 + \frac{1}{\mathrm{N}} x_2 + ... + \frac{1}{\mathrm{N}} x_n$$

$$= \frac{1}{\mathrm{N}}\left(x_1 + x_2 + ... + x_{\mathrm{N}}\right)$$

$$= \mu, \; for\ each\ i\ \ from\ 1\ to\ n.$$

Thus, $\mathrm{E}(x_1) = \mathrm{E}(x_2) = ... \mathrm{E}(x_n) = \mu$

$$\therefore (1) \Rightarrow \mathrm{E}(x) = \frac{1}{n}\left[\mu + \mu + ... \; to\ n\ terms\right]$$

$$= \frac{1}{n} . n\mu$$

$$= \mu$$

which shows that the mean of the sampling distribution of sample mean is the population mean .

Again,

$$= \frac{1}{n^2}\left[\text{Var}(x_1) + \text{Var}(x_2) + ... + \text{Var}(x_n)\right] \longrightarrow (2)$$

the covariance terms vanish since the sample observations are independent of each other.

Now, $\text{Var}(x_i) = \text{E}\left[x_i - \text{E}(x_i)\right]^2$

$$= \text{E}\left[x_i - \mu\right]^2 \qquad \therefore E(x_i) = \mu$$

$$= \frac{1}{\text{N}}\left[(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_N - \mu)^2\right]$$

$$= \sigma^2, \qquad \textit{for each } i \textit{ from 1 to } n$$

$$\therefore (2) \Rightarrow \text{Var}(\bar{\text{x}}) = \frac{1}{\text{n}^2}\left[\sigma^2 + \sigma^2 + ... \textit{ to } n \textit{ terms}\right]$$

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

$$\frac{\sigma^2}{n} \quad \textit{Var}(\bar{x}) = Var\left(\frac{x_1 + x_2 ... + x_n}{n}\right)$$

$$\therefore \text{S.E.}(\bar{x}) = \sqrt{\text{Var}(\bar{x})} = \frac{\sigma}{\sqrt{\text{n}}}$$

Check Your Progress :

1. Define standard Normal Variate.
2. What is the assertion of the statistic under Central Limit Theorem?
3. Name the Particular cases of Central Limit Theorem.

2.5 Summing Up

The central limit theorem statis that for a random sample of size n drawn from a non-normal population with mean and variance, the sample mean approximately follows a noraml distribution with mean and variance. The larger the value of the size of the sample, the better will be the approximation to the normal.

The mean of the sampling distribution of sample mean (()) is the population mean ((((())))) and variance of the sample mean is ((((())). Further, we calculate S.E. ((((((((((.

2.6 References and Suggested Readings

1. Hogg, Tanis, Rao; Probability and Statistical Inference; Pearson.
2. Bhuyan K.C.; Probability Distribution Theory and Statistical Inference; New Central Book Agency (P) Ltd.
3. Gupta S.C. & Kapoor V.K.; Fundamentals of Mathematical Statistics; Sultan Chand & Sons.
4. Hazarika P.; Essential Statistics for Economics and Commerce; Akansha Publishing House.
5. Rao Radhakrishna C., Linear Statistical Inference and its Applications; Wiley Eastern Limited.

2.7 Model Questions

1. Prove that the expectation of sample mean ((())) is the population mean (((()))) and the variance of sample mean is (((()))), where (((()))) is population variance and ((())) is the sample size.
2. For a distribution with unknown mean ((()))) has variance equal to ((())). Use central limit theorem to find how large a sample should be taken from the distribution in order that the probability will be at least 0.95 that the sample mean will be withing 0.5 of the population mean.
3. The life time of a certain brand of an electric bulb may be considered a random variable with mean 1200 hours and standard deviation 250 hours. Find the probability using central limit theorem, that the average life-time of 60 bulbs exceeds 1400 hours.
4. State the Lindberg-Levy Central Limit Theorem.
5. Define Central Limit Theorem. Write few applications of Central Limit Theorem.

Block-1

UNIT 3:

Unit Structure:

3.1 INTRODUCTION

The theory of statistical inference is based on sampling theory for making inferences about a population. The primary aim of sampling is to study the features of a population or to estimate the values of its parameter(s). It may be pointed out that it is possible to get reliable information about a population on the basis of sample information even if nothing is known about it.

Estimation of population parameters by means of sample statistic is one of the important problems of statistical inference. This is often unavoidable and economic also for business dicisions and research studies. Thus, we can define the term estimation as follows -

Estimation : It is a procedure by which sample information is used to estimate the numerical magnitude of one or more parameters of the population. A function of sample values is called on estimator (or statistic) which its numerical value is called an estimate. For example $\bar{x}$ is an estimator of population mean . On the other hand, if =50 for example, the estimate of population mean is said to be 50.

3.2 OBJECTIVES

This unit is an attempt to have the basic ideas of Estimation. After going through this unit you will be able to –

* know the concept of estimation
* have the knowledge of point estimation and interval estimation
* explain the characteristics of a good estimator
* understand the techniques of solving practical problems

3.3 THEORY OF ESTIMATION

Let X be a random variable with probability density function or probability mass function , where $\theta_1, \theta_2, ... \theta_k$ are k parameters the population.

Suppose, a random sample $(x_1, x_2, ..., x_n)$ of size n is drawn from the population and we are to estimate the k parameters $\theta_1, \theta_2, ... \theta_k$. In order to be specific, let x be a normal variate so that its probability density function can be written as $N(x : \mu, \sigma)$. Here, we may be interested to estimate the value of μ or .

It should be noted that, there may exist several estimators of a parameter, e.g., we can have any of the sample mean, median, mode, geometric mean, harmonic mean etc., as an estimator of population mean . Similarly, we can use either or $S = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$ as an estimator of population standard deviation σ.

This technique of estimation, where a single state like mean, median, Standard Deviation etc., is used as an estimator of population parameter, is known as Point Estimation. On the other hand, if an interval is estimated in which the value of the parameter is expected to lie, the procedure is termed as Confidence Interval.

Point Estimation :

There can be more than one estimators of a population parameter. So, it is necessary to determine a good estimator out of a number of available estimators. We know that, a function of random variables , is a random variable. Therefore, a good estimator is one whose distribution is

more concentrated around the population parameter. Thus, we may define point estimation as follows :

A particular value of a statistic which is used to estimate a given parameter is known as point estimate or estimator of the parameter.

According to R. A. Fisher, the founder of the theory, the following are some of the criteria of a good estimator :

(i) Unbaiasedness

(ii) Consistency

(iii) Efficiency

(iv) Sufficiency

(i) **Unbaiasedness** :

A statistic $t = t$ is said to be an unbiased estimator of a parameter θ if . If $E(t) \neq \theta$, then it is said to be a biased estimator of θ. The magnitude of bias = .

We have seen that, , where is the population mean, is said to be an unbiased estimator of the population mean μ. But, since , where $S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$ is an unbiased estimator of σ^2. On the other hand, since $E(s^2) = \sigma^2$, where $S^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$ is an unbiased estimator of σ^2. However, since $S^2 = \frac{n-1}{n}S^2$ and $\frac{n-1}{n}$ approximates 1 when n is large, say $n \geq 30$ for large sample, i.e., samples with size greater than or equal to 30, S can be taken as an estimator of σ.

It should be noted that, a statistic t is sai to be positively or negatively biased according as or <0, i.e.,

One should observe that the bias of an estimator usually decreases as the size of the simple increases.

(ii) **Consistency :**

A statistic $t_n - t_n(x_1, x_2, ..., x_n)$ is said to be a consistent estimator of a parameter θ if converges to θ in probability, i.e.,

i.e., $P\{|t_n - \theta| > E\} \longrightarrow 0$ as $n \longrightarrow \alpha$ for every $\in > 0$

It should be noted that, consistency is essentially a large sample property and strictly speaking it concerns not just one statistic, but a sequence of statistics.

We may note that is a consistent estimator of population mean because and $Var(\bar{x}) = \frac{\sigma^2}{n} \longrightarrow 0$ as $n \longrightarrow \alpha$

Note : An unbiased estimator is necessarily a consistent estimator.

(iii) Efficiency :

It is possible to get many unbiased consistent estimators of a parameter. In such a situation efficiency is the criterion that decides the goodness of an estimator. If there exist several consistent estimators for a parameter θ, then the one whose sampling variance is minimum is known as the *most efficient estimator*.

Let us consider, and t_2 be two estimators of a population parameter θ such that both are either unbiased or consistent. Now, is said to be more efficient estimator than t_2 if Var (t_1) < Var (t_2).

For example, the sample mean $\bar{x}$ and sample median M_e both can be used as an estimator of the population mean . We have seen that,

and $E(M_e) = \mu$, for large sample only. Again, $V(\bar{x}) = \frac{\sigma^2}{n}$ and $V(M_e) = \frac{\pi\sigma^2}{2n}$.

But, $V(\bar{x}) < V(M_e)$

So, sample mean $\bar{x}$ is more efficient than the sample median M_e.

(iv) Sufficeincy :

An estimator is said to be a sufficient estimator if it utilises all the information given in the sample about the parameter, i.e., a statistic

based on a sample drawn from a population having probability density function (p.d.f.) $f(x,\theta)$ is said to be a sufficient estimator of θ if it contains all information about the parameter , i.e., if the conditional distribution of for a given value of *t* is independent of θ, i.e., if does not depend on θ.

It is easy to observe that is a sufficient estimator of .

Sufficient estimators are the most desirable but are not very commonly available. The following points must be noted about sufficient estimators :

1. A sufficient estimator is always consistent.

2. A sufficient estimator is most efficient if an efficient estimator exists.

3. A sufficient estimator may or may not be unbiased.

Method of Points Estimation :

There are several methods of obtaining a point estimator of the population parameter. We shall, however, use the most popular method of maximum likelihood.

Let be a randowm sample of n independent observations from a population with probability density function (p.m.f) f (x; θ), where is unknown parameter for which we desire to find an estimator.

Since are independent random variable, their joint probability function or the probability of obtaining the given sample, termed as likelihood function, is given by

$$L = f(x_1;\theta).f(x_2;\theta).....f(x_n;\theta)$$

$$= \prod_{i=1}^{n} f(x_i;\theta)$$

We have to find the value of θ for which L is maximum. The conditions for maxima of L are :

.

The value of θ satisfying these conditions is known as Maximum Likelihood Estimator (MLE).

Generalising the above, if L is a function of k parameters , the first order conditions for maxima of L are :

$$\frac{\partial L}{\partial \theta_1} = \frac{\partial L}{\partial \theta_2} = ... = \frac{\partial L}{\partial \theta_k} = 0$$

From above, we will get k simultaneous equations in k parameters $\theta_1, \theta_2, ..., \theta_k$, and can be solved to get k maximum likelihood estimators.

In most cases, it is convenient to work using logarithm of L. Since log L is a monotonic transformation of L, the maxima of L and maxima of log L occur at the same value.

Example : Let a random sample of n observations $x_1, x_2, ..., x_n$ be drawn from a normally distributed population.

(i) If mean is unknown and the variance is known, find the maximum likelihood estimate of the mean,

(ii) if the mean is known but the variance is unknown, find the maximum likelihood estimate of the variance.

Solution : (i) Here $f(x_i, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

We have,

$$L = f(x_1, \mu) \bullet f(x_2, \mu) \bullet ..... \bullet f(x_n, \mu)$$

$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\sum \frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\Rightarrow \log_e L = -\frac{n}{2} \log_e (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

Differentiating partially w.r.t. μ , we get

$$\frac{\partial L}{\partial \mu} = \frac{\partial \theta_1}{\partial \theta_1} \frac{\partial \theta_2}{\partial \theta_2} \ldots \frac{\partial \theta_k}{\partial \theta_k} \frac{1}{\sigma^2} \sum (x_i - \mu)$$

The likelihood equation for estimating μ is

$$\Rightarrow \sum (x_i - \mu) = 0$$
$$\Rightarrow \mathrm{E}x_i - n\mu = 0$$
$$\Rightarrow \mu = \frac{\sum x_i}{n}$$
$$\Rightarrow \mu = \bar{x}$$

Thus, the maximum likelihood estimate of μ is the sample mean.

(ii)

Now, $\mathrm{L} = f(x_1, \sigma^2) f(x_2, \sigma^2), \ldots, f(x_n, \sigma^2)$

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

$$\Rightarrow \log_e \mathrm{L} = -\frac{n}{2}\log_e(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum (x_i - \mu)^2$$

Differentiating partially w.r.t. σ^2, we get

$$\frac{1}{\mathrm{L}}\frac{\partial \mathrm{L}}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum (x_i - \mu)^2$$

The likelihood equation for estimating σ^2 is

$$\frac{\partial}{\partial \sigma^2}(\log_e \mathrm{L}) = 0$$
$$\Rightarrow \frac{1}{\mathrm{L}}\frac{\partial \mathrm{L}}{\partial \sigma^2} = 0$$
$$\Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum (x_i - \mu)^2 = 0$$
$$\Rightarrow \sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

Thus, sample variance defined by $S^2 = \frac{\sum (x_i - \mu)^2}{n}$ is an estimator of σ^2.

Example : A random sample of size 5 is taken from a population containing 100 units. If the sample observations are 10, 12, 13, 7, 18, find

(i) an estimate of the population mean

(ii) an estimate of the standard error of sample mean

i.e. $\hat{SE}_{\bar{x}} = \frac{S}{\sqrt{n-1}}$ *for* SRSWR $= \frac{S}{\sqrt{n-1}}\sqrt{\frac{N-n}{N-1}}$ *for* SRSWOR

Now, let us prepare the following table

x	x^2
10	100
12	144
13	169
7	49
18	324
$\sum x = 60$	$\sum x^2 = 786$

$\hat{\mu} = 12$ $\quad \hat{SE}_{\bar{x}} = \frac{3.633}{\sqrt{5-1}}$

$$\therefore \bar{x} = \frac{\sum x}{n} = \frac{60}{5} = 12$$

$$S^2 = \frac{1}{n}\sum x^2 - \bar{x}^2 = \frac{786}{5} - 12^2$$

$$= 157.20 - 144$$

$$= 13.20$$

$$= (3.633)^2$$

Hence we have

for SRSWR

$$= \frac{3.633}{\sqrt{5-1}} \cdot \sqrt{\frac{100-5}{100-1}} \text{ for SRSWOR}$$

Example : A random sample of size 65 was taken to estimate the man annual income of 100 lower income families and the mean and standard deviation were found to be Rs. 6300 and Rs. 9.50 respectively. Find the standard error of the sample mean if sampling was done without replacement.

Solution : Since population is finite and sampling is drawn without replacement hence S.E. of $\bar{x}$ is given by

$$\hat{S.E.}_{\bar{x}} = \frac{\hat{\theta}}{S.E.} = \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Here S = 9.5, N = 100, n = 65

$$\therefore \hat{S.E.}_{\bar{x}} = \text{Rs.}\left[\frac{9.5}{\sqrt{65}} \times \sqrt{\frac{1000-65}{1000-1}}\right]$$

$$= \text{Rs.}\left[\frac{9.5}{8.06} \times \frac{30.58}{31.61}\right]$$

$$= \text{Rs.}1.14$$

Interval Estimation :

Instead of estimating a parameter θ by a single value, we may consider an interval of values which is supposed to contain the parameter . An interval estimate is always expressed by a pair of unequal real values and the unknown parameter lies between these two values. Hence, an interval estimation may be defined as specifying two values that contains the unknown parameter on the basis of a random sample drawn from the population in all probability.

On the basis of random sample drawn from the population characterised by an unknown parameter , let us find two statistics and t_2 such that

$$p(t_1 < \theta) = \alpha_1$$
$$p(t_2 > \theta) = \alpha_2$$

for any two small positive quantities α_1 and α_2.

Combining these two conditions, we may write

$p(t_1 \le \theta \le t_2) = 1 - \alpha$ where $\alpha = \alpha_1 + \alpha_2$

where α is called the level of significance. The interval [, t_2] within which the unknown value of the parameter θ is expected to lie is called the confidence interval, the limits and t_2 so determined are known as confidence limits 1– α is called the confidence level of confidene coefficient. The term 'confidence interval' has its origin in the fact that if we select =0.05 , then we feel confident that the interval [, t_2], would contain the parameter θ in (1–) % or (1–0.005)% or 95% of cases and the amount of confidence is 95%. This further means that if repeated samples of a fixed size are taken from the population with the unknown parameter , then in 95% of the cases, the interval [, t_2] would contain θ and in the remaining 5% of the cases, it would fail to contain

Computation of confidene interval:

Let us assume that we have taken a random sample of size n from a normal population with mean and standard deviation . We assume farther that the population standard deviation is known i.e. its value is specified. We know that the sample mean is normally distributed with mean and standard deviation = .

$$P(\bar{x} - p \times \text{S.E.}_{\bar{x}} \le \mu \le \bar{x} + p \times \text{S.E.}_{\bar{x}}) = 1 - \alpha$$

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

If the assumption of normality is not tenable, then also the sample mean follows normal distribution approximatily, statistically known as asymptotically, with population mean μ and standard diviation as , provided the sample size n is sufficiently large. If the sample size n>30, then the asymptotic normality assumption holds. In order to select the appropriate confidence interval to the population mean, we need to determine a quantity p, say, such that

which finally leads to

$$\phi(p) = 1 - \frac{\alpha}{2}$$

Choosing as 0.05, we have

$$\phi(p) = 1 - \frac{0.05}{2} = 0.975 = \phi(1.96)$$

$$\Rightarrow p = 1.96$$

Hence 95% confidence interval to μ is given by

Similarly, 99% confidence interval to μ is given by

Below we mention the confidence limits of some important statistics for large random samples.

* Confience limits for population proportion P: 95% confidence limits are : p $\pm$ 1.96 S.E.(p) 99% confidence limits are : p 2.58 S.E.(p)

* Confidence limits for that difference of two population means μ_1 and μ_2 :

95% confidence limits are : $(\bar{x}_1 - \bar{x}_2) \pm 1.96 \times \text{S.E.}(\bar{x}_1 - \bar{x}_2)$

99% confidence limits are : $(\bar{x}_1 - \bar{x}_2) \pm 2.58 \times \text{S.E.}(\bar{x}_1 - \bar{x}_2)$

* Confience limits for the difference $P_1 - P_2$ of two population proportion :

95% confidence limits are : $(p_1 - p_2) \pm 1.96$ S.E.$(p_1 - p_2)$

99% confidence limits are : $(p_1 - p_2) \pm 2.58$ S.E.$(p_1 - p_2)$

Example : Construct 95% and 99% confidence intervals for mean of a normal population.

Solution : Let $x_1, x_2, ..., x_n$ be a random sample of size n from a normal population with mean μ and standard deviation .

We know that sampling distribution of is normal with mean

and standard error .

will be a standard normal variate.

From the table of areas under standard normal curve, we can write

$$p(-1.96 \le z \le 1.96) = 0.95$$

or ———————— (A)

or $p\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{x}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}\right)=0.95$

Now, $-1.96\frac{\sigma}{\sqrt{n}} \le \bar{x}-\mu$

or $\mu \le \bar{x}+1.96\frac{\sigma}{\sqrt{n}}$ ———————— (1)

Similarly, $\bar{x}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}$

or $\mu \ge \bar{x}-1.96\frac{\sigma}{\sqrt{n}}$ ———————— (2)

Combining (1) and (2), we have

$$\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1.96\frac{\sigma}{\sqrt{n}}$$

Thus, we can write equation (A) as

$$p\left(\bar{x}-19.6\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)=0.95$$

$$\mu\, p\left(\bar{x}-2.58\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+2.58\frac{\sigma}{\sqrt{n}}\right)=0.99$$

$$p\left(-1.96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1.96\right)=0.95$$

This gives us a 95% confidence interval for the parameter μ.

Similarly, we can construct a 99% confidence interval for as

Example : A pharmaceutical company wants to estimate the mean life of a particular drug under typical weather conditions. A simple random sample of 81 bottles yields the following information :

Sample mean = 23 months

Population variance = 6.25 $(\text{months})^2$

Find an interval estimate with a confidence level of (i) 90% and (ii) 98%

Solution : Since the sample size n = 81 large, the mean life of the drug under consideration ($\bar{x}$) is asymptotically normal with population mean

and Standard Error = Standard diviation =

(i) Consulting Biometrika table, we find that

$$\phi(p) = 1 - \frac{\alpha}{2}$$
$$\Rightarrow \phi(p) = 1 - \frac{0.10}{2} = 0.95$$
$$\Rightarrow \phi(1.645) = 0.95$$
$$\Rightarrow p = 1.6450$$

$\therefore 90\%$ confidence interval for is

$$= \left[23 - 1.645 \times 0.2778,\ 23 + 1.645 \times 0.2778\right]$$
$$= \left[22.5430, 23.4570\right]$$

Example : A random sample of 100 days shows an average daily sale of Rs. 1000 with a standard deviation of Rs. 250 in a particular shop. Assuming a normal distribution, find the limits which have a 95% chance of including the expected sales per day.

Solution : As given, n = 100

= sample average sales =

s = sample standard deviation = Rs. 250

95% confidence interval to the expected sales per day ( ) is given by

Check Your Progress

1. What do you mean by estimation?
2. What are the criteria of a good estimator?
3. Distinguish between point estimation and interval estimation.

3.4 Summing Up

The theory of estimation is divided into two approaches namely point estimation and Interval estimation. In point estimation a single value of the statistic is used to provide an estimate of the parameter. On the other hand, in interval estimation, a range is specified within which the value of the parameter is most likely to lie with a known probability.

The characteristics of a good estimator under point estimation are unbiasedness, consistency, efficiency and sufficiency. There are several methods for obtaining the point estimates.

In interval estimation, we obtain the probable interval within which the unknown value of the parameter is expected to lie is called the confidence interval.

3.5 References and Suggested Readings

1. Hogg, Tanis, Rao; Probability and Statistical Inference; Pearson.
2. Bhuyan K.C.; Probability Distribution Theory and Statistical Inference; New Central Book Agency (P) Ltd.
3. Gupta S.C. & Kapoor V.K.; Fundamentals of Mathematical Statistics; Sultan Chand & Sons.
4. Hazarika P.; Essential Statistics for Economics and Commerce; Akansha Publishing House.
5. Rao Radhakrishna C., Linear Statistical Inference and its Applications; Wiley Eastern Limited.

3.6 Model Questions

1. The following observations constitute a random sample from an unknown population. Estimate the mean and S.D. of the population. Also find the S.E. of sample means : 14,19,17,20,25.
2. A random sample of the heights of 100 students from a large population of students in a university having S.D. of 0.75 ft. has an average height of 5.6 ft. Find (i) 95% and (ii) 99% confidence limits for the average height of all the students of the university.
3. What do you understand by point Estimation? When would you say that estimate of a paraneter is good? Explain briefly.
4. State and explain the principle of maximum likelihood (M.L.) for estimation of population parameter.
5. Discuss the concept of interval estimation and provide suitable example.
6. Explain the following terms:
   (i) Sufficient estimator
   (ii) Efficient estimator
   (iii) Maximum Likelihood estimator

BLOCK II : Unit-1
Testing of Hypothesis-an Introduction

Unit Structure:

1.1 Introduction

In many situations, we have to make decisions consulting only the sample observations. That is, inferences regarding popution characteristics are deduced on the basis of sample survey or sample information. But, inferences drawn such way are not free from the risk of errors due to sampling. Since a sample is only a part of the population, it may not be able to truly represent the entire population. In such situations, the estimated value of the populaton characteristic calculated from the sample may differ from the true value of that population characteristic. This difference is generally termed as sampling error. The problem of decision making arises here when one has to make decision on the basis of sample results even after knowing about sampling error. The modern theory of probability plays a vital role in this kind of decision making and the branch of Statistics that helps us in arriving at the criterion for decisions is known as testing of hypothesis. The theory of testing of hypothesis, initiated by J. Neyman and E.S. Pearson, involves the employment of different Statistical tools to arrive at decisions in certain situations where there is an element of uncertainty on the basis of a fixed size sample.

1.2 Objectives

After completion of the chapter, one will

- have the concept of testing of hypothesis
- know the definition of testing of hypothesis
- understand the steps involved in the process of hypothesis testing

1.3 Concept of Testing of Hypothesis

The testing of hypothesis is a technique by which we test the validity of a given statement about a population. Generally, this is done based on a random sample drawn from the parent population. In other words, it is a rule or procedure for deciding whether to accept or reject the hypothesis within an optimum risk.

Steps involved in the Process of Testing of Hypothesis

The logical steps involved in the process of Testing of Hypothesis are as follows :

1. Making a Formal Statement : The first step in the process of Testing of Hypothesis is to set up a formal statement regarding the characteristic of the concerned population. The statement should be so made that it relates to the nature of the research problem. Such formal statements are also called Null Hypothesis. A complementary statement (called Alternative Hypothesis) should also be set up against the Null Hypothesis.

2. Selection of the level of significance : The null hypothesis are tested for a specific value of level of significance. The level of significance is affected by number of factors like the size of the samples, difference between the sample means, the variability of measurements within samples, etc. In most cases, the level of significance is adopted either as 5% or 1% level.

3. Deciding the Sampling Distribution : After selection of the level of significance, we are to determine the appropriate sampling distribution for testing of hypothesis. The procedure of selecting the correct distribution are similar to those which are used in the context of estimation.

4. Random Sampling and computing the value of the Test Statistic : A sample should be drawn by using Random Sampling technique from the population under study. Further, required sample values are obtained from the drawn sample. The Test Statistic is calculated on the basis of these sample values.

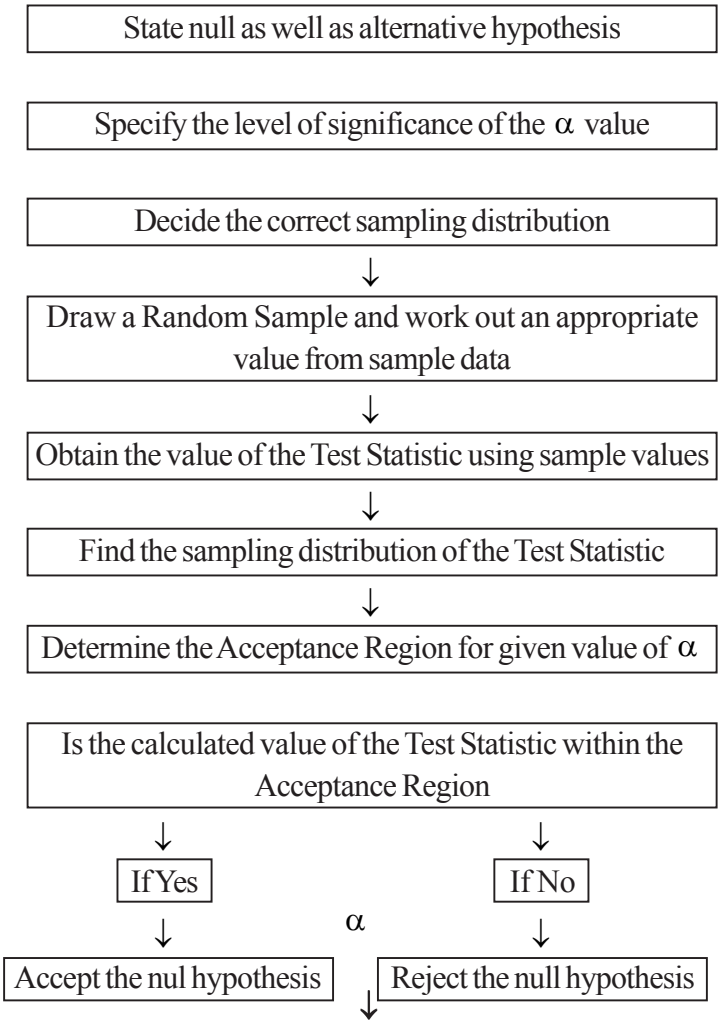
5. Finding the Sampling Distribution : After calculating the value of the test statistic, we find the sampling distribution of it under the null hypothesis (i.e. considering the null hypothesis true).

6. Finding the Acceptance Region : After determining the sampling distribution of the test statistic, we find the acceptance region for it considering the type of the test-one tailed test or two tailed test.

7. Drawing of Conclusion : If the value of the calculated test statistic falls within the acceptance region (determined for a certain level of significance) we consider that the calculated value is insignificant and the test has provided no evidences against the null hypothesis.

If the calculated value falls outside the acceptance region (i.e. in the rejection region) then it will be significant and then we reject our null hypothesis at the level of significance adopted.

Flowchart for Testing of Hypothesis



Check your progress

1. What do you mean by hypothesis testing?
2. Define level of significance
3. What are the errors in testing of hyposthesis ?

1.4 Summing Up:

In testing of hypothesis , we test an assumption regarding a population paramenter. For testing such an assumption sample data are used and decisions are made about population on the basis of sample information. The theory of testing of hypothesis is based on the concept of level of significance and the test statistic. Finanlly decision rules are followed for interpretation.

1.5 References and suggested reading:

1. Hogg, Tanis,Rao; Probability and statistical inference; Pearson
2. Bhuyan K.C; Probability Distribution Theory and Statistical Inference; New Cemtral Book Agency(P) Ltd.
3. Elhance D.N, Elhance Veena, Agarwar B.M,;Fundamentals of Statistics; Kitab Mahal
4. Gupta S.C.,Kapoor V.K.; Fundamentals of Mathematical Statistics ; Sultan Chand & Sons

1.6 Model Questions:

1. Describe the concept of testing of hypothesis.
2. Explain the procedure of testing of hypothesis briefly.
3. How do you set up a suitable significance level? Explain.
4. What are the decision rules to be followed during testing of hypothesis.
5. Distinguish between one tailed and two tailed test.

BLOCK II : Unit-2
TYPES OF HYPOTHESES

Unit Structure

2.1 Introduction

When a researcher observes some known facts and takes up a problem for analysis, he has to start with some assumptions regarding the population under study. Such assumption are often called hypothesis. The researcher has to proceed further on the basis of this hypothis and the facts that are already known. Formulation of an appropriate hypothesis is so crucial for any kind of research. Without having a valid hypothesis, investigations can't be carried out in right direction. Generally, such hypotheses are tested to evaluate possible intepretation about the population characteristics. Since the hypotheses reflect a generalised proposition regarding the population, we must take adaquate care in framing the hypothesis scientifically.

In this unit we will discuss different kinds of hypotheses with examples.

2.2 Objectives

After going through this unit, you will

- know the different types of hypotheses
- have the concept of different types of hypothesis with their definitions
- understand the examples of different types hypotheses

2.3 Types of Hypotheses

There are several types of hypotheses used for research activities. They are discussed below–

2.3.1 Statistical Hypothesis

A hypothesis is a preconceived idea or assumption or statement about the nature of a population or about the value of its parameters. Such a hypothesis (which may or may not be true) about a population parameter that is testable on the basis of the evidence from a random sample is called a statistical hypothesis. The procedure by which we test the validity of a given statistical hypothesis is termed as Testing of Hypothesis or Tests of Hypothesis. Following are different types of hypothesis :

2.3.2 Null Hypothesis

The hypothesis to be tested is termed as Null Hypothesis and it is denoted by the symbol H_0. This hypothesis asserts that there is no (significant) difference between the statistic and the population

parameter under consideration. For example, if we want to test the mean of a particular population, i.e. , for a specified value of a statistic , then the null hypothesis will be $H_o : \mu =$, which means that there is no difference between the population mean μ and the specified value .

2.3.3 Alternative Hypothesis

Any hypothesis other than null hypothesis is known as Alternative Hypothesis and it is denoted by the symbol H_1. For example, if the null hypothesis is $H_o: \mu =$, the possible alternative hypothesis may be as follows :

$H_1: \mu$

or, $H_1 : \mu >$

or, $H_1 : \mu <$

2.4 Simple and Composite Hypothesis :

Any statistical hypothesis that completely specifies tha distribution of the concerned population is called a simple hypothesis. Otherwise if the hypothesis doesnot completely specify the concerned distribution, then it is known as composite hypothesis. For example, for a normal distribution N (μ ,) with δ^2 known, the two hypothesis Ho : $\mu = 50$, against the alternative H_1: =20 are simple hypothesis. But, it alternative hypotheses are given like as follows–

H_1: >20 or H_1: <20

then such hypotheses are termed as composite hypothesis as they are not specifying the value of the population mean .

Check Your Progress

1. What do you mean by Statistical hypothesis ?
2. How are the null hypothesis constructed?
3. Distinguish between null and alternative hypothesis.

2.5 Summing Up

There can be several types of hypothesis. But in the context of testing of hypothesis we use only the Statistical Hypothesis. The two hypothesis in a statistical test are normally termed as - Null hypothesis and Alternative hypothesis. Both the hypothesis are very useful tool in testing the significance of difference between the statistics and the population parameter under consideration. After testing the null hypothesis, we make decision whether to accept or reject the null hypothesis

2.6 References and suggested reading:

1. Hogg, Tanis,Rao; Probability and statistical inference; Pearson
2. Bhuyan K.C; Probability Distribution Theory and Statistical Inference; New Cemtral Book Agency(P) Ltd.
3. Elhance D.N, Elhance Veena, Agarwar B.M,;Fundamentals of Statistics; Kitab Mahal
4. Gupta S.C.,Kapoor V.K.; Fundamentals of Mathematical Statistics ; Sultan Chand & Sons

2.7 Model Questions:

1. What is Statistical hypothesis? Describe briefly
2. State the similarities and differences between null and alternative hypothesis.
3. How to construct the null and alternative hypothesis. Describe with example
4. Distinguish between simple and composite hypothesis
5. Define Statistical Hypothesis. State the different types of statistical hypothesis

BLOCK II : Unit-3
Errors in Testing of Hypotheses

Unit Structure:

3.1 Introduction

The decision of acceptance or rejection of a null hypothesis regarding a population parameter is made on the basis of sample drawn from the parent population. That is why an element of risk of taking wrong decisions, i.e. an element of uncertainly is always involved in making such decisions. Such probable errors are, therefore, categorised and tried to keep minimum during the process of testing of hypothesis. The two distinct types of errors are Type I error and Type II error. In deciding whether to accept or reject a null hypothesis, both the errors play a vital role. In this unit we shall discuss the definitions of these errors and related terms used in testing of hypothesis.

3.2 Objectives

After completion of this unit, you will

- know the definitions of different types of errors
- have the concept of level of significance and critical region
- understand the distinction between parametric and non parametric test.

3.3 Type I error and Type II error :

When we reject a true null hypothesis, the error is called as the type I error. The probability of committing such an error is denoted by the symbol α. Thus,

$= Pr$ (Type I error)

$= Pr$ (Rejecting Ho when Ho is true)

Again, when we accept a false null hypothesis, the error is called as the type II error. The probability of committing this error is denoted by the symbol . Thus,

$=Pr$ (Type II error)

$=Pr$ (Accepting Ho when Ho is false)

In Statistical Quality Control (SQC) terminology Type I error is known as producer's risk whereas Type II error is known as concumer's risk.

3.4 Level of Significance

The maximum probability with which a true null hypothesis (Ho) is rejected (i.e. committing a type I error) in the test procedure is termed as the Level of Significance. Generally, it is denoted by the symbol of type I error, i.e. by . Thus,

Level of Significance = Pr (Type I error)
= Pr (Rejecting Ho when Ho is true)
=

Usually the value of is taken as either 5% (i.e. =0.05) or 1% (i.e. =0.01).

=5%=0.05 means that in 5% of the total number of samples, each of the same fixed size, that can be drawn from a population we are likely to reject a correct Ho. In other words, when we consider level of significance i.e. =5%, there are 5 cases in 100 that we would reject the correct null hypothesis. This implies that we are 95% confident that we have made the right decision in rejecting the null hypothesis and accepting the alternative hypothesis. Similarly, we can interpret other levels of significance.

3.5 Critical Region

The set of values of the test statistic that lead to the rejection of the hypothesis is called the critical region or the rejection region or the region of significance. On the other hand, the set of values which lead to the acceptance of the hypothesis is termed as the acceptance region. Generally, the critical region corresponds to the predetermined value of the level of significance and the acceptance region corresponds to 1– .

3.6 One-Tailed and Two-Tailed Test

In the procedure of testing hypothesis, the test will be a one-tailed or two-tailed depends entirely on the nature of the alternative hypothesis H_1. Supp
In this case, if the alternative hypothesis adopted is
tailed test. More specifically, for $H_1: \mu >$ the test i
it will be termed as a left-tailed test. In this case, t
shown as the shaded area in the following figure :

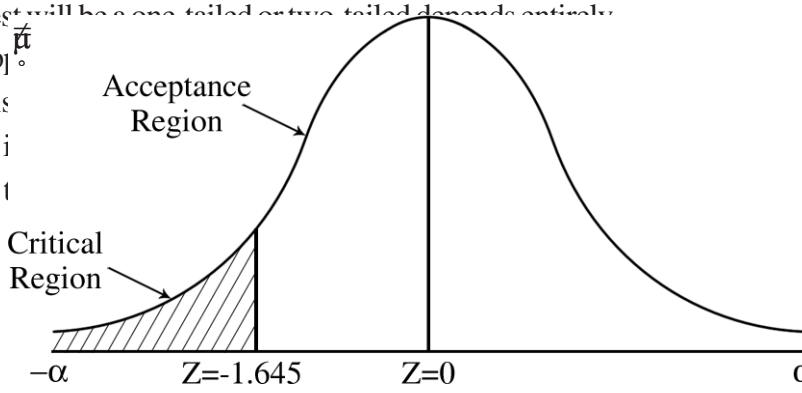


Fig : Left-tailed test

Again, for the null hypothesis Ho: $\mu =$, if the alternative hypothesis is $H_1: \mu$, then the test will be called a two-tailed test, Here, for α =0.05 and for the standard normal variate Z.

Suppose we have considered the level of significance =0.05. Then for a standard normal variate Z, under right tailed test we have, P[Z 1.645]=1–0.05=0.95. In this case we reject Ho if the computed value of Z from a sample lies outside the interval (1.645,) and accept it otherwise. The region for the interval (1.645,) is known as the critical region and is shown in the following figure :

Similarly, for a left-tailed test we have, P[–1.645 ≤ Z)=1–0.05=0.95 we have.

P[–1.96 Z<1.96]=1–0.05=0.95. For such a test we reject Ho if the computed value of Z from a sample lies outside the interval (–1.96, 1.96). This the critical region or rejection rejion will lie both side of the tail as shown in the following figure :

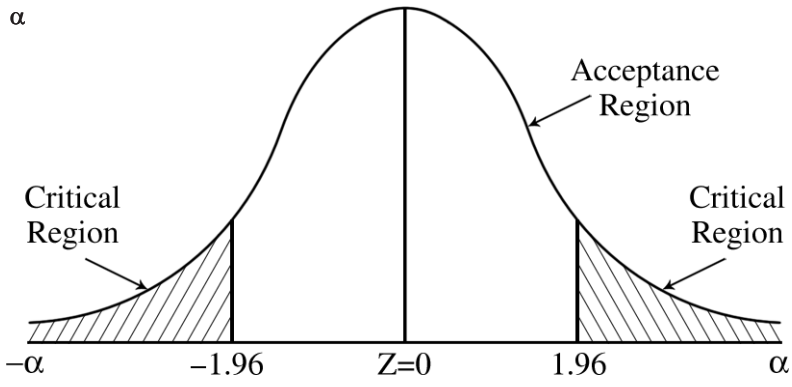


Fig : Two-tailed test

Critical values of Z for various levels of significance are summarized in the following table :

Level of Significance	1%	5%
Critical value for two tailed test	$\|Z_\alpha\|=2.58$	$\|Z \ \|=1.96$
Critical value for one tailed test	$\|Z \ \|=2.33$	$\|Z \ \|=1.645$

3.7 Test of Hypothesis

There are different types of tests of hypotheses for the purpose of testing the hypotheses. The tests of hypotheses can be classified into two categories. They are :

1. Parametric tests or standard tests of hypotheses
2. Non-Parametric tests of distribution-free test of hypotheses

Parametric Test

Any statistical test that makes assumption about the parameters (defining properties) of the parent population distribution(s) from which we draw samples is usually called parametric test. The typical assumptions made are :

- Normality : Data have drawn from a normal distribution
- Homogeneity of variances : Data drawn from different populations have the same variance.
- Linearity : Data have a linear relationship
- Independence : Data are independent each other

Almost all of the most commonly used statistical tests are based on the above assumptions. The following are some important parametric tests :

- Z-test or Large Sample test
- t-test
- –test
- F-test

Non-Parametric Test

In contrast to parametric tests, non-parametric tests do not require any assumptions about the parameters or about the nature of population. In other words, a statistical test used in the case of non-metric independent variabables, is called non-parametric test. When an investigator has no idea regarding the population parameter and the data concerned are strongly non-normal, such tests are adopted to test statistical hypothesis. Moreover, most of the non-parametric tests are applicable to data measured in an ordinal or nominal scale. Following are some of the important non parametric tests:

- The Runs test for Randomness
- The Median test for Randomness
- Wilcoxon Signed Rank test
- The Matched-Pairs Sign test
- Wilcoxon Matched-Pairs Signed Rank-sum test
- Mann-whitney Wilcoxon test

Check Your Progress

1. What are the different types of errors in testing of hypothesis ?
2. What are producer's risk and consumer's risk ?
3. Define Critical Region .

3.8 Summing Up:

In testing of hypothesis, it is aimed to reduce both the types of error,i.e. Type-I and Type -II. But, it is not possible to reduce both the errors together due to fixed samples size. The reduction of one type of error leads to an increase in the other types of error. As the Type -I error is considered to be more dangerous ,we keep the probability of committing Type-I error at a certain level , which is called the level of significance. Generally , the level of significance is denoted by ….. Further, critical values of the

statistic for one -tailed or two- tailed test for a certain level of significance are obtained for any decision procedure. In this process of obtaining the critical values , the concept of critical Region and Acceptance Region arises.

3.9 References and suggested reading:

1. Hogg, Tanis,Rao; Probability and statistical inference; Pearson
2. Bhuyan K.C; Probability Distribution Theory and Statistical Inference; New Cemtral Book Agency(P) Ltd.
3. Elhance D.N, Elhance Veena, Agarwar B.M,;Fundamentals of Statistics; Kitab Mahal
4. Gupta S.C.,Kapoor V.K.; Fundamentals of Mathematical Statistics ; Sultan Chand & Sons
5. Bhardwaj R.S.; Business Statistics ; Excel Books
6.Choudhury L, Sarma R, Deka M, Gogoi S.J.; An Introduction to Statistics; L. Choudhury.

3.10 Model Questions:

1.What are the two types of errors associated with testing of hypothesis . Explain them with examples
2. Define and discuss the Level of Significance.
3. Distinguish between one tailed and two tailed test.
4. What do you mean by Level of Significance? Explain
5. Clarify the underlying concept of Critical Region anD Acceptance Region.

BLOCK II : Unit-4
Parametric Test

Unit Structure:

4.1 Introductions

Parametric tests are usually based on certain properties of the parent population from which samples are to be drawn. The basis assumptions are–

- Normality : Under this assumption we assume that the data drawn from a normal distribution.
- Homoscedasticity : In case of Parametric tests involving more than one population we assume that each population has equal variance.
- Independence : Here we assume that data in each group are randomly and independently drawn from the population.

The following are some important parametric tests :

- Large Sample Test or Z-test
- Student's t-test or t-lest
- Snedecore's F-test or F-test
- Chi-square test

All these test are based on the assumption of normality, i.e. the source of data is considered to be normally distributed. In some cases, the population may not follow normal distribution, yet the tests are applicable on account of the fact that most of the samples drawn and their sampling distributions closely approach to normal distribution.

4.2 Objectives

After completion of this unit you will

- have the concept of parametric test
- understand the definitions of different parametric test
- learn the technique of solving practical problems using different parametric test

4.3 Large Sample Test or Z test

This test is based on the normal probability distribution and is used for large samples i.e. the samples that are greater than equl to 30. The assumptions made in such test are

1. The sampling distribution of the sample statistic follows normal distribution, and
2. the sample values are sufficiently close to the population value and hence can be used in its place for calculating the estimate of the standard error.

If t is a statistic calculated from any sample, then the test statistic under this test is given as :

Some of different types of Z-test are discussed below :

Z-test for a specified mean

Suppose a sample is drawn from a normal population. To test the null hypothesis, Ho: $\mu = \mu_o$ (specified) against the alternative H_1 $\mu \neq \mu_o$ (two-tailed test) the test statistic used is

$$Z_{cal} = \left| \frac{\bar{x} - \mu}{\delta / \sqrt{n}} \right| \sim N(0,1)$$

The decision rule would be :

Reject Ho at 5% (say) level of significance if $Z_{cal} > Z_{tab}$ i.e. 1.96. Otherwise, there is no evidence against Ho at this level significance.

The tabulated value or critical value of Z i.e. Z_{tab} may be different according to the nature of the test (two-tailed test or one-tailed test) and the level of significance considered during the test procedure.

Example : The mean life of 100 electric bulbs produced by a company is found to be 1570 hours with a standard deviation of 120 hours. If μ is the mean life time of all the bulbs produced by the company, test whether =1600 hours or not. Test your hypothesis at 5% level of significance.

Solution : Here the null hypothesis to be tested is

Ho: =1600 hours

against the alternative

H_1: 1600 hours

We are given,

n= sample size = 100

= sample mean = 1570 hours

s= sample s.d.=120 hours

Now, under the null hypothesis, the test statistic is

$$= \left| \frac{1570 - 1600}{120 / \sqrt{100}} \right|$$, considering sample s.d. as the estimated value of population s.d. δ

=

The table value of Z at 5% level of significance is 1.96.

Conclusion : Since the calculated Z>1.96, the tabulated value of Z, we may reject the null hypothesis at 5% level of significance and conclude that probably the population mean is different from 1600 hours.

Z-test for specified proportion

Let us consider a random sample of size n($n \geq 30$) out of which x number of observations possessing a certain attribute.

Now, is the sample proportion of observations possessing the attribute.

If we want to test the hypothesis that the population proportion P has a specified value P_o, the null hypothesis to be tested would be

H_o: $P = P_o$

against the alternative

H_1 : $P \neq P_o$

The test statistic under this null hypothesis is given as

$$|Z| = \left|\frac{p - E(p)}{S.E.(p)}\right| = \left|\frac{p - P}{\sqrt{\frac{PQ}{n}}}\right| \sim N(0,1)$$

Where Q=1–P

Comparing the calculated value of Z with the tabulated value we make the decisions for different types test and for different level of significance as in the previous case.

Example : A coin is tossed 900 times and heads appear 490 time. Does the result test the hypothesis that the coin is unbiased? Test the hypothesis at 1% level of significance.

Solution :

Let P denotes the population proportion of heads.

Here the null hypothesis to be tested is,

H_o : the coin is unbiased, i.e. $P=\frac{1}{2}=0.5$ against the alternative.

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

We are given,

n= sample size = 900

x= No. of heads appeared = 490

p= sample proportion of heads = $\frac{490}{900}$ = 0.54

Now, considering the null hypothesis to be true (or under the null hypothesis), the test statistic is

$$|Z| = \left|\frac{p - E(p)}{S.E.of\ (p)}\right|$$

$$= \left| \frac{p - P}{\sqrt{\frac{P(1-P)}{n}}} \right| \sim N(0,1)$$

$$= \left| \frac{0.54 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{900}}} \right| = \frac{0.04}{\sqrt{0.00278}}$$

=2.39

The table value of Z at 1% level of significance (for two tailed test) is 2.58 i.e. $Z_{0.01}$=2.58.

Conclusion : Since the calculated value of Z

=2.39<2.58, the table value of Z, we may accept the null hypothesis at 1% level of significance and conclude that probably the coin is unbiased.

Z-test for difference of two Means :

Let a random sample of size n_1 is drawn from a population having population mean μ_1 and standard deviation . Again, let an another sample of size n_2 is drawn from a population having population mean and same standard deviation . If and $\overline{x_2}$ are the sample means of the two random samples respectively and we want to test the null hypothesis, Ho: $\mu_1 = \mu_2$, against the alternative H_1: $\mu_1 \neq \mu_2$, the test statistic under the null hypothesis is given as

$$Z = \left| \frac{\left(\overline{x_1} - \overline{x_2}\right) - E\left(\overline{x_1} - \overline{x_2}\right)}{S.E.\left(\overline{x_1} - \overline{x_2}\right)} \right| \sim N(0,1)$$

$$= \left| \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}} \right| \text{ (after simplification)}$$

The decision rule is same as the previous cases of Z test.

Note : (i) If $\delta_1^2 = \delta_2^2 = \delta^2$ (say) then

$$Z = \left| \frac{\overline{x_1} - \overline{x_2}}{\delta\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \sim N(0,1)$$

(ii) If δ_1^2 and δ_2^2 are unknown they can be replaced by the sampling variance s_1^2 and s_2^2 respectively (since n_1 and n_2 are large). In that case

$$Z = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right| \sim N(0,1)$$

Example : In order to compare the scooters of two well-known companies, say, A and B in respect of efficiency in petrol mileage, a random sample of 50 scooters has been selected from each company and following results are obtained :

Company A : Mean mileage = 32.5 mile/litre
S.D. = 4.5 mile/litre
Company B : Mean mileage = 34.8 mile/litre
S.D. = 5.6 mile/litre

Test whether there is any significant difference between the two brands in respect of mean mileage.

Solution : Let μ_1 and denote the mean mileage (population mean) of all scooters of companies A and B respectively.

Here the null hypothesis to be tested is

Ho: =

against the alternative

H_1=

We are given,

Mean mileage of company A= =32.5
S.D. of company A=s_1=4.5
Mean mileage of company B=$\bar{x}_2$ =34.8
S.D. of company B=s_2=5.6
Sample size of company A=sample size of
Company B, i.e. $n_1=n_2=50$
Under the null hypothesis Ho, the test statistic is

$$|Z| = \left| \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E.}(\bar{x}_1 - \bar{x}_2)} \right| = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}} \right| \sim N(0,1)$$

$$= \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right|$$, considering sample S.D.S are estimated value of the population S.D.S

$$= \left| \frac{32.5 - 34.8}{\sqrt{\frac{(4.5)^2}{50} + \frac{(5.6)^2}{50}}} \right| = \frac{2.3}{1.016} = 2.26$$

The table value of Z at 5% level of significance is 1.96.

Conclusion : Since the calculated value of Z=2.26>1.96, the table value of Z, we may reject the null hypothesis at 5% level of significance and conclude that probably the mean mileage of the two brangs of scooters are different.

Z-test for difference of two Proportions :

Let n_1 and n_2 be the number of individuals in two samples selected from populations I and II respectively. Let x_1 and x_2 be the number of individuals possissing a certain attribute in the samples drawn.

Let, P_1 and P_2 be the population proportion of individuals possessing the certain attribute in population I and II respectively.

Now, suppose we want to test the null hypothesis, Ho:$P_1=P_2=P$ (say against the alternative $H_1: P \neq P_2$.

Under the null hypothesis the test statistic given as

Z= $\left| \frac{p_1 - p_2}{S.E.(p_1 - p_2)} \right|$, where p_1 = sample proportion of the first sample

$$= \frac{x_1}{n_1}$$

$$= \left| \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \sim N(0,1)$$ p2 = sample proportion of the 2nd sample

$$= \frac{x_2}{n_2}$$

Where Q =1–P

The conclusion is done in usual manner of comparing the calculated and tabulated value of Z.

Note : If the value of P is not known, it is estimated by

$$\hat{P} = p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

and $\hat{P} = 1 - p = q$

In this case,

$$Z = \left| \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \sim N(0,1)$$

Example : A company has the head office at Kolkata and a branch at Guwahati. The personnel director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for the purpose. Out of a sample of 500 workers at Kolkata 62% favoured the new plan. At Guwahati out of a sample of 400 workers 41% were against the new plan. Is there a significant difference between the two group in their attitude towards the new plan at 5% level of significance?

Solution : Let P_1 and P_2 be the population proportion in Kolkata and Guwahati respectively who prefer the new plan.

Here the null hypothesis to be tested is,

Ho:$P_1=P_2$

against the alternative

$H_1: P_1 \neq P_2$

We are given,

n_1= sample size of workers of Kolkata=500

n_2= sample size of workers of Guwahati= 400

p_1= sample proportion of workers of Kolkata who favours the new plan = 62%=0.62

p_2= sample propertion of workers of Guwahati who favours the new plan = (100–41%)
= 59%=0.59

Now, P =

$$= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{500 \times 0.62 + 400 \times 0.59}{500 + 400} = \frac{546}{900} = 0.607$$

q = 1–p=1–0.6079=0.393

Now, under the null hypothesis Ho, the test statistic is

$$|z| = \left| \frac{p_1 - p_2}{S.E.(p_1 - p_2)} \right| = \left| \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \sim N(0,1)$$

$$= \left| \frac{0.62 - 0.59}{\sqrt{0.607 \times 0.393\left(\frac{1}{500} + \frac{1}{400}\right)}} \right|$$

$$= \frac{0.03}{0.0327} = 0.917$$

The table value of Z at 5% level of significance is 1.96.

Conclusion : Since the calculated value of Z=0.917<1.96, the table value of Z, we may accept the null hypothesis at 5% level of significance and conclude that probably there is no significant difference between the two groups of workers in their attitude towards the new plan.

Student t-test (small sample Test)

In case of small samples i.e. when n<30, we have to use the concept of a new distribution known as students-t distribution. Here population is considered to follow normal distribution whose S.D. δ is not known. The distribution is basically based on the concept of degrees of freedom.

Th degrees of freedom of a set of observations is the number of values which could be chosen independently with the specification of the system. For example, if a variable x assumes n different values and k different linear restrictions are imposed on the values of x, then the degrees of freedom, denoted by , will be =n–k.

In general, k is considered to be 1 in most cases due to the linear restriction. In that case, degrees of freedom =n–1.

It should be noted that the critical value (or table value) of the statistic t at a specified level of significance vary with the degrees of freedom of the distribution.

Application of t-distribution :

Student-t distribution has a large number of applications in statistics some of which are enumerated below :

(i) to test for a single population mean.

(ii) to test the difference between two population means.

(iii) to test the significance of observed correlation co-efficient.

(iv) To test the significance of observed regression co-efficient.

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

t-test for the single population mean :

Suppose a random sample of size n (n<3) is drawn from a Normal population with unknown mean and variance and $\bar{x}$ is the sample mean. Now, to test the null hypothesis that the population mean has a specified mean when population S.D. is unknown, i.e.

Ho: =

against the alternative

H_1=

the test statistic under the null hypothesis is given by,

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} xi$

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

Which is an unbiased estimate of the population variance δ^2

We reject Ho if calculated value of t greater than the tabulated value of t. Otherwise we accept the Ho at the specified level of significance.

Example : A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025cm. A random sample of 10 washers was found to have an average thickness of 0.024cm. with a S.D. of 0.002cm. Test the significance of deviation.

(Given, table value of t at 5% level of significance for 9 degrees of freedom is 2.262 i.e. $t_{0.05,9} = 2.262$)

Solution : Let μ denotes the average thickness of population of washers.

Here the null hypothesis to be tested is,

Ho: =0.025cm

against the alternative

H_1: 0.025cm

We are given,

n=sample size=10

= sample mean = 0.024cm

s=sample s.d.= 0.002cm

Now, under the null hypothesis the test statistic is

$$t = \left|\frac{\bar{x} - \mu}{s/\sqrt{n-1}}\right| \sim t_{0.0\alpha,(n-1)} = \left|\frac{0.024 - 0.025}{0.002/\sqrt{10-1}}\right|$$

$$= \frac{0.003}{0.002} = \frac{3}{2} = 1.5$$

Degrees of freedom=n–1=10–1=9

The table value of t at 5% level of significance for 9 d.f. i.e. $t_{.05,9} = 2.262$

Conclusion : Since the calculated value of t=1.5<2.262, the table value of t, we may accept the null hypothesis at 5% level of significance and conclude that probably the deviation is not significant.

t-test for the difference of two population Means

Let $\bar{x}_1$ be the sample mean of a random sample of size n_1(<30) drawn from a population with mean μ_1 and vareance .

Let $\bar{x}_2$ be an another sample mean of a random sample of size n_2(<3) drawn from a population with mean μ_2 and variance .

If we want to test the null hypothesis that the two population means are equal i.e. Ho: $\mu_1 = \mu_2$

against the alternative

H_1:

then the test statistic (assuming that) under this Ho is given by

$$t = \left| \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \sim t_{0.0\alpha,(n_1+n_2-2)}$$

Where, $s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$

and s_1^2, s_2^2 are the sample variances of the first and second sample respectively

Here, s^2 is considered as an unbiased estimate of δ^2, i.e. $E(s^2) = \delta^2$

The table value of t is obtained for (n_1+n_2-2) d.f. for α% of level of significance.

If the calculated value of t is greater than the tabulated one, we reject the null hypothesis at % level of significance. Otherwise we accept the Ho.

Example : Two samples of 6 and 5 items respectively gave the following results :

Mean of the first sample=40

S.D. of the first sample=8

Mean of the second sample=50

S.D. of the second sample=10

Is the difference of the means significant? (Given)

Solution : Let μ_1 and denote the population means of the first and second population respectively from which samples are drawn.

Here the null hypothesis to be tested is

Ho: =

against the alternative

H_1:

We are given,

$$n_2 = 5, \bar{x}_2 = 50, s_2 = 10$$

Now, $s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$

$$= \sqrt{\frac{6 \times 8^2 + 5 \times 10^2}{6 + 5 - 2}} = \sqrt{98.22} = 9.91$$

Under the null hypothesis, the test statistic is

$$|t| = \left| \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \text{ where } s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

$$= \left| \frac{40 - 50}{9.91\sqrt{\frac{1}{6} + \frac{1}{5}}} \right|$$

$$= \frac{10}{6} = 1.67$$

The degrees of freedom=$n_1+n_2-2=5+6-2=9$

The table value of t at 5% level of significance for 9 d.t. is 2.26.

Conclusion : Since the calculated value of t=1.67<2.26, the table value of t, we may accept the null hypothesis at 5% level of significance and conclude that probably the population means are same.

Chi-Square (λ^2) test

Chi-Square test, also called pearson's chi-square test, is a statistical test applied to sets of categorical data to test whether there is any significant difference between two sample results. The test statistic of this test follows a chi-square distribution that tends to approach normality as the number of degrees of freedom increases.

Application of Chi-Square (λ^2) test

Some of the application of Chi-square test are enumerated below :

(i) to test the hypothetical value of the population variance

(ii) to test the goodness of fit,

(iii) to test the independence of attributes

Chi-Square (λ^2) test to test the hypothetical value of the population variance

Suppose a random sample of size n is drawn from a normal population. If we want to test the null hypothesis that population variance $\delta^2 = \delta_0^2$ (say) i.e. $Ho : \delta^2 = \delta_0^2$ (say) against the alternative,

$H_1 : \delta^2 \neq \delta_0^2$

then the test statistic, under this Ho is given by,

$$\lambda^2 = \sum_{i=1}^{n} \frac{(x_1 - \bar{x})^2}{\delta_0^2} \sim \lambda^2_{0.0\alpha,(n-1)}$$

We reject Ho if calculated λ^2 is greater then tabulated one. Otherwise we accept Ho.

Chi-Square test for test the goodness of fit

The chi-square goodness of fit test is used to find out how the observed value of a given phenomena is significantly different from the expected value.

Let $0_1, 0_2, - - - 0_n$ be a set of observed frequencies and $E_1, E_2, ----, E_n$ be the corresponding set of expected frequencies. Here the null hypothesis to be tested is,

Ho: there is no significant difference between observed and expected frequencies.

The alternative hypothesis would be.

H_1: there is significant difference between observed and expected frequencies.

Under the null hypothesis Ho, the test statistic is given as

$$\lambda^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \lambda^2_{(n-1)}$$

To apply Chi-Square test, the following conditions should be satisfied :

(1) N, the total frequency should be reasonably large, say, greater than 50

(2) $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i = N$

(3) No expected cell frequency should be less than 5. If any expected frequency is less than 5, it should be pooled with the adjacent frequency. The degrees of freedom lost due to pooling should be adjusted accordingly.

Note : The test statistic may be simplified as follows :

$$\lambda^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$= \sum_{i=1}^{n} \frac{O_i^2 - 2O_iE_i + E_i^2}{E_i}$$

$$= \sum_{i=1}^{n} \frac{O_i^2}{E_i} - 2\sum_{i=1}^{n} O_i + \sum_{i=1}^{n} E_i$$

$$= \sum_{i=1}^{n} \frac{O_i^2}{E_i} - N \left[\because \sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i = N \right]$$

Example : Following values of observed and expected frequencies obtained for a distribution.

x :	5	10	15	20	25	30	35	40
Observed freq :	2	2	6	13	15	23	16	13
Expected freq. :	1	3	6	12	17	20	17	13
x :	45	50						
Observed freq. :	6	4						
Expected freq. :	7	4						

Test if the fit is good.

Solution : Here the null hypothesis to be tested is,

Ho : there is no significant difference between the observed and the expected frequency against the alternative.

H_1 : there is significant difference between the observed and the expected frequency.

Let us now prepare the following table :

X	O_i	E_i	O_i^2 / E_i
5	2	1	
10	2 } 10	3 } 10	10.00
15	6	6	
20	13	12	14.08
25	15	17	13.23
30	23	20	26.45
35	16	17	15.06
40	13	13	13.00
45	6	7	5.14
50	4	4	4.00

$$\sum O_i = 100 \quad \sum E_i = 100 \quad \sum \frac{O_i^2}{E_i} = 100.96$$

$$\therefore \lambda^2 = \sum_{i=1}^{n} \frac{O_i^2}{E_i} - N$$

$$= 100.96 - 100 = 0.96$$

The table value of λ^2 at 5% level of significance for 10–2–2=6 d.f. is 12.592.

Conclusion : Since the calculated $\lambda^2 < 12.592$, the tabulated value we may accept the null hypothesis at 5% level of significance and conclude that probably there is no significant difference between observed and expected frequency.

Chi-Square test for independence of attributes :

Suppose the observations be classified according to two attributes A and B where A is divided into 'm' classes namely $A_1, A_2, -----, A_m$ and. B is divided into 'n' classes namely $B_1, B_2, ----, B_n$. Let $(A_i B_{j)}$ denote the number of persons possessing the attribute A_i and B_j. [i=1, 2, ----m; j=1, 2, ----n]. Now, the table showing saved frequencies in different categories having m-rows and n-columns is called a $m \times n$ contingency table.

A \ B	B_1	B_2	------	B_j	------	B_n	Total
A_1	(A_1B_1)	(A_1B_2)	------	(A_1B_j)	------	(A_1B_n)	(A_1)
A_2	(A_1B_1)	(A_2B_2)	------	(A_2B_j)	------	(A_2B_n)	(A_2)
⋮	⋮	⋮		⋮		⋮	⋮
A_i	(A_iB_1)	(A_iB_2)	------	(A_iB_j)	------	(A_iB_n)	(A_i)
⋮	⋮	⋮		⋮		⋮	⋮
A_m	(A_mB_1)	(A_mB_2)	------	(A_mB_j)	------	(A_mB_n)	(A_m)
Total	(B_1)	(B_2)	------	(B_j)	------	(B_n)	N

Here, (A_i) = the number of persons possessing the attribute A_i

= $(i=1, 2, -----, m)$

(B_j) = the number of persons possessing the attribute B_j

$= \sum_{i=1}^{n}(A_iB_j)$ $(j=1, 2, ----, n)$

$\sum_{i=1}^{m}(A_i) = \sum_{j=1}^{n}(B_j) = N$, total frequency $\sum_{j=1}^{n}(A_iB_j)$

Now, we are to test whether the attributes A and B are independent or not. The null hypothesis to be tested would be

Ho : the two attributes are independent against the alternative

H_1: the two attributes are not independent under the null hypothesis, the test statistic is given as

$$\lambda^2 = \sum \frac{(O-E)^2}{E} \sim \lambda^2_{(m-1)(n-1)}$$

$$= \sum \frac{O^2}{E} - N \text{ (simplified form)}$$

Where the expected frequency (E) of any cell is obtained as

$$E = \frac{\text{Row total} \times \text{column total}}{N}, \text{ N=Grand total}$$

The table value of λ^2 is obtained for specified level of significance with d.f. $\gamma = (m-1)\times(n-1)$

If the calculated value of is greater than the tobulated one, we reject the null hypothesis at the specified level. Otherwise we accept the Ho.

Example : Two sample poles votes for two candidates A and B for a public office are taken, one from among residents of urban areas, and the other from residents of rural areas. The results are given below. Examine whether the nature of the area is related to voting preference in this election.

Votes for / Area	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

(Given, $\lambda^2_{0.05,1} = 3.84$)

Solution : Here the null hypothesis to be tested is

Ho : there is no association between the nature of area and voting preference against the alternative

H_1: there is association between the nature of area and voting preference.

Now, we prepare the following table :

Observed frequency O	Expected frequency E	$(O-E)^2$		$\frac{(O-E)^2}{E}$
620	585	1225	χ^2	$\frac{1225}{585} = 2.10$
550	585	1225		$\frac{1225}{585} = 2.10$
380	415	1225		$\frac{1225}{415} = 2.95$

$\sum O = 2000 \qquad \sum E = 2000 \qquad \lambda^2 = \sum \frac{(O-E)^2}{E} = 10.10$

$$\left[\begin{array}{l}\text{Formula to obtain exp ected frequency:} \\ \text{Expected frequency} = \dfrac{\text{Row total} \times \text{column total}}{\text{Grand total}}\end{array}\right]$$

Under the null hypothesis, the test statistic is

$$\lambda^2 = \sum \frac{(O-E)^2}{E} = 10.10$$

Degrees of freedom= $(r-1) \times (c-1) = (2-1) \quad (2-1) = 1$

The table value of for 1 d.f. at 5% level of significance, i.e. $\lambda^2_{0.05,1} = 3.84$

Conclusion : Since the calculated value of $\lambda^2 > 3.84$, the table value of λ^2, we may reject the null

hypothesis at 5% level of significance and conclude that probably the nature of area and voting preference are associated.

Example :

A sample survey conducted in a region regarding educational qualification of 100 men and women reveals the following information :

Sex \ Education	Middle school	High School	College	Total
Male	10	15	25	50
Female	25	10	15	50
Total	35	25	40	100

Test if the educational qualification depend on sex.

Solution : Here the null hypothesis to be tested is,

Ho : educational qualification is independent of sex against the alternative

H_1 : educational qualification is dependent of sex

Now we prepare the following table :

Observed frequency O	Expected frequency E	$\frac{O^2}{E}$
10	$\frac{50\times35}{100}=17.5$	
25	$\frac{50\times35}{100}=17.5$	
15	$\frac{50\times25}{100}=12.5$	
10	$\frac{50\times25}{100}=12.5$	
25	$\frac{50\times40}{100}=20$	
15	$\frac{50\times40}{100}=20$	
$\sum O=100$	$\sum E=100$	$\sum\frac{O^2}{E}=109.93$

$\therefore \lambda^2 = \sum \frac{O^2}{E} - N$, N=total frequency

= 109.93–100

= 9.93

Degrees of freedom=(3–1)(2–1)=2

The tabulated value of λ^2 at 5% level of significance for 2 d.f. is 5.99.

Conclusion : Since the calculated value of $\lambda^2 > 5.99$, the tabulated value, we may reject the null hypothesis at 5% level of significance and conclude that probably educational qualification depend on sex.

F-test (F-Distribution)

The F-distribution is a continuous probability distribution. An F-variate is defined by the ratio of two independent chi-square variables divided by their respective degrees of freedom. If λ_1^2 and λ_2^2 are two independent chi-square variates with n_1 and n_2 degrees of freedom respectively, then F is defined as

$$F = \frac{\lambda_1^2 / n_1}{\lambda_2^2 / n_2}$$

Which follows F-distribution with (n_1, n_2) degrees of freedom.

Application of F-test

Some of the application of F-test are enumerated below :

(1) to test the equality of two population variances

(2) to test the equality of several population means

(3) to test the linearity of regression

(4) to test the significance of observed correlation ratio

(5) to test the significance of observed multiple correlation coefficient

F-test for the equality of two population variances

Let a random sample of size n_1 is drawn from a normal population with variance δ_1^2. Let an another sample of size n_2 is drawn from a normal population with variance δ_2^2. To test whether the two population variances δ_1^2 and δ_2^2 are equal, we construct the null hypothesis as

Ho : $\delta_1^2 = \delta_2^2 = \delta^2$ (say)

against the alternative

H_1 : $\delta_1^2 \neq$

Under this Ho, the test statistic is given as

$$F=\frac{S_1^2}{S_2^2};\left(S_1^2 > S_2^2\right) \sim F(n_1-1, n_2-1)$$

Here S_1^2 and S_2^2 are unbiased estimates of the common population variance δ^2 and are given by

$$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\bar{x})^2$$

and $S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(y_i-\bar{y})^2$

If the calculated F is greater than the tabulated value, we reject the null hypothesis at the specified level of significance. Otherwise we accept the null hypothesis.

Note : The greater of the two mean squares S_1^2 and S_2^2 is taken in the numerator of the F-statistic. This implies that if $S_2^2 > S_1^2$, the F-statistic would be

$$F=\frac{S_2^2}{S_1^2} \sim F_{(n_2-1,\ n_1-1)}$$

Example : Following results are obtained from two independent samples :

	Size	Mean	Sum of squares of deviation from mean
Sample =	9	68	36
Sample =	10	69	42

Test if the population variances are equal.

Solution : Here, the null hypothesis to be tested is

Ho : the population variances are equal i.e.

$\delta_1^2 = \delta_2^2$

against the alternative

$\delta_1^2 \neq \delta_2^2$

We are given,

$n_1=9, \quad \bar{x}=68,$

$n_2=10, \quad \bar{y}=69,$

$$\therefore \quad = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\bar{x})^2 = \frac{36}{8} = 4.50$$

$$S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(y_j-\bar{y})^2$$

$$= \frac{42}{9} = 4.66$$

Under the Ho, the test statistic is given by

$$F = \frac{S_2^2}{S_1^2} \quad \left[\because S_2^2 > S_1^2\right]$$

$$= \frac{4.66}{4.50} = 1.03$$

Here the degrees of freedom is (n_2-1, n_1-1)

$= (10-1=9, 9-1=8) = (9, 8)$

The table value of F at 5% level of significance for (9, 8) d.f. is 3.4.

Conclusion : Since the calculated value of F<3.4, the tabulated value of F, we may accept the null hypothesis at 5% level of significance and conclude that probably the population variances are same.

Example : Two sources of raw materials are under consideration by a company. Both sources sum to have similar characteristics, but the company is not sure about their respective uniformity. A sample of ten lots from source A yeilds a variance of 225, and a sample of eleven lots from source B yeilds a variance of 200. Is it likely that the variance of source A is significantly greater than the variance of source B?

(Consider α=0.01)

Solution : Here, the null hypothesis to be tested is $\delta_1^2 = \delta_2^2$

Ho : the variance of sources A and B are equal i.e.

against the alternative

$H_1 : \delta_1^2 > \delta_2^2$

We are given,

$n_1 = 10,\ S_1^2 = 225$

$n_2 = 11,\ S_2^2 = 200$

$$\therefore \quad = \frac{n_1}{n_1 - 1} S_1^2 = \frac{10}{10-1} \times 225 = 250$$

and $S_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{11}{11-1} \times 200 = 220$

Under Ho, the test statistic is

$$F = \frac{S_1^2}{S_2^2} \qquad \left[\because S_1^2 > S_2^2\right]$$

$$= \frac{250}{220} = 1.14$$

Here, the degrees of freedom is (n_1-1, n_2-1)

$= (10-1, 11-1)=(9, 10)$

The table value of F at 1% level of significance for (9, 10) d.f. is 4.94.

Conclusion : Since the calculated value of $F<4.94$, the tabulated value of F, we may accept the null hypothesis at 1% level of significance and conclude that probably the two population variances are equal.

BLOCK II : Unit-5
Non-Parametric Tests

Unit Structure:

5.1 Introduction

The tests based on the assumption that the samples are drawn from a normally distributed population and the population parameters are partially known or at least estimable have been discussed in the previous unit of Parametric Tests. But there are some situations where such assumptions are not tenable. To over come this limitation we have to adopt an another technique of hypothesis testing called Non-Parametric method.

In this method no assumptions about the parameters or about the nature of the population require. That is why, sometimes this method is also called distribution free method. Non-Parametric methods of estimation and tests often depend on ordered sample and also order statistics.

A random sample $(x_1, x_2, ...x_n)$ is said to be an ordered sample if $x_1<x_2<.....<x_n$. Any statistic defined on the basis of the order of a sample is called an order statistic. e.e. median, quartile, semi-inter quartile range etc.

5.2 Objectives

After going through this unit, one will

- know the definition of Non-Parametric test
- understand the different types of Non-Parametric test
- learn the technique of solving practical problems using different Non-Parametric tests.

5.3 Non-Parametric Tests of Hypothesis

As defined earlier, non-parametric tests donot depend on the distribution of the sampled population. That is why, such tests are also called 'distribution-free tests'. Also, non-parametric methods focus on the location of the probability distribution of the parent population, rather than on specific parameters of the population. Before discussing the different non-parametric tests, let us discuss the advantages and disadvantages of non parametric tests.

Advantages

1. Non-parametric tests require less restrictive assumptions as compared to parametric test.
2. These tests often require very few arithmetic computations.
3. There is no alternative to use a non parametric test if the data are available in ordinal or nominal scale.
4. Non-parametric tests are useful with small samples.

Disadvantages

1. Parametric tests are more powerful than non parametric tests. This means that there is a greater risk of accepting a false hypothesis.
2. Non parametric methods deal with test of hypothesis only. No non parametric method of estimation is available.

The Paired-Sample Sign Test

This test is used to test for consistent differences between pairs of observations. The test has very important application in problems involving paired data such as data relating to the responses of mother and daughter towards ideal family size, weight of patients before and after treatment, etc. In such problems, each pair of sample values can be replaced with a + (plus) sign if the first value is greater than the second, a – (minus) sign if the first value is smaller than the second or be discarded if the two values are same.

Let, P be the proportion of plus signs, i.e.

$$P = \frac{\text{Number of plus signs}}{\text{Total number of pairs}}$$

Let, P be thi proportion of plus sign in the population.

Then the null hypothesis to be tested is

Ho : P=0.5

If the difference is due to chance effects the probability of a + sing for any particular pair is $\frac{1}{2}$, as is the probability of a – sign. If S is the number of times the less frequent sign occurs, then S has the binomial distribution with P=$\frac{1}{2}$.

$$Z = \frac{k - np}{\sqrt{npq}} \sim N(0,1)$$

The table value for a two-tailed test at 5% level of significance can be conveniently found by the expression

$$K = \frac{(n-1)}{2} - (0.98)\sqrt{n}$$

Where n is the sample size minus number of tied pairs.

Ho is rejected if $S \leq K$ for the sign test

In case of large samples (generally considered n>25) then the binomial distribution can be approximate with normal distribution with

Where K= the number of most frequently occuring signs

q=1–p

We compare the calculated value of Z with the table value of Z and draw conclusion as before. The approximation becomes better when a correction for continuity is employed and then

$$Z = \frac{(k \pm 0.5) - 0.5n}{0.5\sqrt{n}} \sim N(0,1)$$

Here, k+0.5 is used when $k < \frac{n}{2}$ and k–0.5 is used when $k > \frac{n}{2}$.

Conclusions are drawn in the similar manner.

Example : A group of 30 people have started physical exercise to reduce their body weight. Their body weights (in kg) are recorded before (x) and after (y) the physical exercise. The data are given below :

x : 56, 58, 62, 57, 56, 60, 64, 66, 67, 62, 64,

y : 55, 58, 60, 56, 56, 58, 62, 60, 63, 60, 63,

x : 67, 68, 55, 70, 62, 61, 64, 60, 58, 57, 64,

y : 65, 66, 58, 70, 63, 62, 62, 61, 55, 56, 63,

x : 52, 56, 58, 57, 60, 62, 64, 67

y : 54, 57, 55, 60, 58, 61, 63, 66

Is there any change in body–weights due to physical exercise?

Solution : Here, the null hypothesis to be tested is,

Ho : there is no change in body-weights due to physical exercise. i.e. P=0.5

against the alternative,

$H_1 : P \neq 0.5$

Here, $D_i = x_i - y_i$

$= + 0 + + 0 - + + + + + + + - 0 - -$

$+ - + + - - + - + + + +$

$n = 30-3=27$

and k =

$\therefore$ Under the null hypothesis, the test statistic is

$Z =$

$=$

$= \dfrac{(19 - 0.5) - 0.5 \times 27}{0.5\sqrt{27}}$

$=$

$= 1.92$

The table value of Z at 5% level of significance is 1.96

Conclusion : Since calculated value of Z<1.96, the tabulated value, we may accept the null hypothesis at 5% level of significance and conclude that probably there is no change in body-weight due to physical exercise.

Wilcoxon signed Rank-Sum test

The Wilcoxon signed rank sum test is an another kind of non-parametric test and can be used to test the null hypothesis that the median of distribution is equal to some value. Moreover, the test can be used in lieu of a one-sample t-test or a paired t-test.

We perform the following steps to carry-out the test.

Case I : Paired Data

1. State the null hypothesis– in this case
   Ho : the median difference, M=O
2. Calculate each paired difference,
   $d_i=x_i-y_i$, where (x_i, y_i) are the pairs of observations. (i=1, 2, ..., n)
3. Rank these differences in ascending order ignoring the signs (i.e. assign rank 1 to the smallest |di|, rank 2 to the next, etc)
4. The cases of tied ranks are assigned ranks by the average method.
5. Label each rank with its sign, according to the sign of di.
6. Calculate T_+, the sum of the ranks of the positive d_is, and T_-, the sum of the ranks of the negative d_is. (As a check the total, $T_+ + T_-$, should be equal to , where n is the number of pairs of observations in the sample)

Case II : Single of observations

1. State the null hypothesis – the median value is equal to some value M, i.e.
   Ho : Median = M
2. Calculate the difference between each observation and the specified value of median M, i.e.
   $d_i=x_i-M$
3. Apply steps 3–6 as above.

Under the null hypothesis, the test statistic, to be denoted by T, is obtained as follows :

6. Choose T=min (T_-, T_+)
7. Use tables of critical values for the Wilcoxon signed rank sum test to find the probability of observcing a value of T or more extreme. Most tables give both one-sided and two-sided p-values. If not, double the one-sided p-value to obtain the two-sided value.
   It can be shown that the distribution of T is approximately normal (when n>20) with mean

   and standard error

   Thus, the test statistic under the Ho would be,

$$Z = \left| \frac{T - \mu_T}{\delta_T} \right| \sim Z(0,1)$$

   We follow the same decision rule as before.

Dealing with ties :

These are two types of tied observations that may arise during the test procedure :

- Observations in the sample may be exactly equal to M (i.e. O in the case of paired differences). Ignore such observations and adjust n accordingly.
- Two or more observations/differences may be equal. If so, average the ranks across the tied observations and reduce the variance by $\frac{t^3 - t}{48}$ for each group of t tied ranks.

Example :

The following table shows the hours of relief provided by two analgesic drugs in 12 patients suffering from arthritis. Is there any evidence that the drug B provides longer relief than the drug A?

Patient :	1	2	3	4	5	6	7	8	9	10	11	12
Drug A :	2.0	3.6	2.6	2.6	7.3	3.4	14.9	6.6	2.3	2.0	6.8	8.5
Drug B :	3.5	5.7	2.9	2.4	9.9	3.3	16.7	6.0	3.8	4.0	9.1	20.9

Solution :

Here the null hypothesis to be tested is

Ho : there is no significant difference between the two drugs with respect to relief hour against the alternative.

Now, the differences between the two drugs with respect to relief hour are :

di ; (Drug B–Drug A) :

+1.5, +2.1, +0.3, –0.2, +2.6, –0.1, +1.8, –0.6, +1.5, +2.0, +2.3, +12.4

The ascending order of these differences (ignoring the sign) is given as,

0.1, 0.2, 0.3, 0.6, 1.5, 1.5, 1.8, 2.0, 2.1, 2.3, 2.6, 12.4

Now, let us rank these differences and label each rank with its sign according to the sign of di.

|di|	0.1	0.2	0.3	0.6	1.5	1.5
Rank	1	2	3	4	5.5	5.5
Sign	–	–	+	–	+	+

|di|	1.8	2.0	2.1	2.3	2.6	12.4
Rank	7	8	9	10	11	12
Sign	+	+	+	+	+	+

$\frac{n(n+1)}{2} = \frac{12 \times 13}{2} = 78$

$\therefore$ T_+ = sum of the ranks of the positive dis

= 3+5.5+5.5+7+8+9+10+11+12

= 71

and T_- = sum of the ranks of the negative dis

= 1+2+4

= 7

$T_+ + T_- = 71+7=78$

and

i.e. $T_+ + T_- =$

Now, $T=\min(T_-, T_+) = 7$

We can use a normal approximation in this case. Here we have 2 tied ranks, so we must reduce the variance by

$\therefore$ We get,

$$Z$$

$$= \left| \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{t^3 - t}{48}}} \right|$$

$$= \left| \frac{7 - \frac{12 \times 13}{4}}{\sqrt{\frac{12 \times 13 \times 25}{24} - 0.125}} \right|$$

$$= \left| \frac{7 - 39}{\sqrt{162.5 - 0.125}} \right| = 2.511$$

$$\frac{t^3 - t}{48} = \frac{2^2 - 2}{48} = 0.125$$

The table value of Z at 5% level of significance for one-tailed test is 1.64

Conclusion : Since the calculated value of Z>1.64, the table value of Z, we may reject the null hypothesis at 5% level of significance and conclude that probably the drug B provides longer relief than the drug A.

Mann–Whitney U-test

Mann-Whitney U-test is the non-parametric test that helps us to determine whether two random samples have come from identical population or not. Usually this test is used when the data are ordinal and when the assumptions of the t-test are not met. The basic assumption of the test is that the distribution of the two populations are continuous with equal standard deviation.

Let n_1 and n_2 be the sizes of the samples taken from population I and population II respectively. The steps to be carry out for the test are as follows :

1. State the null hypothesis– in this case, Ho : the population means are equal,

   i.e. $\mu_1 = \mu_2$

against the alternative

$$H_1 :$$

2. Rank all the n_1+n_2 observations, and arrange in ascending order.
3. Find R_1 and R_2, where R_i denotes the sum of ranks of the i^{th} sample (i=1, 2)

Now, for the sample sizes n_1 and n_2, the sum of R_1 and R_2 is simply the sum of first n_1+n_2 positive integers, which is given as

This formula enables us to find R_2 if we know R_1 and vice versa. We obtain the following two statistics to make decisions :

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

For small samples, if both n_1 and n_2 are less than 10 special tables (statistical tables for Mann–Whitney U test) must be used. If U is smaller than the critical value, Ho can be related to the standard normal curve by the statistic

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \sim N(0,1)$$

Where, $\frac{n_1 n_2}{2} = \mu_U$

$$\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \delta_U$$

$$U_1 = n_1 n_2 + \frac{(n_1 + n_2)(n_1 + n_2 - 1)}{2} - R_1$$

We will follow same decision rule as before.

Example :

Twenty-four applicants for a position are interviewed by three administrators and rated on a scale of 5 as to suitability for the position. Each applicant is given a 'suitability' score which is the sum of the three numbers. Although college education is not a requirement for the position, a personnel director felt that it might have some bearing on suitability for the position. Raters made their ratings on the basis of individual interviews and were not told the educational background of the applicants. Twelve of the applicants had completed at least two years of college. Use the Mann Whitney U-test to determine whether there was a difference in the scores of the two groups. Use a 0.05 level of significance, Group A had an educational background of less than two years of college, while group B had completed at least two years of college.

"Suitability" Scores

Group A	Group B
7	8
11	9
9	13

4	14
8	11
6	10
12	12
11	14
9	13
10	9
11	10
11	8

Conclusion : Here, the null hypothesis to be tested is

Ho : There is no significant difference in the scores of the two groups against the alternative there is significant difference in the scores of the two groups.

Common Ranks		Observations
1	————	4
2	————	6
3	————	7
5	————	8
5	————	8
5	————	8
8.5	————	9
8.5	————	9
8.5	————	9
8.5	————	9
12	————	10
12	————	10
12	————	10
16	————	11
16	————	11
16	————	11
16	————	11
16	————	11
19.5	————	12

$\frac{4+5+6}{3} = 5$

$\frac{14+8+9+10+12+14+13+9+10+11+11+8}{12} = \frac{129}{12} = 10.75$

$\frac{1+2+3+5+5+5+8.5+8.5+8.5+8.5+12+12+12+16+16+16+16+19.5}{18} = \frac{288}{18} = 16$

19.5 ———— 12

21.5 ———— 13

21.5 ———— 13
23.5 ———— 14
23.5 ———— 14

Now, we allocate the ranks to the corresponding observations of the two groups as follows :

Group A	Corresponding Ranks	Group B	Corresponding Ranks
7	3	8	5
11	16	9	8.5
9	8.5	13	21.5
4	1	14	23.5
8	5	11	16
6	2	10	12
12	19.5	12	19.5
11	16	14	23.5
9	8.5	13	21.5
10	12	9	8.5
11	16	10	12
11	16	8	5
	R_1=123.5		R_2=176.5

$$\frac{n_1(n_1+n_2+1)}{2} + \frac{n_2(n_1+n_2+1)}{2} = \frac{(12+12)(12+12+1)}{2} = 300 = R_1 + R_2$$

R_1+R_2=123.5+176.5=300

Again,

This serves as a check for internal consistency.

Now, for n_1=12, n_2=12, R_1=123.5 & R_2=176.5

We have,

$U_1 = n_1n_2 +$ $\qquad$ $-R_1$

$= 12 \quad 12 +$ $\qquad$ -123.5

$= 144+78-123.5$

$= 98.5$

and U_2 $= n_1 n_2 +$ $- R_2$

$= 12 \quad 12 +$ $- 176.5$

$= 144 + 78 - 176.5$

$= 45.5$

If the table value of U is available for $n_1=12$ and $n_2=12$, we may use it for decision making (Table : Critical Values of the Mann–Whitney U),

Otherwise, we may use the test statistic

Where $U = \min(U_1, U_2)$

$$= \left| \frac{45.5 - \frac{12 \times 12}{2}}{\sqrt{\frac{12 \times 12(12+12+1)}{12}}} \right|$$

$$= \left| \frac{45.5 - 72}{17.3205} \right|$$

$= 1.53$

$$Z = \left| \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 . n_2 (n_1 + n_2 + 1)}{12}}} \right| \sim N(0,1)$$

The tabulated value of Z at 5% level of significance is 1.96.

Conclusion : Since the calculated value of Z <1.96, the tabulated value of Z, we may accept the null hypothesis at 5% level of significance and conclude that probably there is no significant difference in the scores of the two groups.

Runs Test for Randomness

Run test of randomness is a non-parametric test that is used to know the randomness in data. It is alternative test to test auto-correlation in the data. Run test of randomness is basically based on the run. A run is defined as a sequence of identical symbols which are preceded and followed by different or no symbols at all. Run test of randomness assumes that the mean and variance are constant and the probability is independent.

Test Procedure

The first step in the runss test is to count the number of runs in the data sequence. For example, suppose that a sequence of two symbols, A and B, occured as follows :

ABAABABBBAAABBA

The number of runs in the above sequence are 9.

In should be noted that when there are n observation, where each is denoted by either symbol A

or by B, the possible number of runs would lie between and including 2 to n.

Now to test randomness of a sample, we set the null hypothesis as :

Ho : the sequence was produced in a random manner, i.e. the sample is a random.

against the alternative

H_1 : the sample is not random.

As per the null hypothesis, if we get a significantly large or small number of runs, Ho is rejected.

Now, to construct the test statistic, let us assume that n_1 be the number of symbol of one type and n_2 be the number of symbols of other type, so that $n=n_1+n_2$ is the total number of observations in the sequence.

Again, let R be the number of runs in the sequence.

Using theory of algebra, it can be shown that R is a random variable having mean $\mu_R = \frac{2n_1n_2}{n} + 1$ and standard error

For large sample runs test (when $n_1>10$ and $n_2>10$), the distribution of R can be approminated by a normal distribution.

In this case, the test statistic would be

$$Z = \left| \frac{R - \mu_R}{\delta_R} \right| \sim N(0,1)$$

$$\delta_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n}{n^2(n-1)}}$$

We reject Ho, if calculated Z>tabulated Z at α% of level of significance. Otherwise we accept Ho.

For a small sample runs test, there are tables to determine critical values that depend on values of n_1 and n_2.

Example :

The weights (gms) of 31 apples have been collected from a consignment and are as follows :

106, 107, 76, 82, 106, 107, 115, 93, 187, 95, 123, 125, 111, 92, 86, 70, 127, 68, 130, 129, 139, 119, 115, 128, 100, 186, 84, 99, 113, 204, 111

Test the hypothesis that the sample is random.

Solution : Let us denote the increase in the successive observation by a plus (+) sign and the decrease of successive observation by a minus (–) sing. From the given data, we can have the following sequence of plus (+) and minus (–) signs.

+ – + + + + – + – + + – – – – + – + – +

– – + – + – + + + –

Let, n_1 = the number of plus sign

n_2 = the num ber of minus sign

R = the number runs

Here, the null hypothesis to be tested is

Ho : the sample is random

against the alternative

H_1 : the sample is not random

From, the sequence of plus and minus sign, we have

$n_1=16, \quad n_2=14$

and R=20

$n = n_1+n_2=16+14=30$

Thus,

$$= \frac{2\times16\times14}{n}+1 = 15.93$$

and δ_R

$\therefore$ the test statistic is given as

$$= \left|\frac{20-15.93}{2.68}\right| = 1.52$$

$$Z = \left|\frac{R-\mu_R}{\delta_R}\right| \sim N(0,1)$$

$$\delta_R = \sqrt{\frac{2\times16\times14(2\times16\times14-30)}{900\times29}} = 2.68$$

The table value of Z at 5% level of significance is 1.96.

Conclusion : Since the calculated value of Z<1.96, the table value of Z, we may accept the null hypothesis at 5% level of significance and conclude that probably the sample is random.

5.4 Median test for randomness

Median test for randomness is a kind of non-parametric test. It is a test based upon the number of runs above and below the median of the sample. If the sample is random, the successive observations of the sample are expected to be above or below median and consequently the number of runs (R) would be large.

The test of hypothesis in this case would be a one tailed test. The null hypothesis to be tested is,

Ho : the sample is random i.e. number of runs $R \geq \mu_R$

against the alternative

H_1 : the sample is not random i.e. number of runs $R < \mu_R$

It can be shown that R follows normal distribution with mean = and standard error

, where n denotes the number of observations (excluding the observations that are equal to median value) in the sequences.

The test statistic is given as,

$$Z = \left| \frac{R - \mu_R}{\delta_R} \right| \sim N(0,1)$$

We follow the same decision rule as before,

Example :

The weights (gms) of 31 apples have been collected from a consignment and are as follows :
106, 107, 76, 82, 106, 107, 115, 93, 187, 95, 123,
125, 111, 92, 86, 70, 127, 68, 130, 129, 139, 119,
115, 128, 100, 186, 84, 99, 113, 204, 111
Test the hypothesis that the sample is random.

Solution : Let us arrange the observations in ascending order as follows :
68, 70, 76, 82, 84, 86, 92, 93, 95, 99, 100, 106, 106,
107, 107, 111, 111, 113, 115, 115, 119, 123, 125, 127,
128, 129, 130, 139, 186, 187, 204

$\therefore$ the median of the sequence is 111.

Let, L denotes that an observation is lower than median and H denotes that an observation is higher than it.

Then, the given sequence may be written in L and H in the following way :

L L L L L L H L H L H H L L L H L H H H H H H L H L L H H (ignoring the observations that are equal to median)

Thus, we have, R = the number of runs = 14

n = number of observations in the sequence (ignoring the observation that are equal to median)

= 29

Here, the null hypothesis to be tested is,

Ho : the sample is random i.e. against the alternative

Now, $\mu_R = \frac{n+2}{2} = \frac{29+2}{2} = 15.5$

and

$$\delta_R \cong \sqrt{\frac{n(n-2)}{4(n-1)}} = \sqrt{\frac{29 \times 27}{4 \times 28}} = 2.64$$

$\therefore$ the test statistics is given as,

Z

$$= \left| \frac{14 - 15.5}{2.64} \right|$$

$$= 0.568$$

The table value of Z at 5% level of significance for one tailed test is 1.645.

Conclusion : Since calculated Z<1.645, the tabulated value of Z, we may accept the null hypothesis at 5% level of significance and conclude that probably the sample is random.

Check Your Progress

1. What is Ordered sample?
2. Non Paramatic Tests are called distribution free tests. Why?
3. Point out the uses of the paired sample sign test and wilcoxon signed rank sum test.

5.5 Summing Up:

Non Paramatric tests are methods of statistical analysis thatb do not require the nature of the population. Moreover, no assumptions are made about the parameters and that is why, such tests are also termed as 'distribution-free tests'.Thee are several advantages and disadvantages of using such tests. The most popular non parametric test are namely ,the Paired-sample sign test, Wilcoxon Signed Rank-Sum Test, Mann Whitney U-Test, Runs Test for Randomness and Median test for randomness.

$$= \left| \frac{R - \mu_R}{\sigma_R} \right| \sim N(0,1)$$

The Paired Sample Sign Test is used to test for consistent differences between pairs of observations. Again the Wilcoxon Signed rank sum test is applied to test whether the median of a distribution is equal to a specific value or not. On the other hand, Mann-Whitney U-Test helps us to determine whether two random samples have the same population or not .Another two tests namely- Runs Test for randomness and Median Test for randomness are used to know the randomness in data.

5.6 References and suggested reading

1. Hogg, Tanis,Rao; Probability and statistical inference; Pearson
2. Bhuyan K.C; Probability Distribution Theory and Statistical Inference; New Cemtral Book Agency(P) Ltd.
3. Elhance D.N, Elhance Veena, Agarwar B.M,;Fundamentals of Statistics; Kitab Mahal
4. Gupta S.C.,Kapoor V.K.; Fundamentals of Mathematical Statistics ; Sultan Chand & Sons
5. Bhardwaj R.S.; Business Statistics ; Excel Books
6. Choudhury L, Sarma R, Deka M, Gogoi S.J.; An Introduction to Statistics; L. Choudhury.

5.7 Model Questions

1. What are non parametric tests? Explain Briefly
2. Point out the advantages and disadvantages of non parametric tests.

3. Define the paired sample sign test
4. Write the steps to carry out the Wilcoxon signed Rank Sum Test
5. What is Mann Whitney U -Test. How the test is carried out?
6. Discuss the tests used for randomness in data.

BLOCK III : Unit-1

Basic Concepts of Partial and Multiple Correlation and Regression

Unit Structure:

1.1 Introduction

So far we have confined our discussion to univariate distribution which involve only one variable . However, in business, the key to decision-making often lies in the understanding of the relationships between two or more variables. A distribution where each unit of the series assumes two values is called a Bivariate Distribution. For example, a company in the distribution business may determine that there is a relationship between the price of crude oil and its own transportation costs. A marketing executive might want to know how strong the relationship is between advertising dollars and sales dollars for a product or a company. In this chapter, we will study the concept of correlation and how it can be used to estimate the relationship between two variables using regression.

1.2 Objective

After going through this unit, you will be able to
Understand the concept of partial correlations
Understand the concept of multiple correlations
Understand the significance of multiple regression

1.3: Basic Concepts of partial correlation

1.3.1 Partial Correlation

The partial correlation coefficient describes the relationship between one of the independent variables and the dependent variable, given that the other independent variables are held constant

statistically. In partial correlation, the linear association between a dependent variable and one particular independent variable is studied, holding other independent variables constant.

If a dependent variable X_1 and two independent variables X_2 and X_3 are included in the partial correlation analysis, then the partial correlation between X_1 and X_2holding X_3 constant is denoted by $r_{12.3}$. Similarly, partial correlation betweenX_1 and X_3 holding X_2constant is denoted by $r_{13.2}$

Depending upon the number of independent variables which are held constant partial correlation coefficients are often called as zero-order, first-order, second-order correlation coefficients.

The partial correlation between X_1 and X_2 holding X_3 constant is determined by

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} \quad ;$$

Similarly partial correlation between X_1 and X_3 holding X_2 constant $r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}}$

and

Partial correlation between X_2 and X_3 holding X_1 constant is given by -

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

Again we have the following relations:

(a) $r_{12.3} = \sqrt{R^2_{12.3}} = \sqrt{1-\frac{S^2_{1.23}}{S^2_{1.3}}}$ *and* $r_{13.2} = \sqrt{R^2_{13.2}} = \sqrt{1-\frac{S^2_{1.23}}{S^2_{1.2}}}$

(b) $r_{12.3} = \sqrt{b_{12.3}}xb_{21.3}$;
$r_{13.2} = \sqrt{b_{13.2}}xb_{31.2}$ *and*
$r_{23.1} = \sqrt{b_{23.1}}xb_{32.1}$

Using these relations the partial correlation between X_1 and X_2 holding X_3 constant can also be determined as:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{13}^2}\left(\frac{S_1}{S_2}\right) \times \frac{r_{12} - r_{13}r_{23}}{1 - r_{13}^2}\left(\frac{S_2}{S_1}\right)$$

$$= \frac{\left(r_{12} - r_{13}r_{23}\right)^2}{\left(1 - r_{13}^2\right)\left(1 - r_{13}^2\right)}$$

Stop to consider

The partial correlation coefficient describes the relationship between one of the independent variables and the dependent variable, given that the other independent variables are held constant statistically. For e.g. the relationship between the yield of wheat and fertiliser when all other variables such as nature of soil, irrigation, climate, seed and techniques of cultivation are kept constant is termed as partial correlation.

Note

(1) $r_{xy}=r_{yx}$
(2) The value of the partial correlation coefficient lies between -1 and 1.

1.3.2 Partial Correlation Coefficient in four Variables

If there are four variables X_1,X_2,X_3 and X_4 under consideration for the joint study, then the partial correlation coefficient between X_1 and X_2 eliminating the influence of X_3 and $X_4$is given by

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{1 - r_{13.4}^2}\sqrt{1 - r_{23.4}^2}} \qquad (1)$$

Or alternatively

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}} \qquad (2)$$

Formula (1) and (2) are identical and hence whatever expression we take for $r_{12.34}$we shall get the same value . In the similar manner formula for other partial correlation coefficient can be calculated.$\mathbf{r}_{12.34}$is known as the second order partial correlation coefficient .

1.3.3 Partial Correlation Coefficient in four Variables

If there are four variables X_1,X_2,X_3 and X_4 under consideration for the joint study, then the partial correlation coefficient between X_1 and X_2 eliminating the influence of X_3 and $X_4$is given by

$$r_{12.34} = \frac{r_{12.4} - r_{13.4} r_{23.4}}{\sqrt{1 - r_{13.4}^2}\sqrt{1 - r_{23.4}^2}} \qquad (1)$$

Or alternatively

$$r_{12.34} = \frac{r_{12.3} - r_{14.3} r_{24.3}}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}} \qquad (2)$$

Formula (1) and (2) are identical and hence whatever expression we take for $r_{12.34}$we shall get the same value . In the similar manner formula for other partial correlation coefficient can be calculated.$\mathbf{r}_{12.34}$is known as the second order partial correlation coefficient .

.

Example 1 In a trivariate distribution it is found that r_{12}=0.70 , r_{13}=0.61 , r_{23}=0.40. Find the values of $r_{23.1}$, $r_{13.2}$ and $r_{12.3}$.

Solution: The partial correlation between X_2 and X_3 holding X_1 constant is determined by

$$r_{23.1} = \frac{r_{23} - r_{21} r_{31}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{31}^2}}$$

Substituting the given values we get

$$r_{23.1} = \frac{0.40 - 0.70 \times 0.61}{\sqrt{1 - (0.70)^2}\sqrt{1 - (0.61)^2}}$$
$$= \frac{0.40 - 0.427}{\sqrt{0.51}\sqrt{0.6279}}$$
$$= \frac{0.027}{0.714 \times 0.7924}$$
$$= \frac{0.027}{0.5657}$$
$$= 0.0477$$

Self Assessment Questions

1. Given the following values: r_{23}=0.4

r_{13}=0.61 , r_{12}=0.7

Find the partial correlation coefficients: $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

2. In a tri-variate distribution. it was foundr_{12}=0.75 , r_{13}= 0.9 , r_{23}=0.6

Find the values of $r_{12.3}$, $r_{23.1}$ and $r_{1.23}$

Exercise 2. The correlation between intelligence test scores and school achievement in a group of school students is 0.80. The correlation between intelligence test scores and age in the same age group is 0.70 and the score between intelligence test scores and age is 0.60. Find out the correlation between intelligence test scores and school achievement in children of the same age. Comment on the result.

Solution: Let x_1= intelligence test scores; x_2=school achievement; x_3= age of children

Given that $r_{12}=0.80$, $r_{13}=0.70$, $r_{23}=0.60$

Then the correlation between intelligence test scores and school achievement, keeping the influence of age as constant, is given by

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$$
$$= \frac{0.8 - 0.7 \times 0.6}{\sqrt{1-(0.7)^2}\sqrt{1-(0.6)^2}}$$
$$= \frac{0.8 - 0.42}{\sqrt{0.51}\sqrt{0.64}} = \frac{0.38}{0.57} = 0.667$$

Hence it can be concluded that intelligence test scores and school achievement are associated to each other to the extent of of $r_{12.3}=0.667$ while the influence of childrens' age is held constant.

Exercise 3. Given $r_{12.4}=0.60$, $r_{13.4}=0.50$, $r_{23.4}=0.70$ find $r_{12.34}$ and $r_{13.24}$

Solution:

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{1-r^2_{13.4}}\sqrt{1-r_{23.4}^2}}$$
$$= \frac{0.6 - (0.5 \times 0.7)}{\sqrt{\{1-(0.5)^2\}}\sqrt{\{1-(0.7)^2\}}}$$
$$= \frac{0.6 - 0.35}{\sqrt{0.75 \times 0.51}}$$
$$= \frac{0.25}{\sqrt{0.3825}}$$
$$= \frac{0.25}{0.62} = 0.403$$

$$r_{13.24} = \frac{r_{13.4} - r_{12.4}r_{23.4}}{\sqrt{1-r^2_{12.4}}\sqrt{1-r_{23.4}{}^2}}$$
$$= \frac{0.5-(0.6\times 0.7)}{\sqrt{\{1-(0.6)^2\}}\sqrt{\{1-(0.7)^2\}}}$$
$$= \frac{0.5-0.42}{\sqrt{0.64\times 0.51}}$$
$$= \frac{0.08}{0.57}$$
$$= 0.14$$

Example 4. Suppose a computer has found for a given set of variables X_1,X_2 and X_3 the correlation coefficients are $r_{12=0.91}$, $r_{13=0.33}$ and r_{23}= 0.81. Explain whether these computations may be said to be free from error.

Solution : For determining whether the given computations are correct or not , we calculate the value of the partial correlation coefficient $r_{12.3}$ for variables 1 and 2 keeping the influence of variable 3 constant . If the value of $r_{12.3}$ is less than one , then the computation may be said to be free from error .

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{12}{}^2}\sqrt{1-r_{23}{}^2}}$$
$$= \frac{0.91-(0.33\times 0.81)}{\sqrt{1-(0.33)^2}\sqrt{1-(0.81)^2}}$$
$$= \frac{0.91-0.2673}{\sqrt{1-0.1089}\sqrt{1-0.6561}}$$
$$= \frac{0.6427}{\sqrt{0.8911\times 0.3439}}$$
$$= 1.161$$

Since the calculated value of $r_{12.3}$ is more than one, the computation given in the question are not free from error.

1.4 Multiple Correlation – The aim of the theory of multiple correlation is to study the joint effect of a group of variables not included in the group. In general, the coefficient of multiple correlation measures the extent of the association between the dependent variable and several independent variables taken together. Thus while studying multiple correlation, the effect of certain independent factors on a dependent factor is studied without treating any factor constant.

A linear multiple correlation is denoted by the symbol R and the necessary subscripts are added to it. The subscript before the decimal represents the dependent variable and the subscripts after the decimal represent the independent variables which affect the dependent variable. For example $R_{1.23}$ indicates that the variable X_1 is associated with the variables X_2 and X_3 and the formula for it is:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

By symmetry we may also write

$$R_{1.23} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \text{ and } \quad R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

Note:

(1) The value of multiple correlation coefficient always lie between 0 and 1.
(2) If $r_{12}=r_{13}=0$ then $R_{1.23}=0$ implying no linear relationship between the variables.
(3) $R_{1.23}=R_{1.32}$
(4) If $R_{1.23}=1$, the correlation is called perfect

Exercise 5 In a three variate multiple correlation analysis, the following results were obtained $r_{12}=0.59$ $r_{13}=0.46$ and $r_{23}=0.77$. Find $R_{1.23}$

Solution: Multiple Correlation Coefficient is defined as

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.50)^2 + (0.46)^2 - 2(0.59 \times 0.46 \times 0.77)}{1 - (0.77)^2}}$$

$$= \sqrt{\frac{0.3481 + 0.2116 - 0.418}{0.4071}}$$

$$= \sqrt{\frac{0.5597 - 0.418}{0.4071}}$$

$$= \sqrt{\frac{0.1417}{0.4071}} = 0.589$$

Stop to consider

The multiple correlation may be defined as a statistical tool designed to measure the degree of relationship existing among three or more variables. it is denoted by the symbol R

$R_{1.23}$ stands for the coefficients of multiple correlation between X_1, X_2 and X_3. The subscript to the left stands for dependent variable while the subscripts to the right of the dot represent independent variables.

1.5 Multiple Regression:

Simple regression analysis is bivariate linear regression in which one dependent variable, y, is predicted by one independent variable, x. Examples of simple regression applications include models to predict retail sales by population dens.

Simple regression analysis (discussed in Chapter 12) is bivariate linear regression in which one dependent variable, y, is predicted by one independent variable, x. Examples of simple regression applications include models to predict retail sales by population density. Regression analysis with two or more independent variables or with at least one nonlinear predictor is called multiple regression analysis.

Regression analysis with two or more independent variables or with at least one nonlinear predictor is called multiple regression analysis. Regression analysis helps in developing a regression equation by which the value of a dependent variable can be estimated given a value of an independent variable. If a regression model characterises the relationship between a dependent y and only one independent x, then such a regression model is called a simple regression model given by

$$\hat{Y} = a + bX \qquad (1)$$

But if more than one independent variable is associated with a dependent variable then such a regression model is called a multiple regression model. In multiple regression analysis, the dependent variable, y, is sometimes referred to as the response variable

A multiple regression equation is an equation which is used for estimating a dependent variable say X_1 from a set of independent variables X_2 , X_3,... and is called the regression equation of X_1 on X_2 , X_3,.... For two independent variables X_2 , X_3 we have the following simplest regression equation of X_1 on X_2 and X_3 and is of the form :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \qquad (2)$$

Where a, $b_{12.3}$ and $b_{13.2}$ are the parameters to be estimated by the Principle of Least Square. The equations so obtained are known as the normal equations

In equation (1) , a is called the X- intercept , $b_{12.3}$ indicates the slope of the regression line of X_1 on X_2called partial regression coefficient ofX_1 on $X_2$when X_3 is held constant . $b_{13.2}$ indicates the slope of the regression line of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3keeping X_2constant.

If we take the deviations of the variables from their respective means and denote these deviations by x_1 , x_2 andx_3 i.e. if

$$x_1 = X_1 - \bar{X}_1$$
$$x_2 = X_2 - \bar{X}_2$$
$$x_3 = X_3 - \bar{X}_3$$

Note :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$
$$\bar{X}_1 = a + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3$$
$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

We have

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad (3)$$

The normal equations in (2) reduces to

$$\Sigma x_2 x_1 = b_{12.3}\Sigma\ x_2^2 + b_{13.2}\Sigma\ x_3 x_2 \qquad (4)$$
$$\Sigma x_2 x_3 = b_{12.3}\Sigma x_3 x_2 + b_{13.2}\Sigma\ x_3^2$$

Solving the two normal equations in (4) the partial regression coefficients $b_{12.3}$ and $b_{13.2}$ can be obtained as

$$b_{12.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{s_1}{s_2} \quad \text{and} \quad b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{s_1}{s_3} \qquad (5)$$

Where S_1 , S_2 and S_3 are the standard deviations of X_1, X_2 and X_3 respectively. Thus using equation (3) and equation (5) we get the multiple regression equation of X_1 on X_2 and X_3 as

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2_{23}}\right)\frac{S_1}{S_2}(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{23}}\right)\frac{S_2}{S_1}\left(X_1 - \bar{X}_1\right) \qquad (6)$$

Similarly the multiple regression equation of X_2 on X_3 and X_1 as

$$X_2 - \bar{X}_2 = \left(\frac{r_{23} - r_{12}r_{13}}{1 - r^2_{13}}\right)\frac{S_2}{S_3}(X_3 - \bar{X}_3) + \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2_{13}}\right)\frac{S_2}{S_1}\left(X_1 - \bar{X}_1\right) \qquad (7)$$

Exercise 6 Find the multiple linear regression equation of X_1 on X_2 and X_3 from the data relating to three variables given below:

X_1	4	6	7	9	13	15
X_2	15	12	8	6	4	3
X_3	30	24	20	14	10	4

Solution:

The regression equation of X_1 on X_2 and X_3 is

$$\Sigma X_1 = a + b_{12.3}\Sigma\ x_2 + b_{13.2}\Sigma\ x_3$$
$$\Sigma x_2 x_1 = b_{12.3}\Sigma\ x_2^2 + b_{13.2}\Sigma\ x_3 x_2$$
$$\Sigma x_2 x_3 = b_{12.3}\Sigma x_3 x_2 + b_{13.2}\Sigma\ x_3^2$$

Calculating the required values we get:

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_2^2	X_3^2	X_1^2
4	15	30	60	120	450	225	900	16
6	12	24	72	144	288	144	576	36
7	8	20	56	140	160	64	400	49
9	6	14	54	126	64	36	196	81
13	4	10	52	130	40	16	100	169
15	3	4	45	60	12	9	16	225
$\sum X_{1=54}$	$\sum X_{2=48}$	$\sum X_3=102$	$\sum X_1X_{2=339}$	$\sum X_1X_{3=}720$	$\sum X_2X_{3=}1034$	$\sum X_2^{2=}494$	$\sum X_3^{2=}2188$	$\sum X_1^{2=}576$

Substituting the values in the normal equations

$48a+494\ b_{12.3}+1034b_{13.2}=54$ (2)

$102a+1034\ b_{12.3}+2188\ b_{13.2}=720$ (3)

Multiplying equation (1) by 8 we have

$48a+384\ b_{12.3}+816\ b_{13.2}=432$ (4)

Subtracting Equation (2) from equation (4) we get

$110\ b_{12.3}+218\ b_{13.2}=93$ (5)

Multiplying Equation (1) by 17 we get

$102a+816\ b_{12.3}+1734\ b_{13.2}=918$ (6)

Subtracting Equation (3) from equation (4) we get

$218\ b_{12.3}+454\ b_{13.2}=-198$ (7)

Multiplying equation (5) by 109 we obtain

$11990\ b_{12.3}+24970\ b_{13.2}=10137$ (8)

Multiplying equation (7) by 55 , we get

$11990\ b_{12.3}+23762\ b_{13.2}=-10890$ (9)

Subtracting equation (8) from equation (9) we get

$1208\ b_{12.3}=-753$

$b_{12.3}=-0.623$

Substituting the value of $b_{12.3}$ in equation (5) we get

$110\ b_{12.3}=218\ (-0.623)=-93$

$110\ b_{12.3}=135.814-93$

$b_{12.3}=42.814/110=0.389$

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in equation (1) we get

$6a+48\ (0.389)+102\ (-0.623)=54$

$6a=54+63.546-18.672=98.874$

Self Assesment Questions

3. The simple correlation coefficients between temperature temperature(x_1) , yield of corn (x_2) and rainfall (x_3) are $r_{12}=0.59$, $r_{13}=0.46$ and $r_{23}=0.77$. Calculate the partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$

a =16.479

Thus the required equation is

X_1= 16.479 +0.389X_2-0.623 X_3

Exercise 7. Given r_{12}=0.28 r_{23}=0.49 r_{31}=0.51

$\sigma_1 = 2.7$ $\sigma_2 = 2.4$ $\sigma_3 = 2.7$

Solution The required equation of X_3 on X_1 and X_2 is

$X_{3=}$ $b_{31.2}X_1$ +$b_{31.2}$ X_2 (1)

Where $b_{31.2} = \frac{r_{31} - r_{32}r_{12}}{1 - r_{12}^2} . \frac{s_3}{s_1}$ and $b_{32.1} = \frac{r_{32} - r_{13}r_{12}}{1 - r_{12}^2} . \frac{s_3}{s_2}$

The symbols s_1, s_2 and s_3 are the standard deviations of X_1 , X_2 and X_3respectively. However this represents the sample standard deviation and the population standard deviation and is given by symbol σ. But some authors, without making this distinction use the symbol σ for both sample s.d and population S.D

$$b_{31.2} = \frac{0.51 - 0.49x0.28}{1 - (0.28)^2} x \frac{2.7}{2.7}$$

$$= 0.405$$

$$b_{32.1} = \frac{0.49 - 0.51x0.28}{1 - (0.28)^2} x \frac{2.7}{2.4}$$

$$= 0.424$$

Equation (1) gives

X_3= 0.405 X_1+0.424 X_2

1.6 Summary

- Partial and multiple correlations describe the relationship between two variables when influence of one or more other variables is held constant or involved.

- Multiple regressions establish an equation for estimating value of a dependent variable given the value of two or more independent variables.

- The coefficient of partial correlation between X_1 and X_2 keeping X_3 constant is denoted by the symbol $r_{12.3}$

- In case of four variables the partial correlation coefficient between X_1 and X_2 eliminating the influence of X_3 and X_4 is denoted by $r_{12.34}$

- The coefficient of multiple correlation is denoted by the symbol $R_{1.23}$

- The coefficient of multiple determination is denoted by the symbol R^2

- The value of partial correlation coefficient lies between -1 and +1

- The value of partial multiple correlation coefficient lies between 0 and 1

1.7: Key words

- Partial correlation coefficient:
  It is a measure of the degree of linear association between a dependent variable and one particular independent variable and one particular independent variable when all other independent variables are kept constant.

- Multiple correlation coefficient:
  It gives the effects of all the independent variables on a dependent variable.

- A multiple regression equation:

It is an equation which is used for estimating a dependent variable from a set of independent variables

1.8: Answers to 'Check Your Progress'

1. Partial coefficient of correlation describes the relationship between one of the independent variables and dependent variable, given that the other independent variables are held constant statistically.

2. Through multiple correlation analysis we can attempt to measure the degree of association between a dependent variable y and two or more independent variables.

3. A multiple regression equation is an equation which estimates an average relationship between a dependent variable and two or more independent variables.

4. Depending upon the number of independent variables held constant, we often call partial correlation coefficient as zero order, first order, second order correlation coefficient.

5. If the value of a multiple correlation coefficient is one, it indicates perfect correlation.

6. The main limitation of multiple correlation coefficient is that like partial correlation coefficient it also assumes that the simple or zero order correlation on which it depends are of linear type.

1.9. Questions and Answers

Questions and Exercises

Self Assessment Questions:

Multiple Choice Questions

1. The range of partial correlation coefficient $r_{12.3}$ is
   (a) -1 to 1
   (b) 0 to ∞
   (c) 0 to 1
   (d) none of these

2. If multiple correlation coefficient $R_{1.23}$ =1, then it implies a
   (a) lack of linear relationship
   (b) perfect relationship
   (c) reasonably good relationship
   (d) none of these

Answer 1.(a) 2.(b)

Fill in the blanks:

1. A coefficient which examines the association between a dependent variable and an independent variable after factoring out the effect of other independent variables is known as _____.

2. A statistical technique that develops an equation that relates a dependent variable to one or more independent variables is called _____.

Answer: 1. Partial correlation coefficient, 2. Regression analysis,

State whether True or False:

1. The closer the coefficient of multiple correlation is to 1, the better the relationship between the variables.
2. The coefficient of multiple correlation is the square root of coefficient of multiple determination.

Answer: 1.True, 2.True

Short answer questions:

1. What is the importance of partial correlation analysis?
2. Define multiple correlation.
3. What are partial regression coefficients?
4. Given $r_{12} = 0.5$, $r_{13} = 0.4$ and $r_{23} = 0.1$, find the values of $r_{12.3}$.
5. If $r_{12} = 0.9$, $r_{13} = 0.75$ and $r_{23} = 0.7$, find the values of $R_{1.23}$.

Long answer questions:

1. Define the following terms:
   a) Coefficient of partial and multiple correlation.
   b) Coefficient of multiple determination.
2. What are normal equations and how are they used in multiple regression analysis?

   Distinguish between partial and multiple correlation and point out their usefulness in statistical analysis

3. The simple correlation coefficients between profits (X_1), sales (X_2) and advertising expenditure (X_3) of a factory are $r_{12} = 0.69$, $r_{13} = 0.45$ and $r_{23} = 0.58$ find the partial correlation coefficients $r_{12.3}$ and $r_{13.2}$ and interprete them.

4. In a trivariate distribution:

$\sigma_1 = 3, \sigma_2 = 4, \sigma_3 = 5$, $r_{23} = 0.4$, $r_{31} = 0.6$, $r_{13} = 0.7$

Determine the regression equation of X_1 on X_2 and X_3 if the variates are measures from their means.

5. The following data are given:

   $x_1 = 6$, $x_2 = 7$, $x_3 = 8$; $s_1 = 1 s_1 = 2$, $s_1 = 3$; $r_{23} = 0.8$, $r_{12} = 0.6$, $r_{13} = 0.7$

   a) Find the regression equation of x_3 on x_1 and x_2
   b) Estimate the value of x_3 when $x_1 = 4$ and $x_2 = 5$

2.0 Further Reading

3. Essential statistics for Economics and Business Studies: Padmalochan Hazarika, Akansha Publishing House, New Delhi.
4. Business Statistics: S. Saha, New Central Book Agency.
5. Basic Statistics:B. L. Agarwal, New Age International Limited.
6. Statistics for Management: Prentice Hall of India Limited, Levin Richard & Rubin David.
7. Quantitative Techniques for Decision Making:Anand Sharma, Himalaya Publishing House.
6. Methods of Statistical Analysis : P.S Grewal, Sterling Publishers Private Limited

BLOCK III : Unit 2
Various Formulae, Problems, Uses and Limitation of Partial and Multiple Correlations and Regressions

2.1: Introduction

In this unit we describe about partial correlation coefficient which describes the relationship between one of the independent variables and the dependent variable, given that the other independent variables are held constant statistically. In partial correlation, the linear association between a dependent variable and one particular independent variable is studied, holding other independent variables constant.

2.2 Objective

After going through this unit, you will be able to
Understand the formulas and problems of partial correlations
Understand the formulas and problems multiple correlations
Understand the problems of multiple regression

2.3: Various formulae and Problems of:

2.3.1: Partial Correlation

The partial correlation between X_1 and X_2 holding X_3 constant is determined by

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} \quad ;$$

Similarly partial correlation between X_1 and X_3 holding X_2 constant $r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}}$

and

Partial correlation between X_2 and X_3 holding X_1 constant is given by -

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

Again we have the following relations:

(a) $r_{12.3} = \sqrt{R^2_{12.3}} = \sqrt{1-\frac{S^2_{1.23}}{S^2_{1.3}}}$ *and* $r_{13.2} = \sqrt{R^2_{13.2}} = \sqrt{1-\frac{S^2_{1.23}}{S^2_{1.2}}}$

(b) $r_{12.3} = \sqrt{b_{12.3}}xb_{21.3}$;
$r_{13.2} = \sqrt{b_{13.2}}xb_{31.2}$ *and*
$r_{23.1} = \sqrt{b_{23.1}}xb_{32.1}$

Using these relations the partial correlation between X_1 and X_2 holding X_3 constant can also be determined as:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{13}^2}\left(\frac{S_1}{S_2}\right) \times \frac{r_{12} - r_{13}r_{23}}{1 - r_{13}^2}\left(\frac{S_2}{S_1}\right)$$

$$= \frac{\left(r_{12} - r_{13}r_{23}\right)^2}{\left(1 - r_{13}^2\right)\left(1 - r_{13}^2\right)}$$

Note

(1) $r_{xy}=r_{yx}$
(2) The value of the partial correlation coefficient lies between -1 and 1.

2.3.2: Multiple Correlations

A linear multiple correlation is denoted by the symbol R and the necessary subscripts are added to it. The subscript before the decimal represents the dependent variable and the subscripts after the decimal represent the independent variables which affect the dependent variable. For example $R_{1.23}$ indicates that the variable X_1 is associated with the variables X_2 and X_3 and the formula for it is:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

By symmetry we may also write

$$R_{1.23} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \text{ and } \quad R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

Note:

(1) The value of multiple correlation coefficient always lie between 0 and 1.
(2) If $r_{12}=r_{13}=0$ then $R_{1.23}=0$ implying no linear relationship between the variables.
(3) $R_{1.23}=R_{1.32}$
(4) If $R_{1.23}=1$, the correlation is called perfect

2.3.3: Regressions

A multiple regression equation is an equation which is used for estimating a dependent variable say X_1 from a set of independent variables X_2 , X_3,... and is called the regression equation of X_1 on X_2 , X_3,.... For two independent variables X_2 , X_3 we have the following simplest regression equation of X_1 on X_2 and X_3 and is of the form :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \qquad (2)$$

Where a, $b_{12.3}$ and $b_{13.2}$ are the parameters to be estimated by the Principle of Least Square. The equations so obtained are known as the normal equations

In equation (1) , a is called the X- intercept , $b_{12.3}$ indicates the slope of the regression line of X_1 on X_2called partial regression coefficient ofX_1 on $X_2$when X_3 is held constant . $b_{13.2}$ indicates the slope of the regression line of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3keeping X_2constant.

If we take the deviations of the variables from their respective means and denote these deviations by x_1 , x_2 andx_3 i.e. if

$$x_1 = X_1 - \overline{X}_1$$
$$x_2 = X_2 - \overline{X}_2$$
$$x_3 = X_3 - \overline{X}_3$$

Note :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$
$$\overline{X}_1 = a + b_{12.3}\overline{X}_2 + b_{13.2}\overline{X}_3$$
$$X_1 - \overline{X}_1 = b_{12.3}(X_2 - \overline{X}_2) + b_{13.2}(X_3 - \overline{X}_3)$$

We have

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad (3)$$

The normal equations in (2) reduces to

$$\Sigma x_2 x_1 = b_{12.3}\Sigma\ x_2^2 + b_{13.2}\Sigma\ x_3 x_2 \qquad (4)$$
$$\Sigma x_2 x_3 = b_{12.3}\Sigma x_3 x_2 + b_{13.2}\Sigma\ x_3^2$$

Solving the two normal equations in (4) the partial regression coefficients $b_{12.3}$ and $b_{13.2}$ can be obtained as

$$b_{12.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \right) \frac{s_1}{s_2} \quad \text{and} \quad b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) \frac{s_1}{s_3} \qquad (5)$$

Where S_1 , S_2 and S_3 are the standard deviations of X_1, X_2 and X_3 respectively. Thus using equation (3) and equation (5) we get the multiple regression equation of X_1 on X_2 and X_3 as

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2 23} \right) \frac{S_1}{S_2} (X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2 23} \right) \frac{S_2}{S_1} \left(X_1 - \bar{X}_1 \right) \qquad (6)$$

Similarly the multiple regression equation of X_2 on X_3 and X_1 as

$$X_2 - \bar{X}_2 = \left(\frac{r_{23} - r_{12}r_{13}}{1 - r^2 13} \right) \frac{S_2}{S_3} (X_3 - \bar{X}_3) + \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2 13} \right) \frac{S_2}{S_1} \left(X_1 - \bar{X}_1 \right) \qquad (7)$$

2.4: Uses and limitation of:

2.4.1: Partial Correlation

Advantages

(1) Partial correlation is especially useful in the analysis of interrelated variables where the linear relation between a dependent and independent variable can be studied by eliminating the influence of other independent variables.

(2) In partial correlation, relationships are expressed concisely in a few well defined coefficients.

(3) Partial correlation is of greatest value when used in conjunction with gross and multiple correlation in the analysis of factors affecting variations in many kind of phenomenon.

Limitation

(1) While calculating partial correlation coefficient it is assumed that the simple correlation or zero order correlation from which partial correlation is studied have linear relationship between the variables. In actual practice, particularly in social science this assumption is not desirable because linear relationship doesnot generally exist in such situations.

(2) The effects of the independent variables are studied additively not jointly. The various independent variables considered in the study are assumed to be independent of each other, however in actual practice this may not be true and there may be a relationship among the variables.

(3) The computation of partial correlation coefficients especially beyond first order is laborious and time consuming.

2.4.2: Multiple Correlations

The advantages and limitations of multiple correlation :

(1) Through multiple correlation analysis we intend to measure the degree of association between a single dependent variable and a number of independent variables taken together as a group.
(2) It expresses the type and degree of relationship in a few concise coefficients. For example $R_{1.234}^2$ =0.85 means that three factors X_2, X_3, and X_4 explain 85%of the squared variability in X_1.
(3) It gives the best predictions that can be computed linearly from the independent variables

Limitations

(1) In multiple correlation it is assumed that the relationships between variables are linear which does not hold good .For example in the field of agriculture most relationships are non-linear, therefore, linear regression coefficients cannot describe non -linear relations accurately.
(2) Also it is assumed that the effects of the independent variables upon the dependent variable are separate, distinct and additive. But it is not possible if interrelation exist among the variables.
(3) The calculation involved in multiple correlation is quite tedious and therefore should be interpreted with caution.

2.4.3: Regressions

Regression analysis with two or more independent variables or with at least one nonlinear predictor is called multiple regression analysis. Regression analysis helps in developing a regression equation by which the value of a dependent variable can be estimated given a value of an independent variable. If a regression model characterises the relationship between a dependent y and only one independent x, then such a regression model is called a simple regression model given by

$$\hat{Y} = a + bX \qquad (1)$$

But if more than one independent variable is associated with a dependent variable then such a regression model is called a multiple regression model. In multiple regression analysis, the dependent variable, y, is sometimes referred to as the response variable

A multiple regression equation is an equation which is used for estimating a dependent variable say X_1 from a set of independent variables X_2 , X_3,... and is called the regression equation of X_1 on X_2 , X_3,.... For two independent variables X_2 , X_3 we have the following simplest regression equation of X_1 on X_2 and X_3 and is of the form :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \qquad (2)$$

Where a, $b_{12.3}$ and $b_{13.2}$ are the parameters to be estimated by the Principle of Least Square. The equations so obtained are known as the normal equations

In equation (1) , a is called the X- intercept , $b_{12.3}$ indicates the slope of the regression line of X_1 on X_2called partial regression coefficient ofX_1 on $X_2$when X_3 is held constant . $b_{13.2}$ indicates the slope of the regression line of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3keeping X_2constant.

If we take the deviations of the variables from their respective means and denote these deviations by x_1 , x_2 andx_3 i.e. if

$$x_1 = X_1 - \bar{X}_1$$
$$x_2 = X_2 - \bar{X}_2$$
$$x_3 = X_3 - \bar{X}_3$$

Note :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$
$$\bar{X}_1 = a + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3$$
$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

We have

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad (3)$$

The normal equations in (2) reduces to

$$\Sigma x_2 x_1 = b_{12.3}\Sigma\ x_2^2 + b_{13.2}\Sigma\ x_3 x_2 \qquad (4)$$
$$\Sigma x_2 x_3 = b_{12.3}\Sigma x_3 x_2 + b_{13.2}\Sigma\ x_3^2$$

Solving the two normal equations in (4) the partial regression coefficients $b_{12.3}$ and $b_{13.2}$ can be obtained as

$$b_{12.3} = \left(\frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \right) \frac{s_1}{s_2} \quad \text{and} \quad b_{13.2} = \left(\frac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} \right) \frac{s_1}{s_3} \qquad (5)$$

Where S_1 , S_2 and S_3 are the standard deviations of X_1, X_2 and X_3 respectively. Thus using equation (3) and equation (5) we get the multiple regression equation of X_1 on X_2 and X_3 as

$$X_1 - \overline{X}_1 = \left(\frac{r_{12} - r_{13} r_{23}}{1 - r^2 23} \right) \frac{S_1}{S_2} (X_2 - \overline{X}_2) + \left(\frac{r_{13} - r_{12} r_{23}}{1 - r^2 23} \right) \frac{S_2}{S_1} \left(X_1 - \overline{X}_1 \right) \qquad (6)$$

Similarly the multiple regression equation of X_2 on X_3 and X_1 as

$$X_2 - \overline{X}_2 = \left(\frac{r_{23} - r_{12} r_{13}}{1 - r^2 13} \right) \frac{S_2}{S_3} (X_3 - \overline{X}_3) + \left(\frac{r_{12} - r_{13} r_{23}}{1 - r^2 13} \right) \frac{S_2}{S_1} \left(X_1 - \overline{X}_1 \right) \qquad (7)$$

2.5: Summary

Partial and multiple correlations describe the relationship between two variables when influence of one or more other variables is held constant or involved.

Multiple regressions establish an equation for estimating value of a dependent variable given the value of two or more independent variables

2.6: Key words

The value of partial correlation coefficient lies between -1 and +1

The value of multiple correlation coefficient lies between 0 and 1

Multiple regression equation: It is an equation which is used for estimating a dependent variable from a set of independent variables

2.7: Answers to 'Check Your Progress'

Partial correlation coefficient:

It is a measure of the degree of linear association between a dependent variable and one particular independent variable and one particular independent variable when all other independent variables are kept constant.

Multiple correlation coefficient:

It gives the effects of all the independent variables on a dependent variable.

A multiple regression equation:

It is an equation which is used for estimating a dependent variable from a set of independent variables

2.8. Questions and Answers

Example 1 In a trivariate distribution it is found that $r_{12}=0.70$, $r_{13}=0.61$, $r_{23}=0.40$. Find the values of $r_{23.1}$, $r_{13.2}$ and $r_{12.3}$.

Solution: The partial correlation between X_2 and X_3 holding X_1 constant is determined by

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

Substituting the given values we get

Given $r_{12.4}=0.60$, $r_{13.4}=0.50$, $r_{23.4}=0.70$ find $r_{12.34}$ and $r_{13.24}$

Solution:

$$r_{12.34} = \frac{r_{12.4} - r_{13.4} r_{23.4}}{\sqrt{1 - r^2_{13.4}}\sqrt{1 - r_{23.4}{}^2}}$$

$$= \frac{0.6 - (0.5 \times 0.7)}{\sqrt{\{1 - (0.5)^2\}}\sqrt{\{1 - (0.7)^2\}}}$$

$$= \frac{0.6 - 0.35}{\sqrt{0.75 \times 0.51}}$$

$$= \frac{0.25}{\sqrt{0.3825}}$$

$$= \frac{0.25}{0.62} \quad = 0.403$$

$$r_{13.24} = \frac{r_{13.4} - r_{12.4} r_{23.4}}{\sqrt{1 - r^2_{12.4}}\sqrt{1 - r_{23.4}{}^2}}$$

$$= \frac{0.5 - (0.6 \times 0.7)}{\sqrt{\{1 - (0.6)^2\}}\sqrt{\{1 - (0.7)^2\}}}$$

$$= \frac{0.5 - 0.42}{\sqrt{0.64 \times 0.51}}$$

$$= \frac{0.08}{0.57}$$

$$= 0.14$$

Ex 2: In a three variate multiple correlation analysis, the following results were obtained r_{12}=0.59 r_{13}=0.46 and r_{23}=0.77. Find $R_{1.23}$

Solution: Multiple Correlation Coefficient is defined as

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.50)^2 + (0.46)^2 - 2(0.59 \times 0.46 \times 0.77)}{1 - (0.77)^2}}$$

$$= \sqrt{\frac{0.3481 + 0.2116 - 0.418}{0.4071}}$$

$$= \sqrt{\frac{0.5597 - 0.418}{0.4071}}$$

$$= \sqrt{\frac{0.1417}{0.4071}} = 0.589$$

Stop to consider

The multiple correlation may be defined as a statistical tool designed to measure the degree of relationship existing among three or more variables. it is denoted by the symbol R

$R_{1.23}$ stands for the coefficients of multiple correlation between X_1, X_2 and X_3. The subscript to the left stands for dependent variable while the subscripts to the right of the dot represent independent variables.

Short answer questions:

1. Given $r_{12} = 0.5$, $r_{13} = 0.4$ and $r_{23} = 0.1$, find the values of $r_{12.3}$.
2. If $r_{12} = 0.9$, $r_{13} = 0.75$ and $r_{23} = 0.7$, find the values of $R_{1.23}$.

Long answer questions:

1. In a trivariate distribution:
   $\sigma_1 = 3, \sigma_2 = 4, \sigma_3 = 5$, $r_{23} = 0.4$, $r_{31} = 0.6$, $r_{13} = 0.7$
   Determine the regression equation of X_1 on X_2 and X_3 if the variates are measured from their means.
2. The following data are given:
   $x_1 = 6$, $x_2 = 7$, $x_3 = 8$; $s_1 = 1 s_1 = 2$, $s_1 = 3$; $r_{23} = 0.8$, $r_{12} = 0.6$, $r_{13} = 0.7$
   a) Find the regression equation of x_3 on x_1 and x_2
   b) Estimate the value of x_3 when $x_1 = 4$ and $x_2 = 5$

1.0 Further Reading

2. Essential statistics for Economics and Business Studies: Padmalochan Hazarika, Akansha Publishing House, New Delhi.
3. Business Statistics: S. Saha, New Central Book Agency.
4. Basic Statistics:B. L. Agarwal, New Age International Limited.
5. Statistics for Management: Prentice Hall of India Limited, Levin Richard & Rubin David.
6. Quantitative Techniques for Decision Making:Anand Sharma, Himalaya Publishing House.

6. Methods of Statistical Analysis : P.S Grewal, Sterling Publishers Private Limited

2.9: Further Reading

1. Essential statistics for Economics and Business Studies: Padmalochan Hazarika, Akansha Publishing House, New Delhi.
2. Business Statistics: S. Saha, New Central Book Agency.
3. Basic Statistics:B. L. Agarwal, New Age International Limited.
4. Statistics for Management: Prentice Hall of India Limited, Levin Richard & Rubin David.
5. Quantitative Techniques for Decision Making:Anand Sharma, Himalaya Publishing House.
6. Methods of Statistical Analysis : P.S Grewal, Sterling Publishers Private Limited

BLOCK III : Unit 3

Coefficient of Multiple Determinations, Association of Attributes: Concept, Order of a Class, Class Frequency, Consistency of Data

Unit Structure:

3.1: Introduction

In this unit you will learn about the concept of multiple determination and adjusted coefficient of determination. . In multiple regressions, the coefficient of multiple determinations represents the proportion of the variation in Y that is explained by the set of independent variables selected. The Adjusted Coefficient of Determination (Adjusted R-squared) is an adjustment for the Coefficient of Determination that takes into account the number of variables in a data set. The unit also discusses the concept of association of attributes and discusses about consistency of data**.** Qualitative variables are called attributes and theory of attributes deals with the measurement of data whose magnitude cannot be directly measured numerically. Though, the qualitative data can be quantified but for the sake of clear understanding and convenience, the statistical methodologies for the analysis of qualitative data have been separately developed. Association of two attributes measure the degree of relationship between two phenomena whose sizes cannot be measured but one can only determine the presence or absence of a particular attribute or quality. The unit also discusses about the consistency of the data, the conditions for consistency and the independence of attributes. A data is said to be consistent if no class frequency turns out to be negative.

3.2: Objective

After going through this unit, you will be able to

- Learn about the coefficient of multiple determination
- Learn about the association of attributes
- Describe the class and class frequency
- Discuss about the consistency of data

3.3: Coefficient of Multiple Determination (R^2):

The coefficient of determination can be thought of as a percent which gives an idea of how many data points fall within the line formed by the regression equation. It is represented by the symbol R and represents the proportion of the total variation in the dependent variable y , accounted for or explained by the independent variables in the multiple regression model . The value of R^2 lies between 0 and 1. It is given by

$$R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

The higher the coefficient, the higher percentage of points pass through the line when the data points and the line are plotted. If the coefficient is 0.80, then 80% of the points should fall within the regression line. A higher coefficient is an indicator of a better goodness of fit for the observations.

3.4: Adjusted Coefficient of Determination

Sometimes additional independent variables added to multiple regression model increase R^2 while their contribution is insignificant in explaining the variation in the dependent variable. To correct for this defect, R^2 is adjusted by taking into account the degrees of freedom which decreases with inclusion of additional independent or explanatory variables in the model.

The following equation shows the relationship between adjusted $\bar{R}^2$ and R^2 .

$$\bar{R}^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right]$$

where $n = sample\,size$

k = *the number of independent* variables *in the regression equation*

One major difference between R-square and adjusted $\bar{R} - square$ is that R-square supposes that every independent variable in the model explains the variation in the dependent variable. It gives the percentage of explained variation as if all independent variables in the model affect the

dependent variable, whereas adjusted R-square gives the percentage of variation explained by only those independent variables that in reality affect the dependent variable.

Example 3.1 Suppose R^2= 0.944346527, k=2 and n=8. Calculate Adjusted R^2.

Solution:

$$\bar{R}^2 = 1-(1-R^2)\left[\frac{n-1}{n-(k+1)}\right]$$

where n = *sample size*

k = *the number of independent variables in the regression equation*

$$= 1-(1-0.944346527)\left[\frac{8-1}{8-(2+1)}\right]$$

$$= 0.922085138$$

Based on the value of R^2, the proportion of variation explained by the estimated regression line is approximately 0.922 or 92.2 percent.

3.5: Concept of Association of Attributes:

Correlation as a statistical tool is used to measure the degree of relationship between two phenomenon which are capable of direct quantitative measurement. Data regarding such phenomenon are known as statistics of variables. On the other hand, certain phenomenon like blindness, deafness etc. are not capable of direct quantitative measurements are called attributes. Such phenomenon is studied only on the presence or absence of the attribute. Thus, method of association of attributes is employed to measure the degree of relationship between two phenomena which we cannot measure and where we can only determine the presence or absence of a particular attribute.

3.6 Classes And Class Frequencies

While dealing with statistics of attributes, the data has to be classified .The classification is done on the basis of presence or absence of a particular attribute or characteristic. When we are studying only one attribute two classes are formed- one possessing that attribute and another not possessing it. When two attributes are studied four classes are formed. The number of observations assigned to any class is termed as 'class frequency'. Class frequencies are represented by enclosing the corresponding class symbols in brackets. Hence (A) stands for the number of items or individuals who possess the attribute A. (AB) stands for the number of items who possess both the attributes A and B , (Aβ) denotes the number of items possessing attribute

A but not possessing attribute B and so on . The classes having one or more positive attributes are called positive classes and classes having one or more negative attributes are called negative classes.

Number of Classes: The total number of classes comprising of the various attributes are given by 3^n , n representing the number of attributes . If one attribute is studied , then thre will be 3^1 =3 classes . Thus if literacy is studied , the presence of literacy is represnted by A, its absence by α and total by N. Therefore there will be 3 classes i.e A, α and N. If two attributes are studied , the number of classes will be 3^2 =9 . The 9 classes are A, α , B, β , AB, α B, A β , α β and N.

If three attributes are studied, the number of classes will be 3^3=27 classes.

Class Frequencies

The number of observations assigned to any class is called the '**frequency of the class' or 'class frequency'. Class frequencies are generally denoted by enclosing the corresponding class symbols in small brackets ( ).**

Let 'A' be an attribute say 'Honesty'

(A) Dnotes the number of prsons possessing 'A' attribute

i.e (A) =10 means that there are ten persons who are honest

Also 10 is the frequency of the class

If 'B' stands for male , then β would mean female

If 'A' stands for honest , then α would mean non-honest

Therefore ,

(AB)= 30, means that there are 30 males who are honest

(Aβ) = 10 , means there are 10 females who are honest

(α β) =15 , means there are 15 females who ar not honest and so on

3.6 .1 Order of a class

The order of a class depends upon the number of attributes under study . A class having one attribute is known as the class of first order. A class having two attributes is called the class of second order and so on.

The total number of observations denoted by the symbol N is called class or frequency of zero order. Since no attributes are specified. Thus we have:

N : Class (or Frequency) of Zero Order

(A), (B), (α), (β) : Class (or Frequency) of First Order

(AB), (A β), (αB) , (α β) : Class (or Frequency) of Second Order

3.6.2 Ultimate Class Frequencies

The four groups (AB), (Aβ) , (αB) , (α β) are called ultimate class frequencies . They are represented by enclosing the corresponding class symbols viz, AB, α B, A β and α β in the small brackets ( ).

The number of ultimate classes is determined by 2^n, where n is the number of attributes. Thus for one attribute , there will be 2^1=2 ultimate classes; for two attributes, there will be 2^2=4 ultimate classes ; and for three attributes , there will be 2^3= 8 ultimate classes.

The total number of observations is equal to the positive and negative frequencies of the same class of the first order i.e ,

N=(A) +(α) or N= (B) + (β)

Moreover,

(A)= (AB) +(Aβ) ; (B) = (AB)+(αB)

(α) = (αB) +(α β) ; (β) = (Aβ) +(α β)

The frequencies of the positive , negative and ultimate classes can be known from the following table which is known as the the contingency table of order (2x2) for two attributes A and B is shown below:

	A	α	Total
B	(AB)	(α B)	(B)
β	(A β)	(α β)	(β)
Total	(A)	(α)	N

3.6.1 Relationship between the class frequencies

The frequency of a lower order class can always be expressed in terms of the higher order class frequencies as follows:

$N=(A)+(\alpha)$

$(A)=(AB)+(A\beta)$

$(B)=(AB)+(\alpha B)$

$(\alpha)=(\alpha B)+(\alpha\beta)$

$(\beta)=(A\beta)+(\alpha\beta)$

If the number of attributes is n, there will be 3^n classes and

2^n class frequencies.

Stop to consider

- Phenomenon like blindness, deafness etc. not capable of direct quantitative measurements are called attributes.
- The number of observations assigned to any class is termed as 'class frequency'.
- Method of association of attributes is employed to measure the degree of relationship between two phenomena which we cannot measure and where we can only determine the presence or absence of a particular attribute.

Ex1. From the following data find the missing frequencies

$(AB)= 50$, $(\alpha\beta)=25$, $(\alpha)= 100$, $N=250$

Solution : The missing frequencies are $(A\beta)$, (β), (αB) , (A) and (B).

Putting these values in the contingency table , we have

	A	α	Total
B	50	75	125
β	100	25	125
Total	150	100	250

Working

$(\alpha)= (\alpha B)+(\alpha\beta)$

$100= (\alpha B)+25$

$(\alpha B)=100-25=75$

$(B)= (AB)+(\alpha B)$

= 50+75

= 125

N=(B)+(β)

250= 125+ (β)

(β)= 250-125 ==125

(α)= (αB)+(αβ)

(β)=(Aβ)+(αβ)

125= (Aβ) + 25

(Aβ)=125-25=100

(A)= (AB)+(Aβ)

=50+100

=150

Hence (A)= 150 , (Aβ)=100 , (β)= 125, (αB)=75 and (B)= 125

3.7: Consistency of data.

Consistency of Data: A given set of data is said to be consistent if none of the class frequencies is negative. If any class frequency is negative, the given set of data is said to be inconsistent.

Exercise 4.0 Test the consistency in the data given below:

(i) N=1200 (A)=600 (AB)=400 (B)= 500
(ii) N=1000 (AB)=200 (A)=150 (B)=300

Solution:

Case I We are given (A) =600 (AB) =400 (B) =500 and N=1200

We substitute these values in the following contingency table:

	A	α	Total
B	(AB) =400	(α B)=100*	(B)=500
B	(A β) =200*	(α β) =500*	(β) =700*
Total	(A) =600	(α)=600	N =1200

From the table

$(A\ \beta) = (A) - (AB) = 600-400=200$

$(\alpha\ B) = (B)-(AB) = 500-400= 100$

$(\alpha\ \beta) = (\alpha)- (\alpha\ B) = 600=100=500$

Since all the ultimate class frequencies are positive we conclude that the given daa is consistent.

Case II:

We are given (A)=150 (AB)=200 (B)= 300 and N=1000

We substitute these values in the following contingency table:

	A	α	Total
B	(AB) =200	$(\alpha\ B)=100^{*}$	(B)=300
B	$(A\ \beta) =-50$	$(\alpha\ \beta) =150$	$(\beta) =700$
Total	(A) =150	(α)=850	N =1000

From the table

$(A\ \beta) = (A) - (AB) = 150-200= -50$

$(\alpha\ B) = (B)-(AB) = 300-200= 100$

$(\alpha\ \beta) = (\alpha)- (\alpha\ B) = 850-100 =750$

We find that one of the ultimate class frequencies i.e. (Aβ) is negative and hence the given data is inconsistent.

3.8: Summary

- In statistics, an attribute means a quality on characteristic which are not related to quantitative measurements and hence cannot be measured directly.
- Association of attributes relates to the study of relationship between two or more attributes.
- Class frequency denotes the number of individuals who possess that attribute.

- A given set of data are said to be consistent if none of the class frequencies is negative.

- Two attributes A and B are said to be independent in case the presence or absence of one attribute has no relationship with the presence or absence of other attribute.

3.9: Key words

The coefficient of determination is multiple regression represents the proportion of the total variation in the multiple values of the dependent variable y accounted for or explained by the independent variables in the multiple regression model.

Attribute: In statistics, an attribute means a quality or characteristic which are not related to quantitative measurements and hence cannot be measured directly.

Classes refer to different attributes, their subgroups and combinations. The number of observations assigned to them is called their class frequencies.

The attributes are said to be independent when the presence or absence of one attribute does not affect the presence or absence of the other.

In order to find the degree of association between two or more sets of attributes, the coefficient of association is used.

3.10: Answers to 'Check Your Progress'

1. The coefficient of determination is multiple regression represents the proportion of the total variation in the multiple values of the dependent variable y accounted for or explained by the independent variables in the multiple regression model.
2. In statistics, an attribute means a quality or characteristics which are not related to quantitative measurements and hence cannot be measured directly.

3. Classes refer to different attributes, their subgroups and combinations. The number of observations assigned to them is called their class frequencies.

4. The attributes are said to be independent when the presence or absence of one attribute does not affect the presence or absence of the other.

3.11: Questions and Answers

Self Assessment Questions:

Multiple Choice Questions

1. A quality on characteristic which are not related to qualitative measurement is called
   (a) attribute
   (b) class frequency
   (c) classes
   (d) none of these
2. The coefficient is used to
   (a) find the degree of association between two or more sets of attributes.
   (b) find the degree of association between two subgroups.
   (c) find the degree of association between the frequencies
   (d) none of these

Answer: 1(a) 2(b)

Fill in the blanks:

1. A coefficient which examines the association between a dependent variable and an independent variable after factoring out the effect of other independent variables is known as _____.

2. A statistical technique that develops an equation that relates a dependent variable to one or more independent variables is called _____.

3. The total variation explained by a regression model is given by _____.

4. In statistics, _____ means a quality on characteristic which are not related to quantitative measurements and hence cannot be measured directly.

Answer: 1. Partial correlation coefficient, 2. Regression analysis,
3. R^2, 4. Attribute,

State whether True or False:

1. The closer the coefficient of multiple correlation is to 1, the better the relationship between the variables.

2. The coefficient of multiple correlation is the square root of coefficient of multiple determination.
3. Classes refer to different attributes, their sub-groups and combinations.
4. The number of attributes attached to any class is termed its class frequency.

Answer: 1.True, 2.True, 3.True, 4. True.

3.9 Short answer questions:

1. Define association of attributes.
2. How is the coefficient of association calculated?
3. When is a set of data said to be consistent?

Long answer questions:

1. Under what condition is it important to use the adjusted multiple coefficient of determination ?.

2. Test the consistency of the data given below:
   Case I : (AB) = 200, (A) = 300, (α) = 200, (B) = 250, (αβ) = 150, (N) = 500
   Case II : (AB) = 250, (A) = 150, (α) = 1000, (B) = 500, (αβ) = 600, (N) = 1750
3. From the following ultimate frequencies , find the frequencies of the positive and negative classs, and the total number of observations
   (AB)=100 ; (αB)=8 , (Aβ)=50 , (αβ)=40

4.0 Further Reading:

1. Essential statistics for Economics and Business Studies: Padmalochan Hazarika, Akansha Publishing House, New Delhi.
2. Business Statistics: S. Saha, New Central Book Agency.
3. Basic Statistics:B. L. Agarwal, New Age International Limited.
4. Statistics for Management: Prentice Hall of India Limited, Levin Richard & Rubin David.
5. Quantitative Techniques for Decision Making:Anand Sharma, Himalaya Publishing House.
6. Methods of Statistical Analysis : P.S Grewal, Sterling Publishers Private Limited

BLOCK III : Unit 4

Kinds of Association of Attributes, Methods of Measuring Association Between Two Attributes, Partial Association.

Unit Structure:

4.1: Introduction:

In this unit you will learn about the association of attributes. Two attributes are said to be associated if they are not independent but are related .The method of association of attributes is employed to measure the degree of relationship between two phenomena which cannot be measured but its presence or absence can only be determined. The phenomena's which cannot be measured quantitatively, i.e beauty, honesty, insanity, deafness etc. The observations possessing a particular quality, say , honesty are grouped together , counted and given numerical shape i.e they are quantified. There are different kinds of association between attributes and a quantitative measure for nature and the degree of association can be found using different methods .

4.2 Objective

After going through this unit, you will be able to

Learn about the association of attributes

Learn about the kinds of association of attributes

Explore the different methods of association of attributes

4.3 Kinds of Association of attributes

There can be three kinds of association between attributes

(1) Positive Association **:** When two attributes are present or absent together in the data, it is known as positive association. Such association is found between literacy and employment, smoking and cancer, vaccination and immunity from a disease, etc.

(2) Negative Association**:** When the presence of an attribute is associated with the absence of the other attribute, it is called negative association. Such association is found between vaccination and attack of a disease, cleanliness and ill-health, etc.

(3) Independence**.** When there exists no association between two attributes or when they have no tendency to be present together or the presence of one attribute does not affect the other attribute, the two attributes are regarded as independent.

Thus two attributes A and B may be (i) positively associated (ii) negatively associated or (iii) independent based on the following conditions :

Thus

If $(XY) > \frac{(X)(Y)}{N}, then\ X\ and\ Y\ are\ positively\ associated$

$(XY) < \frac{(X)(Y)}{N}, then\ X\ \ and\ \ Y\ are\ \ negatively\ associated$

$(XY) = \frac{(X)(Y)}{N}, then\ X\ and\ Y\ are\ independent$

N being the number of items

Exercise 4.1 Show whether A and B are independent, positively associated or negatively associated for the following values:

(AB) =128 (α B) = 384 (A β)=24 (α β)=72

Solution: Using the given values we have

(A)= (AB)+(Aβ)

= 128+24

= 152

(B) = (AB) + (α B)

= 128+384

= 512

(α) = (αB) +(αβ)

= 384+72

= 456

(N) =(A)+(α)=152+456=608

Thus

$$\frac{(A)\times(B)}{N}=\frac{152\times152}{608}=128$$

(AB) =128

Therefore $(AB)=\frac{(A)\times(B)}{N}$

Hence A and B are independent.

Self Assessment Questions

4. When is a set of data said to be consistent ?

5. Define positive and negative association.

6. When are two attributes independent ?

4.4 Methods of studying Association between Two Attributes

The nature and degree of association between two or more attributes can be studied by using the following methods

(1) Proportion Method : This method consists in comparing the presence or absence of a given attribute in the other

Two attributes A and B are said to be :

Positively associated if $\frac{(AB)}{(B)}>\frac{(A\beta)}{(\beta)}$ or $\frac{(AB)}{(A)}>\frac{(\alpha B)}{(\alpha)}$

Negatively associated if $\frac{(AB)}{(B)}<\frac{(A\beta)}{(\beta)}$ or $\frac{(AB)}{(A)}<\frac{(\alpha B)}{(\alpha)}$

However , if $\frac{(AB)}{(B)}=\frac{(A\beta)}{(\beta)}$ or $\frac{(AB)}{(A)}=\frac{(\alpha B)}{(\alpha)}$ $, then\ A\ and\ B\ are\ independent$

This method can only determine the nature of association between attributes that is whether it is positive or negative or no association but it does not study the degree of association whether it is high or low.

(2) Comparison of Observed and Expected Frequencies

Two attributes A and B are said to be

Positively associated if $(AB) > \frac{(A)(B)}{N}$

Negatively associated if $(AB) < \frac{(A)(B)}{N}$

And independent if $(AB) = \frac{(A)(B)}{N}$

Similar expressions can be obtained for other ultimate class frequencies such as (Aβ) , (αB) and (αβ)

The main limitation of this method is that with the help of this method we can only find out the nature of association between the attributes ,whether the association between them is Positive , Negative or Independent. We cannot determine the degree of association.

(3) Yule's Coefficient of Association In order to understand properly the significance of association or the relationship between two or more attributes ,it is necessary to find the degree of association between them.

The nature and degree of association between two attributes can be found out by applying the following formula given by Prof.G.V Yule

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha\beta)}{AB)(\alpha\beta) + (A\beta)(\alpha\beta)}$$

Where Q denotes coefficient of association. The value of Q lies between -1 and 1.

If the attributes are independent of each other, the coefficient of association will be zero.

If the attributes are perfectly or positively associated, the coefficient will be +1.

If they are completely negatively associated or disassociated , the coefficient will be -1. thus the value of coefficient of association ranges from -1 to +1.

The degree of association is measured by the coefficient of association given by Prof. YULE is as follows:

$$Q = \frac{(AB) \times (\alpha\beta) - (A\beta) \times (\alpha\beta)}{(AB) \times (\alpha\beta) + (A\beta) \times (\alpha\beta)}$$

Where : Q is coefficient of Association.

CHARACTERISTIS OF YULE'S COEFFICIENT OF ASSOCIATION

1) If Q = 0 there is no association.

   Q = +1 the association is positive and perfect.

   Q = - 1 the association is negative and perfect.

   Generally Q lies between +1 and -1.

   Yule's coefficient is independent of the relative proportion of A's and α's in the data. The value of the coefficient remains the same if all the terms containing A, α, B,β are multiplied by a constant.

   Yule's coefficient of association is superior because it provides information not only on the nature , but also on the degree of association .

(4) Coefficient of Colligation

   Prof. YULE has given another important coefficient which is also independent of the relative proportion of A's and α's is known as coefficient of colligation and is denoted by **ϒ (gamma)** which can be calculated with the help of following formula:

   This measure of association is given by Prof. Yule. The measure is denoted by the symbol Y and is given by

$$Y = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha B)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha B)}}}$$

4.5 Partial association

Partial association is also known as association in a sub universe. If two attributes A & B are associated with each other it is likely that this association may be due to the association of attributes A with C and attributes of B with C. Thus association of A & B in the sub population C is known as Partial Association.

4.6 Summary

Positive Association**:** When two attributes are present or absent together in the data, it is known as positive association.

Negative Association**:** When the presence of an attribute is associated with the absence of the other attribute, it is called negative association.

Independence**.** When there exists no association between two attributes or when they have no tendency to be present together or the presence of one attribute does not affect the other attribute, the two attributes are regarded as independent.

In order to find the degree of association between two or more sets of attributes, the coefficient of association is used. Yule's coefficient of association is superior because it provides information not only on the nature, but also on the degree of association.

Generally Q lies between +1 and -1.

4.7: Key words

- Attribute: In statistics, an attribute means a quality or characteristic which are not related to quantitative measurements and hence cannot be measured directly.
- There can be three kinds of association between attributes.
- Yule's Coefficient of Association measures the degree of association between attributes.

4.8: Answers to 'Check Your Progress'

1. In statistics, an attribute means a quality or characteristics which are not related to quantitative measurements and hence cannot be measured directly.
2. Two attributes A and B may be (i) positively associated (ii) negatively associated or (iii) independent.
3. Positive Association**:** When two attributes are present or absent together in the data, it is known as positive association.
4. Negative Association**:** When the presence of an attribute is associated with the absence of the other attribute, it is called negative association.
5. The attributes are said to be independent when the presence or absence of one attribute does not affect the presence or absence of the other.
6. Yule's Coefficient of Association measures the degree of association between attributes.
7. Yule's coefficient is independent of the relative proportion of A's and α's in the data
8. Generally Q lies between +1 and -1

4.9. Questions and Answers

Self Assessment Questions:

Multiple Choice Questions

1. The coefficient is used to
   (a) find the degree of association between two or more sets of attributes.
   (b) find the degree of association between two subgroups.
   (c) find the degree of association between the frequencies
   (d) none of these

2. The value of Q lies between
   (a) 0 and 1
   (b) -1 and 1
   (c) -1 and 0
   (d) None of the above

Answer 1(a) 2(b)

Fill in the blanks:

1. In statistics, _____ means a quality or characteristic which are not related to quantitative measurements and hence cannot be measured directly.

2. When observed and expected frequencies are equal, then both attributes are _____

Answer: 1. Attribute 2. Independent

State whether True or False:

1. Classes refer to different attributes, their sub-groups and combinations.
2. The number of attributes attached to any class is termed its class frequency.

Answer 1.True, 2. True

Short answer questions:

1. Define association of attributes.
2. How is the coefficient of association calculated?
3. What is partial association?

Long answer questions:

1. Compute the following table and find Yule's coefficient of association

N = 800, (A) = 470, (B) = 450 and (AB) = 230

2. Test the consistency of the data given below:
   Case I : (AB) = 200, (A) = 300, (α) = 200, (B) = 250, (αβ) = 150, (N) = 500
   Case II : (AB) = 250, (A) = 150, (α) = 1000, (B) = 500, (αβ) = 600, (N) = 1750
3. A teacher examined 280 students in Economics and Auditing and found that 160 failed in Economics, 140 failed in Auditing and 80 failed in both the subjects. Is there any association between failure in Economics and Auditing?
4. Eighty –eight residents of an Indian city, who were interviewed during a sample survey and classified below according to their smoking and tea drinking habits. Calculate Yule's coefficient of Association and comment on its value.

	Smokers	Non-Smokers
Drinking Tea	40	33
Non- Drinking	3	12

4.10 Further Reading:

1. Essential statistics for Economics and Business Studies: Padmalochan Hazarika, Akansha Publishing House, New Delhi.
2. Business Statistics: S. Saha, New Central Book Agency.
3. Basic Statistics: B. L. Agarwal, New Age International Limited.
4. Statistics for Management: Prentice Hall of India Limited, Levin Richard & Rubin David.
5. Quantitative Techniques for Decision Making: Anand Sharma, Himalaya Publishing House.
6. Methods of Statistical Analysis : P.S Grewal, Sterling Publishers Private Limited

Block IV : Unit-1

Measures of Inequality, Standard Deviation and Variance, Coefficient of Variation

Unit Structure:

1.0 Introduction

In this study, you will learn about the measures of inequality, concepts of variance and coefficient of variation. In addition, you will learn about the standard deviation of logarithm. It is considered to be the most useful measure of dispersion or standard deviation or root mean square deviation about the mean. Although inequality has long been topic of intense interest to sociologists, few have bothered to carefully specify what they mean by the term. Inequality can be viewed from different perspectives, all of which are related. Most common metric is *Income Inequality*, *Inequality of Wealth, and Inequality of Opportunity, lifetime Inequality etc.* Each of these inequality theories is connected to the others and offers unique yet complementary perspectives on the origins and effects of inequality. As a result, governments are better guided when developing particular policies to combat inequality.

1.1 Unit Objectives

After going through this unit, you will be able to

- Learn about the significance of range and mean deviation
- Learn about the significance of variance and coefficient of variance

1.2 Certain Measures of Inequality

1.2.1 Range

Let us consider the distribution of income of n persons. Let y_i denote the income of the i^{th}person,i=1,2...n .Then the range R is defined as the gap between the highest and the lowest income levels as a ratio of mean income . Thus the range R is given by

$$R = \frac{\text{Max } Y_i - \text{Min } Y_i}{\overline{Y}} \text{ Where } \overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

If income is divided equally, then R=0. The main limitation of range as a measure of income inequality is that it ignores the distribution in between the extremes.

Advantages

It can be easily understood

It is easy to calculate and it is the simplest method of measuring dispersion

Disadvantages

1. It is too indefinite to be used as a practical measure of dispersion because it depends entirely upon extreme values.
2. It is not based on all observations
3. It is affected by sampling fluctuations

Uses

It is used in quality control

1.2.2 The relative mean deviation

Let y_i denote the income of the i^{th} person, i=1,2...n .The relative mean deviation M is given by

$$M = \frac{\sum_{i=1}^{n} |y_i - \bar{y}|}{n\bar{y}}$$

With perfect equality M=0 and with all income going to one person $M = \frac{2(n-1)}{n}$

The main problem with the relative mean deviation is that it is not at all sensitive to income transfers whatsoever unless they cross the dividing line of μ on the way. It is therefore, rather arbitrary.

Advantages

I is easy to understand and compute.

Mean deviation about an arbitrary point is least when the point is median.

Disadvantages

In mean deviation the signs of all deviations are taken as positive and therefore it is not suitable for further algebraic treatments .

It is rarely used in social .

It is often not useful for statistical inferences.

Uses

Mean deviation and its coefficients are used in studying economic problems such as distribution of income and wealth in a society.

1.2.3 The Standard Deviation

In the present measurement concept, both the precision concept and the concept of uncertainty are expressed in terms of standard deviation or times standard deviation. Among the various measures of dispersion the standard deviation is considered to be the most useful as the other

measures lack adequacy and accuracy. The range is not satisfactory as its magnitude is determined by most extreme cases in the entire group. Mean deviation method is also an unsatisfactory measure as it ignores the algebraic signs of deviation. To some extent standard deviation is one such measure which helps to get rid of the negative sign without committing algebraic violence. The most widely accepted answer for a concise expression to understand the dispersion of data is to square the difference of each value from the group mean which will in fact give positive values. When these squared deviations are added up and then divided by the the number of values in the data, the result is the variance, which is always a positive number, but in different units than the mean. This inconvenience is removed by using the square root of the variance which is the population standard deviation or S.D. In other words S.D is the square root of the averaged squared deviations from the mean . S.D is also sometimes referred to as "root –mean –square deviation."The measure has precise mathematical significance and eliminates the disregard of signs in the average deviation, by squaring the deviations. Squaring of deviations provides added weight to the extreme items and also the deviations are recorded from the arithmetic mean. The measure is defined as the root-mean square measure of dispersion, a quadratic mean of deviations about the arithmetic mean. Thus if x_i denotes the income of the i^{th}

person, i=1,2, …,..n and $\bar{x}$ denotes their average or mean income then variance is denoted by V and is defined by

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

Now the square root of the mean of the squares of the deviations of individual items from their arithmetic mean defined by standard deviation is given by

$$\sigma = \sqrt{V} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \qquad ...(1)$$

For grouped data (discrete variable)

Stop to consider

1. What purpose does a measure of inequality serve ?
2. Why is standard deviation considered to be a good measure?
3. What is the mean and variance of a standard normal variate?

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{n} f_i}} \qquad ...(2)$$

And for grouped data (continuous variable)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i (M - \bar{x})^2}{\sum_{i=1}^{n} f_i}} \qquad ...(3)$$

Where M is the mid -value of the group.

If we take deviations from assumed mean then we get the following two alternative methods of calculating standard deviation:

a) Assumed Mean Method : In case of individual series we have the following formula:

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} \quad , d = x - A \qquad ...(4)$$

In case of frequency distribution

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \qquad ...(5)$$

b) Step deviation method: If the deviations from assumed mean A have some common factors then these deviations are divided by the highest common factor h (say) and the step deviations denoted by 'd' are obtained. Thus $d' = d/h$ where h is the H.C.F of the deviations d . In this method we have the following formula:

For individual series ,

$$\sigma = \sqrt{\frac{\sum d'^2}{n} - \left(\frac{\sum d'}{n}\right)} \times h \qquad ...(6)$$

For frequency distribution

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times h \qquad ...(7)$$

Where $d' = \frac{d}{h} = \frac{x - A}{h}$, A = Assumed mean

Since the normal distribution generally is stated in terms of standardized deviations, an innumerable number of standards exist for determining the pattern of dispersion compared to the theoretical case. Three often used standards are as follows:

$\bar{x} \pm 1\sigma$ includes 68.3% of the frequencies

$\bar{x} \pm 2\sigma$ includes 95.5% of the frequencies

$\bar{x} \pm 3\sigma$ includes 99.7% of the frequencies

In the absence of dispersion , the value of the standard deviation is zero . The size of the standard deviation varies directly in relation to the amount of dispersion. The greater the dispersion the larger its value, and the converse. Although the size of each observation affects the value of the standard deviation, extreme deviations, because of the squaring process, exert an undue weight upon the size of the standard deviation.
Some of the advantages and limitations of standard deviation are

Advantages:

1. The value of standard deviation is based on every observation in a set of data. It is the only measure of variation capable of algebraic treatment and less affected by fluctuations of sampling as compared to other measures of inequality.
2. It is possible to calculate the combined standard deviation of two or more sets of data.
3. S.D is useful in further statistical investigation. For example it plays a vital role in comparing skewness, correlation, and widely used in sampling theory.

Limitations

1. In comparison to other measures of variation, calculation of standard deviations is difficult.
2. The main demerit of variance is, that its unit is the square of the unit of measurement of variate values . For clarity , say, the variable X is measured in cms, the unit of variance is cm^2.Generally , this value is large and makes it difficult to decide about the magnitude of variation .
3. While calculating standard deviation , more weight is given to extreme values and less to those near mean . This is because of the fact that while calculating S.D , the deviations from the mean are squared , therefore large deviations when squared are proportionately more than small deviations .

Uses

It is most widely used as a measure of dispersion

It is widely used in biological studies

It is used in fitting a normal curve to a frequency distribution

1.2.4 Mathematical Properties of Standard Deviation

1. Combined Standard deviation

Let a distribution having n_1 observations has mean $\bar{x}_1$ and standard deviation σ_1 and let another distribution having n_2 observations has mean $\bar{x}_2$ and standard deviation σ_2 . Then the standard deviation σ of the distribution obtained by combining the two distributions having altogether $(n_{1+}n_{2)}$ observations is given by:

$$\sigma = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2}{n_1 + n_2}}$$

Where $d_1 = x_1 - \bar{x}$, $d_2 = x_2 - \bar{x}$ and $\bar{x} = \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$

Where $\bar{x}$ is the mean of the combined distribution.

In a similar way standard deviation of the combined distribution of two or more distributions can be obtained.

2. Standard deviation of the first n natural numbers is given by

$$\sigma = \sqrt{\frac{1(n^2 - 1)}{12}}$$

3. Standard deviation is independent of change of origin but not of scale .

1.2.5 The Standard Deviation of Logarithm

In contrast with taking of variance or standard deviation of actual values , if we take the variance and standard deviations of logarithms of the actual values then each of the variance and standard deviation gives greater importance to income transfers at the lower end. The standard deviation of logarithms is denoted by H and it is defined by

$$H = \sqrt{\frac{\sum_{i=1}^{n} (\log y_i - \log \bar{y})^2}{n}}$$

4. When is standard deviation of logarithm recommended?

One advantage of the use of logarithm is that it eliminates the arbitrariness of the units and therefore of absolute values , since a change of units, which takes the form of a multiplication of the absolute values, comes out in the logarithmic form as an addition of a constant , and therefore goes out in the wash when pairwise differences are being taken. It is ,therefore ,no wonder that the standard deviation of the logarithm has frequently cropped up as a suggested measure of inequality."[Amartya Sen , "On Economic Inequality",pp.28-29]

1.2.6 Coefficient of Variation

The standard deviation cannot be the sole basis for comparing two distributions. If we have a standard deviation of 10 and a mean of 5, the values vary by an amount twice as large as the mean itself. On the other hand, if we have a standard deviation of 10 and a mean of 5000, the

variation relative to the mean is insignificant. Therefore we cannot know the dispersion of a set of data until we know the standard deviation, the mean and how the standard deviation compares with the mean. Also if two series differ in their units of measurements , their variability cannot be measurd by any measure discussd so far. What we need is a relative measure that will give us a fuel for the magnitude of the deviation relative to the magnitude of the mean. Hence in situations where either the two series have different units of measurements , or their means differ sufficiently in size , the coefficient of variation should be used as a measure of dispersion .

The coefficient of variation is one such relative measure of dispersion. It is a simple statistic using the mean and standard deviation. It relates the standard deviation and the mean by expressing the standard deviation as a percentage of the mean. It is a unitless measure of dispersion. The unit of measure then is "percent" rather than the same unit as the original data. The formula for the coefficient of variation is

$$\text{Coefficient of variation} = \frac{s\tan dard\ \ deviation}{mean} \times 100$$

$$= \frac{\sigma}{\bar{x}} \times 100$$

Stop to consider

5. What is the utility of coefficient of variation?

6. Write a short note on coefficient of variation.

Example 1 The following nine measurements are the heights in inches in a sample of nine soldiers. Compute the standard deviation .

Height(X) : 69 66 67 69 64 63 65 68 72

Solution : Here $\sum_{i=1}^{n} xi$ = 69+66+….+72= 603

Mean ,$\bar{x}$=603/9= 67 inches

The standard deviation is given by

$$\sigma = \sqrt{V} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \qquad ...(1)$$

$$\sigma = \sqrt{\frac{64}{9}}$$

Example 2 Calculate the M.D from the A.M of the series 20,22,27,30,31,32,35,40,45,48.

Solution : We can tabulate the result as follows:

Serial No	1	2	3	4	5	6	7	8	9	10	Total
Marks (X)	20	22	27	30	31	32	35	40	45	48	330
$\lvert X-\overline{X}\rvert$	13	11	6	3	2	1	2	7	12	15	72

Here Mean $\overline{X} = \frac{\sum X}{n} = \frac{330}{10} = 33$

Therefore M.D from mean $= \overline{X} = \frac{\sum \lvert X-\overline{X}\rvert}{n} = \frac{72}{10} = 7.2$

Example 3 Compute the standard deviation for the following data:

11,12,13,14,15,16,17,18,19,20,21

Solution : Here first we calculate the mean as

$$\overline{X} = \frac{\sum X}{n} = \frac{176}{11} = 16$$

And then calculate the standard deviation as follows:

x	$(x-\overline{x})$	$(x-\overline{x})^2$
11	-5	25
12	-4	16
13	-3	9
14	-2	4
15	-1	1

16	0	0
17	1	1
18	2	4
19	3	9
20	4	16
21	5	25

Thus by formula (1)

$$\sigma = \sqrt{\frac{110}{11}} = \sqrt{10} = 3.16$$

Example 4 Find the standard deviation from the following record of number of scooter accidents in a street:

No. of accidents	1	2	4	5	6
No. of days	2	3	3	1	1

Solution : The solution is worked out using Assumed Mean Method

x	f	d=x-4	fd	fd^2
1	2	-3	-6	18
2	3	-2	-6	12
4	3	0	0	0
5	1	1	1	1
6	1	2	2	4
	N=10		$\sum fd = -9$	$\sum fd^2 = 35$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

$$= \sqrt{\frac{35}{10} - \left(\frac{-9}{10}\right)^2}$$

$$= \sqrt{3.5 - 0.81}$$

$$= 1.64$$

Example 5 Calculate standard deviation from the following data:

Age (year)	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of persons	3	61	132	153	140	51	2

Solution : Standard deviation is calculated from the following table :

Age	Mid value(x)	Frequency (f)	d =x-55	$d^/$=d/10	$fd^/$	$fd^{/2}$
20-30	25	3	-30	-3	9	27
30-40	35	61	-20	-2	-122	244
40-50	45	132	-10	-1	-132	132
50-60	55	153	0	0	0	0
60-70	65	140	10	1	140	140
70-80	75	51	20	2	102	204
80-90	85	2	30	3	6	18
		N=542			$\sum fd^/$=-15	$\sum fd^{/2}$=765

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

$$= \sqrt{\frac{765}{542} - \left(\frac{-15}{542}\right)^2} \times 10$$

$$= \sqrt{1.4114 - 0.0009} \times 10$$

$$= 11.876$$

i.e. S.D=11.876 years

Example 6. From the prices of shares X and Y below, state which is stable or more consistent in value:

X	55	54	52	53	56	58	52	50	51	49
Y	108	107	105	105	106	107	104	103	104	101

Solution : We have to calculate coefficient of variation for two series as follows:

X	$\lvert X-\overline{X}\rvert$	$\lvert X-\overline{X}\rvert^2$	Y	$\lvert Y-\overline{Y}\rvert$	$\lvert Y-\overline{Y}\rvert^2$
55	2	4	108	4	16
54	1	1	107	3	9
52	-1	1	105	1	1
53	0	0	105	1	1
56	3	9	106	2	4
58	5	25	107	3	9
52	-1	1	104	0	0
50	-3	9	103	-1	1
51	-2	4	104	0	0
49	-4	16	101	-3	9
Total		70			50

For X, Mean $(\overline{X}) = \frac{\sum_{1=1}^{n} X_i}{n} = \frac{530}{10} = 53$ and for Y, Mean $\overline{Y}) = \frac{\sum_{1=1}^{n} Y_i}{n} = \frac{1050}{10} = 105$

S.D $((\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$

$= \sqrt{\frac{70}{10}} = \sqrt{7} = 2.65$

S.D $((\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n}}$

$= \sqrt{\frac{50}{10}} = \sqrt{5} = 2.23$

For the series X

$C.V = \frac{\sigma}{\overline{x}} \times 100$

$= \frac{2.65}{53} \times 100$

$= 5$

For the series Y

$C.V = \frac{\sigma}{\overline{x}} \times 100$

$= \frac{2.23}{105} \times 100$

$= 2.123$

Since C.V for Y <C.V for X, therefore the share Y is more consistent or stable in value.

Example 7 Following is the statement of marks obtained by two students: A and B in 10 examination papers. Comment on whose marks are more consistent , A or B?

Marks scored by A	44	80	76	48	52	72	68	56	60	54
Marks scored by B	48	75	54	60	63	69	72	51	57	66

Left as an exercise for students.

Example 8 The following datagive the number of passengers travelling by Jet Airways from Guwahati to Delhi in one week .

115 122 129 113 119 124 132 120 110 116

Calculate the mean and standard deviation and determine the percentage of class that lie between (i) μ±δ (ii) μ±2δ .What percentage of cases lie outside these limits?

Solution : The calculation for mean and standard deviation are shown in the following table

x	$(x-\bar{x})$	$(x-\bar{x})^2$
115	-5	25
122	2	4
129	9	81
113	-7	49
119	-1	1
124	4	16
132	12	144
120	0	0
110	-10	100
116	-4	16
1200	0	436

$$\mu = \frac{\sum x}{N} = \frac{1200}{10} = 120 \quad and \quad \sigma^2 = \frac{\sum (x-\bar{x})^2}{N} = \frac{436}{10} = 43.6$$

$$Therefore \quad \sigma = \sqrt{\sigma^2} = \sqrt{43.6} = 6.60$$

The percentage of cases that lie between a given limit are as follows :

Interval	Values within interval	Percentage of Population	Percentage falling outside
$\mu \pm \sigma$ =120±6.60 =113.4 and 126.6	113,115,116,119,120,122,124	70%	30%
$\mu \pm \sigma$ = 120±2(6.60)	110,113,115,116,119,120,122,124,129,132	100%	nil

=106.80 and 133.20			

Example 9 The superintendent of a Civil hospital wanted to see the number of days patients stay in the hospital after surgery and for this he chose randomly 200 patients . The data are given below:

Hospital stay(in days)	1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24
Number of patients	18	90	44	21	9	9	4	5

(a) Calculate the mean number of days patients stay in the hospital along with standard deviation of the same .

(b) How many patients are expected to stay between 0 and 17 days.

Example 10 For a group of 50 male workers, the mean and the standard deviation of their monthly wages are Rs 6300 and Rs 900 respectively. For a group of 40 female workers, these are Rs 5400 and Rs 600 respectively. Find the standard deviation of monthly wages for the combined group of workers.

Stop to consider

7. Define coefficient of variation. Calculate standard deviation for the following data:

Class 1-3 3-5 5-7 7-9 9-11 11-13 13-15

Frequency 1 9 25 35 17 10 3

Solution : Given that

$n_1=50 \quad , \bar{x}_1=6300 \quad , \sigma_1=900$

$n_2=40 \quad , \bar{x}_2=5400 \quad , \sigma_2=600$

The combined mean

$$\bar{x}=\frac{n_1x_1+n_2x_2}{n_1+n_2}=\frac{50\times 6300+40\times 5400}{50+40}=\frac{315000+216000}{90}=\frac{531000}{90}=\text{Rs } 5900$$

The combined standard deviation

$$\sigma = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2}{n_1 + n_2}}$$

where $d_1 = x_1 - \bar{x} = \text{Rs.}(6300 - 5900) = \text{Rs}400$

$d_2 = x_2 - \bar{x} = \text{Rs.}(5400 - 5900) = \text{Rs} - 500$

$$\sigma = \sqrt{\frac{50 \times (900)^2 + 40 \times (600)^2 + 50 \times (400)^2 + 40 \times (-500)^2}{50 + 40}}$$

$= \text{Rs}\,900$

Example 11 Mr. Goswami wants to invest Rs 10,000 in one of the two companies A or B. Average return in a year from company A is Rs 16,000 with a standard deviation of Rs. 125, while in company B, the average return in a year is Rs .20,000 with a standard deviation of Rs 200.

Which company will you recommend to Mr. Goswami for investment? Justify your answer .

Solution: Coefficient of variation for company $A = \frac{15}{16{,}000} \times 100 = 0.78\%$

Coefficient of variation for company $B = \frac{200}{20{,}000} \times 100 = 1\%$

Since the coefficient of variation for company A is less, company A is more consistent and Mr. Goswami should invest in company A.

Uses of different measures of inequality

Though range is a very crude measure of dispersion it is mployed for a number of purposes like quality control , fluctuations in share prices, variations in money rats and rates of exchange , weather forecast etc. For example, in industry it is used for quality control . In stock markt range is used to study the fluctuations in share prices . The meteorological department uses rang to determine the difference between maximum temperature and the minimum temperature. This is on account of the fact that the public is generally interested in knowing the limits within which the temperature fluctuates on a particular day .

The mean deviation has found favour particularly with conomists and business statisticians because of the fact that the other measure –standard deviation – gives more weightage to the extreme values . Mean deviation is extensively employed in studies on the distribution of personal wealth and in studies relating to forecasting business cycles . It is the standard deviation that is most extensively employed. It forms the basis of a number of statistical techniques and is used in many other techniques as well. The theory of sampling , regression and correlation , analysis of variance etc all use standard deviation extensively.

1.3 Summary

Summary

- Range is one of the most basic measures of variation. It is the difference between the smallest data item in the set and the largest.
- Mean deviation is the average of the absolute deviations taken from a central value, generally the mean or median .
- Standard deviation (σ) is defined as the square root of the variance. It measures the variability about the mean of a data set: the closer to the mean, the lower the standard deviation. Its symbol is **σ** (the Greek letter sigma)
- Standard Deviation is a statistical tool that is used widely by statisticians, economists, financial investors, mathematicians, and government officials. It allows these experts to see how variable a collection of data is.
- The standard deviation σ or its square, the variance, cannot be very useful in comparing two series where either the units are different or the mean values are different.
- The disadvantage of the standard deviation lies in the amount of work involved in its calculation and the large weight it attaches to extreme values because of the process of squares involved in its calculation.

1.4 Key words

Range It is the difference between the largest and the smallest observations in a set **.**

Mean deviation It is the average of the absolute deviations taken from a central value, generally the mean or median

Variance It is the average of the squares of the deviations taken from mean

Standard deviation The positive square root of the variance is called standard deviation

Coefficient of variation-It is a relative measure of dispersion which relates the standard deviation and the mean by expressing the standard deviation as a percentage of the mean

1.5 Answers to 'Check Your Progress'

1. Standard deviation σ(sigma) is defined as the square root of the mean of the squares of the deviations of individual items from their arithmetic mean.
2. Coefficient of variation is defined as the square root of the variance divided by the mean income

1.6. Questions and Answers

Multiple Choice Questions

1. Which of the following is a method of measuring deviations from the average?
   (a) Root mean square (b) Lorenz Curve
   (c) Gini Coefficient (d) None of the above
2. Coefficient of variation is given by
   (a) $\frac{\sigma}{\bar{x}}$ (b) $\frac{\bar{x}}{\sigma}$ (c) $\frac{\bar{x}}{\sigma}\times 100$ (d) $\frac{\sigma}{\bar{x}}\times 100$
3. Which of the following is a unit free number
   (a) S.D (b) variance (c) M.D (d) C.V
4. Root mean square deviation from mean is
   (a) Standard deviation (b) Coefficient of variation
   (c) both (d) none

Answer: 1(b) 2(d) 3(d) 4(a)

Fill in the blanks

1. The square of standard deviation , namely σ^2 is termed as _____

2. If in a series , coefficient of variation is 64 and mean is 1, the standard deviation shall be _____

Answer 1. Variance 2. 68.26

State whether true or false

1. The range is the easiest measure to measure inequality
2. variance and coefficient of variance are the same

Answer 1. True 2. False

Match Column A with Column B

Column A	Column B
1. Square root of variance	A. 1905
2. Coefficient of variation	B. 1912

Answer: 3(A) 4(B)

Short –Answer Questions :

1. Define mean deviation.
2. Define standard deviation
3. Write two advantages of standard deviation .
4. What is coefficient of variation?
5. Why is standard deviation preferred to mean deviation?

Long Answer Questions

1. Discuss the various measures of Inequality.
2. Discuss the variance and the coefficient of variation as measures of inequality.
3. Find out the standard deviation and variance from the following frequency distribution.

Marks:	0-4	4-8	8-12	12-16
No of students	4	8	2	1

For a group of 50 male workers, the mean and standard deviation of their daily wages are Rs 72 and Rs 9 respectively. For another group of 40 female workers these are Rs 54 and Rs 6 respectively. Find the standard deviation for the combined group of 90 workers

4. From the data given below, state which series is more variable?

Variable	Series A	Series B
10-20	10	18
20-30	18	22
30-40	32	40
40-50	40	32
50-60	22	18
60-70	18	10

5. The following are the scores of two batsmen P and Q in a series of innings.

A	12	15	106	173	117	10	19	136	184	29
B	147	112	6	42	14	0	37	148	13	101

Who is the better run-getter? Who is more consistent?

6. In two factories A and B , engaged in the same activity , the average weekly wages and standard deviation are as follows:

Factory	Average weekly wages (Rs)	S.D of weekly wages (Rs)	No. of wage earners
A	460	50	100

B	490	40	80

(i) Which factory pays larger amount as weekly wages?

(ii) Which factory shows greater variability in the distribution of wages?

(iii) What is the mean and standard deviation of all workers in these two factories taken together?

7. The first of two samples has 50 items with mean 54.4 and standard deviation 8. If the whole group has 150 items with mean 51.7 and S.D 7.6, find the mean and standard deviation of the second group.

8. Explain the procedure of calculating the mean deviation from grouped frequency distribution . How standard deviation is superior than mean deviation ?

9. Weights of the students in kgs . are recorded by a machine as under:

   50 55 57 49 54 61 64 59 58 56

   If the weighing machine shows weight less by 5 kg, find correct values of range , standard deviation and coefficient of variation without calculating the correct weights.

10. A group of 100 selected students average 163.8 cm in height with a coefficient of variation of 3.2%, what was the standard deviation of their height.

1.7 Further Reading

Further Reading

Hazarika, P.L. Essential Statistics for Economics and Business Statistics .New Delhi. Akansha Publishing House,2012

Sharma. J.K Business Statistics .Pearson Education , New Delhi,2007

Gupta, S.C Fundamental of Statistics , New Delhi : S. Chand and sons ,2005

Gupta , S.P .Statistical Methods .New Delhi . S. Chand and Sons ,2005

Hooda, R.P. Statistics for Business and Economics. New Delhi: Macmillan India Ltd 2002

BLOCK IV : Unit-2

Lorenz curve and Gini coefficient

Unit Structure:

2.0 Introduction

In this study you will learn about the Lorenz Curve and Gini Coefficient . Lorenz Curve is a graphic method for studying dispersion or variation in income inequality. It was developed by Max. O. Lorenz in 1905 for representing inequality in the wealth distribution. The Gini Coefficient is a measure of statistical dispersion and it is the most popular method for operationalizing income inequality in the public health literature.

2.1 Unit Objectives

After going through this unit, you will be able to

- Understand about the significance of Lorenz Curve
- Study about the function of Gini Coefficient

2.2 The Lorenz Curve

The Lorenz Curve is a graphic method of studying variation. It was developed by Max O Lorenz in 1905 for representing inequality of the wealth distribution i.e the variability of the distribution of income and wealth is quantified using Lorenz curves. Consequently, the Lorenz Curve is a measurement of how far a statistical series' actual distribution deviates from the line

of equal distribution. The Lorenz Coefficient measures the magnitude of this deviation. There is greater inequality or unpredictability in the series if the Lorenz Curve is farther away from the line of equal distribution, and vice versa. However, it can be applied with equal advantage for comparing the distribution of profits amongst different groups of business and such other things.

Construction of Lorenz Curve

The steps involved in the construction of a Lorenz Curve are as follows:

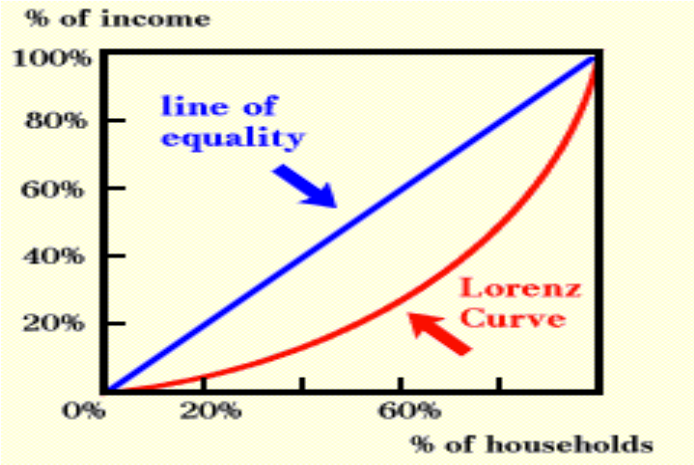
Step 1: The first step is to convert the given series into a cumulative frequency series. Then, the various items in the series are converted into percentages of the cumulative sum using the assumption that the cumulative sum of the items (or the mid-values of the class intervals) equals 100.

Step 2: The cumulative frequencies and items are plotted on a graph's X- and Y-axes, respectively, in the second step. For drawing Lorenz curve, the percentage of the population arranged from the poorest to the richest are represented on the horizontal axis (x-axis). The share of total income received by each percentage of population is represented along the (y-axis).This is also cumulated to 100. The zero percentage on the x-axis must be joined with the 100% along the y-axis. This is called the Line of Equal Distribution. This line makes an angle of 45^0 with the X-axis. Thus it is obvious that 0% of the population enjoys 0% of the income and 100% of the population enjoys 100% of income. If the wealth is equally distributed among the people then the Lorenz curve is a straight line or simply the diagonal. The further the Lorenz Curve is away from the line of equal distribution, the more unequal is the distribution of income.

Step 3: In the last step plot the actual data on the graph and obtain a curve joining the plotted points. This curve shows the actual distribution of the given statistical series.

The actual distribution curve is known as **Lorenz Curve.** If there is closeness in the Lorenz Curve to the Equal Distribution Line, it means that there is lesser variation in the distribution. However, if there is larger gap between the Lorenz Curve and the Equal Distribution Line, it means that there is greater variation in the distribution.

Besides, if two Lorenz Curves are drawn on the same graph paper, then the one which is further away from the equal distribution line shows greater variation.



Application of Lorenz Curve

A graphic measure of dispersion in a statistical series is known as Lorenz Curve. It provides the user with an immediate glimpse of the degree of variation in the given statistical distribution from its mean value; hence, is a simple measure. Prof. Lorenz first used this measure of dispersion for the measurement of economic inequality related to the distribution of income and wealth across different nations or different time periods for the same nation. Since then, the application of the Lorenz Curve has spread widely for the measurement of disparity of distribution related to various parameters of wages and profits.

The parameters in which the Lorenz Curve is now applied for the measure of dispersion are as follows:

- Distribution of Income
- Distribution of Wages
- Distribution of Wealth
- Distribution of Profits
- Distribution of Production
- Distribution of Population

Government agencies are especially interested in Lorenz curves, especially for net worth and income distributions within their country. Lorenz curves inform

governments of how public policy is working (or not working). It may also be an indicator of how a government should establish its tax brackets based on gaps or ranges of income.

Merits of Lorenz Curve

The merits of the Lorenz Curve are as follows:

1. It is attractive and gives a rough idea of the extent of dispersion.
2. Lorenz Curve makes it easy to compare two or more series.

Demerits of Lorenz Curve

Lorenz curve suffer from a serious limitation:

1. With the help of the Lorenz Curve, one can only get a relative idea of the dispersion of the given distribution as compared with the line of equal distribution. Also, it does not provide the user with any numerical value of the variability for the given distribution. The inequality measured is not expressed in quantitative terms. It merely gives a picture of the extent to which an income distribution is pulled away from the line of equal distribution which ensures that everyone has the same income.

2. It is difficult to draw Lorenz Curve.

Example Draw the Lorenz curve for the data relating to the profit of 50 manufacturing firms and show the extent of inequality present in the profits:

Profit (Rs'000)	10-20	20-30	30-40	40-50	50-60
No of firms	5	13	18	10	4

Solution : Computation of Lorenz Curve

Profit(Rs'000)	Mid Point (m)	Frequency (No. of firms ,f)	Cumulative frequency	% cumulative frequency(X)	Total value (mxf)	Cumulative total values	% cumulative total values (Y)
10-20	15	5	5	10	75	75	4.4
20-30	25	13	18	36	325	400	23.5
30-40	35	18	36	72	630	1,030	60.6
40-50	45	10	46	92	450	1,480	87.1
50-60	55	4	50	100	220	1,700	100.0

Now the Lorenz curve is drawn taking the percentage cumulative frequency (X) on X-axis and percentage cumulative total values (Y) on Y-axis . Look at figure carefully and study how it is drawn .

In recent studies, the Lorenz curve technique is used as a tool to inter distributional considerations in economic analysis. The concept of the Lorenz curve has been extended and generalized to study the relationships amogst distributions of different economic variables. The generalized Lorenz curve is called concentration curves and the Lorenz curve is only a special case to curves, viz, the concentration curve for i.

Stop to consider

8. What is a Lorenz curve ?

9. Explain Lorenz curve with the help of a diagram .

2.3 Gini Coefficient : A measure that has been widely used to represent the extent of inequality is the Gini Coefficient developed by the Italian statistician and sociologist Corrada Gini and published in his paper 'Variability and Mutability' in the year 1912. It is also known as Gini index or Gini ratio.

Gini coefficient is based on the Lorenz curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half square in which the curve lies. The

Gini coefficient is usually defined mathematically based on the Lorenz curve, which plots the proportion of the total income of the population (y-axis) that is cumulatively earned by the bottom X% of the population as illustrated by the figure. The line at 45^0 represents perfect equality of incomes. The Gini coefficient can be considered as the ratio of the area that lies between the line of equality and Lorenz curve (say A) over the total area under the line of equality (say A and B)

$$\text{i.e } G = \frac{A}{A+B}$$

Theoretically Gini coefficient can lie between the two extreme values of 0 to 1. A Gini coefficient of zero expresses perfect equality where all values are the same i.e. where everyone has an exactly equal income. A Gini coefficient of '1' expresses maximal inequality among values where only one person has all the income. However a value greater than 1 may occur if some persons have negative income or wealth. The application areas of Gini coefficient are in the study of inequalities in discipline as diverse as sociology, economics, health science , ecology , chemistry, engineering and agriculture . In fact, Gini coefficient lies between 0.5 and 0.7 for countries having highly unequal income distributions and between 0.2 and 0.35 for countries having relatively equitable distributions. The World Bank is the main organization that provides the Gini index data. However, data is only available for 130 countries. Numerous other organizations provide statistics on income inequality and the ranking of countries using the World Bank's Gini index data.

Inspite of being popular with economists and statisticians, Gini Coefficient suffers from few drawbacks:

1) The Gini index is a relative measure that fails to capture absolute differences in income. It is possible for the Gini index of a country to rise due to increasing income inequality while the number of people living in absolute poverty is actually declining. This is because the Gini index violates the Pareto improvement principle, which says income inequality can increase with an increase in all incomes in a given society.

2) The measurement of the area between the diagonal line OP and the Lorenz Curve is at times difficult .The area between the diagonal and the Lorenz curve can be obtained by using integral calculus provided we know the functional form of the Lorenz curve.

3) Two countries could have different income distributions but the same Gini index. For example, in a country where 50% of the people have no income and the other 50% of the people have equal income, the Gini index is 0.5. In another scenario, where 75% of people with no income account for 25% of a country's total income, and the top 25% of people with an income account for 75% of the country's total income, the Gini index will also be 0.5. Consequently, as a basis for ranking the differences in income inequality between countries, the Gini index could be misleading.

4) The index does not capture social benefits or interventions that bridge inequality between rich and poor.

5) The Gini index also does not capture social benefits or other interventions aimed at bridging inequality between rich and poor. Subsidised housing, healthcare, education and social grants for the vulnerable are measures that subsidise household incomes, reducing income inequality to someextent.

6) Demographic changes or characteristics of the population are not reflected by the Gini index. Countries with high ratios of elderly people whose main sources of income are pensions, or countries with high student ratios are likely to have higher levels of income inequality as measured by the Gini index.

2.4 Summary

- The Lorenz Curve is a graphic method of studying variation. It was developed by Max O Lorenz in 1905 for representing inequality of the wealth distribution.
- The closer the Lorenz curve is to the line of perfect equality, the less the inequality and smaller the Gini coefficient.
- If the wealth is equally distributed among the people then the Lorenz curve is a straight line or simply the diagonal. This line is called the line of equal distribution.

- Gini coefficient is based on the Lorenz curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half square in which the curve lies.
- Gini coefficient can lie between the two extreme values of 0 to 1.
- A Gini coefficient of zero expresses perfect equality where all values are the same i.e. where everyone has an exactly equal income.
- A Gini coefficient of '1' expresses maximal inequality among values where only one person has all the income.
- The application areas of Gini coefficient are in the study of inequalities in discipline as diverse as sociology, economics, health science, ecology, chemistry, engineering and agriculture.
- The World Bank is the main organisation that provides the Gini index data.

2.5 Key Terms

- The Lorenz Curve – The Lorenz Curve is a graphic method of studying deviations from the average.
- Gini coefficient - Gini coefficient is based on the Lorenz curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half square in which the curve lies.

2.6 Answer to '**Check Your Progress'**

1. The Lorenz Curve is a graphic method of studying deviations from the average. It was developed by Max O Lorenz in 1905 for representing inequality of the wealth distribution.
2. The Gini Coefficient is a measure of statistical dispersion.
3. Lorenz curve Visually depicts inequality across a population in a manner easy to understand and analyze
4. Is used to help calculate the Gini coefficient, a primary mathematical mean of calculating inequality

2.7 Questions and Exercise

Multiple Choice Questions

1. ___ is a measure of statistical dispersion

   (a) Standard deviation (b) Lorenz Curve

   (b) Gini Coefficient (d) None of the above

   Answer: 1 (c)

2. ___depicts inequality across a population

   (a) Standard deviation (b) Lorenz Curve

   (c) Gini Coefficient (d) None of the above

   Answer 2 (b)

Fill in the blanks

1. The Lorenz curve was developed by Max O Lorenz in _____
2. Gini Coefficient of) indicates _____

Answer 1. 1905 2. Perfect equality

State whether true or false

1. A Gini Coefficient of 1 measures perfect equality.
2. The Lorenz curve is a graphic method of studying variation.

Answer : 1.False 2. True

Match Column A with Column B

Column A	Column B
1. Lorenz curve was developed by	A. Standard deviation

2. Gini Coefficient was developed in B. Unit free number

Answer: 1(C) 2(D)

Short –Answer Questions :

1. Why was Lorenz curve developed?
2. Give one limitation of Lorenz Curve.
3. Who developed Gini Coefficient?
4. What is Gini Index?
5. What purpose does a Gini Coefficient serve ?
6. Give one limitation of Gini Coefficient.

Long Answer Questions

1. Draw a Lorenz curve of the data given below:

Income:	100	200	400	500	800
No of persons:	80	70	50	30	20

2. Discuss the importance of Gini Coefficient.
3. Why is the Lorenz Curve Important?
4. How Does the Lorenz Curve Measure Inequality?

2.8 Further Reading

Hazarika, P.L. Essential Statistics for Economics and Business Statistics .New Delhi. Akansha Publishing House,2012

Sharma. J.K Business Statistics .Pearson Education , New Delhi,2007

Gupta, S.C Fundamental of Statistics , New Delhi : S. Chand and sons ,2005

Gupta , S.P .Statistical Methods .New Delhi . S. Chand and Sons ,2005

Block IV : Unit-3

Pareto's Law of Income Distribution, deprivation index

Unit Structure:

3.0 Introduction

You will learn about the Pareto's Law of Income Distribution and Deprivation Index. The Pareto Distribution is used in describing social , scientific and geophysical phenomena in society. Pareto created a mathematical formula in the early 20^{th} century that described the inequalities in wealth distribution that existed in his native country of Italy. Deprivation index was associated with the construction of Human Development index (HDI) in Human Development Report, 1990 published by the United Nations Development Programme (UNDP, 1990).

3.1 Objective

- Understand Pareto's Law of Income Distribution
- Learn about Deprivation Index

3.2 Pareto Law of Income Distribution

In the 1890s, Vilfredo Pareto (1848-1923), who was an Italian Economist, Sociologist and Engineer studied income tax data from England, Ireland, several Italian cities , German states, and Peru. He wanted to study the problem of distribution of income among the citizens of a state. He plotted the number of people earning an income above a certain threshold against the respective threshold on double logarithmic

paper and revealed a linear relationship. It is sometimes referred to as the Pareto Principle or the 80-20 Rule. Pareto felt that he had discovered a new type of "universal law" that was the result of underlying economic mechanisms. Since then, Pareto's discovery has been confirmed and generalized to the distribution of firm size (Axtell (2001)) and wealth, which also follow Pareto distributions, at least in the upper tail. Pareto observed that 80% of the country's wealth was concentrated in the hands of only 20% of the population. The theory is now applied in many disciplines such as incomes, productivity, populations, and other variables. The Pareto distribution serves to show that the level of inputs and outputs is not always equal.

Pareto's Law of income distribution can precisely be stated as follows:

"In all places and at all times , the distribution of income in a stable economy is given approximately by the empirical formula

$y=A(x-a)^{-\beta}$

where y is the number of people having income x or greater,'a' is the lowest income at which the curve begins and A and β are certain parameters"

The graph of the equation approaches the line x=a as x tends to a and it approaches the X-axis when x tends to infinity. Thus the Pareto curve is asymptotic to the lines x=a and y=0 and it looks like a hyperbola.

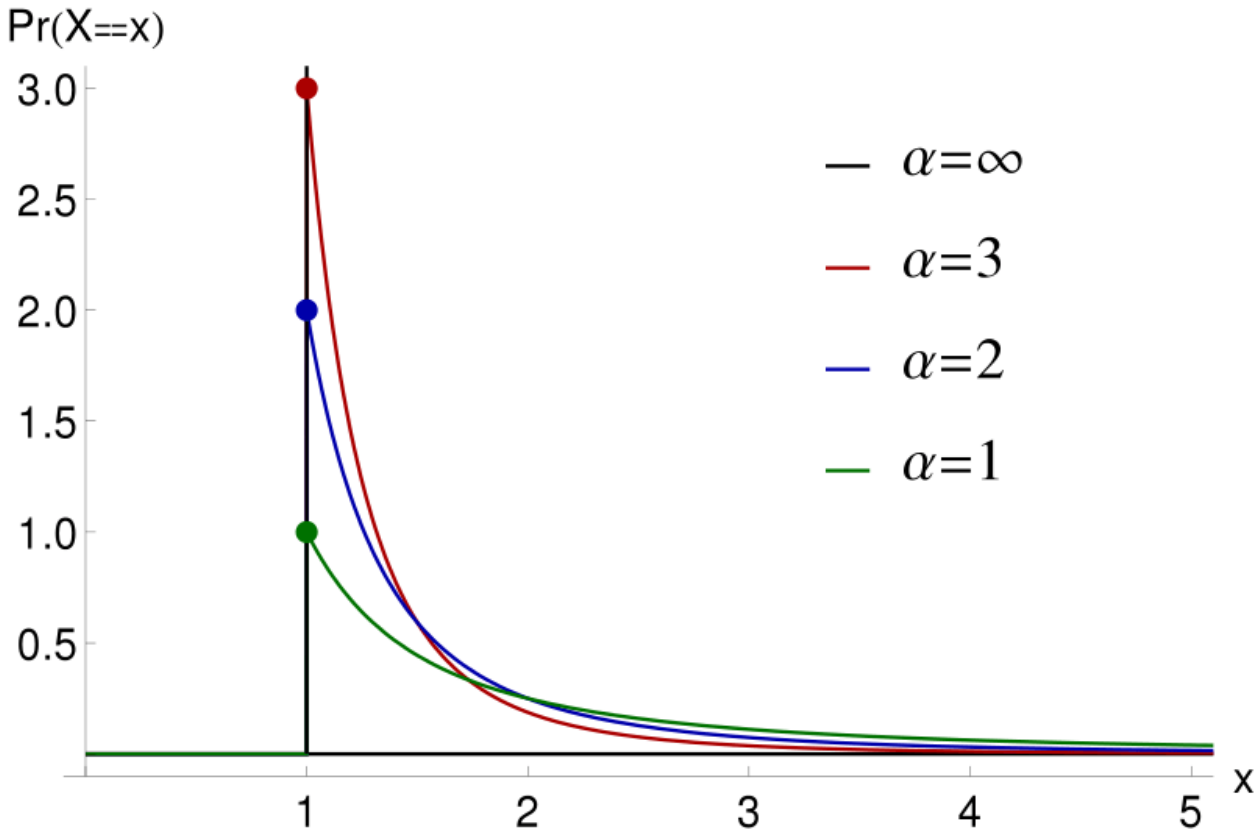
If we shift the origin to the point (a,0) then the Pareto curve takes the form

$\mathbf{Y=Ax^{-\beta}}$

Pareto observed that in many countries the value of β varied from 1.2 to 1.9 and consequently the value of gamma can be taken approximately as 1.5

Thus, gamma can be interpreted as the income elasticity of y i.e as the elasticity of the decrease in the number of persons when passing to a higher income class. Thus, the value of gamma provides a certain measure of the inequality of distribution of income. An increase in the value of gamma implies a corresponding increase in the difference between the incomes of various classes of people.

On a chart, the Pareto distribution is represented by a slowly declining tail, as shown below. Pareto chart is also called a **Pareto diagram** and **Pareto analysis**



The chart is defined by the variables α and x. It provides two main applications. One of the applications is to model the distribution of wealth among individuals in a country. The chart shows the extent to which a large portion of wealth in any country is owned by a small percentage of the people living in that country. The second application is to model the distribution of city populations, where a large percentage of the population is concentrated in the urban centers and a lower amount in the rural areas. The population in urban centers continues to increase while the rural population continues to decline as younger members of the population migrate to urban centers.

Applications of Pareto Chart

Pareto charts are the best chart to do the analysis of the bulk of data. In business industries, these charts are used very often. Let us see some of its more applications.

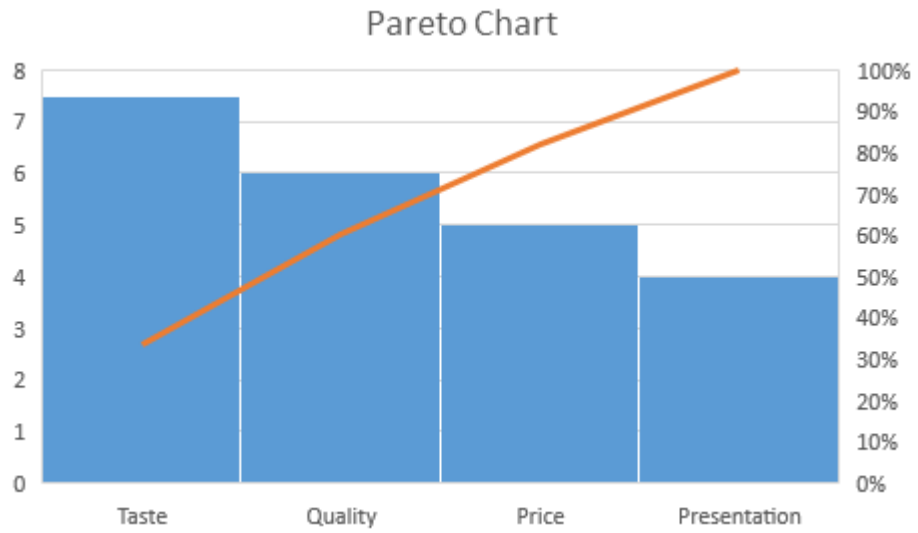
- For the analysis of the revenue growth of the organisation with respect to the time period.
- To choose for any specific data and work on it, in a broad set of data available.
- To explain to other people the set of data that one has.
- For the analysis of population growth in a city or country or all over the world every year.
- To check the global problems and focus on resolving the major one.
- To check the major complaints coming from the public and resolve them on priority

Pareto Chart Example

Let us take an example, where we need to prepare a chart of feedback analysis for XYZ restaurant, as per the reviews and ratings received from the customers. Here the customers are given a checklist of four points based on which they have to rate the restaurant out of 10. The four points are:

1. Taste of the Food
2. Quality of the food
3. Price
4. Presentation

Now, let us draw the Pareto chart for the Feedback of XYZ restaurant as per the data received



Thus, Pareto chart considers the percentage of frequency (or measure) and cumulative percentage of measures to draw a line along with bars. Also, the cumulative percentage adds up to 100.

Limitations of Pareto's Law

While the 80-20 Pareto distribution rule applies to many disciplines, it does not necessarily mean that the input and output must be equal to 100%. For example, 20% of the company's customers could contribute 70% of the company's revenues. The ratio brings a total of 90%. It shows that the Pareto concept is merely an observation that suggests that the company should focus on certain inputs more than others.

Pareto's curve does not seem to fit better for lower incomes although it fits better for higher incomes. Pareto's law of income distribution is not effective for all types of economies. It is usually relevant to incomes in the capitalist countries and also the countries with feudal and early capitalist conditions . The definition is very rigid and stated in a very general form. Because of this Pareto's Law has been subjected to serious criticism by a number of scientific investigations. However, the distribution of income in every stable economy is found to follow approximately Pareto pattern.

3.3 Deprivation Index

Deprivation index was associated with the construction of Human Development index (HDI) in Human Development Report, 1990 published by the United Nations Development Programme (UNDP, 1990). Deprivation index is a composite index. One minus the deprivation index is the development index.

There are various components of deprivation. Usually the following components of deprivation in respect of a particular geographical region are taken into consideration.

(i) Life expectancy at birth(in years)

(ii) Adult literacy rate (in per cent)

(iii) Combined primary, secondary and tertiary enrolment ratio

(iv) Per capita Gross Domestic Product(GDP)

The deprivation indicator in respect of a particular above-mentioned component is obtained by using the following formula

Where maximum value and minimum value are specified for a particular geographical region and actual value denotes the value of the indicator at a particular point of time in the given region.

The average of the deprivation indicators in respect of all the four components of deprivation as mentioned above is taken. This is the composite deprivation index. One minus this composite deprivation index is the development index for the region

3.4 Summary

Pareto's Law of Income Distribution forms the basis of the well-known, but often overlooked, 'eighty-twenty' rule that a small proportion of customers (or donors) are accountable for a very large share of sales turnover or income.

Pareto Law of Income Distribution is used to model the distribution of wealth among individuals in a country

Deprivation index is a composite index

One minus the deprivation index is the development index

3.5 Key words

Pareto Law of Income Distribution : Pareto created a mathematical formula in the early 20^{th} century that described the inequalities in wealth distribution that existed in his native country of Italy

Pareto Curve: On a chart, the Pareto distribution is represented by a slowly declining tail called Pareto Curve

Deprivation Index: It is an index used in the construction of human development index

Development Index: One minus the deprivation index is the development index

3.6 Answers to 'Check Your Progress'

What is a Pareto chart used for?

What is the 80/20 rule of Pareto charts?

Give example of a Pareto Chart.

What is development index?

3.7 Questions and Answers

How do you create a Pareto chart?

What are the applications of Pareto Chart?

What is the deprivation index ?

What are the components of the deprivation index?

3.8 Further Reading

Hazarika, P.L. Essential Statistics for Economics and Business Statistics .New Delhi. Akansha Publishing House,2012

Sharma. J.K Business Statistics .Pearson Education , New Delhi,2007

Gupta, S.C Fundamental of Statistics , New Delhi : S. Chand and sons ,2005

Gupta , S.P .Statistical Methods .New Delhi . S. Chand and Sons ,2005

Hooda, R.P. Statistics for Business and Economics. New Delhi: Macmillan India Ltd 2002

BLOCK V : Unit-1

Decision Theory Concepts

Unit Structure

1.0 Introduction

In this unit, you will learn about decision theory also known as decision analysis. The basic function of any business manager is decision making for various purposes. The quality of decision making depends on the quantity and quality of information available to him or utilised by him. Decision theory is a systematic approach for such a function. Elements of the decision process are alternative courses of action, states of nature, pay-off and pay-off table.

1.1 Unit Objectives

After going through this unit, you will be able to

- Learn about the elements of a decision process
- Learn about payoff matrix and loss table

1.2 Decision Theory Concepts

Statistical decision theory provides an analytical and systematic approach to the study of decision making wherein data concerning the occurrence of different outcomes (consequence) may be evaluated to enable the decision maker to identify a suitable decision alternative (or

course of action). Irrespective of the type of decision model, there are certain essential characteristics which are common to all as listed below:

1. Acts- A decision maker has to first determine the alternative course of action so called, acts of strategies from which he has to choose one considered to be the best.
2. States of nature – The circumstances, which affect the outcome of a decision problem but are beyond the control of the decision maker are known as states of nature or events.
3. Payoff – A numerical value resulting from each possible combination of alternatives and states of nature is called payoff. The payoff values are always conditional values because of an unknown states of nature.

Self Assessment Questions

1. What is a pay off table ?
2. What is opportunity loss?

A tabular arrangement of these conditional outcomes (payoff) values is known as payoff matrix as shown in Table 8.1

Table 1.1 General form of Payoff Matrix

	Courses of action (Alternatives)				
States of Nature	S_1	S_2	...	...	S_n
N_1	P_{11}	P_{12}			P_{1n}
N_2	P_{21}	P_{22}			P_{2n}
...					
...					
N_n	P_{m1}	P_{m2}			P_{mn}

4. Opportunity Loss table – The opportunity loss is defined to be the difference between the highest possible profit for a state of nature and the actual profit obtained for the particular decision taken i.e. an opportunity loss is the loss incurred due to the failure of not adopting the best possible course of action or strategy. The pay off table which represents

the cost or loss incurred because of failure to take the best possible action is called the opportunity loss table. If for a given state of nature N_i , the n^{th} pay offs corresponding to the n courses of action are given by $p_{i1},p_{i2},...,p_{in}$ and if M stands for the maximum of these quantities the respective opportunity losses asre computed as shown in the table.

Table 1.2 Regret (Opportunity loss table)

	Courses of action (Conditional Opportunity Loss)				
States of Nature	S_1	S_2	...	...	S_n
N_1	M_1-P_{11}	M_1-P_{12}			M_1-P_{1n}
N_2	M_2-P_{21}	M_2-P_{22}			M_2-P_{2n}
...					
...					
N_n	M_m-P_{m1}	M_m-P_{m2}			M_m-P_{mn}

1.3 Summary .

A decision may be defined as the process of choosing an alternative course of action , given that atleast two alternatives exist.

State of nature refers to a future event not under the control of the decision maker.

Payoff is the benefit that accrues from a given combination of a decision alternative and a state of nature.

Opportunity loss refers to the profit that could have been earned if stock had been sufficient to supply a unit that was demanded.

Alternative courses of action

Conditional profit values are the profit that would result from a given combination of decision alternatives and state of nature.

1.4 Key Terms

Decision: It is a conclusion reached after due consideration.

Alternative courses of action: It refers to a choice of two or more possibilities of things, propositions, the selection of which preclude any other possibility:

1.5 Check Your Progress

Decision theory is a systematic approach for decision making depending on the quality and quantity of information available to him or utilised by him.

Various elements of a decision-making process include decision alternatives, state of nature, payoff, pay-off table and opportunity loss table.

1.6 Questions and Exercises

Self Assessment Questions

Multiple Choice Questions

1. Elements of the decision process include

(a) pay-off (b) pay-off table (c) regret table (d) all of the above

2. Benefit that accrues from a given combination of a decision alternative and a state of nature.

(a) pay-off (b) loss (c) alternatives (d) all of the above

Answer : 1(d) 2(a)

Fill in the blanks

1. A _____may be defined as the process of choosing an alternative course of action .

2. _____refers to a choice of two or more possibilities of things.

Answer : 1. Alternative 2. Alternative course of action

State whether True or False

1. A payoff table cannot include the probability value for each event .

2. Acts are referred to as the strategies available to the decision maker

Answer: 1.False 2. True

Match Column A with Column B

	Column A		Column B
1	It is an element of decision process.	A	Strategies available to the decision maker
2	Acts are	B	Opportunity Loss Table

Answer: 1(B) 2(A)

Short Answer Questions

1. Define a pay off table.

2. Define opportunity loss table.

3. What are the elements of decision making process?

Long –Answer Questions

1. Discuss the various elements of the decision making process.

2. Calculate the loss table from the following pay-off table.

Action	Events			
	E_1	E_2	E_3	E_4
A_1	50	300	-150	50
A_2	400	0	100	0
A_3	-50	200	0	100
A_4	0	300	300	0

3. What is a pay-off table?

4. A baker produces a certain type of special pastry at an average cost of Rs 3 and sells it a price of Rs 5. This pastry is produced over the weekend and is sold during the following week, such pastry being produced but not sold during a week's time are totally spoiled and have to be thrown away. According to past experience the weekly demand for these pastries is never less than 78 or greater than 80 . Formulate action space, pay-off table and loss table

1.7 Further Reading

Hazarika, P.L. Essential Statistics for Economics and Business Statistics .New Delhi. Akansha Publishing House,2012

Sharma.J.K Business Statistics .Pearson Education , New Delhi,2007

Gupta,S.C Fundamental of Statistics , New Delhi : S.Chand and sons ,2005

Gupta, S. P .Statistical Methods .New Delhi . S.Chand and Sons ,2005

Hooda, R.P . Statistics for Business and Economics. New Delhi: Macmillan India Ltd 2002

Sharma Anand. Statistics for Management, Himalaya Publishing House, Geetanjali Press Pvt . Ltd, Nagpur

BLOCK V : Unit-II

Steps in Decision Making Process

Unit Structure:

2.0 Introduction

In this unit, you will learn about decision theory also known as decision analysis. . A decision may be defined as the process of choosing an alternative course of action given that at least two alternatives exist. A decision is taken under three situations, viz (a) certainty (b) uncertainty and (c) risk. It is trivial to take a decision under certainty because a decision-maker knows what will be the result when a particular decision is taken. A decision under uncertainty means the choice of action out of the many courses of action at hand when the outcome of any action is unknown . In decision making, under risk, more than one state of nature and decision-maker has sufficient information to assign probabilities to each of these states.

2.1 Unit Objectives

After going through this unit, you will be able to

- Learn about the steps in the decision making process
- Explain about the various types of decision –making environment

2.2 Steps in the Decision Making Process

The decision making process involves the following steps :

1. Identifying and defining the problem
2. Listing of all possible future events , called states of nature , which can occur in the context of decision problem
3. Identification of all the course of action which are available to the decision maker.
4. Construct a pay off table for each possible combination of alternatives course of action and state of nature.
5. Choose the criterion that results in the largest pay –off.

2.3 Decision making Environment

Decisions are made under three types of environments:

1. Decision making under condition of certainty: In this case the decision maker has complete knowledge (perfect information) of the consequence of every decision choice (alternative) with certainity. Obviously, he will select an alternative that yields the largest payoff.
2. Decision making under uncertainity : In this case the decision maker is unable to specify the probabilities with which the various states of nature (future) will occur.
3. Decision making Under conditions of Risk –As in case of decision making under conditions of uncertainity here also more than one states of nature exist. But here the decision maker has sufficient information to allow him to assign probabilities to the various states of nature.

   Under conditions of uncertainity , the decision maker has knowledge about states of nature that happens but lacks the knowledge about the probabilities of their occurrence.

   Under conditions of uncertainity , a few decision criterions are available which could be of help to the decision maker.

   (1) Maximax Criteron or Criterion of Optimism: This criterion provides the decision maker with optimistic criterion . The working method is summarised as follows:
   - Locate the maximin payoff value corresponding to each alternative
   - Select an alternative with maximin payoff

(2) Maximin Criterion or Criterion of Pessimism: This criterion provides the decision maker with pessimistic criterion. The working method is summarised as follows:

- Locate the minimum payoff value corresponding to each alternative
- Select an alternative with maximum pay-off value

(3) Minimax Criterion or Minimum Regret Criterion : This criterion is also known as opportunity loss decision criterion or minimax regret criterion . The working method is summarised as follows:

Determine the amount of regret corresponding to each alternative for each state of nature.

- The regret for j^{th} event corresoponding to the i^{th} alternative is given by
  i^{th}regret =(maximum payoff –ith payoff) for j^{th} event
- Determine the maximum regret amount for each alternative
- Choose the alternative which corresponds to the minimum of the maximum regrets.

(4) Hurwicz Criterion or Criterion of Realism : Also called weighted average criterion , it is a compromise between the maximax (optimistic) and minimax (pessimistic) decision criterion . This concept allows the decision maker to take into account both maximum and minimum for each alternative and assign them weights according to his degree of optimism (0r pessimism). The working method is summarised as follows:

- Choose an appropriate degree of optimism ,α so that $(1-\alpha)$ represents degree of pessimism
- Determine the maximum as well as minimum of each alternative and obtain
  $P= \alpha$. Maximum $+(1-\alpha)$ Minimum for each alternative
- Choose the alternative that yields the maximum value of P

(5) Laplace Criterions or Criterion of Rationality – Also known as equal probabilities criterions or criterion of rationality . Since the probability of states of nature are not

known , it is assumed that all states of nature will occur with equal probability . The working method is summarised as follows:

- Determine the expected value for each alternative ; if n denotes the number of events and P_i's denote the payoffs , then expected value is given by

  $\frac{1}{n}[P_1 + P_2 + ... + P_n]$

- Choose the alternative that yields the maximum value of P.

Stop to consider

Payoff – A numerical value resulting from each possible combination of alternatives and states of nature is called payoff.

Decisions are made under three types of environments :

Decision making under condition of certainity

Decision making under condition of uncertainity

Decision making under condition of risk

Example 1 A company has to choose one of the three types of biscuits , Marie Gold , Good day , Oreo. Sales expected during next year are highly uncertain .marketing Department estimates the profits considering manufacturing cost , promotional efforts and distribution set up as given in the table below.

Types of biscuits	Profits on estimated level of sales (in lakhs) for quantities		
	5,000	10,000	20,000
Marie Gold(M)	15	25	45
Good Day(G)	20	55	65
Oreo(O)	25	40	70

Solve using decision making under uncertainity.

Solution : Maximin criterion : In using the maximin criterion , the decision maker adopts a pessimistic approach and tries to maximise his security in the face of a highly uncertain situation. For worst situation , the payoffs are 15, 20 and 25 for M,G and O respectively. Even at the pessimistic level , the manager tries to make the best of the situation reaching the decision as Oreo , pay off being the best amongst the worst. This maximises the minimum pay-off. Hence, the company will launch Oreo(O).

MinimaxCriterion : Here, the maximum pay-offs are 45,65 and 70 in three cases. In order to gain atleast the minimum of these maximams , minimum pay off is 45 for Marie gold (M). Hence the strategy will be to launch Marie Gold(M).

MaximaxCriterion : In this case , the decision maker become totally optimistic and chooses the strategy that makes the best of the best . The largest payoff for each type are 45, 65 and 70. The maximum payoff being 70, the company choose to launch Oreo(O).

Laplace Criterion : when decision maker has no definite information about the probability of occurrence of various states of nature , he makes simple assumption that each is equally likely. Therefore , the probability of each to occur is 1/3.

Expected pay-offs are :

$$E(M) = \frac{1}{3} \times 15 + \frac{1}{3} \times 25 + \frac{1}{3} \times 45 = 28.33$$
$$E(G) = \frac{1}{3} \times 20 + \frac{1}{3} \times 55 + \frac{1}{3} \times 65 = 46.67$$
$$E(O) = \frac{1}{3} \times 25 + \frac{1}{3} \times 40 + \frac{1}{3} \times 70 = 45$$

The strategy for Good Day expects the maximax pay-off. Hence , the decision would be to launch Good Day biscuit.

Hurwicz Criterion : Maximin and Maximax are two extremes on the scale of optimism . It would be pragmatic to assume that a business manager's attitude would fall somewhere in between rather than at either extremes. Hurwicz ,therefore, propounds a combination of the two criteria in what is known as Hurwicz Alpha criterion.

In this case , the decision maker's degree of optimism is represented by α, the coefficient of optimism , varying between 0 and 1 ; α=0 denoting total pessimism and α=1 , total optimism .

A decision index D_i is defined by

$D_i=\alpha M_i+(1-\alpha)m_i$ where

M_i=Max. pay-offs from any of the outcomes resulting from the i^{th} strategy

m_i= min payoff from any of the outcomes resulting from the i^{th} strategy

For each strategy , the value of decision index is found and strategy with the highest values of outcome is chosen .

Let us assume α=0.6 in this example

$D_i(M)=(0.6x45)+(1-0.6)x15=33$

$D_i(G)=(0.6x65)+(1-0.6)x20=47$

$D_i(O)=(0.6x70)+(1-0.6)x25=52$

The strategy chosen , therefore , would be to launch Good day (G) producing the best outcome i.e. 52

Regret Criterion : Loss of opportunity is a common phenomenon in the business world. The Regret Criterion , the dissatisfaction associated with not having got the best that would have been possible if the state of nature of occurrence were known in advance.

A measure of regret of an outcome is the opportunity cost computed as the difference in pay-off of the outcome and the largest pay-off which could have been obtained under the corresponding state of nature . This table so obtained is called Opportunity Loss Table (OL table)

Revised pay-off (subtracting pay-off from highest of that event)i.e. regret pay-off

	5000	10,000	20,000	Max. regret
Marie Gold(M)	25-15	55-25	70-45	30
Good Day(G)	25-20	55-55	70-65	5

Oreo(O)	25-25	55-40	70-70	15

Minimum of maximum regret is 5 corresponding to Good day biscuits .Hence it is decided to launch Good Day(G).

The choice of the decision maker is the reason for inconsistency in the above results . The personality of the decision maker plays an important role in these decisions.

Stop to consider

Criterions for decision making under uncertainity-

Maximax Maximin Hurwicz

Minimax Laplace Regret

Example 2 Consider a bottling company that is thinking of various alternatives to increase its production to meet the increasing market demand . Use the various criterions of decision making under uncertainity to arrive at a solution .

Alternatives	State of nature (Product Demand)			
	High	Moderate	Low	Nil
Expand	50,000	25,000	-25,000	-45,000
Construct	70,000	30,000	-40,000	-80,000
Subcontract	30,000	15,000	-1,000	-10,000

Solution : (1) Maximin Criterion

Alternatives	State of nature (Product Demand)				Maximum of Rows
	High	Moderate	Low	Nil	
Expand	50,000	25,000	-25,000	-45,000	50,000
Construct	70,000	30,000	-40,000	-80,000	70,000
Subcontract	30,000	15,000	-1,000	-10,000	30,000

Thus the maximax payoff is rs 70,000 corresponding to the alternative " construct"

(ii) Maximin Criterion

Alternatives	State of nature (Product Demand)				Minimum of Rows
	High	Moderate	Low	Nil	
Expand	50,000	25,000	-25,000	-45,000	-45,000
Construct	70,000	30,000	-40,000	-80,000	-80,000
Subcontract	30,000	15,000	-1,000	-10,000	-10,000

Thus the maximin payoff is Rs -10,000 corresponding to the alternative 'subcontract'

(iii)) Minimax Criterion or Minimum Regret Criterion

Alternatives	State of nature (Product Demand)			
	High	Moderate	Low	Nil
Expand	50,000	25,000	-25,000	-45,000
Construct	70,000	30,000	-40,000	-80,000
Subcontract	30,000	15,000	-1,000	-10,000

Calculation of Regret

Alternatives	State of nature (Product Demand)				Maximum of Rows
	High	Moderate	Low	Nil	
Expand	20,000	5000	24,000	35,000	-45,000
Construct	0	0	39000	70000	-80,000
Subcontract	40000	15000	0	0	-10,000

This table shows that the company will minimize its regret to Rs 35,000 by selecting alternative 'expansion'

(iv) Hurwicz Criterion

Let $\alpha=0.8$

Alternatives	State of nature (Product Demand)				Max of rows	Min of rows	H=α max+(1-α) min
	High	Moderate	Low	Nil			
Expand	20,000	5000	24,000	35,000	50,000	-45,000	31000
Construct	0	0	39000	70000	70,000	-80,000	40000
Subcontract	40000	15000	0	0	30,000	-10,000	22000

Thus according to the HurwiczCriterions , the company will choose alternative ' construct'

(v) Laplace Criterion

Alternatives	State of nature (Product Demand)				Expected Pay Off
	High	Moderate	Low	Nil	
Expand	20,000	5000	24,000	35,000	1250
Construct	0	0	39000	70000	-5000
Subcontract	40000	15000	0	0	8500

Note : $(EP_1) = 1/4[50,000+25,000+(-25000)+(-45000)]=1250$ etc.

Decision making Under Risk :

Here more than one state of nature exists and the decision maker has sufficient information to assign probabilities to each of these states. These probabilities could be obtained from the past records or simply the subjective judgement of the decision maker. Under conditions of risk , a number of decision criterions are available which could be of help to the decision maker.

(1) Expected Value Criterions – The expected monetary value for a given course of action is the weighted sum of possible payoffs for each alternative . It is obtained by summing the payoffs for each course of action multiplied by the probabilities associated with state of nature. It consist of the following steps:

Stop to Consider

Three criterions for decision making under risk are :

Expected value Criterion

Expected Opportunity Loss

Expected Value of Perfect Information

Construct a payoff table listing the alternative decisions and the various state of nature . Enter the conditional profit for each decision event combination along with the associated probabilities.

Calculate the EMV for each decision alternative by multiplying the conditional profits by assigned probabilities and adding the resulting conditional values.

- Select the alternative that yields the highest EMV.

(2) Expected Opportunity Loss (EOL) Criterion : In this approach , first construct a conditional profit table for each decision –event combination along with the associated probabilities . for each event , compute the conditional opportunity loss by subtracting the corresponding payoff from the maximum payoff for that event . Calculate the expected opportunity loss for each alternative by multiplying the conditional opportunity losses by the assigned probabilities and summing up their product . The alternative that yields the lowest EOL is selected.

(3) Expected Value of Perfect Information (EVPI) Perfect information means complete and accurate information about the future demand and that remove all the uncertainity for future.

EVPI represents the maximum amount of money the decision maker has to pay to get this additional information about the occurrence of various state of nature before a decision has to be made . The procedure to calculate the expected value of perfect information is as follows:

- Construct conditional profit table with perfect information

- Construct expected profit table with perfect information
- Determine EVPI from the following relation

EVPI = EPPI- max EMV

Example 3 A newspaper boy has the following probabilities of selling a magazine

No of copies sold	Probability
10	0.10
11	0.15
12	0.20
13	0.25
14	0.30

Cost of the copy is 30 paisa and sale price is 50 paise. He cannot return the unsold copies. How many should he order?

Solution

(1) EMV Criterion

CP =30 p

SP = 50 p

Profit = SP-CP=20 p

We construct the Conditional Profit table

Conditional Profit = Profit *S.P =20 S.P when $D \geq S$

= 50 D-30 S, when $D < S$

Possible Demand	Probability	Possible Stock				
		10	11	12	13	14
10	0.10	200	170	140	110	80
11	0.15	200	220	190	160	130
12	0.20	200	220	240	210	180
13	0.25	200	220	240	260	230
14	0.30	200	220	240	260	280

The news boy must , therefore , order 12 copies to earn the highest possible average daily profit of 222.50 paise

(2) Expected Opportunity Loss

Using the conditional profit table , the conditional loss table is prepared by the following relation

Row maximum – other elements of row

Possible Demand	Probability	Possible Stock				
		10	11	12	13	14
10	0.10	200	170	140	110	80
11	0.15	200	220	190	160	130
12	0.20	200	220	240	210	180
13	0.25	200	220	240	260	230
14	0.30	200	220	240	260	280

Opportunity Loss table

Possible Demand	Probability	Possible Stock				
		10	11	12	13	14
10	0.10	0	3	6	9	12
11	0.15	3	0	4.5	9	13.5

12	0.20	8	4	0	6	12
13	0.25	15	10	5	0	7.5
14	0.30	24	18	12	6	0
EOL		50	35	27.5	30	45

The optimum stock action is the one which minimizes the expected opportunity loss . Therefore the newspaper boy should keep a stock of 12 copies where there is a minimum expected loss of 27.5 paise.

(iii) Expected profit table With Perfect Information

Possible Demand	Probability	Conditional profit Under Certainity	Expected profit With Perfect Information
10	0.10	200	20
11	0.15	220	33
12	0.20	240	48
13	0.25	260	65
14	0.30	280	84

The expected value of perfect information is given by

EVPI= EPPI-max EMV

= 250-222.5

= 27.5 paise

Thus this is the maximum amount which the newspaperboy is willing to pay per day for perfect information .

Example 4 A manufacturer is faced with a problem of fast change of technology and hence , fast change in the product line . At this point of time , the research and development wing of the organization has suggested an improved new product line with easy acceptance . It will cost the manufacturer Rs 60,000 for the pilot testing and development testing before establishing the

product in the market . The organisation has 100 customers and each customer , might purchase , at the most ,one unit of the product , , due to its cost and newness . The selling price suggested is Rs 6,000 for each unit and selling estimate is Rs 2,000 for each unit. The probability distribution for proportion of customers buying the product is estimated as follows:

Proportion of Customers	Probability
0.04	0.1
0.08	0.1
0.12	0.2
0.16	0.4
0.20	0.2

Work out the expected opportunity losses and suggest whether the manufacturer should develop the product or not .

Solution : Let p be the proportion of customers , who purchase the new product . The conditional profit would be governed by the relationship as

(6,000-2,000)x100p-(60,000)=Rs (40,000p-60,000)

The conditional profit and opportunity loss are given as under :

Nature of state (proportion of customers)	Probability	Conditional Profit		Opportunity Loss (Rs.)	
		A_1(develop the product)	A_2(donot develop the product)	A_1	A_2
0.04	0.1	-44,000	0	44,000	0
0.08	0.1	-28,000	0	28,000	0
0.12	0.2	-12,000	0	12,000	0
0.16	0.4	4,000	0	0	4,000

0.20	0.2	20,000	0	0	20,000

Hence , EOL (A_1)=44,000x0.1+28,000x0.1+12,000x0.2+ 0x0.04+0x0.02

= Rs 9,600

To seek minimisation of opportunity loss, the manufacturer should not develop the product.

2.4 Summary

Maximax criterion- For uncertain conditions, a decision made based on best of the best alternatives

Maximin criterion-A decision under uncertainity using the best opportunity for the most pessimistic outcomes.

Laplace criterion –Under uncertain conditions, when we use equal probability for all opportunities it is , called Laplace criterion of decision making.

Expected Opportunity Loss- It is the minimum value expected outcome of the situation from all the available alternatives. It is the product of the conditional pay-off with the corresponding value of the probability.

EVPI- In probabilistic situations, the improvement in the expected value of the outcome due to better available information.

2.5 Key Terms

Decision: It is a conclusion reached after due consideration.

Alternative courses of action: It refers to a choice of two or more possibilities of things, propositions, the selection of which preclude any other possibility:

Decision under uncertainity: Here more than one state of nature exists but the decision maker lacks the knowledge about the probabilities of their occurrence.

Decision under risk: Here more than one state of nature exists and the decision-maker has sufficient information to assign probabilities to each of these states.

2.6 Check Your Progress

Decision theory is a systematic approach for decision making depending on the quality and quantity of information available to him or utilised by him.

There are three types of decision -- making environments: certainty, uncertainty and risk.

Various elements of a decision-making process include decision alternatives, state of nature, payoff, pay-off table and opportunity loss table.

2.7 Questions and Exercises

Self Assessment Questions

Multiple Choice Questions

1. A type of decision making environment is

(a) certainty (b) uncertainty (c) risk (d) all of these

2. Elements of the decision process include

(a) pay-off (b) pay-off table (c) regret table (d) all of the above

3.Which of the following needs the coefficient of optimism (α)

(a) equally likely (b) maximin (c) realism (d) minimax

4. Which of the following is not used for decision making under uncertainty?

(a) maximin (b) maximax (c) minimax (d) minimize expected loss

5. The value of the coefficient of optimism (α) is needed while using the criterion of

(a) equally likely (b) maximum

(c) realism (d) minimax

Answer : 1(d) 2(d) 3(d) 4(c)

Fill in the blanks

1. Decision making under _____ refer to situations where more than one outcome can result from any single decision.

2. Coefficient of _____ is needed while using the criterion of realism .

3. _____has more than one state of nature exists and the decision-maker has sufficient information to assign probabilities to each of these states.

Answer : 1. Uncertainity 2.Optimism 4. Decision under risk

State whether True or False

1. Hurwicz criterion is also called the weighted average criterion .

2. Maximax criterion comes under decision making under risk.

Answer : 1. True 2. False

Match Column A with Column B

	Column A		**Column B**
1.	It is an element of decision process.	A.	Laplace Criterion
2.	It is helpful for decision maker	B.	Decision making under certainty
3.	It is based on what is known as the principle of insufficient reason.	C.	Opportunity loss table
4.	Whenever there exists only one outcome for a decision, we are dealing with this category	D.	Decision tree

Answer : 1(C) 2(D) 3(A) 4(B)

Short Answer Questions

1. What are the elements of decision making process?
2. Name some criterion that helps in decision making under uncertainty.
3. Name some decision criterion under condition of risk.
4. Define Hurwicz criterion.
5. Define EOL.
6. What is EVPI?
7. Give example of a decision taken under certainty.

Long –Answer Questions

1. Discuss the various elements of the decision making process.

2. Discuss the difference between decision making under certainty, uncertainty and risk .

3. Briefly explain 'expected value of perfect information' with examples.

4. A product of boats has estimated the following distribution of demand for a particular kind of boat:

No. demanded	0	1	2	3	4	5	6
Probability	0.14	0.27	0.27	0.18	0.09	0.04	0.01

Each boat costs him Rs 7000 and he sells them for Rs 10,000 each . Boats left unsold at the end of the season must be disposed of for Rs 6000 each . How many boats should be in stock so as to maximise his expected profit ?

5. Calculate the loss table from the following pay-off table.

Action	Events			
	E_1	E_2	E_3	E_4
A_1	50	300	-150	50
A_2	400	0	100	0
A_3	-50	200	0	100

A_4	0	300	300	0

Suppose that the probabilities of the events in this table are

$P(E_1)=0.15$ $P(E_2)=0.45$ $P(E_3)=0.25$ $P(E_4)=0.15$

Calculate the expected pay-off and expected loss to each action.

6. The estimated sales of proposed types of perfumes are as under:

Types of perfumes	Estimated Sales		
	Rs 20,000	Rs 10,000	Rs 2,000
A	25	15	10
B	40	20	5
C	60	25	3

Help the manager to take effective decisions under minimax and Laplace method.

2.8 Further Reading

Hazarika, P.L. Essential Statistics for Economics and Business Statistics .New Delhi. Akansha Publishing House,2012

Sharma.J.K Business Statistics .Pearson Education , New Delhi,2007

Gupta,S.C Fundamental of Statistics , New Delhi : S.Chand and sons ,2005

Gupta, S. P .Statistical Methods .New Delhi . S.Chand and Sons ,2005

Hooda, R.P . Statistics for Business and Economics. New Delhi: Macmillan India Ltd 2002

Sharma Anand. Statistics for Management, Himalaya Publishing House, Geetanjali Press Pvt . Ltd, Nagpur

BLOCK V: Unit-3

Decision Tree

Unit Structure:

3.0 Introduction

You will also learn about decision trees. Decision tree analysis involves the construction of a diagram showing all the possible courses of action, states of nature, and the probabilities associated with the state of nature. The decision diagram looks very much like the drawing of a tree , therefore also called a decision –tree. The basic advantage of decision tree approach is that it structures the decision process and helps decision making in an orderly , systematic and sequential manner.

3.1 Unit Objectives

After going through this unit, you will be able to

- Describe about the importance of decision tree

3.2 Decision Tree analysis

A decision tree is a graphic display of various decision alternatives and the sequence of events as if they were branches of a tree. In decision tree analysis probabilities can be introduced into the analysis of complex decisions involving many alternatives and future conditions which are

unknown but can be specified in terms of a set of discrete probabilities or a continuous probability distribution .It is a useful tool in making decisions concerning investments, the acquisition or disposal of physical property, project management, personnel and new product strategies.

The term *decision tree* is derived from the physical appearance of the usual graphic representation of this technique. A decision tree contains not only the probabilities of outcomes, but also the conditional monetary value (or utility) values attached to those outcomes . This is the reason why decision trees are used to indicate the expected values of different actions .

Decision trees make use of standard symbols:

- Squares symbolize decision points, where the decision maker must choose among several possible actions. From these decision nodes, one branch is drawn for each of the possible actions.
- Circles represent chance events, where some state of nature is realised. These chance events are not under the decision maker's control. From these chance nodes one branch is drawn for each possible outcome.

Chance node

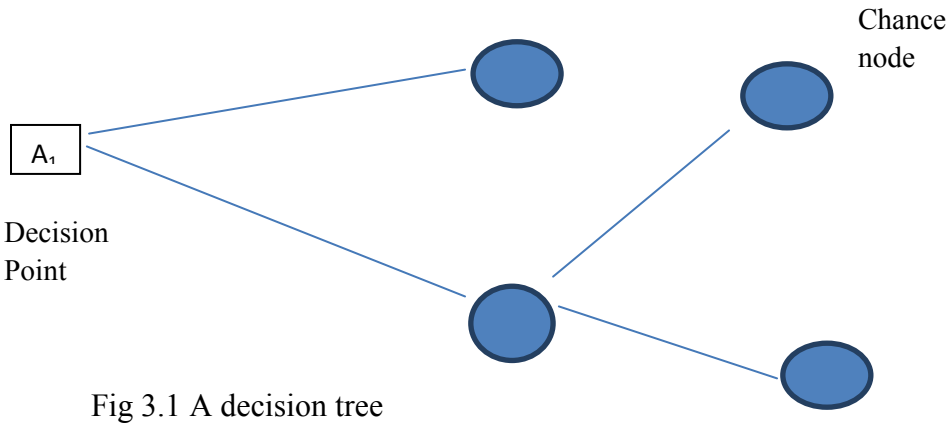


Fig 3.1 A decision tree

Self Assessment Questions

1. What is a decision tree?
2. State two advantages of decision tree .
3. Give two limitations of decision tree.

Various elements of the decision process like decision alternatives, states of nature, probabilities attached to the states of nature are indicated with the help of a decision tree. Branches coming out of a decision point represents the alternative courses of action. At the end of each decision branch, there is a state of nature node from which chance events arise in the form of sub branches. The respective payoffs and the probabilities associated with alternative courses and the chance events are shown alongside these branches.

Decision trees are useful for representing the interrelated, sequential and multi-dimensional aspects of a decision-making problem. By drawing a decision tree, one is in a position to visualize the entire complexity of the decision problem in all its dimensions.

Since it is impossible to evaluate an immediate decision act without first considering all future outcomes that result from this decision, one begins the analysis at the end of the tree. The last decision point is of primary importance to us. This point is analysed and decision taken which yields optimal EMV and then roll back to the last but one decision point , make the same EMV analysis for decision and roll back to the preceding decision point . The rolling back process continues till the initial decision point is reached.

3.3:Advantages and Disadvantages of Decision Trees: The following points highlights some advantages and disadvantages of Decision Trees

Advantages of the Decision Tree Approach

1. It structures the decision process and helps decision making in an orderly, systematic and sequential manner.
2. It displays the logical relationship between the parts of a complex decision and identifies the time sequence in which various actions and subsequent events would occur.

3. It communicates the decision-making process to others in an easy and clear manner, illustrating each assumption about the future.

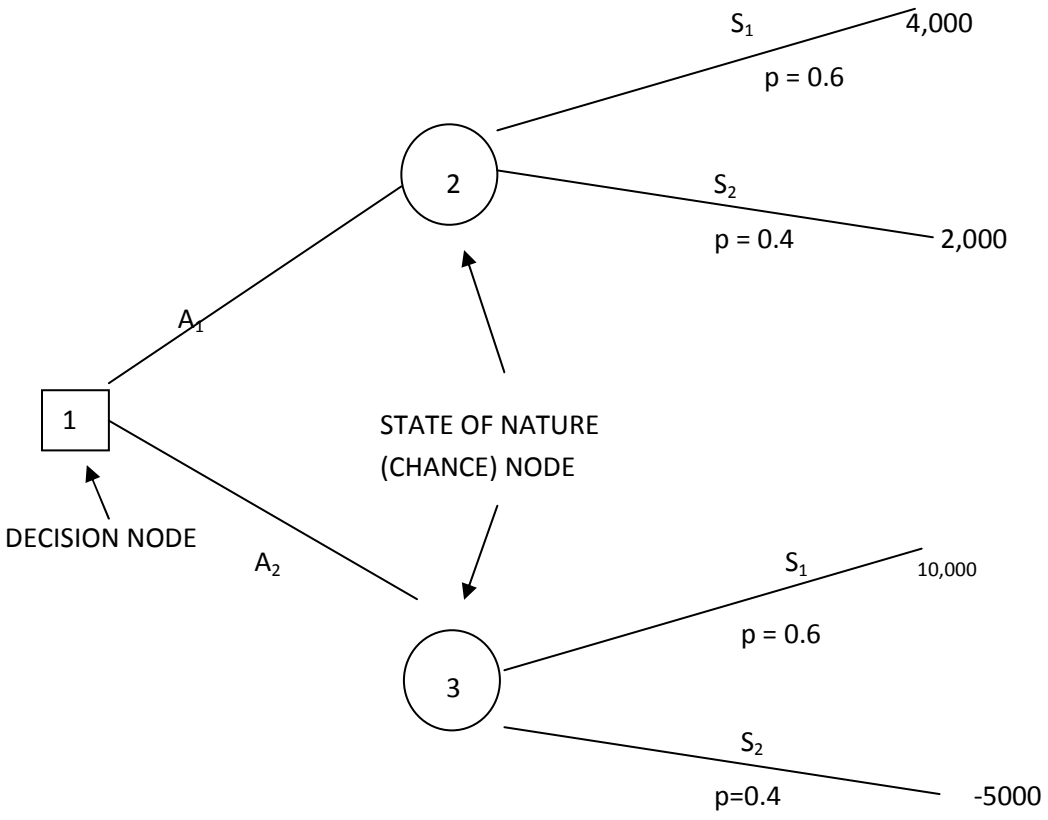
Limitations of Decision Tree Approach

1. The tree diagram become more and more complicated as the number of decision alternatives increases and more variables are introduced.
2. It analyses the problem in terms of expected values and thus A gives an average valued solution.
3. Quite often inconsistency arise in assigning probabilities for different events.

Example 1 Suppose a decision making problem is represented by the following table.

States of Nature	Probability	Alternative Actions	
		A_1(produce 25 units)	A_2(produce 75 units)
S_1(High demand)	0.6	4,000	10,000
S_2(low demand)	0.4	2,000	-5000

The decision tree for the above problem is shown in Fig 3.2



Example 2 A trading company of Mumbai is considering expansion of its activities and planning to open a marketing office at Kolkata to boost the sales in North East. It is to be decided whether to operate from the existing office at Mumbai and cover the area by frequent travelling or else establishing the office at Kanpur. The connected pay-offs and probabilities of two alternatives are as under:

Alternatives	States of nature	Probability	Pay-off(Rs in lakhs)
A. Operate from Mumbai	(i) Increase in demand by 30%	0.60	50
	(ii) No appreciable change	0.40	5
B. Open office at Kanpur	(i) Increase in demand by 30%	0.70	40
	(ii) No appreciable change	0.30	-10

Help the company to take proper decision.

Solution: Expected pay-off for alternative A = (0.60x50)+(0.40x5)=Rs 32 lakhs

: Expected pay-off for alternative B = (0.70x40)+(0.30x(-10))=Rs 25 lakhs

The expected pay-off for alternative A being higher, it is advisable to operate from Mumbai .

As seen from the pay-offs new office at Kolkata would entail certain expenditures which may not be justified if the sales do not pick up . Hence calculated risk may be taken to operate from Mumbai only.

Example 3. A businessman has two options, either to invest in project A or B but due to the paucity of capital he is unable to undertake both of them simultaneously. Project A requires a capital of Rs 30,000and project B 50,000. Survey reports show high, medium and low demands with corresponding probabilities of 0.4, 0.4, and 0.2 respectively for project A and 0.3, 0.4 and

0.3 for project B . Net profit from investment A are Rs 75,000, Rs 55,000 and Rs 35,000 and corresponding figures for project B are likely to be Rs 100,000, Rs 80,000 and Rs 70,000 for high , medium and low demand respectively. What decision should the businessman take? Decide by constructing an appropriate decision –tree.

Solution .

Expected net profit for project A=0.4x75, 000+0.4x55, 000+0.2x35, 000-30,000

= Rs 29,000

Expected net profit for project B= 0.3x100, 000+0.4x80, 000+0.3x70, 000-50,000

= Rs 33,000

Since the expected net profit for project B is more than A, the businessman should invest in project B.

3.3 Summary

A decision tree – A pictorial representation of a decision process, indicating alternatives and their associated probabilities; drawn in form of a tree , indicating root(the decision point) and branches (courses open , strategies, available and chances obtained with conditional payoff)

Decision Nodes represent Features/Attributes: Decision nodes represent the features/attributes based on which data is split into children nodes.

A decision tree has three key parts: a root node, leaf node and branches

3.4 Key Terms

Decision: It is a conclusion reached after due consideration.

Decision tree: A graphic display of various decision alternatives and sequence of events as if they were branches of a tree.

3.5 Check Your Progress

Decision theory is a systematic approach for decision making depending on the quality and quantity of information available to him or utilised by him.

There are three types of decision -- making environments: certainty, uncertainty and risk.

Various elements of a decision-making process include decision alternatives, state of nature, payoff, pay-off table and opportunity loss table.

Decision tree is the graphical presentation for displaying acts and events in a decision problem in the form of a tree diagram.

3.6 Questions and Exercises

1. What is a decision tree?

2. What are the three key parts of a decision tree?

Self Assessment Questions

Multiple Choice Questions

	ColumnA		**Column B**
1.	Decision tree	A.	Laplace Criterion
2.	It is helpful for decision maker	B.	Is a graphic display of various decision alternatives
3.	It is based on what is known as the principle of insufficient reason.	C.	Opportunity loss table

Answer 1. (B) 2. (C) 3. (A)

Fill in the blanks

1. _____is the graphical presentation for displaying acts and events in a decision problem in the form of a tree diagram

2._____ coming out of a decision point represents the alternative courses of action.

3. _____represent the features/attributes based on which data is split into children nodes.

Answer 1. Decision tree 2. Branches 3. Decision nodes

State whether True or False

1. A decision tree is highly useful to a decision maker in multi –stage situations.

2. A decision tree has three key parts.

Answer 1. True 2. True

Match Column A with Column B

	Column A		**Column B**
1.	It is an element of decision process.	A.	The alternative courses of action.
2.	Square symbolizes	B.	Decision making under certainty
3.	Branches coming out of a decision point represents	C.	Decision points

Answer : 1. (B) 2. (C) 3. (A)

Short Answer Questions

1. What are decision nodes?

2. What is a decision tree?

3. What do square represent?

4. What do circles represent?

Long –Answer Questions

1. Write a note on the different symbols used in decision tree.

2. What is a decision tree? Explain its advantages and limitations.

3. Give example of a decision tree.

3.7 Further Reading

Hazarika, P.L. Essential Statistics for Economics and Business Statistics .New Delhi. Akansha Publishing House, 2012

Sharma, J.K Business Statistics .Pearson Education, New Delhi, 2007

Gupta, S.C Fundamental of Statistics, New Delhi: S.Chand and Sons ,2005

Gupta ,S.P .Statistical Methods .New Delhi . S. Chand and Sons , 2005

Hooda , R.P . Statistics for Business and Economics. New Delhi: Macmillan India Ltd 2002

Sharma, Anand. Statistics for Management, Himalaya Publishing House, Geetanjali Press Pvt . Ltd, Nagpur