GOVERNMENT OF TAMIL NADU

HIGHER SECONDARY SECOND YEAR

STATISTICS

A publication under Free Textbook Programme of Government of Tamil Nadu

Department Of School Education

Untouchability is Inhuman and a Crime

Government of Tamil Nadu

First Edition - 2019
Revised Edition - 2020, 2022
(Published under New Syllabus)

NOT FOR SALE

Content Creation

State Council of Educational
Research and Training

Printing & Publishing

Tamil NaduTextbook and Educational
Services Corporation

www.textbooksonline.tn.nic.in

CONTENTS

STATISTICS

X0163

E-book

4RDUL

Assessment

KEY FEATURES OF THE BOOK

Profile of a Statistician	Presents a brief history and contribution of a statistician
Learning Objectives	Goals to transform the classroom processes a learner centric
DO YOU KNOW?	Amazing facts, Rhetorical questions to lead students to Statistical inquiry
Note	Additional inputs to content is provided
Activity	Directions are provided to students to conduct activities in order to explore, enrich the concept
Infographics	Visual representation of the lesson to enrich learning
	To motivate the students to further explore the content digitally and take them to virtual world
Success Story	Success Stories given as a source of inspiration
Points to Remember	Summary of each lesson is given at the end
ICT	To enhance digital skills among students
Evaluation	Assess students to pause, think and check their understanding
Glossary	Explanation of scientific terms

Career in Statistics

After completion of Higher Secondary Course, the subject Statistics is an essential part of the curriculum of many undergraduate, postgraduate, professional courses and research level studies. At least one or more papers are included in the Syllabus of the following courses:

Under Graduate Courses

B.A.(Economics)
B.Com
B.B.A
B.C.A
B.Sc.(Maths)
B.Pharm
B.Ed
B.Stat
B.E
Diploma Courses

Post Graduate Courses

M.A.(Economics)
M.Com
M.B.A
M.C.A
M.Sc
M.Pharm
M.Ed
M.Stat
M.E
C.A
I.C.W.A
Actuarial science

Competitive Eaminations

UPSC
TNPSC
Staff Selection Commission Examinations
I.A.S
I.F.S
and many more

Specialized fields in Statistics : Colleges/universities, Indian Statistical Institute(ISI) offer a number of specialisations in statistics at undergraduate, postgraduate and research level. A candidate with bachelor's degree in statistics can also apply for Indian Statistical Services (ISS).

Job Titles

- Statisticians
- Business Analyst
- Mathematician
- Professor
- Risk Analyst
- Data Analyst
- Content Analyst
- Statistics Trainer
- Data Scientist
- Consultant
- Biostatistician
- Econometrician

Job Areas

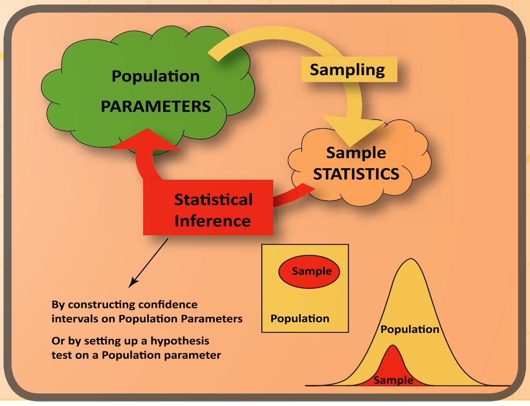
- Census
- Ecology
- Medicine
- Election
- Crime
- Economics
- Education
- Film
- Sports
- Tourism

Skills Required for a statistician

- Strong Foundation in Mathematical Statistics
- Logical Thinking & Ability to Comprehend Key Facts
- Ability to Interact with people from various fields to understand the problems
- Strong Background in Statistical Computing
- Ability to stay updated on recent literature & statistical software
- Versatility in solving problems

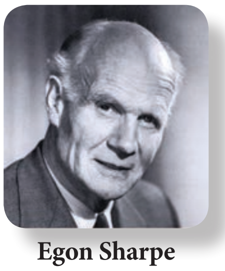
CHAPTER

1 TESTS OF SIGNIFICANCE – BASIC CONCEPTS AND LARGE SAMPLE TESTS


Population
PARAMETERS
Sampling
Sample
STATISTICS
Statistical
Inference
By constructing confidence
intervals on Population Parameters
Or by setting up a hypothesis
test on a Population parameter
Sample
Population
Population
Sample

Jerzy Neyman

Jerzy Neyman (1894-1981) was born into a Polish family in Russia. He is one of the Principal architects of Modern Statistics. He developed the idea of confidence interval estimation during 1937. He had also contributed to other branches of Statistics, which include Design of Experiments, Theory of Sampling and Contagious Distributions. He established the Department of Statistics in University of California at Berkeley, which is one of the preeminent centres for statistical research worldwide.



Egon Sharpe Pearson

Egon Sharpe Pearson (1885-1980) was the son of Prof. Karl Pearson. He was the Editor of *Biometrika*, which is still one of the premier journals in Statistics. He was instrumental in publishing the two volumes of *Biometrika Tables for Statisticians*, which has been a significant contribution to the world of Statistical Data Analysis till the invention of modern computing facilities.

Neyman and Pearson worked together about a decade from 1928 to 1938 and developed the theory of testing statistical hypotheses. *Neyman-Pearson Fundamental Lemma* is a milestone work, which forms the basis for the present theory of testing statistical hypotheses. In spite of severe criticisms for their theory, in those days, by the leading authorities especially Prof.R.A.Fisher, their theory survived and is currently in use.

"Statistics is the servant to all sciences" *– Jerzy Neyman*

LEARNING OBJECTIVES

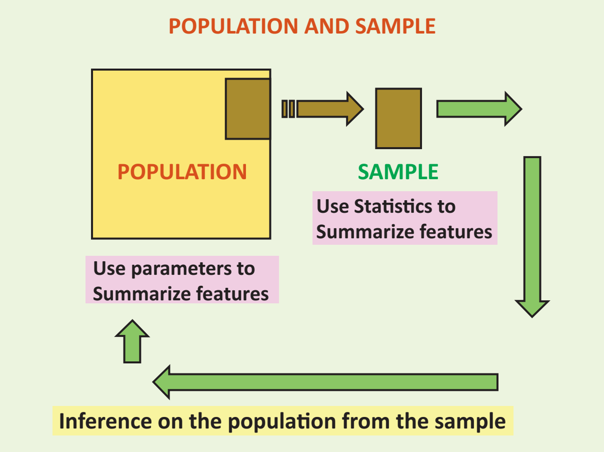
The students will be able to

- understand the purpose of hypothesis testing;
- define parameter and statistic;
- understand sampling distribution of statistic;
- define standard error;
- understand different types of hypotheses;
- determine type I and type II errors in hypotheses testing problems;
- understand level of significance, critical region and critical values;
- categorize one-sided and two-sided tests;
- understand the procedure for tests of hypotheses based on large samples; and
- solve the problems of testing hypotheses concerning mean(s) and proportion(s) based on large samples.

Introduction

In XI Standard classes, we concentrated on collection, presentation and analysis of data along with calculation of various measures of central tendency and measures of dispersion. These kinds of describing the data are popularly known as **descriptive statistics**. Now, we need to understand another dimension of statistical data analysis, which is called **inferential statistics**. Various concepts and methods related to this dimension will be discussed in the first four Chapters of this volume. Inferential Statistics may be described as follows from the statistical point of view:

One of the main objectives of any scientific investigation or any survey is to find out the unknown facts or characteristics of the population under consideration. It is practically not feasible to examine the entire population, since it will increase the time and cost involved. But one may examine a part of it, called **sample**. On the basis of this limited information, one can make decisions or draw inferences on the unknown facts or characteristics of the population.


POPULATION AND SAMPLE
POPULATION
SAMPLE
Use Statistics to Summarize features
Use parameters to Summarize features
Inference on the population from the sample

Thus, inferential statistics refers to a collection of statistical methods in which random samples are used to draw valid inferences or to make decisions in terms of probabilistic statements about the population under study.

Before going to study in detail about Inferential Statistics, we need to understand some of the important terms and definitions related to this topic.

1.1 PARAMETER AND STATISTIC

A **population**, as described in Section 2.4 in XI Standard text book, is a collection of units/objects/numbers under study, whose elements can be considered as the values of a random variable, say, X. As mentioned in Section 9.3 in XI Standard text book, there will be a probability distribution associated with X.

Parameter: Generally, **parameter** is a quantitative characteristic, which indexes/identifies the respective distribution. In many cases, statistical quantitative characteristics calculated based on all the units in the population are the respective parameters. For example, population mean, population standard deviation, population proportion are parameters for some distributions.

Recall: The unknown constants which appear in the *probability density function or probability mass function* of the random variable X, are also called **parameters** of the corresponding distribution/population.

The parameters are commonly denoted by Greek letters. In Statistical Inference, some or all the parameters of a population are assumed to be unknown.

Random sample: Any set of reliazations $(X_1, X_2, ..., X_n)$ made on X under independent and identical conditions is called a **random sample**.

Statistic: Any statistical quantity calculated on the basis of the random sample is called a **statistic**. The sample mean, sample standard deviation, sample proportion *etc*., are called **statistics** (plural form of *statistic*).They will be denoted by Roman letters.

Let $(x_1, x_2, ..., x_n)$ be an observed value of $(X_1, X_2, ..., X_n)$. The collection of $(x_1, x_2, ..., x_n)$ is known as *sample space*, which will be denoted by '**S**'.

Note 1:

A set of n sample observations can be made on X, say, $x_1, x_2, ..., x_n$ for making inferences on the unknown parameters. It is to be noted that these n values may vary from sample to sample. Thus, these values can be considered as the realizations of the random variables $X_1, X_2, ..., X_n$, which are assumed to be independent and have the same distribution as that of X. These are also called independently and identically distributed (*iid)* random variables.

NOTE

The *statistic* itself is a random variable and has a probability distribution.

Note 2:

In Statistical Inference, the sample standard deviation is defined as $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$, where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. It may be noted that the divisor is $n - 1$ instead of n.

Note 3:

The statistic itself is a random variable, until the numerical values of $X_1, X_2, ..., X_n$ are observed, and hence it has a probability distribution.

Notations to denote various population parameters and their corresponding sample statistics are listed in Table 1.1. The notations will be used in the first four chapters of this book with the same meaning for the sake of uniformity.

Table 1.1 Notations for Parameters and Statistics

Statistical measure	Parameter	Statistic	Value of the Statistic for a given sample
Mean	μ	$\overline{X}$	$\overline{x}$
Standard deviation	σ	S	s
Proportion	P	p	p_0

1.2 SAMPLING DISTRIBUTION

The probability distribution of a statistic is called **sampling distribution** of the statistic. In other words, it is the probability distribution of possible values of the statistic, whose values are computed from possible random samples of same size.

The following example will help to understand this concept.

Example 1.1

Suppose that a population consists of 4 elements such as 4, 8, 12 and 16. These may be considered as the values of a random variable, say, *X*. Let a random sample of size 2 be drawn from this population under *sampling with replacement* scheme. Then, the possible number of samples is 4^2.

It is to be noted that, if we take samples of size n each from a finite population of size N, then the number of samples will be N^n *under with replacement scheme and* ${}^{N}C_{n}$ *samples under without replacement scheme.*

In each of the 4^2 samples, the sample elements x_1 and x_2 can be considered as the values of the two *iid* random variables X_1 and X_2. The possible samples, which could be drawn from the above population and their respective means are presented in Table 1.2.

Table 1.2 Possible Samples and their Means

Sample Number	Sample elements (x_1, x_2)	Sample Mean $\bar{x}$
1	4,4	4
2	4,8	6
3	4,12	8
4	4,16	10
5	8,4	6
6	8,8	8
7	8,12	10
8	8,16	12
9	12,4	8
10	12,8	10
11	12,12	12
12	12,16	14
13	16,4	10
14	16,8	12
15	16,12	14
16	16,16	16

The set of pairs (x_1, x_2) listed in column 2 constitute the sample space of samples of size 2 each.

Hence, the sample space is:

$\mathbb{S}$ = {(4,4), (4,8), (4,12), (4,16), (8,4), (8,8), (8,12), (8,16), (12,4), (12,8), (12,12), (12,16), (16,4), (16,8), (16,12), (16,16)}

The sampling distribution of $\overline{X}$, the sample mean, is determined and is presented in Table 1.3.

Table 1.3 Sampling Distribution of Sample Mean

Sample mean: $\bar{x}$	4	6	8	10	12	14	16	Total
Probability: $P(\bar{X}=\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	1

Note 4: The sample obtained under sampling *with replacement* from a finite population satisfies the conditions for a random sample as described earlier.

Note 5: If the sample values are selected under *without replacement scheme*, independence property of X_1, X_2, ... X_n will be violated. Hence it will not be a random sample.

Note 6: When the sample size is greater than or equal to 30, in most of the text books, the sample is termed as a **large sample**. Also, the sample of size less than 30 is termed as **small sample**. However, in practice, there is no rigidity in this number *i.e.*, 30, and that depends on the nature of the population and the sample.

Note 7: The learners may recall from XI Standard Textbook that some of the probability distributions possess the additive property. For example, if $X_1, X_2, ..., X_n$ are *iid* $N(\mu, \sigma^2)$ random variables, then the probability distributions of $X_1 + X_2 + ... + X_n$ and $\bar{X}$ are respectively the $N(n\mu, n\sigma^2)$ and $N(\mu, \sigma^2/\text{n})$. These two distributions, in statistical inference point of view, can be considered respectively as the sampling distributions of the sample total and sample mean of a random sample drawn from the $N(\mu, \sigma^2)$ distribution. The notation $N(\mu, \sigma^2)$ refers to the normal distribution having mean μ and variance σ^2.

1.3 STANDARD ERROR

The standard deviation of the sampling distribution of a statistic is defined as the **standard error** of the statistic, which is abbreviated as *SE.*

For example, the standard deviation of the sampling distribution of the sample mean, $\bar{x}$, is known as the standard error of the sample mean, or *SE* $(\bar{X})$.

If the random variables $X_1, X_2, ..., X_n$ are independent and have the same distribution with mean μ *and variance* σ^2, then variance of $\bar{X}$ becomes as

$$V\left(\bar{X}\right)=V\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}V\left(X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$$

Thus, $SE\left(\bar{X}\right)=\dfrac{\sigma}{\sqrt{n}}$.

Also, note that mean of $\bar{X}=E\left(\bar{X}\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(X_i\right)=\frac{1}{n}\sum_{i=1}^{n}\mu=\frac{n\mu}{n}=\mu$

Example 1.2

Calculate the standard error of $\bar{X}$ for the sampling distribution obtained in *Example 1.*

Solution:

Here, the population is {4, 8, 12, 16}.

Population size (N) = 4, Sample size (n) = 2

Population mean (μ) = (4 + 8 + 12+ 16)/4 = 40/4 = 10

The population variance is calculated as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

$$= \frac{1}{4}\left[\left(4-10\right)^2 + \left(8-10\right)^2 + \left(12-10\right)^2 + \left(16-10\right)^2\right] = \frac{1}{4}\left[36+4+4+36\right] = 20.$$

Hence, $SE(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{20}{2}} = \sqrt{10}$

This can also be verified from the sampling distribution of $\bar{X}$ (*see* Table 1.3)

$$\mathrm{V}(\bar{X}) = \sum(\bar{x}-\mu)^2 P(\bar{X} = \bar{x})$$

where the summation is taken over all values of $\bar{x}$

Thus, $\mathrm{V}(\bar{X}) = (4-10)^2\frac{1}{16} + (6-10)^2\frac{2}{16} + (8-10)^2\frac{3}{16} + (10-10)^2\frac{4}{16}$

$$+(12-10)^2\frac{3}{16} + (14-10)^2\frac{2}{16} + (16-10)^2\frac{1}{16}$$

$$= \frac{1}{16}(36+32+12+0+12+32+36) = 10$$

Hence, the standard deviation of the sampling distribution of $\bar{X}$ is = $\sqrt{10}$.

Standard Errors of some of the frequently referred statistics are listed in Table 1.4.

Table 1.4 Statistics and their Standard Errors

Statistic	Standard error
Sample proportion: p	$\sqrt{\frac{PQ}{n}}$, where P is the population proportion and $Q = 1 - P$.
Difference between the means $\bar{X}$ and $\bar{Y}$ of two independent samples: $\left(\bar{X}-\bar{Y}\right)$	$\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}$ where m and n are the sizes of samples drawn from the populations whose variances are σ_X^2 and σ_Y^2 respectively.
	$\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}$, where σ^2 is the common variance of the populations.
Difference between the proportions p_X and p_Y of two independent samples: $(p_X - p_Y)$	$\sqrt{\frac{P_XQ_X}{m}+\frac{P_YQ_Y}{n}}$, where m and n are sizes of the samples drawn from the populations whose proportions are respectively P_X and P_Y; $Q_X = 1 - P_X$, $Q_Y = 1 - P_Y$.
	$\sqrt{\hat{p}\hat{q}\left(\frac{1}{m}+\frac{1}{n}\right)}$, where $\hat{p} = \frac{mp_X + np_Y}{m+n}$, $\hat{q} = 1-\hat{p}$, m and n are sample sizes, when P_X and P_Y are unknown.

1.4 NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

In many practical studies, as mentioned earlier, it is necessary to make decisions about a population or its unknown characteristics on the basis of sample observations. For example, in bio-medical studies, we may be investigating a particular theory that the recently developed medicine is much better than the conventional medicine in curing a disease. For this purpose, we propose a statement on the population or the theory. Such statements are called hypotheses.

Thus, a **hypothesis** can be defined as a statement on the population or the values of the unknown parameters associated with the respective probability distribution. All the hypotheses should be tested for their validity using statistical concepts and a representative sample drawn from the study population. '*Hypotheses*' is the plural form of '*hypothesis*'.

A **statistical test** is a procedure governed by certain determined/derived rules, which lead to take a decision about the null hypothesis for its rejection or otherwise on the basis of sample values. This process is called **statistical hypotheses testing**.

The statistical hypotheses testing plays an important role, among others, in various fields including industry, biological sciences, behavioral sciences and Economics. In each hypotheses testing problem, we will often find as there are two hypotheses to choose between *viz*., null hypothesis and alternative hypothesis.

Null Hypothesis:

A hypothesis which is to be actually tested ***for possible rejection*** based on a random sample is termed as **null hypothesis**, which will be denoted by H_0.

YOU WILL KNOW

(i) Generally, it is a hypothesis of no difference in the case of comparison.

(ii) Assigning a value to the unknown parameter in the case of single sample problems

(iii) Suggesting a suitable model to the given environment in the case of model construction.

(iv) The given two attributes are independent in the case of *Chi*-square test for independence of attributes.

Alternative Hypothesis:

A statement about the population, which contradicts the null hypothesis, depending upon the situation, is called **alternative hypothesis**, which will be denoted by H_1.

For example, if we test whether the population mean has a specified value μ_0, then the null hypothesis would be expressed as:

$$H_0: \mu = \mu_0$$

The alternative hypothesis may be formulated suitably as anyone of the following:

(i) $H_1: \mu \neq \mu_0$

(ii) $H_1: \mu > \mu_0$

(iii) $H_1: \mu < \mu_0$

The alternative hypothesis in (i) is known as two-sided alternative and the alternative hypothesis in (ii) is known as one-sided (right) alternative and (iii) is known as one-sided (left) alternative.

1.5 ERRORS IN STATISTICAL HYPOTHESES TESTING

A **statistical decision** in a hypotheses testing problem is either of rejecting or not rejecting H_0 based on a given random sample. Statistical decisions are governed by certain rules, developed by applying a statistical theory, which are known as **decision rules**. The decision rule leading to rejection of H_0 is called as **rejection rule**.

The null hypothesis may be either true or false, in reality. Under this circumstance, there will arise four possible situations in each hypotheses testing or decision making problem as displayed in Table 1.5.

Table 1.5 Decision Table

	H_0 is true	H_0 is false
Reject H_0	Type I error	Correct decision
Do not Reject H_0	Correct decision	Type II error

It must be recognized that the final decision of rejecting H_0 or not rejecting H_0 may be incorrect. The error committed by rejecting H_0, when H_0 is really true, is called **type I error**. The error committed by not rejecting H_0, when H_0 is false, is called **type II error**.

Example 1.3

A soft drink manufacturing company makes a new kind of soft drink. Daily sales of the new soft drink, in a city, is assumed to be distributed with mean sales of ₹40,000 and standard deviation of ₹2,500 per day. The Advertising Manager of the company considers placing advertisements in local TV Channels. He does this on 10 random days and tests to see whether or not sales has increased. Formulate suitable null and alternative hypotheses. What would be type I and type II errors?

Solution:

The Advertising Manager is testing whether or not sales increased more than ₹40,000. Let μ be the average amount of sales, if the advertisement does appear.

The null and alternative hypotheses can be framed based on the given information as follows:

Null hypothesis: H_o: $\mu = 40000$

i.e., The mean sales due to the advertisement is not significantly different from ₹40,000.

Alternative hypothesis: H_1: $\mu > 40000$

i.e., Increase in the mean sales due to the advertisement is significant.

(i) If type I error occurs, then it will be concluded as the advertisement has improved sales. But, really it is not.

(ii) If type II error occurs, then it will be concluded that the advertisement has not improved the sales. But, really, the advertisement has improved the sales.

The following may be the penalties due to the occurrence of these errors:

If type I error occurs, then the company may spend towards advertisement. It may increase the expenditure of the company. On the other hand, if type II error occurs, then the company will not spend towards advertisement. It may not improve the sales of the company.

1.6 LEVEL OF SIGNIFICANCE, CRITICAL REGION AND CRITICAL VALUE(S)

In a given hypotheses testing problem, the *maximum probability* with which we would be willing to tolerate the occurrence of type I error is called **level of significance** of the test. This probability is usually denoted by 'α'. Level of significance is specified before samples are drawn to test the hypothesis.

The level of significance normally chosen in every hypotheses testing problem is 0.05 (5%) or 0.01 (1%). If, for example, the level of significance is chosen as 5%, then it means that among the 100 decisions of rejecting the null hypothesis based on 100 random samples, maximum of 5 of among them would be wrong. It is emphasized that the 100 random samples are drawn under identical and independent conditions. That is, the null hypothesis H_0 is rejected wrongly based on 5% samples when H_0 is actually true. We are about 95% confident that we made the right decision of rejecting H_0.

Critical region in a hypotheses testing problem is a subset of the sample space whose elements lead to rejection of H_0. Hence, its elements have the dimension as that of the sample size, say, $n(n > 1)$. That is,

$$\text{Critical Region} = \left\{ \underset{\sim}{x} = (x_1, x_2, ..., x_n) \mid H_0 \text{ is rejected} \right\}.$$

A subset of the sample space whose elements does not lead to rejection of H_0 may be termed as **acceptance region**, which is the complement of the critical region. Thus,

$$\mathbf{S} = \{\text{Critical Region}\} \cup \{\text{Acceptance Region}\}.$$

Test statistic, a function of statistic(s) and the known value(s) of the underlying parameter(s), is used to make decision on $H_{0.}$ Consider a hypotheses testing problem, which uses a **test statistic** $t(\underset{\sim}{X})$ and a constant c for deciding on H_0. Suppose that H_0 is rejected, when $t(\underset{\sim}{x}) > c$. It is to be noted here that $t(\underset{\sim}{X})$ is a scalar and is of dimension one. Its sampling distribution is a univariate probability distribution. The values of $t(\underset{\sim}{X})$ satisfying the condition $t(\underset{\sim}{x}) > c$ will identify the samples in the sample space, which lead to rejection of H_0. It does not mean that $\{t \mid t(\underset{\sim}{x}) > c\}$ is the corresponding critical region. The value 'c', distinguishing the elements of the critical region and the acceptance region, is referred to as **critical value**. There may be one or many critical values for a hypotheses testing problem. The critical values are determined from the sampling distribution of the respective test statistic under H_0.

Example 1.4

Suppose an electrical equipment manufacturing industry receives screws in lots, as raw materials. The production engineer decides to reject a lot when the number of defective screws is one or more in a randomly selected sample of size 2.

$$\text{Define } X_i = \begin{cases} 1, & \text{if } i^{\text{th}} \text{ screw is defective} \\ 0, & \text{if } i^{\text{th}} \text{ screw is not defective}' \end{cases}, \qquad i = 1, 2$$

Then, X_1 and X_2 are *iid* random variables and they have the *Bernoulli* (P) distribution.

Let $H_0 : P = \frac{1}{3}$ and $H_1 : P = \frac{2}{3}$

The sample space is $\mathbb{S}$ = {(0,0),(0,1),(1,0),(1,1)}

If $T(X_1, X_2)$ represents the number of defective screws, in each random sample, then the statistic $T(X_1, X_2) = X_1 + X_2$ is a random variable distributed according to the *Binomial* (2, P) distribution. The possible values of $T(X_1, X_2)$ are 0, 1 and 2. The values of $T(X_1, X_2)$ which lead to rejection of H_0 constitute the set {1,2}.

But, the critical region is defined by the elements of $\mathbb{S}$ corresponding to $T(X_1,X_2) = 1$ or 2. Thus, the critical region is {(0,1), (1,0), (1,1)} whose dimension is 2.

Note 8: When the sampling distribution is continuous, the set of values of $t(\underset{\sim}{X})$ corresponding to the rejection rule will be an interval or union of intervals depending on the alternative hypothesis. It is empahazized that **these intervals identify the elements of critical region, but they do not constitute the critical region.**

When the sampling distribution of the test statistic Z is a normal distribution, the critical values for testing H_0 against the possible alternative hypothesis at two different levels of significance, say 5% and 1% are displayed in Table 1.6.

Table 1.6 Critical values of the Z statistic

Alternative hypothesis	Level of Significance (α)	
	0.05 or 5%	0.01 or 1%
One- sided (right)	$z_\alpha = z_{0.05} = 1.645$	$z_\alpha = z_{0.01} = 2.33$
One- sided (left)	$-z_\alpha = -z_{0.05} = -1.645$	$-z_\alpha = -z_{0.01} = -2.33$
Two-sided	$z_{\alpha/2} = z_{0.025} = 1.96$	$z_{\alpha/2} = z_{0.005} = 2.58$

1.7 ONE-TAILED AND TWO-TAILED TESTS

In some hypotheses testing problem, elements of the critical region may be identified by a rejection rule of the type $t(\underset{\sim}{X}) \geq c$. In this case, P($t(\underset{\sim}{X}) \geq c$) will be the area, which falls at the right end (Figure1.1) under the curve representing the sampling distribution of $t(\underset{\sim}{X})$. The statistical test defined by this kind of critical region is called **right-tailed test**.

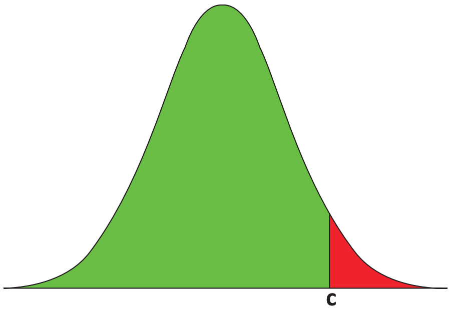


Figure 1.1. Right-tailed Test

On the other hand, suppose that the rejection rule $t(\underset{\sim}{X}) \leq c$ determines the elements of the critical region. Then, P($t(\underset{\sim}{X}) \leq c$) will be the area, which falls at the left end (Figure.1.2) under the curve representing the sampling distribution of $t(\underset{\sim}{X})$. The statistical test defined by this kind of critical region is called **left -tailed test**.

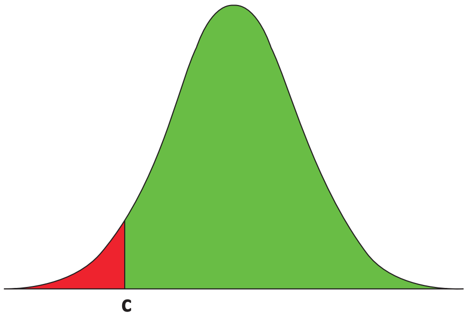


Figure 1.2 Left-tailed Test

The above two tests are commonly known as **one-tailed tests**.

Note 9: It should be noted that the sampling distribution of $t(\underset{\sim}{X})$ need not be with symmetric shape always. Sometimes, it may be positively or negatively skewed.

Example 1.5

Suppose a pizza restaurant claims its average pizza delivery time is 30 minutes. But you believe that the restaurant takes more than 30 minutes. Now, the null and the alternate hypotheses can be formulated as

$H_0 : \mu = 30$ minutes and $H_1 : \mu > 30$ minutes

Suppose that the decision is taken based on the delivery times of 4 randomly chosen pizza deliveries of the restaurant. Let X_1, X_2, X_3, and X_4 represent the delivery times of the such four occasions. Also, let H_0 be rejected, when the sample mean exceeds 31. Then, the critical region is

$$\text{Critical Region} = \left\{(x_1, x_2, x_3, x_4) \mid \bar{x} = \frac{x_1 + x_2 + x_3 + x_4}{4} > 31\right\}$$

In this case, $P\left(\bar{X} > 31\right)$ will be the area, which fall at the right end under the curve representing the sampling distribution of $\bar{X}$. Hence, this test can be categorized as a right-tailed test.

Suppose that H_0 is rejected, when either $t\left(\underset{\sim}{X}\right) \leq a$ or $t\left(\underset{\sim}{X}\right) \geq b$ holds. In this case, P($t\left(\underset{\sim}{X}\right) \leq a$) and P($t\left(\underset{\sim}{X}\right) \geq b$) will be the areas, which fall respectively at left and right ends under the curve representing the sampling distribution of $t(\underset{\sim}{X})$ (Figure 1.3). The statistical test defined with this kind of rejection rule is known as **two-tailed test**.

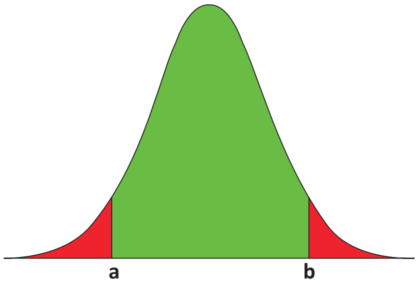


Figure 1.3 Two-tailed Test

CHAPTER 2 TESTS BASED ON SAMPLING DISTRIBUTIONS - I

W.S. Gossett

W Gosset (1876-1937), born in England studied Chemistry and Mathematics at New College , Oxford. Upon graduating in 1899, he joined a brewery in Ireland. Gosset applied his statistical knowledge both in the brewery and on the farm to the selection of the best varieties of Barley. Gosset acquired that knowledge by study, by trial and error, and by spending two terms in 1906–1907 in the biometrical laboratory of Karl Pearson. Gosset and Pearson had a good relationship. Pearson helped Gosset with the mathematics of his research papers. The brewery where he was employed allowed publishing his work under a pseudonym ("Student"). Thus, his most noteworthy achievement is now called Student's *t*, rather than Gosset's, *t*-distribution.

LEARNING OBJECTIVES

PBAHO

The student will be able to

- understand the purpose for using *t*-test and chi-square test .
- understand procedures for tests of hypotheses based on small samples.
- solve problems to test the hypotheses concerning mean(s) using *t*-distribution.
- solve problems to test the hypothesis whether the population has a particular variance using chi-square test.
- solve problems to test the hypotheses relating to independence of attributes and goodness of fit using chi-square test.

Introduction

In the earlier chapter, we have discussed various problems related to tests of significance based on large samples by applying the standard normal distribution. However, if the sample size is small ($n < 30$) the sampling distributions of test statistics are far from normal and the procedures discussed in Chapter-1 cannot be applied, except the general procedure (Section 1.8). But in this case, there exists a probability distribution called *t*-distribution which may be used instead of standard normal distribution to study the problems based on small samples.

2.1 STUDENT'S *t* DISTRIBUTION AND ITS APPLICATIONS

2.1.1 Student's *t*-distribution

If $X \sim N(0,1)$ and $Y \sim \chi^2_n$ are independent random variables, then

$T = \dfrac{X}{\sqrt{Y/n}}$ is said to have *t*-distribution with *n* degrees of freedom. This can be denoted by t_n.

χ^2-distribution and some of its applications are discussed in Section 2.2

Note 1: The degrees of freedom of *t* is the same as the degrees of freedom of the corresponding chi-square random variable.

Note 2: The *t*-distribution is used as the sampling distribution(s) of the statistics(s) defined based on random sample(s) drawn from normal population(s).

i) If $X_1, X_2, \ldots, X_n$ is a random sample drawn from $N(\mu, \sigma^2)$ population then

$$X = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ and } Y = \frac{\sum\left(X_i - X\right)^2}{\sigma^2} \sim \chi^2_{n-1} \text{ are independent.}$$

Hence,

$$T_1 = \frac{X}{\sqrt{Y/n-1}}$$

$$= \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \times \frac{\sigma\sqrt{(n-1)}}{\sqrt{\sum\left(X_i - \overline{X}\right)^2}}$$

$$= \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \qquad \text{where } S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{(n-1)}}.$$

ii) If $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$ are independent random samples drawn from $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$ populations respectively, then

$$\frac{\left(\overline{X} - \overline{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0,1) \text{ and } \frac{\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2 + \sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2}{\sigma^2} \sim \chi^2_{m+n-2} \text{ are independent.}$$

Then,

$$T_2 = \frac{\left(\overline{X} - \overline{Y}\right) - \left(\mu_X - \mu_Y\right)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

$$\text{where } S_p^2 = \frac{\sum_{i=1}^{m}\left(X_i-\overline{X}\right)^2+\sum_{j=1}^{n}\left(Y_j-\overline{Y}\right)^2}{m+n-2}$$

3. If $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is a random sample of n paired observations drawn from a bivariate normal population, then $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$ is a random sample drawn from $N(\mu_D, \sigma_D{}^2)$. Here $\mu_D = \mu_X - \mu_Y$.

Hence,

$$T_3 = \frac{\overline{D}-\mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$$

2.1.2 Properties of the Student's *t*-distribution

1. *t*–distribution is symmetrical distribution with mean zero.
2. The graph of *t*-distribution is similar to normal distribution except for the following two reasons:
   i. The normal distribution curve is higher in the middle than *t*-distribution curve.
   ii. *t*–distribution has a greater spread sideways than the normal distribution curve. It means that there is more area in the tails of *t*-distribution.
3. The *t*-distribution curve is asymptotic to *X*-axis, that is, it extends to infinity on either side.
4. The shape of *t*-distribution curve varies with the degrees of freedom. The larger is the number of degrees of freedom, closeness of its shape to standard normal distribution (fig. 2.1).
5. Sampling distribution of *t* does not depend on population parameter. It depends on degrees of freedom (n–1).

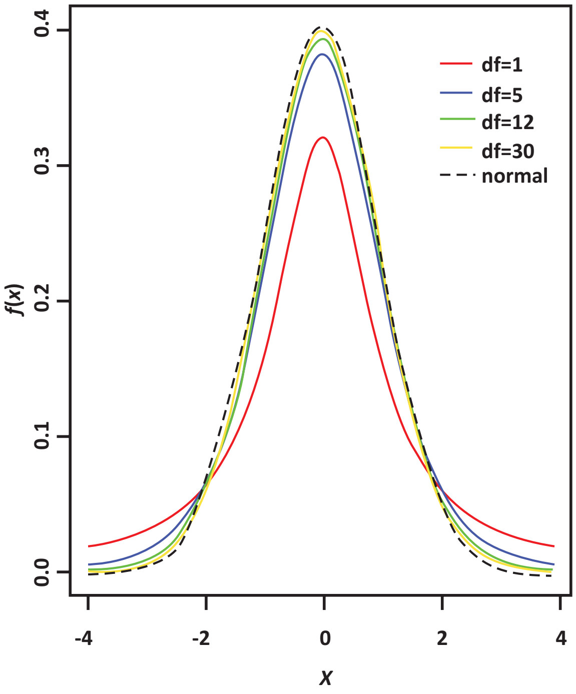


Figure 2.1. Student's *t*-distribution

2.1.3 Applications of *t*-distribution

The *t*-distribution has the following important applications in testing the hypotheses for small samples.

1. To test significance of a single population mean, when population variance is unknown, using T_1.

2. To test the equality of two population means when population variances are equal and unknown, using T_2.
3. To test the equality of two means – paired t-test, based on dependent samples, T_3.

YOU WILL KNOW

The t-distribution has few more applications but they are not considered in this Chapter. You will study these applications in higher classes.

2.1.4 Test of Hypotheses for Normal Population Mean (Population Variance is Unknown)

Procedure:

Step 1 : Let μ and σ^2 be respectively the mean and variance of the population under study, where σ^2 is unknown. If μ_0 is an admissible value of μ, then frame the null hypothesis as

H_0: $\mu = \mu_0$ and choose the suitable alternative hypothesis from

(i) H_1: $\mu \neq \mu_0$ (ii) H_1: $\mu > \mu_0$ (iii) H_1: $\mu < \mu_0$

Step 2 : Describe the sample/data and its descriptive measures. Let $(X_1, X_2, \ldots, X_n)$ be a random sample of n observations drawn from the population, where n is small ($n < 30$).

Step 3 : Specify the level of significance, α.

Step 4 : Consider the test statistic $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$ under H_0, where $\overline{X}$ and S are the sample mean and sample standard deviation respectively. The approximate sampling distribution of the test statistic under H_0 is the t-distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculate the value of t for the given sample $(x_1, x_2, \ldots x_n)$ as $T = \dfrac{\overline{x} - \mu}{s/\sqrt{n}}$.

here $\overline{x}$ is the sample mean and $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ is the sample standard deviation.

Step 6 : Choose the critical value, t_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Critical Value (t_e)	$t_{n-1\,,\alpha/2}$	$t_{n-1,\alpha}$	$-t_{n-1,\alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Rejection Rule	$\lvert t_0 \rvert \geq t_{n-1,\alpha/2}$	$t_0 > t_{n-1,\alpha}$	$t_0 < -t_{n-1,\alpha}$

Example 2. 1

The average monthly sales, based on past experience of a particular brand of tooth paste in departmental stores is ₹ 200. An advertisement campaign was made by the company and then a sample of 26 departmental stores was taken at random and found that the average sales of the particular brand of tooth paste is ₹ 216 with a standard deviation of ₹ 8. Does the campaign have helped in promoting the sales of a particular brand of tooth paste?

Solution:

Step 1 : Hypotheses

Null Hypothesis H_0: $\mu = 200$

i.e., the average monthly sales of a particular brand of tooth paste is not significantly different from ₹ 200.

Alternative Hypothesis H_1: $\mu > 200$

i.e., the average monthly sales of a particular brand of tooth paste are significantly different from ₹ 200. It is one-sided (right) alternative hypothesis.

Step 2 : Data

The given sample information are:

Size of the sample (n) = 26. Hence, it is a small sample.

Sample mean $(\bar{x})$ = 216, Standard deviation of the sample = 8.

Step 3 : Level of significance

α = 5%

Step 4 : Test statistic

The test statistic under H_0 is $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

Since n is small, the sampling distribution of T is the t-distribution with $(n–1)$ degrees of freedom.

Step 5 : Calculation of test statistic

The value of T for the given sample information is calculated from

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \text{ as}$$

$$t_0 = \frac{216 - 200}{8/\sqrt{26}} = 10.20$$

Step 6 : Critical value

Since H_1 is one-sided (right) alternative hypothesis, the critical value at α =0.05 is

$$t_e = t_{n\text{-}1,\,\alpha} = t_{25,0.05} = 1.708$$

Step 7 : Decision

Since it is right-tailed test, elements of critical region are defined by the rejection rule $t_0 > t_e = t_{n\text{-}1,\,\alpha} = t_{25,0.05}$ = 1.708. For the given sample information t_0 =10.20 > t_e =1.708. It indicates that given sample contains sufficient evidence to reject H_0. Hence, the campaign has helped in promoting the increase in sales of a particular brand of tooth paste.

Example 2.2

A sample of 10 students from a school was selected. Their scores in a particular subject are 72, 82, 96, 85, 84, 75, 76, 93, 94 and 93. Can we support the claim that the class average scores is 90?

Solution:

Step 1 : Hypotheses

Null Hypothesis H_0: $\mu = 90$

i.e., the class average scores is not significantly different from 90.

Alternative Hypothesis $H_1 : \mu \neq 90$

i.e., the class means scores is significantly different from 90.

It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

Size of the sample (n) = 10. Hence, it is a small sample.

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

Since n is small, the sampling distribution of T is the t - distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculation of test statistic

The value of T for the given sample information is calculated from $t_0 = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$ as under:

x_i	$u_i = x_i - A; (A = 85)$	u_i^2
72	–13	169
82	–3	9
96	11	121
85	0	0
84	–1	1
75	–10	100
76	–9	81
93	8	64
94	9	81
93	8	64
	$\sum_{i=1}^{10} u_i = 0$	$\sum_{i=1}^{10} u_i^2 = 690$

Sample mean

$$\overline{x} = A + \frac{\sum_{i=1}^{10} u_i}{n} \text{ where } A \text{ is assumed mean}$$

$= 85 + 0 = 85$

Sample standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{10} u_i^2}$$

$$= \sqrt{\frac{1}{9} \times 690}$$

$$= \sqrt{76.67}$$

$= 8.756$

Hence,

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

$$= \frac{85-90}{8.756/\sqrt{10}} = \frac{-5}{2.77}$$

$= -1.806$ and

$|t_0| = 1.806$

Step 6 : **Critical value**

Since H_1 is two-sided alternative hypothesis, the critical value at $\alpha = 0.05$ is $t_e = t_{n-1,\frac{\alpha}{2}} = t_{9,0.025} = 2.262$

Step 7 : **Decision**

Since it is two-tailed test, elements of critical region are defined by the rejection rule $|t_0| > t_e = t_{n-1,\frac{\alpha}{2}} = t_{9,0.025} = 2.262$. For the given sample information $|t_0| = 1.806 < t_e = 2.262$. It indicates that given sample does not provide sufficient evidence to reject H_0. Hence, we conclude that the class average scores is 90.

2.1.5 Test of Hypotheses for Equality of Means of Two Normal Populations (Independent Random Samples)

Procedure:

Step 1 : Let μ_X and μ_Y be respectively the means of population-1 and population-2 under study. The variances of the population-1 and population-2 are assumed to be equal and unknown given by σ^2.

Frame the null hypothesis as $H_0 : \mu_X = \mu_Y$ and choose the suitable alternative hypothesis from (i) $H_1 : \mu_X \neq \mu_Y$ (ii) $H_1 : \mu_X > \mu_Y$ (iii) $H_1 : \mu_X < \mu_Y$

Step 2 : Describe the sample/data. Let $(X_1, X_2, \ldots, X_m)$ be a random sample of m observations drawn from Population-1 and $(Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations drawn from Population-2, where m and n are small (*i.e.*, $m < 30$ and $n < 30$). Here, these two samples are assumed to be independent.

Step 3 : Set up level of significance (α)

Step 4 : Consider the test statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{under } H_0 \; (i.e., \mu_X = \mu_Y)$$

where S_p is the "pooled" standard deviation (combined standard deviation) given by

$$S_p = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}} \, ;$$

and

$$s_X^{\;2} = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$$

$$s_Y^{\;2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

The approximate sampling distribution of the test statistic

$$T = \frac{(\bar{X} - \bar{Y})}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \qquad \text{under } H_0$$

is the t-distribution with $m+n-2$ degrees of freedom *i.e.*, $t \sim t_{m+n-2}$.

Step 5 : Calculate the value of T for the given sample $(x_1, x_2, \ldots x_m)$ and $(y_1, y_2, \ldots y_n)$ as

$$t_0 = \frac{(\bar{x} - \bar{y})}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Here $\bar{x}$ and $\bar{y}$ are the values of $\bar{X}$ and $\bar{Y}$ for the samples. Also $s_x^{\;2} = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$, $s_y^{\;2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ are the sample variances and $s_p = \sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}}$.

Step 6 : Choose the critical value, t_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Critical Value (t_e)	$t_{n-1, \frac{\alpha}{2}}$	$t_{n-1, \frac{\alpha}{2}}$	$-t_{(n-1), \alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Rejection Rule	$\|t_0\| \geq t_{n-1,\frac{\alpha}{2}}$	$t_0 > t_{n-1,\alpha}$	$t_0 < -t_{n-1,\alpha}$

Example 2.3

The following table gives the scores (out of 15) of two batches of students in an examination.

Batch I	6	7	9	2	13	3	4	8	7	11
Batch II	5	6	5	7	1	7	2	7		

Test at 1% level of significance the average performance of the students in Batch I and Batch II are equal.

Solution:

Step 1 : **Hypotheses:** Let μ_X and μ_Y denote respectively the average performance of students in Batch I and Batch II. Then the null and alternative hypotheses are :

Null Hypothesis $H_0 : \mu_X = \mu_Y$

i.e., the average performance of the students in Batch I and Batch II are equal.

Alternative Hypothesis $H_1 : \mu_X \neq \mu_Y$

i.e., the average performance of the students in Batch I and Batch II are not equal.

Step 2 : **Data**

The given sample information are:

Sample size for Batch I : m =10

Sample size for Batch II : n = 8

Step 3 : **Level of significance**

$\alpha = 1\%$

Step 4 : **Test statistic**

The test statistic under H_0 is

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

The sampling distribution of T under H_0 is the t-distribution with m+n–2 degrees of freedom *i.e.*, $t \sim t_{m+n-2}$

Step 5 : Calculation of test statistic

To find sample mean and sample standard deviation:

x_i	$u_i = x_i - \overline{x}$ $(\overline{x} = 7)$	u_i^2	y_i	$v_i = y_i - \overline{y}$ $(\overline{y} = 5)$	v_i^2
6	-1	1	5	0	0
7	0	0	6	1	1
9	2	4	5	0	0
2	-5	25	7	2	4
13	6	36	1	-4	16
3	-4	16	7	2	4
4	-3	9	2	-3	9
8	1	1	7	2	4
7	0	0			
11	4	16			
$\sum_{i=1}^{10} x_i = 70$	$\sum_{i=1}^{10} u_i = 0$	$\sum_{i=1}^{10} u_i^2 = 108$	$\sum_{i=1}^{8} y_i = 40$	$\sum_{i=1}^{8} v_i = 0$	$\sum_{i=1}^{8} v_i^2 = 38$

To find sample means:

Let $(x_1, x_2, \dots, x_{10})$ and $(y_1, y_2, \dots, y_8)$ denote the scores of students in Batch I and Batch II respectively.

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{70}{10} = 7$$

$$\overline{y} = \frac{\sum_{i=1}^{8} y_i}{8} = \frac{40}{8} = 5$$

To find combined sample standard deviation:

$$s_X^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2 = \frac{1}{9}\sum_{i=1}^{10} u_i^2 = \frac{108}{9} = 12$$

$$s_Y^2 = \frac{1}{7}\sum_{i=1}^{8}(y_i - \overline{y})^2 = \frac{1}{7}\sum_{i=1}^{8} v_i^2 = \frac{38}{7} = 5.4$$

Pooled standard deviation is:

$$S_p = \sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}} = \sqrt{\frac{108+38}{10+8-2}} = \sqrt{9.125} = 3.021$$

The value of T is calculated for the given information as

$$t_0 = \frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{7-5}{3.021\sqrt{\frac{1}{10} + \frac{1}{8}}} = 1.3957$$

Step 6 : Critical value

Since H_1 is two-sided alternative hypothesis, the critical value at $\alpha = 0.01$ is

$t_e = t_{m+n-2,\frac{\alpha}{2}} = t_{16,0.005} = 2.921$

Step 7 : Decision

Since it is two-tailed test, elements of critical region are defined by the rejection rule $|t_0| < t_e = t_{m+n-2,\frac{\alpha}{2}} = t_{16,0.005} = 2.921$. For the given sample information $|t_0| = 1.3957 < t_e = 2.921$. It indicates that given sample contains insufficient evidence to reject H_0. Hence, the mean performance of the students in these batches are equal.

Example 2.4

Two types of batteries are tested for their length of life (in hours). The following data is the summary descriptive statistics.

Type	Number of batteries	Average life (in hours)	Sample standard deviation
A	14	94	16
B	13	86	20

Is there any significant difference between the average life of the two batteries at 5% level of significance?

Solution:

Step 1 : Hypotheses

Null Hypothesis $H_0 : \mu_X = \mu_Y$

i.e., there is no significant difference in average life of two types of batteries *A* and *B*.

Alternative Hypothesis $H_0 : \mu_X \neq \mu_Y$

i.e., there is significant difference in average life of two types of batteries *A* and *B*. It is a two-sided alternative hypothesis

Step 2 : Data

The given sample information are :

m = number of batteries under type A = 14

n = number of batteries under type B = 13

$\overline{x}$ = Average life (in hours) of type A battery = 94

$\overline{y}$ = Average life (in hours) of type B battery = 86

s_X = standard deviation of type A battery =16

s_Y = standard deviation of type B battery = 20

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

The sampling distribution of T under H_0 is the t-distribution with $m+n-2$ degrees of freedom *i.e.*, $t \sim t_{m+n-2}$

Step 5 : Calculation of test statistic

Under null hypotheses H_0:

$$t_0 = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

where s is the pooled standard deviation given by,

$$s_p = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}}$$

$$= \sqrt{\frac{(14-1)(16)^2 + (13-1)(20)^2}{14+13-2}} = \sqrt{325.12} = 18.03$$

The value of T is calculated for the given information as

$$t_0 = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{94-86}{18.03\sqrt{\frac{1}{14} + \frac{1}{13}}} = \frac{8}{6.944} = 1.15$$

Step 6 : Critical value

Since H_1 is two-sided alternative hypothesis, the critical value at $\alpha = 0.05$ is $t_e = t_{m+n\text{-}2, \frac{\alpha}{2}} = t_{25,\, 0.025} = 2.060$.

Step 7 : Decision

Since it is a two-tailed test, elements of critical region are defined by the rejection rule $|t_0| < t_e = t_{m+n\text{-}2, \frac{\alpha}{2}} = t_{25,\, 0.025} = 2.060$. For the given sample information $|t_0| = 1.15 < t_e = 2.060$. It indicates that given sample contains insufficient evidence to reject H_0. Hence, there is no significant difference between the average life of the two types of batteries.

2.1.6 To test the equality of two means – paired *t*-test

Procedure:

Step 1 : Let X and Y be two correlated random variables having the distributions respectively $N(\mu_X, \sigma_X^2)$ (Population-1) and $N(\mu_Y, \sigma_Y^2)$ (Population-2). Let $D = X–Y$, then it has normal distribution $N(\mu_D = \mu_X - \mu_Y, \sigma_D^2)$.

Frame null hypothesis as

$H_0 : \mu_D = 0$

And choose alternative hypothesis from

(i) $H_1 : \mu_D \neq 0$ (ii) $H_1 : \mu_D > 0$ (iii) $H_1 : \mu_D < 0$

Step 2 : Describe the sample/data. Let $(X_1, X_2, \ldots, X_m)$ be a random sample of m observations drawn from Population-1 and $(Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations drawn from Population-2. Here, these two samples are correlated in pairs.

Step 3 : Set up level of significance (α)

Step 4 : Consider the test statistic

$$T = \frac{\bar{D}}{\frac{S}{\sqrt{n}}} \text{ under } H_0 .$$

where $\bar{D} = \dfrac{\sum_{i=1}^{n} D_i}{n}$; $D_i = X_i - Y_i$ and $S = \sqrt{\dfrac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}}$.

The approximate sampling distribution of the test statistic T under H_0 is t - distribution with $(n–1)$ degrees of freedom.

Step 5 : Calculate the value of T for the given data as

$$t_0 = \frac{\bar{d}}{\frac{s}{\sqrt{n}}}$$

where $\bar{d} = \dfrac{\sum_{i=1}^{n} d_i}{n}$; $d_i = x_i - y_i$ (sample mean) and

$$s = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$$ (sample standard deviation)

Step 6 : Choose the critical value, t_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu_D \neq 0$	$\mu_D > 0$	$\mu_D < 0$
Critical Value (t_e)	$t_{n-1, \frac{\alpha}{2}}$	$t_{n-1, \alpha}$	$-t_{n-1, \alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_D \neq 0$	$\mu_D > 0$	$\mu_D < 0$
Rejection Rule	$\lvert t_0 \rvert \geq t_{n-1, \frac{\alpha}{2}}$	$t_0 > t_{n-1, \alpha}$	$t_0 < -t_{n-1, \alpha}$

Example 2.5

A company gave an intensive training to its salesmen to increase the sales. A random sample of 10 salesmen was selected and the value (in lakhs of Rupees) of their sales per month, made before and after the training is recorded in the following table. Test whether there is any increase in mean sales at 5% level of significance.

Salesman	1	2	3	4	5	6	7	8	9	10
Before	15	22	6	17	12	20	18	14	10	16
After	17	23	16	20	14	21	18	20	10	11

Solution:

Step 1 : **Hypotheses**

Null Hypothesis $H_0 : \mu_D = 0$

i.e., there is no significant increase in the mean sales after the training.

Alternative Hypothesis $H_1 : \mu_D > 0$

i.e., there is significant increase in the mean sales after the training. It is a one-sided alternative hypothesis.

Step 2 : **Data**

Sample size $n = 10$

Step 3 : **Level of significance**

$\alpha = 5\%$

Step 4 : **Test statistic**

Test statistic under the null hypothesis is

$$T = \frac{\overline{D}}{\frac{S}{\sqrt{n}}}$$

The sampling distribution of T under H_0 is t - distribution with (10–1) = 9 degrees of freedom.

Step 5 : Calculation of test statistic

To find $\bar{d}$ and s:

Let x denote sales before training and y denote sales after training

Salesmen	x_i	y_i	$d_i = y_i - x_i$	$d_i - \bar{d}$	$(d_i - \bar{d})^2$
1	15	17	2	0	0
2	22	23	1	-1	1
3	6	16	10	8	64
4	17	20	3	1	1
5	12	14	2	0	0
6	20	21	1	-1	1
7	18	18	0	-2	4
8	14	20	6	4	16
9	10	10	0	-2	4
10	16	11	-5	-7	49
		Total	$\sum_{i=1}^{n} d_i = 20$	$\sum_{i=1}^{n}(d_i - \bar{d}) = 0$	$\sum_{i=1}^{n}(d_i - \bar{d})^2 = 140$

Here instead of $d_i = x_i - y_i$ it is assumed $d_i = y_i - x_i$ for calculations to be simpler.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{20}{10} = 2$$

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(d_i - \bar{d})^2} = \sqrt{\frac{140}{9}} = \sqrt{15.56} = 3.94$$

The calculated value of the statistic is

$$t_0 = \frac{\bar{d}}{\frac{s}{\sqrt{n}}} = \frac{2}{\frac{3.94}{\sqrt{10}}} = 1.6052$$

Step 6 : Critical value

Since H_0 is a one-sided alternative hypothesis, the critical value at 5% level of significance is $t_e = t_{n\text{-}1,\,\alpha} = t_{9,0.05} = 1.833$

Step 7 : Decision

It is a one-tailed test. Since $|t_0| = 1.6052 < t_e = t_{n\text{-}1,\,\alpha} = t_{9,0.05} = 1.833$, H_0 is not rejected. Hence, there is no evidence that the mean sales has increased after the training.

2.2 CHI-SQUARE DISTRIBUTION AND ITS APPLICATIONS

Karl Pearson

Karl Pearson (1857-1936) was a English Mathematician and Biostatistician. He founded the world's first university statistics department at University College, London in 1911. He was the first to examine whether the observed data support a given specification, in a paper published in 1900. He called it 'Chi-square goodness of fit' test which motivated research in statistical inference and led to the development of statistics as separate discipline.

Karl Pearson chi-square test the dawn of Statistical Inference - C R Rao.

***Karl Pearson's famous chi square paper appeared in the spring of 1900, an auspicious beginning to a wonderful century for the field of statistics** - B. Efron*

2.2.1 Chi-Square distribution

The square of standard normal variable is known as a chi-square variable with 1 degree of freedom (d.f.). Thus

If $X \sim N(\mu, \sigma^2)$, then it is known that $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$. Further Z^2 is said to follow χ^2 – distribution with 1 degree of freedom (χ^2 – pronounced as chi-square)

Note: i) If $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ are n *iid* random variables, then

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(X_i - \mu\right) / \sigma^2 \text{ follows } \chi^2 \text{ with } n \text{ } d.f \text{ (additive property of } \chi^2\text{)}$$

ii) If μ is replaced by $\overline{X} = \dfrac{1}{n}\sum_{i-1}^{n} X_i$ then $\dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2}$ follows $\chi^2_{n\text{-}1}$

2.2.2 Properties of χ^2 distribution

- It is a continuous distribution.
- The distribution has only one parameter *i.e. n d.f.*
- The shape of the distribution depends upon the *d.f, n.*
- The mean of the chi-square distribution is n and variance $2n$
- If U and V are independent random variables having χ^2 distributions with degree of freedom n_1 and n_2 respectively, then their sum $U + V$ has the same χ^2 distribution with $d.f$ $n_1 + n_2$.

2.2.3 Applications of chi-square distribution

- To test the variance of the normal population, using the statistic in note (ii) (sec. 2.2.1)
- To test the independence of attributes. (sec. 2.2.5)
- To test the goodness of fit of a distribution. (sec. 2.2.6)
- The sampling distributions of the test statistics used in the last two applications are approximately chi-square distributions.

2.2.4 Test of Hypotheses for population variance of the normal population (Population mean is assumed to be unknown)

Procedure

Step 1 : Let μ and σ^2 be respectively the mean and the variance of the normal population under study, where σ^2 is known and μ unknown. If $\sigma_0{}^2$ is an admissible value of σ^2, then frame the

Null hypothesis as H_0: $\sigma^2 = \sigma_0{}^2$

and choose the suitable alternative hypothesis from

(i) H_1: $\sigma^2 \neq \sigma_0{}^2$ (ii) H_1: $\sigma^2 > \sigma_0{}^2$ (iii) H_1: $\sigma^2 < \sigma_0{}^2$

Step 2 : Describe the sample/data and its descriptive measures. Let $(X_1, X_2, \ldots, X_n)$ be a random sample of n observations drawn from the population, where n is small ($n < 30$).

Step 3 : Fix the desired level of significance α.

Step 4 : Consider the test statistic $\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$ under H_0. The approximate sampling distribution of the test statistic under H_0 is the chi-square distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculate the value of the of χ^2 for the given sample as $\chi_0{}^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

Step 6 : Choose the critical value of χ_e^2 corresponding to α and H_1 from the following table.

Alternative Hypothesis (H_1)	$\sigma^2 \neq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\sigma^2 < \sigma_0^2$
Critical value (χ_e^2)	$\chi^2{}_{n-1,\frac{\alpha}{2}}$ and $\chi_0{}^2 \leq \chi^2{}_{n-1,1-\frac{\alpha}{2}}$	$\chi^2_{n-1,\alpha}$	$\chi^2_{n-1,1-\alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\sigma^2 \neq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\sigma^2 < \sigma_0^2$
Rejection Rule	$\chi^2{}_{n-1,\frac{\alpha}{2}}$ and $\chi_0{}^2 \leq \chi^2{}_{n-1,1-\frac{\alpha}{2}}$	$\chi_0{}^2 > \chi^2{}_{n-1,\alpha}$	$\chi_0{}^2 < \chi^2{}_{n-1,\,1-\alpha}$

NOTE

If the population mean μ is known then for testing $H_0 : \sigma^2 = \sigma_0^2$ against any of the alternatives, we use $\chi_0^2 = \dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma_0^2}$ with n *d.f.*

Example 2.6

The weights (in kg.) of 8 students of class VII are 38, 42, 43, 50, 48, 45, 52 and 50. Test the hypothesis that the variance of the population is 48 kg, assuming the population is normal and μ is unknown.

Solution:

Step 1 : **Null Hypothesis** H_0: $\sigma^2 = 48$ kg.

i.e. Population variance can be regarded as 48 kg.

Alternative hypothesis H_1: $\sigma^2 \neq 48$ kg.

i.e. Population variance cannot be regarded as 48 kg.

Step 2 : The given sample information is
Sample size (n)= 8

Step 3 : **Level of significance**
$\alpha = 5\%$

Step 4 : **Test statistic**

Under null hypothesis H_0

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

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

Step 5 : Calculation of test statistic

The value of chi-square under H_0 is calculated as under:

To find $\bar{x}$ and sample variance s^2, we form the following table.

x_i	$(x_i - 46)$	$(x_i - 46)^2$
38	-8	64
42	-4	16
43	-3	9
50	4	16
48	2	4
45	-1	1
52	6	36
50	4	16
$\sum_{i=1}^{8} x_i = 368$	0	$\sum_{i=1}^{8} (x_i - \bar{x})^2 = 162$

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{n} = \frac{368}{8} = 46$$

$$s^2 = \frac{\sum_{i=1}^{8}(x_i - \bar{x})^2}{(n-1)} = \frac{\sum_{i=1}^{8}(x_i - 46)^2}{(8-1)} = \frac{162}{7} = 23.143\cdot$$

The calculated value of chi-square is $\chi_0^{\;2} = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{8}(x_i - \bar{x})^2}{\sigma_0^2} = \frac{162}{48} = 3.375$

Step 6 : Critical values

Since H_1 is a two sided alternative, the critical values at α =0.05 are $\chi^2_{7,\,0.025} = 16.01$ and $\chi^2_{7,0.975} = 1.69\cdot$

Step 7 : Decision

Since it is a two-tailed test, the elements of the critical region are determined by the rejection rule $\chi_0^{\;2} \geq \chi^{\;2}_{n-1,\frac{\alpha}{2}}$ or $\chi_0^{\;2} \leq \chi^{\;2}_{n-1,1-\frac{\alpha}{2}}$.

For the given sample information, the rejection rule does not hold, since

$1.69 = \chi^2_{7,\,0.975} < \chi_0^{\;2}\,(=3.375) < \chi^2_{7,0.025} = 16.01.$

Hence, H_0 is not rejected in favour of H_1. Thus, Population variance can be regarded as 48 kg.

Example 2.7

A normal population has mean μ (unknown) and variance 9. A sample of size 9 observations has been taken and its variance is found to be 5.4. Test the null hypothesis H_0: $\sigma^2 = 9$ against H_1: $\sigma^2 > 9$ at 5% level of significance.

Solution:

Step 1 : Null Hypothesis H_0: $\sigma^2 = 9$.

i.e., Population variance regarded as 9.

Alternative hypothesis H_1: $\sigma^2 > 9$.

i.e. Population variance is regarded as greater than 9.

Step 2 : Data

Sample size (n) = 9

Sample variance (s^2) = 5.4

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

Under null hypothesis H_0

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

follows chi-square distribution with (n-1) degrees of freedom.

Step 5 : Calculation of test statistic

The value of chi-square under H_0 is calculated as

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{8\times 5.4}{9} = 4.8$$

Step 6 : Critical value

Since H_1 is a one-sided alternative, the critical values at α =0.05 is $\chi_e^2 = \chi^2_{8,\,0.05} = 15.507$.

Step 7 : Decision

Since it is a one-tailed test, the elements of the critical region are determined by the rejection rule $\chi_0^2 > \chi_e^2$.

For the given sample information, the rejection rule does not hold , since $\chi_0^2 = 4.8 < \chi^2_{8,\,0.05} = 15.507$. Hence, H_0 is not rejected in favour of H_1. Thus, the population variance can be regarded as 9.

Example 2.8

A normal population has mean μ (unknown) and variance 0.018. A random sample of size 20 observations has been taken and its variance is found to be 0.024. Test the null hypothesis H_0: σ^2 = 0.018 against H_1: σ^2 < 0.018 at 5% level of significance.

Solution:

Step 1 : Null Hypothesis H_0: σ^2 = 0.018.

i.e. Population variance regarded as 0.018.

Alternative hypothesis H_1: σ^2 < 0.018.

i.e. Population variance is regarded as lessthan 0.018.

Step 2 : Data

Sample size (n) = 20

Sample variance (s^2) = 0.024

Step 3 : Level of significance

α = 5%

Step 4 : Test statistic

Under null hypothesis H_0

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

follows chi-square distribution with (n–1) degrees of freedom.

Step 5 : Calculation of test statistic

The value of chi-square under H_0 is calculated as

$$\chi_0^{\,2} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{19 \times 0.024}{0.018} = 25.3$$

Step 6 : **Critical value**

Since H_1 is a one-sided alternative, the critical values at α =0.05 is $\chi_e^{\,2} = \chi^2_{19,\,0.95} = 10.117$.

Step 7 : **Decision**

Since it is a one-tailed test, the elements of the critical region are determined by the rejection rule $\chi_0^{\,2} < \chi_e^{\,2}$

For the given sample information, the rejection rule does not hold, since

$\chi_0^{\,2} = 25.3 > \chi_e^{\,2} = \chi^2_{19,\,0.95} = 10.117$.

Hence, H_0 is not rejected in favour of H_1. Thus, the population variance can be regarded as 0.018.

2.2.5 Test of Hypotheses for independence of Attributes

Another important application of χ^2 test is the testing of independence of attributes.

Attributes: Attributes are qualitative characteristic such as levels of literacy, employment status, *etc*., which are quantified in terms of levels/scores.

Contigency table: Independence of two attributes is an important statistical application in which the data pertaining to the attributes are cross classified in the form of a two – dimensional table. The levels of one attribute are arranged in rows and of the other in columns. Such an arrangement in the form of a table is called as a contingency table.

Computational steps for testing the independence of attributes:

Step 1 : **Framing the hypotheses**

Null hypothesis H_0: The two attributes are independent

Alternative hypothesis H_1: The two attributes are not independent.

Step 2 : **Data**

The data set is given in the form of a contigency as under. Compute expected frequencies E_{ij} corresponding to each cell of the contingency table, using the formula

$$E_{ij} = \frac{R_i \times C_j}{N}; \quad i = 1, 2, \ldots m;\ j = 1, 2, ..n$$

where,

N = Total sample size

R_i = Row sum corresponding to i^{th} row

C_j = Column sum corresponding to j^{th} column

NOTE

The contingency table consisting of m rows and n columns.
The observed data is presented in the form of a contingency table :

		Attribute B						Total
		B_1	B_2	…	B_j	…	B_n	
Attribute A	A_1	O_{11}	O_{12}	…	O_{1j}	…	O_{1n}	R_1
	A_2	O_{21}	O_{22}	…	O_{2j}	…	O_{2n}	R_2
	:	:	:	:	:	:	:	:
	.	.	.	.	.	.	.	.
	A_i	O_{i1}	O_{i2}	…	O_{ij}	…	O_{in}	R_i
	:	:	:	:	:	:	:	:
	.	.	.	.	.	.	.	.
	A_m	O_{m1}	O_{m2}	…	O_{mj}	…	O_{mn}	R_m
Total	C_1	C_2	…	C_j	…	C_n	$N = m{\times}n$	

Step 3 : Level of significance

Fix the desired level of significance α

Step 4 : Calculation

Calculate the value of the test statistic as

$$\chi_0^{\ 2} = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Step 5 : Critical value

The critical value is obtained from the table of χ^2 with $(m-1)(n-1)$ degrees of freedom at given level of significance, α as $\chi^2_{(m-1)(n-1),\,\alpha}$.

Step 6 : Decision

Decide on rejecting or not rejecting the null hypothesis by comparing the calculated value of the test statistic with the table value. If $\chi_0^{\ 2} \geq \chi^2_{(m-1)(n-1),\,\alpha}$ reject H_0.

Note:

- N, the total frequency should be reasonably large, say greater than 50.
- No theoretical cell-frequency should be less than 5. If cell frequencies are less than 5, then it should be grouped such that the total frequency is made greater than 5 with the preceding or succeeding cell.

Example 2.9

The following table gives the performance of 500 students classified according to age in a computer test. Test whether the attributes age and performance are independent at 5% of significance.

Performance	Below 20	21-30	Above 30	Total
Average	138	83	64	285
Good	64	67	84	215
Total	202	150	148	500

Solution:

Step 1 : **Null hypothesis** H_0: The attributes age and performance are independent.

Alternative hypothesis H_1: The attributes age and performance are not independent.

Step 2 : **Data**

Compute expected frequencies E_{ij} corresponding to each cell of the contingency table, using the formula

$$E_{ij} = \frac{R_i \times C_j}{N} \quad i = 1,2;\ j = 1,2,3$$

where,

N = Total sample size

R_i = Row sum corresponding to i^{th} row

C_j = Column sum corresponding to j^{th} column

Performance	Below average	Average	Above average	Total
Average	$\frac{285\times 202}{500} = 115.14$	$\frac{285\times 150}{500} = 85.5$	$\frac{285\times 148}{500} = 84.36$	285
Good	$\frac{215\times 202}{500} = 86.86$	$\frac{215\times 150}{500} = 64.5$	$\frac{215\times 148}{500} = 63.64$	215
Total	202	150	148	500

Step 3 : **Level of significance** α = 5%

Step 4 : **Calculation**

Calculate the value of the test statistic as

$$\chi_0{}^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Step 5 : **This chi-square test statistic is calculated as follows:**

$$\chi_0^2 = \frac{(138-115.14)^2}{115.14} + \frac{(83-85.50)^2}{88.50} + \frac{(64-84.36)^2}{84.36} + \frac{(64-86.86)^2}{86.86} + \frac{(67-64.50)^2}{64.50} + \frac{(84-63.64)^2}{63.64}$$

= 22.152 with degrees of freedom (3–1)(2–1) = 2

Step 6 : **Critical value**

From the chi-square table the critical value at 5% level of significance is $\chi^2{}_{(2-1)(3-1),0.05} = \chi^2{}_{2,0.05} = 5.991$.

Step 7 : Decision

As the calculated value $\chi_0^2 = 22.152$ is greater than the critical value $\chi^2_{2,0.05} = 5.991$, the null hypothesis H_0 is rejected. Hence, the performance and age of students are not independent.

NOTE

If the contigency table is 2 x 2 then the value of χ^2 can be calculated as given below:

	A	not *A*	Total
B	*a*	*b*	*a*+*b*
not *B*	*c*	*d*	*c*+*d*
Total	*a*+*c*	*b*+*d*	*N*=*a*+*b*+*c*+*d*

$$\chi_0^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi_\alpha^2(1d.f)$$

The following example will illustrate the procedure

Example 2.10

A survey was conducted with 500 female students of which 60% were intelligent, 40% had uneducated fathers, while 30 % of the not intelligent female students had educated fathers. Test the hypothesis that the education of fathers and intelligence of female students are independent.

Solution:

Step 1 : Null hypothesis H_0: The attributes are independent *i.e.* No association between education fathers and intelligence of female students

Alternative hypothesis H_1: The attributes are not independent *i.e* there is association between education of fathers and intelligence of female students

Step 2 : Data

The observed frequencies (*O*) has been computed from the given information as under.

	Intelligent females	Not intelligent females	Row total
Educated fathers	$300-120=180$	$\frac{30}{100}\times 200=60$	240
Uneducated fathers	$\frac{40}{100}\times 300=120$	$200-60=140$	260
Total	$\frac{60}{100}\times 500=300$	$500-300=200$	N= 500

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Calculation

Calculate the value of the test statistic as

$$\chi_0{}^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Step 5 : Find the value of x^2

where, a = 180, b = 60, c = 120, d = 140, N = 500

$$\chi_0{}^2 = \frac{500(180\times140-60\times120)^2}{(180+60)(120+140)(180+120)(60+140)} = 43.269$$

Step 6 : Critical value

From chi-square table the critical value at 5% level of significance is $\chi^2{}_{1,0.05} = 3.841$

Step 7 : Decision

The calculated value $\chi_0{}^2 = 43.269$ is greater than the critical value $\chi^2{}_{1,0.05} = 3.841$, the null hypothesis H_0 is rejected. Hence, education of fathers and intelligence of female students are not independent.

2.2.6 Tests for Goodness of Fit

Another important application of chi-square distribution is testing goodness of a pattern or distribution fitted to given data. This application was regarded as one of the most important inventions in mathematical sciences during 20th century. Goodness of fit indicates the closeness of observed frequency with that of the expected frequency. If the curves of these two distributions do not coincide or appear to diverge much, it is noted that the fit is poor. If two curves do not diverge much, the fit is fair.

Computational steps for testing the significance of goodness of fit:

Step 1 : Framing of hypothesis

Null hypothesis H_0 : The goodness of fit is appropriate for the given data set

Alternative hypothesis H_1 : The goodness of fit is not appropriate for the given data set

Step 2 : Data

Calculate the expected frequencies (E_i) using appropriate theoretical distribution such as Binomial or Poisson.

Step 3 : Select the desired level of significance α

Step 4 : Test statistic

The test statistic is

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

where k = number of classes

O_i and E_i are respectively the observed and expected frequency of i^{th} class such that $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$.

If any of E_i is found less than 5, the corresponding class frequency may be pooled with preceding or succeeding classes such that E_i's of all classes are greater than or equal to 5. It may be noted that the value of k may be determined after pooling the classes.

The approximate sampling distribution of the test statistic under H_0 is the chi-square distribution with k-1-s $d.f$, s being the number of parametres to be estimated.

Step 5 : Calculation

Calculate the value of chi-square as

$$\chi_0^{\ 2} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

The above steps in calculating the chi-square can be summarized in the form of the table as follows:

Step 6 : Critical value

The critical value is obtained from the table of χ^2 for a given level of significance α.

Step 7 : Decision

Decide on rejecting or not rejecting the null hypothesis by comparing the calculated value of the test statistic with the table value, at the desired level of significance.

Example 2.11

Five coins are tossed 640 times and the following results were obtained.

Number of heads	0	1	2	3	4	5
Frequency	19	99	197	198	105	22

Fit binomial distribution to the above data.

Solution:

Step 1 : Null hypothesis H_0: Fitting of binomial distribution is appropriate for the given data.

Alternative hypothesis H_1: Fitting of binomial distribution is not appropriate to the given data.

Step 2 : Data

Compute the expected frequencies:

n = number of coins tossed at a time = 5

Let X denote the number of heads (success) in n tosses

N = number of times experiment is repeated = 640

To find mean of the distribution

x	f	fx
0	19	0
1	99	99
2	197	394
3	198	594
4	105	420
5	22	110
Total	640	1617

Mean : $\bar{x} = \dfrac{\sum fx}{\sum f} = \dfrac{1617}{640} = 2.526$

The probability mass function of binomial distribution is :

$$p(x) = {}^{n}C_{x}\, p^{x}\, q^{n-x},\ x = 0,1,\ldots,n \qquad (2.1)$$

Mean of the binomial distribution is $\overline{x} = np$.

Hence, $\hat{p} = \dfrac{\overline{x}}{n} = \dfrac{2.526}{5} \approx 0.5$

$$\hat{q} = 1 - \hat{p} \approx 0.5$$

For $x = 0$, the equation (2.1) becomes

$P(X = 0) = P(0) = 5c_0\,(0.5)^5 = 0.03125$

The expected frequency at $x = N\,P(x)$

The expected frequency at x =0 : $N \times P(0)$

$$= 640 \times 0.03125 = 20$$

We use recurrence formula to find the other expected frequencies.

The expected frequency at x+1 is

$\dfrac{n-x}{x+1}\left(\dfrac{p}{q}\right) \times$ Expected frequency at x

x	$\dfrac{n-x}{x+1}$	$\dfrac{p}{q}$	$\dfrac{n-x}{x+1}\left(\dfrac{p}{q}\right)$	Expected frequency at $x = N\,P(x)$
0	5	1	5	20
1	2	1	2	100
2	1	1	1	200
3	0.5	1	0.5	200
4	0.2	1	0.2	100
5	0	1	0	20

Table of expected frequencies:

Number of heads	0	1	2	3	4	5	Total
Expected frequencies	20	100	200	200	100	20	640

Step 3 : **Level of significance**

$\alpha = 5\%$

Step 4 : **Test statistic**

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

Step 5 : **Calculation**

The test statistic is computed as under:

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
19	20	–1	1	0.050
99	100	–1	1	0.010
197	200	–3	9	0.045
198	200	–2	4	0.020
105	100	5	25	0.250
22	20	2	4	0.200
			Total	0.575

$$\chi_0{}^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
$$= 0.575$$

Step 6 : **Critical value**

Degrees of freedom = $k - 1 - s = 6 - 1 - 1 = 4$

Critical value for $d.f$ 4 at 5% level of significance is 9.488 *i.e.*, $\chi^2_{4,0.05} = 9.488$

Step 7 : **Decision**

As the calculated $\chi_0{}^2$(=0.575) is less than the critical value $\chi^2_{4,0.05} = 9.488$, we do not reject the null hypothesis. Hence, the fitting of binomial distribution is appropriate.

Example 2.12

A packet consists of 100 ball pens. The distribution of the number of defective ball pens in each packet is given below:

x	0	1	2	3	4	5
f	61	14	10	7	5	3

Examine whether Poisson distribution is appropriate for the above data at 5% level of significance.

Solution:

Step 1 : **Null hypothesis** H_0: Fitting of Poisson distribution is appropriate for the given data.

Alternative hypothesis H_1: Fitting of Poisson distribution is not appropriate for the given data.

Step 2 : **Data**

The expected frequencies are computed as under:

To find the mean of the distribution.

x	f	fx
0	61	0
1	14	14
2	10	20
3	7	21
4	5	20
5	3	15
Total	100	90

$$\overline{x} = \frac{\sum fx}{\sum f} = \frac{90}{100} = 0.9$$

Probability mass function of Poisson distribution is:

$$p(x) = \frac{e^{-m} m^x}{x!}; \; x = 0, 1, \ldots \qquad (2.2)$$

In the case of Poisson distribution mean $(m) = \overline{x} = 0.9$.

At $x = 0$, equation (2.2) becomes

$$p(0) = \frac{e^{-m} \; m^0}{0!} = e^m = e^{0.9} = 0.4066 \cdot$$

The expected frequency at x is $N\,P(x)$

Therefore, The expected frequency at 0 is

$$N \times P(0)$$
$$= 100 \times 0.4066$$
$$= 40.66$$

We use recurrence formula to find the other expected frequencies.

The expected frequency at *x*+1 is

$$\frac{m}{x+1} \times \text{Expected frequency at } x$$

x	$\frac{m}{x+1}$	Expected frequency at $x = N\,P(x)$
0	0.9	40.66
1	$\frac{0.9}{2}$	36.594
2	$\frac{0.9}{3}$	16.4673
3	$\frac{0.9}{4}$	4.94019
4	$\frac{0.9}{5}$	1.1115
5	$\frac{0.9}{6}$	0.20007

Table of expected frequency distribution (on rounding to the nearest integer)

x	0	1	2	3	4	5
Expected frequency	41	37	16	5	1	0

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

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

Step 5 : Calculation

Test statistic is computed as under:

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
61	41	20	400	9.756
14	37	-23	529	14.297
10	16	-6	36	2.250
7	5			
5 } 15	1 } 6	9	81	13.5
3	0			
			Total	39.803

Note: In the above table, we find the cell frequencies 0,1 in the expected frequency column (E) is less than 5, Hence, we combine (pool) with either succeeding or preceding one such that the total is made greater than 5. Here we have pooled with preceding frequency 5 such that the total frequency is made greater than 5. Correspondingly, cell frequencies in observed frequencies are pooled.

$$\chi_0^{\,2} = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}$$

$$=39.803$$

Step 6 : Critical value

Degrees of freedom = $(k - 1 - s) = 4 - 1 - 1 = 2$

Critical value for 2 *d.f* at 5% level of significance is 5.991 *i.e.*, $\chi^2_{2,0.05} = 5.991$

Step 7 : Decision

The calculated $\chi_0^{\,2}$(=39.803) is greater than the critical value (5.991) at 5% level of significance. Hence, we reject H_0. i.e., fitting of Poisson distribution is not appropriate for the given data.

Example 2.13

A sample 800 students appeared for a competitive examination. It was found that 320 students have failed, 270 have secured a third grade, 190 have secured a second grade and the remaining students qualified in first grade. The general opinion that the above grades are in the ratio 4:3:2:1 respectively. Test the hypothesis the general opinion about the grades is appropriate at 5% level of significance.

Step 1 : Null hypothesis H_0: The result in four grades follows the ratio 4:3:2:1

Alternative hypothesis H_1: The result in four grades does not follows the ratio 4:3:2:1

Step 2 : **Data**

Compute expected frequencies:

Under the assumption on H_0, the expected frequencies of the four grades are:

$$\frac{4}{10}\times 800 = 320\ ; \frac{3}{10}\times 800 = 240; \frac{2}{10}\times 800 = 160; \frac{1}{10}\times 800 = 80$$

Step 3 : **Test statistic**

The test statistic is computed using the following table.

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
320	320	0	0	0
270	240	30	900	3.75
190	160	30	900	5.625
20	80	-60	3600	45
			Total	54.375

The test statistic is calculated as

$$\chi_0^{\ 2} = \sum_{i=1}^{k}\frac{\left(O_i - E_i\right)^2}{E_i}$$
$$= 54.375$$

Step 4 : **Critical value**

The critical value of χ^2 for 3 d.f. at 5% level of significance is 7.81 *i.e.*, $\chi^2_{3,0.05} = 7.81$

Step 5 : **Decision**

As the calculated value of $\chi_0^{\ 2}$(=54.375) is greater than the critical value $\chi^2_{3,0.05} = 7.81$, reject H_0. Hence, the results of the four grades do not follow the ratio 4:3:2:1.

Example 2.14

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

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

Test whether the occurence of the digits in the directory are equal at 5% level of significance.

Step 1 : **Null hypothesis** H_0: The occurrence of the digits are equal in the directory.

Alternative hypothesis H_1: The occurrence of the digits are not equal in the directory.

Step 2 : Data

The expected frequency for each digit = $\frac{10000}{10} = 1000$

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test statistic

The test statistic is computed using the following table.

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1026	1000	26	676	0.676
1107	1000	107	11449	11.449
997	1000	3	9	0.009
966	1000	34	1156	1.156
1075	1000	75	5625	5.625
933	1000	67	4489	4.489
1107	1000	107	11449	11.449
972	1000	28	784	0.784
964	1000	36	1296	1.296
853	1000	147	21609	21.609
			Total	58.542

The test statistic is calculated as

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= 58.542$$

Step 4 : Critical value

Critical value for 9 df at 5% level of significance is 16.919 i.e., $\chi^2_{9,0.05} = 16.919$

Step 5 : Decision

Since the calculated χ_0^2 (58.542) is greater than the critical value $\chi^2_{9,0.05} = 16.919$, reject H_0. Hence, the digits are not uniformly distributed in the directory.

POINTS TO REMEMBER

- ❖ If the number of elements in the sample is less than 30, it is called a small sample.
- ❖ For conducting *t*-tests the parent population(s) should be normal and the samples(s) should be small.
- ❖ In case of two sample problems based on *t*-distribution the sizes of both samples must be less than 30.
- ❖ The *t*-distribution is symmetrical about its mean(zero)
- ❖ When the degrees of freedom is large the *t*-distribution converges to *N*(0,1) distribution.

- The degree of freedom represents the number of independent observation in the sample.
- The sampling distribution of the test statistic for testing hypothesis about normal population mean is t_{n-1}, when n is small and σ is unknown.
- The sampling distribution of the test statistic for testing equality of two normal population mean is t_{m+n-2} when m, $n < 30$ and the common population variance σ^2 is unknown.
- If $Z \sim N(0,1)$ then $Z^2 \sim \chi^2$ with 1 d.f.
- The uses of χ^2 – distribution are (i) testing the specified variance of a normal population (ii) testing goodness of fit and (iii) testing independence of attributes.
- When expected frequency for a cell is less than 5, it is should be clubbed with the adjacent cells such that the expected frequency in the resultant cell is greater than 5.
- The degrees of freedom for the χ^2 – statistic used for the independence of attributes is $(m-1) \times (n-1)$, where m and n are respectively the number of rows and columns in a contegency table.
- The expected cell frequency testing independence of attributes is calculated as
  $$\frac{\text{Row total} \times \text{Column total}}{\text{Sample Size}}$$
- The expected cell frequency in testing goodness of fit is calculated as sample size × {probability for the corresponding cell}

EXERCISE 2

I. Choose the best answer.

1. Student's '*t*' distribution was found by

a) Karl Pearson b) Laplace

c) R.A. Fisher d) William S.Gosset

2. Support of student's t random variable is

a) $-\infty < t \le 0$ b) $0 \le t < \infty$

c) $-\infty < t < \infty$ d) $0 \le t \le 1$

3. Paired t-test is applicable when the observations in both the samples are

a) Paired b) Correlated

c) Equal in number d) all the above

4. The number of degrees of freedom for the test statistic $t = \dfrac{(\bar{x}-\mu)}{s/\sqrt{n}}$ is

a) $n - 1$ b) n c) $n - 2$ d) $n + 1$

5. Standard error of difference between two sample means in the case of small samples is

a) $\sigma^2\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}$ b) $\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}$ c) $s_p\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}$ d) $\sqrt{\dfrac{\sigma_1}{m}+\dfrac{\sigma_2}{n}}$

6. If the size of sample is larger than 30, the *t*-distribution tends to

   a) normal distribution　　b) *F*-distribution

   c) chi-square distribution　　d) Poisson distribution

7. If a random sample of 10 observations has variance 324 then standard error is

   a) $18/\sqrt{10}$　　b)18/10

   c) 10/18　　d) $2/\sqrt{5}$

8. A sample of 16 units was taken for testing an hypothesis concerning the mean of a normal population. Then the degrees of freedom of the appropriate test statictic is

   a) 14　　b) 15

   c) 16　　d) 8

9. If $s_1{}^2$ and $s_2{}^2$ are respectively the variance of two independent random samples of sizes '*m*' and '*n*'. Then standard deviation of the combined sample is

   a) $\sqrt{\dfrac{ms_1^2+ns_2^2}{m+n}}$　　b) $\sqrt{\dfrac{(m-1)s_1^2+(n-1)s_2^2}{m+n}}$

   c) $\sqrt{\dfrac{ms_1^2+ns_1^2}{m+n+2}}$　　d) $\sqrt{\dfrac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}}$

10. A company gave an intensive training to its salesman to increase the sales. A random sample of 6 salesmen was selected and the value of their sales made before and after the training is recorded. Which test will be more appropriate to test whether there is an increase in mean sales

    a) normal test　　b) paired *t*-test　　c) χ^2-test　　d) *F*-test

11. If the order of the contigency table is (5 × 4). Then the degree of freedom of the corresponding chi-square test statistic is

    a) 18　　b) 17　　c) 12　　d) 25

12. For testing the hypothesis concerning variance of a normal population ________ is used.

    a) *t*-test　　b) *F*-test　　c) *Z*-test　　d) χ^2-test

13. If σ^2 is the variance of normal population, then the degrees of freedom of the sampling distribution of the test statistic for testing $H_0 : \sigma^2 = \sigma_0^2$ is:

    a) $n-1$　　b) $n+1$　　c) n　　d) $n-2$

14. If *n* is the degree of freedom of chi-square distribution then its variance is

    a) n　　b) $n-1$　　c) $2n$　　d) $n+1$

15. If chi-square is performed for testing goodness of fit to a data with *k* classes on estimating '*s*' parameters then degrees of fredom of test statistic is.

    a) $k-s$　　b) $(k-1)(s-1)$

    c) $k-1-s$　　d) $k-1$

16. The statistic χ^2, with usual notations, in case of contingency table of order (*m* × *n*) is given by

    a) $\chi_0{}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\dfrac{(O_{ij}-E_{ij})^2}{E_{ij}}$　　b) $\chi_0{}^2 = \sum_{i=1}^{k}\left[\dfrac{(O_i-E_i)}{E_i}\right]^2$

    c) $\chi_0{}^2 = \sum_{i=1}^{k}\dfrac{(O_i-E_i)}{E_i}$　　d) $\chi_c{}^2 = \sum_{i=1}^{k}\dfrac{O_i}{E_i}$

II. Give very short answer to the following questions.

17. Define student’s *t*-statistic.
18. Define: degrees of freedom.
19. Define the paired *t*-statistic.
20. When paired *t*-test can be applied?
21. Write the test statistic to test the difference between normal population means.
22. Write the standard error of the difference between sample means.
23. Define chi-square statistic.
24. Write the applications of chi-square distribution.
25. What are the minimum requirements of chi-square test?
26. Define an attribute.
27. Give the recurrence formula for binomial distribution.

III. Give short answer to the following questions.

28. List out the properties of *t*-distribution.
29. Write down the applications of *t*-distribution.
30. Explain the testing procedure to test the normal population mean, when population variance is unknown.
31. Write down the procedure to test significance for equality of means of two normal populations based on small samples.
32. A random sample of ten students is taken and their marks in a particular subject are recorded. The average mark is 60 with standard deviation 6.5. Test the hypothesis that the average mark of students is 55.
33. State the properties of χ^2 distribution.
34. What is a contigency table?
35. Write the procedure to test the population variance.
36. Write the test procedure for testing the independence of attributes.
37. Write down the computational steps for testing the significance of goodness of fit.
38. Give the test statistic for 2 × 2 contingency tables.

IV. Give detailed answer to the following questions.

39. A random sample of 10 packets containing cashew nuts weigh (in grams) 70,120,110,101, 88,83,95,98,107,100 each. Test whether the population mean weight of 100 grams?
40. The average run of cricket player from the past records is 80. The recent scores of the player in 6 test matches are 84, 82, 83, 79, 83 and 85. Test whether the average run is more than 80?
41. The heights (in feet) of 6 rain trees in a town A are 30, 28, 29, 32, 31, 36 and that of 8 rain trees in another town B are 35, 36, 37, 30, 32, 29, 35, 30. Is there any significant difference in mean heights of rain trees?

42. Samples of two types of electric bulbs were tested for life (in hours) and the following data were obtained.

	TYPE-I	TYPE-II
Number of units	8	7
Mean of the samples (in hrs.)	1134	1024
Standard deviation of the samples (in hrs)	35	40

Test the hypothesis that the population means are equal at 5% level of significance.

43. The number of pages typed by 5 DTP – operators for 1 hour in the morning sessions are 10, 12, 13, 8, 9 and the number of pages typed by them in the afternoon are 11, 15, 12, 10, 8. Is there any significant difference in the mean number of pages typed?

44. An IQ test was conducted to 5 persons before and after they were trained. The results are given below:

Candidates	I	II	III	IV	V
IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether any change in IQ at 1% level of significance.

45. The marks secured by 9 students in Statistics and that of 12 students in Business Mathematics are given below:

Marks in Statistics	65	74	64	58	60	67	71	69	75			
Marks in Business Mathematics	52	45	59	47	53	64	58	62	54	61	57	48

Test whether the mean marks obtained by the students in Statistics and Business mathematics differ significantly at 1% level of significance.

46. A test was conducted with 6 students before and after the training programme. Their marks were recorded and tabulated as shown below. Test whether the training was helpful in improving their scores.

Before training	100	160	113	122	120	105
After training	120	155	120	128	115	100

47. An experiment was conducted 144 times with tossing of four coins and the number of heads appeared at each throw are recorded.

No. of heads	0	1	2	3	4
frequency	10	34	56	36	8

Fit binomial distribution to the above data.

48. The distribution of the number of defective blades produced in a single shift in a factory over 100 shifts is given below.

Number of defective blades	0	1	2	3	4
Number of shifts	12	14	23	18	33

Test whether the number of defective blades follows a Poisson distribution with mean = 0.44. Use $\alpha = 0.05$.

49. The quality grade of electric components produced in two factories is given in the table given below.

Factory	Quality of grade				Total
	Poor	Medium	Good	Excellent	
A	136	165	151	148	600
B	31	58	55	36	180
Total	167	223	206	184	780

Test whether there is any association between factories and quality of grades.

50. The eyesight was tested among 2000 randomly selected patients from a city and the following details are obtained.

Gender	Eye-sight		Total
	Poor	Good	
Male	620	380	1000
Female	550	450	1000
Total	1170	830	2000

Can we conclude that there is an association between gender and quality of eye-sight at 5% level of significance?

51. The weights (in kg) of 10 students from a school are 38,40,45,53,47,43,55,48,52,49. Can we say that variance of the distribution of weights of all students from the above school is equal to 20 kg?

52. In a sample of 200 households in a colony, two brands of hair oils A and B are applied by 90 females. Further, 60 females and 70 males are using brand A. To test whether there is any association between the gender and brand of hair oil used, by constructing a contigency table.

ANSWERS

I.	1. d	2. c	3. d	4. a	5. c
	6. a	7. a	8. c	9. d	10. b
	11. c	12.d	13. a	14.c	15. c
	16 a				

III. 32. $t = 2.43$, reject H_0

IV. 39. $t = -0.6202$, do not reject H_0

40. $t = 3.16$, reject H_0

41. $|t| = -1.23443$, we do not reject H_0

42. $t = 5.683$, reject H_0

43. $|t| = 1$, we do not reject H_0

44. $t = 0.8164$, we do not reject H_0

45. $t = 4.4898$,reject H_0

46. $|t| = 0.7304$, we do not reject H_0

47. $\chi_0^2 = 0.407407$, we do not reject H_0 at 5% level with 4 d.f.

48. $\chi_0^2 = 35.10855$, reject H_0 at 5% level with 4 d.f.

49. $\chi_0^2 = 5.79$, do not reject H_0 at 5% level with 3 d.f.

50. $\chi_0^2 = 10.09165$, reject H_0 at 5% level with 2 d.f.

51. $\chi_0^2 =14$, we do not reject H_0 at 5% level with 9 d.f.

52.

Gender	Hair Oil Brands		TOTAL
	A	B	
Male	70	40	110
Female	60	30	90
Total	130	70	200

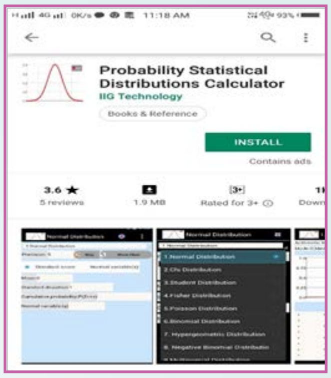
$\chi_0^2 = 0.1998$, we do not reject H_0 at 5% level with 1 d.f.

ICT CORNER

TESTS BASED ON SAMPLING DISTRIBUTIONS I

STATS IN YOUR PALM

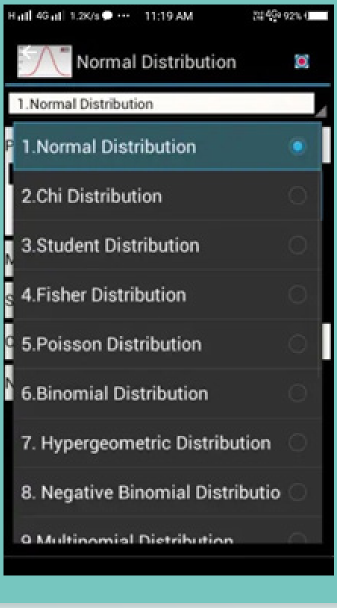
This activity is to calculate Chi distribution, Binomial Distribution, Students Distribution



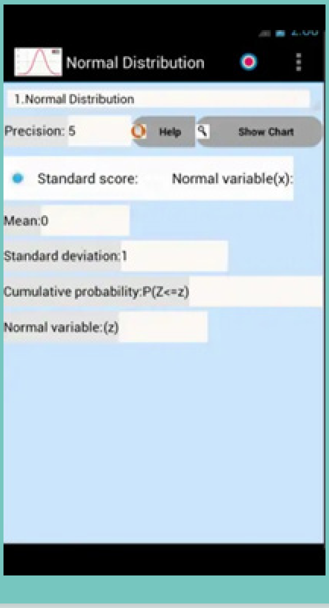
Steps:

- This is an android app activity. Open the browser and type the URL given (or) scan the QR code. (Or) search for Probability Statistical Distributions Calculator in google play store.
- (i) Install the app and open the app, (ii) click "**Menu**", (iii) In the menu page click "**Students Distribution**" menu.
- Input freedom degree and t-store, cumulative probability to get the output.

Step-1



Step-2



Precision: 5

● Standard score: Normal variable(x):

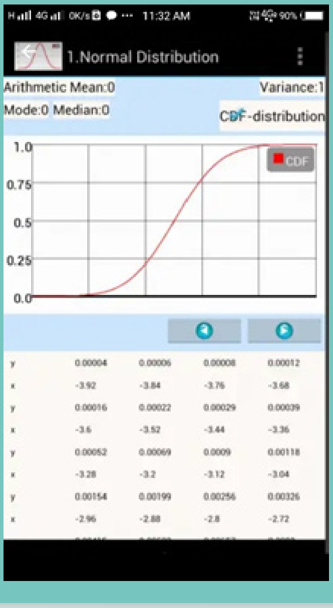
Mean:0

Standard deviation:1

Cumulative probability:P(Z<=z)

Normal variable:(z)

Step-3



Arithmetic Mean:0 Variance:1

Mode:0 Median:0

y	0.00004	0.00006	0.00008	0.00012
x	-3.92	-3.84	-3.76	-3.68
y	0.00016	0.00022	0.00029	0.00039
x	-3.6	-3.52	-3.44	-3.36
y	0.00052	0.00069	0.0009	0.00118
x	-3.28	-3.2	-3.12	-3.04
y	0.00154	0.00199	0.00256	0.00326
x	-2.96	-2.88	-2.8	-2.72

Pictures are indicatives only*

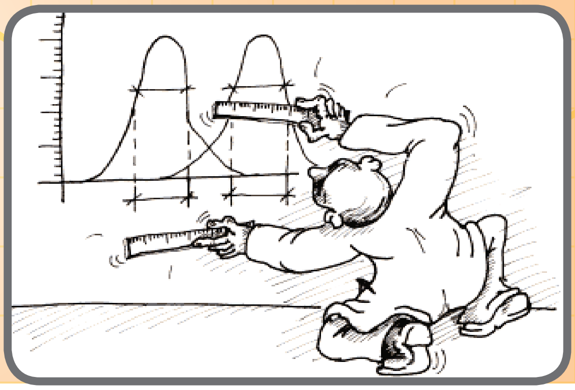
URL:

URL:http://play.google.com/store/apps/details?id=net.eaglepeak.distributions_calculator

https://www.geogebra.org/m/wfencemf

CHAPTER 3

TESTS BASED ON SAMPLING DISTRIBUTIONS – II



R. A. Fisher

Sir Ronald Aylmer Fisher (1890–1962) was a British statistician and geneticist. His work in statistics, made him popularly known as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century Statistics". In Genetics, his work used Mathematics to combine Mendelian Genetics and natural selection and this contributed to the revival of Darwinism in the early 20th century revision of the Theory of Evolution.

"Natural selection is a mechanism for generating an exceedingly high degree of improbability"

"The Best time to plan an experiment is after you have done it"

"The analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic"

LEARNING OBJECTIVES

The students will be able to

- ❖ compare variances of two populations
- ❖ understand the testing of hypothesis for comparing three or more population means.
- ❖ differentiate Treatments and Blocks.
- ❖ differentiate one-way and two-way Analysis of Variance.
- ❖ calculate *F*-ratio for Treatments and Blocks.
- ❖ infer by comparing the estimated and critical values.

WT1LK

Introduction

In the previous chapters, we have discussed various concepts used in testing of hypotheses and problems relating to means of the populations. Although many practical problems involve inferences about population means or proportions, the inference about population variances is important and needs to be studied. In this chapter we will study (i) testing equality of two population variances (ii) one-way ANOVA and (iii) two-way ANOVA, using *F*-distribution.

3.1 *F*-DISTRIBUTION AND ITS APPLICATIONS

F-statistic is the ratio of two sums of the squares of deviations of observations from respective means. The sampling distribution of the statistic is *F*-distribution.

Definition: *F*-Distribution

Let X and Y be two independent χ^2 random variates with m and n degrees of freedom respectively. Then $F = \dfrac{X/m}{Y/n}$ is said to follow *F*-distribution with (m, n) degrees of freedom. This *F*-distribution is named after the famous statistician R.A. Fisher (1890 to 1962).

Definition: *F*-Statistic

Let $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$ be two independent random samples drawn from $N(\mu_X, \sigma_X{}^2)$ and $N(\mu_Y, \sigma_Y{}^2)$ populations respectively.

Then,

$$\frac{1}{\sigma_X{}^2}\sum_{i=1}^{m}\left(X_i-\overline{X}\right)^2 \sim \chi^2_{m-1} \text{ and } \frac{1}{\sigma_Y^2}\sum_{j=1}^{n}\left(Y_j-\overline{Y}\right)^2 \sim \chi^2_{n-1}$$

are independent

(1) Hence, *F*-Statistic is defined as

$$F = \frac{(m-1)S_X^2}{\sigma_X^2} \bigg/ \frac{(n-1)S_Y^2}{\sigma_Y^2} \sim F_{m-1,n-1}$$

where

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_i-\overline{X}\right)^2 \quad \text{and} \quad S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Y_j-\overline{Y}\right)^2$$

CARE

If the populations are not normal, *F* – test may not be used.

Assumptions for testing the ratio of two normal population variances

i) The population from which the samples were obtained must be normally distributed.

ii) The two samples must be independent of each other.

(2) *F*-Statistic is also defined as the ratio of two mean square errors.

Applications of *F*-distribution

The following are some of the important applications where the sampling distribution of the respective statistic under H_0 is *F*–distribution.

(i) Testing the equality of variances of two normal populations. [Using (1)]

(ii) Testing the equality of means of k (>2) normal populations. [Using (2)]

(iii) Carrying out analysis of variance for two-way classified data. [Using (2)]

3.2 TEST OF SIGNIFICANCE FOR TWO NORMAL POPULATION VARIANCES

Test procedure:

This test compares the variances of two independent normal populations, *viz.,* $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$.

Step 1 : **Null Hypothesis** $H_0 : \sigma_X^2 = \sigma_Y^2$

That is, there is no significant difference between the variances of the two normal populations.

The alternative hypothesis can be chosen suitably from any one of the following

(i) $H_1 : \sigma_X^2 < \sigma_Y^2$ (ii) $H_1 : \sigma_X^2 > \sigma_Y^2$ (iii) $H_1 : \sigma_X^2 \neq \sigma_Y^2$

Step 2 : **Data**

Let $X_1, X_2,\ldots, X_m$ and $Y_1, Y_2,\ldots, Y_n$ be two independent samples drawn from two normal populations respectively.

Step 3 : **Level of significance** α

Step 4 : **The test Statistic**

$F = \dfrac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \dfrac{S_1^2}{S_2^2}$ under H_0 and its sampling distribution under H_0 is $F_{(m\text{-}1,\, n\text{-}1)}$.

Step 5 : **Calculation of the Test Statistic**

The test statistic $F_0 = \dfrac{s_X^2}{s_Y^2}$

Step 6 : **Critical values**

H_1	$\sigma_X^2 < \sigma_Y^2$	$\sigma_X^2 > \sigma_Y^2$	$\sigma_X^2 \neq \sigma_Y^2$
Critical value(s) f_e	$f_{(m-1,\, n-1),1-\alpha}$	$f_{(m-1,\, n-1),\, \alpha}$	$f_{(m-1,\, n-1),1-\alpha/2}$ and $f_{(m-1,\, n-1),\, \alpha/2}$

Step 7 : **Decision**

H_1	$\sigma_X^2 < \sigma_Y^2$	$\sigma_X^2 > \sigma_Y^2$	$\sigma_X^2 \neq \sigma_Y^2$
Rejection Rule	$F_0 \leq f_{(m-1,\, n-1),\, 1-\alpha}$	$F_0 \geq f_{(m-1,\, n-1),\, \alpha}$	$F_0 \leq f_{(m-1,\, n-1),\, 1-\alpha/2}$ or $F_0 \geq f_{(m-1,\, n-1),\, \alpha/2}$

Note 1: Since $f_{(m-1,\ n-1),\ 1-\alpha}$ is not avilable in the given F-table, it is computed as the reciprocal of $f_{(n-1,\, m-1),\alpha}$.

i.e., $f_{(m-1,\, n-1),\, 1-\alpha} = \dfrac{1}{f_{(n-1,\, m-1),\, \alpha}}$

Note 2: A F-test is based on the ratio of variances, it is also known as Variance Ratio Test.

Note 3: When μ_X and μ_Y are known, for testing the equality of variances of two normal populations, the test statistic is

$$F = \frac{\frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_X)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_Y)^2}$$ and follows $F_{m,n}$-distribution under H_0

Example 3.1

Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means as 160 inches square and 91 inches square respectively. Test the hypothesis that the variances of the two populations from which the samples are drawn are equal at 10% level of significance.

Solution:

Step 1 : Null Hypothesis: $H_0 : \sigma_X{}^2 = \sigma_Y{}^2$

That is there is no significant difference between the two population variances.

Alternative Hypothesis: $H_1 : \sigma_X{}^2 \neq \sigma_Y{}^2$

That is there is significant difference between the two population variances.

Step 2 : Data

$m = 9, n = 8$

$$\sum_{i=1}^{9}\left(x_i - \bar{x}\right)^2 = 160 \qquad \sum_{j=1}^{8}\left(y_j - \bar{y}\right)^2 = 91$$

Step 3 : Level of significance

$\alpha = 10\%$

Step 4 : Test Statistic $F = \dfrac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \dfrac{S_1^2}{S_2^2}$, under H_0.

Step 5 : Calculation

$$s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2 \text{ and } s_Y^2 = \frac{1}{n\ -1}\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2$$

$$s_X^2 = \frac{160}{8} = 20 \qquad s_Y^2 = \frac{91}{7} = 13$$

$$F_0 = \frac{s_X^2}{s_Y^2} = \frac{20}{13} = 1.54$$

Step 6 : Critical values

Since H_1 is a two-sided alternative hypothesis the corresponding critical values are:

$$f_{(8,\,7),0.05} = 3.73 \text{ and } f_{(8,\,7),0.95} = \frac{1}{f_{(7,8),0.05}} = \frac{1}{3.5} = 0.286$$

Step 7 : Decision

Since $f_{(8,\,7),0.95} = 0.286 < F_0 = 1.54 < f_{(8,\,7),0.05} = 3.73$, the null hypothesis is not rejected and we conclude that there is no significant difference between the two population variances.

Note 4: The critical values of F corresponding to $\alpha = 0.05$ requires table values at 0.025 and 0.975 which are not provided. Hence α is taken as 0.1 in this example.

Example 3.2

A medical researcher claims that the variance of the heart rates (in beats per minute) of smokers is greater than the variance of heart rates of people who do not smoke. Samples from two groups are selected and the data is given below. Using = 0.05, test whether there is enough evidence to support the claim.

Smokers	Non Smokers
$m = 25$	$n = 18$
$s_1^2 = 36$	$s_2^2 = 10$

Solution:

Step 1 : **Null Hypothesis:** $H_0 : \sigma_1{}^2 = \sigma_2{}^2$

That is there is no significant difference between the two population variances.

$H_1 : \sigma_1{}^2 > \sigma_2{}^2$

That is, the variance of heart rates of smokers is greater than that of non-smokers.

Step 2 : **Data**

Smokers	Non Smokers
$m = 25$	$n = 18$
$s_1^2 = 36$	$s_2^2 = 10$

Step 3 : **Level of significance** $\alpha = 5\%$

Step 4 : **Test statistic**

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2}{S_2^2}$$

Step 5 : **Calculation**

$$F_0 = \frac{s_1^2}{s_2^2} = \frac{36}{10} = 3.6$$

Step 6 : **Critical value**

$$f_{(m\text{-}1,n\text{-}1),0.05} = f_{(24,17),0.05} = 2.19$$

Step 7 : **Decision**

Since $F_0 = 3.6 > f_{(24,17),0.05} = 2.19$, the null hypothesis is rejected and we conclude that the variance of heart beats for smokers seems to be considerably higher compared to that of the non-smokers.

Example 3.3

The following table gives the random sample of marks scored by students in two schools, A and B.

School A	63	72	80	60	85	83	70	72	81
School B	86	93	64	82	81	75	86	63	63

Is the variance of the marks of students in school A is less than that of those in school B? Test at 5% level of significance.

Solution:

Let $X_1, X_2, \ldots, X_m$ represent sample values for school A and let $Y_1, Y_2, \ldots, Y_n$ represent sample values for school B.

Step 1 : Null Hypothesis: $H_1 : \sigma_X{}^2 = \sigma_Y{}^2$

That is, there is no significant difference between the two population variances.

Alternative Hypothesis: $H_1 : \sigma_X{}^2 < \sigma_Y{}^2$

That is, the variance of marks in school A is significantly less than that of school B.

Step 2 : Data

$X_1, X_2, \ldots, X_m$ are sample from school A

$Y_1, Y_2, \ldots, Y_n$ are sample from school B

Step 3 : Test statistic

$$F = \frac{s_X^2}{s_Y^2}$$

$$s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2$$

$$s_Y^2 = \frac{1}{n\ -1}\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2$$

Step 4 : Calculations

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
63	-11	121	86	9	81
72	-2	4	93	16	256
80	6	36	64	-13	169
60	-14	196	82	5	25
85	11	121	81	4	16
83	9	81	75	-2	4
70	-4	16	86	9	81
72	-2	4	63	-14	196
81	7	49	63	-14	196
666		628	693		1024

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i}{m} = \frac{666}{9} = 74$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{693}{9} = 77$$

$$s_X^2 = \frac{1}{9-1} \times 628 \quad = \frac{1}{8} \times 628 = 78.5$$

$$s_Y^2 = \frac{1}{9-1} \times 1024 = \frac{1}{8} \times 1024 = 128$$

$$F_0 = \frac{78.5}{128} = 0.613$$

Step 5 : Level of significance

$\alpha = 5\%$

Step 6 : Critical value

$$f_{(9\text{-}1,9\text{-}1),0.95} = \frac{1}{f_{(8,8),0.05}} = \frac{1}{3.44} = 0.291$$

Step 7 : Decision

Since $F_0 = 0.613 > f_{(8,8),0.95} = 0.291$, the null hypothesis is not rejected and we conclude that in school B there seems to be more variance present than in school A.

3.3 ANALYSIS OF VARIANCE (ANOVA)

In chapter 2, testing equality means of two normal populations based on independent small samples was discussed. When the number of populations is more than 2, those methods cannot be applied.

ANOVA is used when we want to test the equality of means of more than two populations. For example, through ANOVA, one may compare the average yield of several varieties of a crop or average mileages of different brands of cars.

ANOVA cannot be used in all situations and for all types of variables. It is based on certain assumptions, and they are listed below:

1. The observations follow normal distribution.
2. The samples are independent.
3. The population variances are equal and unknown.

According to R.A. Fisher ANOVA is the "Separation of variance, ascribable to one group of causes from the variance ascribable to other groups".

The data may be classified with respect to different levels of a single factor/or different levels of two factors.

The former is called one-way classified data and the latter is called two-way classified data. Applications of ANOVA technique to these kinds of data are discussed in the following sections.

3.3.1 One-way ANOVA

ANOVA is a statistical technique used to determine whether differences exist among three or more population means.

In one-way ANOVA the effect of one factor on the mean is tested. It is based on independent random samples drawn from k – different levels of a factor, also called treatments.

The following notations are used in one-way ANOVA. The data can be represented in the following tabular structure.

Data representation for one-way ANOVA

Treatments					Total
Treatment 1	x_{11}	x_{12}	…	x_{1n_1}	$x_{1.}$
Treatment 2	x_{21}	x_{22}	…	x_{2n_2}	$x_{2.}$
⋮	⋮	⋮	…	⋮	⋮
Treatment k	x_{k1}	x_{k2}	…	x_{kn_k}	$x_{k.}$

x_{ij} - the j^{th} sample value from the ith treatment, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$

k - number of treatments compared.

$x_{i.}$ - the sample total of i^{th} treatment.

n_i - the number of observations in the i^{th} treatment.

$$\sum_{i=1}^{k} n_i = n$$

The total variation in the observations x_{ij} can be split into the following two components

i) variation between the levels or the variation due to different bases of classification, commonly known as treatments.

ii) The variation within the treatments *i.e.* inherent variation among the observations within levels.

Causes involved in the first type of variation are known as assignable causes. The causes leading to the second type of variation are known as chance or random causes.

The first type of variation that is due to assignable causes, can be detected and controlled by human endeavor and the second type of variation that is due to chance causes, is beyond the human control.

3.3.2 Test Procedure

Let the observations x_{ij}, $j = 1, 2, \ldots, n_i$ for treatment i, be assumed to come from $N(\mu_i, \sigma^2)$ population, $i = 1, 2, \ldots, k$ where σ^2 is unknown.

Step 1 : **Framing Hypotheses**

Null Hypothesis $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

That is, there is no significant difference among the population means of k treatments.

Alternative Hypothesis

$H_1: \mu_i \neq \mu_j$ for atleast one pair (i,j); $i, j = 1,2,\ldots,k$; $i \neq j$

That is, at least one pair of means differ significantly.

Step 2 : **Data**

Data is presented in the tabular form as described in the previous section

Step 3 : **Level of significance : α**

Step 4 : **Test Statistic**

$F = \dfrac{MST}{MSE}$ which follows $F_{(k-1,\, n-k)}$, under H_0

To evaluate the test statistic we compute the following:

(i) Correction factor: $C.F = \dfrac{G^2}{n}$ where $G = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$

(ii) Total Sum of Squares: $TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - C.F$

(iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^{k}\dfrac{x_{i.}^2}{n_i} - C.F,$

where $x_{i.} = \sum_{j=1}^{n_i} x_{ij},\ i = 1, 2, \ldots, k$

(iv) Sum of Squares due to Error: $SSE = TSS - SST$

Degrees of Freedom (d.f)

Degrees of freedom (d.f.)		d.f.
Total Sum of Squares	Total no. of observations –1	n–1
Treatment Sum of Squares	Total no. of observations –1	k–1
Error of Sum Squares	Total no. of observations –1	n–k

Mean Sum of Squares

Mean Sum of Squares due to treatment: $MST = \dfrac{SST}{k-1}$

Mean Sum of Squares due to Error:

$$MSE = \frac{SSE}{n-k}$$

Step 5 : Calculation of Test statistic

ANOVA Table (one-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	*F*-ratio
Treatments	*SST*	*k*-1	$MST = \frac{SST}{k-1}$	$F_0 = \frac{MST}{MSE}$
Error	*SSE*	*n*-*k*	$MSE = \frac{SSE}{n-k}$	
Total	*TSS*	*n*-1		

Step 6 : Critical value

$f_e = f_{(k\text{-}1,\, n\text{-}k),\alpha}$.

Step 7 : Decision

If $F_0 < f_{(k\text{-}1,\, n\text{-}k),\alpha}$ then reject H_0.

3.3.3 Merits and Demerits of One-Way ANOVA

Merits

- Layout is very simple and easy to understand.
- Gives maximum degrees of freedom for error.

Demerits

- Population variances of experimental units for different treatments need to be equal.
- Verification of normality assumption may be difficult.

Example 3.4

Three different techniques namely medication, exercises and special diet are randomly assigned to (individuals diagnosed with high blood pressure) lower the blood pressure. After four weeks the reduction in each person's blood pressure is recorded. Test at 5% level, whether there is significant difference in mean reduction of blood pressure among the three techniques.

Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	12	8	4

Solution:

Step 1 : Hypotheses

Null Hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3$

That is, there is no significant difference among the three groups on the average reduction in blood pressure.

Alternative Hypothesis: H_1: $\mu_i \neq \mu_j$ for atleast one pair (i, j); $i, j = 1, 2, 3$; $i \neq j$.

That is, there is significant difference in the average reduction in blood pressure in atleast one pair of treatments.

Step 2 : Data

Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	12	8	4

Step 3 : Level of significance $\alpha = 0.05$

Step 4 : Test statistic

$$F_0 = \frac{MST}{MSE}$$

Step 5 : Calculation of Test statistic

						Total	Square
Medication	10	12	9	15	13	59	3481
Exercise	6	8	3	0	2	19	361
Diet	5	9	12	8	4	38	1444
						G = 116	5286

Individual squares

Medication	100	144	81	225	169
Exercise	36	64	9	0	4
Diet	25	81	144	64	16

$$\sum\sum x_{ij}^2 = 1162$$

1. Correction Factor: $CF = \frac{G^2}{n} = \frac{(116)^2}{15} = \frac{13456}{15} = 897.06$

2. Total Sum of Squares: $TSS = = \sum\sum x_{ij}^2 - C.F$

$= 1162 - 897.06 = 264.94$

3. Sum of Squares between Treatments: $SST = \frac{\sum x_i^2}{n_i} - C.F$

$$= \frac{5286}{5} - 897.06$$

$$= 1057.2 - 897.06$$

$$= 160.14$$

4. Sum of Squares due to Error: $SSE = TSS - SST$

$= 264.94 - 160.14 = 104.8$

ANOVA Table (one-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	*F*-ratio
Between treatments	160.14	3 – 1 = 2	80.07	$F_o = \frac{80.07}{8.73} = 9.17$
Error	104.8	12	8.73	
Total	264.94	n – 1 = 15 – 1 = 14		

Step 6 : **Critical value**

$f_{(2,\ 12),0.05} = 3.8853.$

Step 7 : **Decision**

As $F_0 = 9.17 > f_{(2,\ 12),0.05} = 3.8853$, the null hypothesis is rejected. Hence, we conclude that there exists significant difference in the reduction of the average blood pressure in atleast one pair of techniques.

Example 3.5

Three composition instructors recorded the number of spelling errors which their students made on a research paper. At 5% level of significance test whether there is significant difference in the average number of errors in the three classes of students.

Instructor 1	2	3	5	0	8		
Instructor 2	4	6	8	4	9	0	2
Instructor 3	5	2	3	2	3	3	

Solution:

Step 1 : **Hypotheses**

Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$

That is there is no significant difference among the mean number of errors in the three classes of students.

Alternative Hypothesis

$H_1 : \mu_i \neq \mu_j$ for at one pair (i, j); $i,j = 1,2,3;\ i \neq j$.

That is, atleast one pair of groups differ significantly on the mean number of errors.

Step 2 : Data

Instructor 1	2	3	5	0	8		
Instructor 2	4	6	8	4	9	0	2
Instructor 3	5	2	3	2	3	3	

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

$$F_0 = \frac{MST}{MSE}$$

Step 5 : Calculation of Test statistic

								Total	Square
Instructor 1	2	3	5	0	8			18	324
Instructor 2	4	6	8	4	9	0	2	33	1089
Instructor 3	5	2	3	2	3	3		18	324
								69	

Individual squares

Instructor 1	4	9	25	0	64		
Instructor 2	16	36	64	16	81	0	4
Instructor 3	25	4	9	4	9	9	

$$\sum\sum x_{ij}^2 = 379$$

Correction Factor: $CF = \frac{G^2}{n} = \frac{(69)^2}{18} = \frac{4761}{18} = 264.5$

Total Sum of Squares: $TSS = \sum\sum x_{ij}^2 - C.F$

$$= 379 - 264.5 = 114.5$$

Sum of Squares between Treatments: $SST = \frac{\sum x_i^2}{n_i} - C.F$

$$= \left(\frac{324}{5} + \frac{1089}{7} + \frac{324}{6}\right) - 264.5$$

$$= (64.8 + 155.6 + 54) - 264.5$$

$$= (274.4) - 264.5$$

$$= 9.9$$

Sum of Squares due to Error: $SSE = TSS - SST$

$$= 114.5 - 9.9$$

$$= 104.6$$

ANOVA Table

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	*F*-ratio
Between treatments	9.9	3 – 1 = 2	$\frac{9.9}{2} = 4.95$	$F_0 = \frac{4.95}{6.97} = 0.710$
Error	104.6	15	$\frac{104.6}{15} = 6.97$	
Total		n – 1 = 18 – 1 = 17		

Step 6 : **Critical value**

The critical value = $f_{(15,\ 2),0.05}$ = 3.6823.

Step 7 : **Decision**

As F_0 = 0.710 < $f_{(15,\ 2),0.05}$ = 3.6823, null hypothesis is not rejected. There is no enough evidence to reject the null hypothesis and hence we conclude that the mean number of errors made by these three classes of students are not equal.

3.4 TWO-WAY ANOVA

In two-way ANOVA a study variable is compared over three or more groups, controlling for another variable. The grouping is taken as one factor and the control is taken as another factor. The grouping factor is usually known as Treatment. The control factor is usually called Block. The accuracy of the test in two-way ANOVA is considerably higher than that of the one-way ANOVA, as the additional factor, block is used to reduce the error variance.

In two-way ANOVA, the data can be represented in the following tabular form.

Blocks

		1	2	3	…	m	$x_{i.}$
Groups or Treatments	1	x_{11}	x_{12}	$x_{13.}$	…	$x_{1\mathrm{m}}$	$x_{1.}$
	2	x_{21}	x_{22}	$x_{2.}$	…	x_{2m}	$x_{2.}$
	3	x_{31}	x_{32}	$x_{3.}$	…	$\mathrm{x}_{3\mathrm{m}}$	$x_{3.}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	k	x_{k1}	x_{k2}	x_{k3}	…	x_{km}	$x_{k.}$
	$x._j$	$x_{.1}$	$x_{.2}$	$x_{.3}$	…	$x_{.m}$	G

We use the following notations.

x_{ij} - i^{th} treatment value from the j^{th} block, i = 1,2, ..., k; j = 1,2, ..., m.

The i^{th} treatment total - $x_{i.} = \sum_{j=1}^{m} x_{ij},\ i = 1,\ 2,\ ...,\ k$

The j^{th} block total - $x_{.j} = \sum_{i=1}^{k} x_{ij}, \ j = 1, 2, ..., m$

Note that, $k \times m = n$, where m = number of blocks, and k = number of treatments (groups) and n is the total number of observations in the study.

The total variation present in the observations x_{ij} can be split into the following three components:

i) The variation between treatments (groups)

ii) The variation between blocks.

iii) The variation inherent within a particular setting or combination of treatment and block.

3.4.1 Test Procedure

Steps involved in two-way ANOVA are:

Step 1 : In two-way ANOVA we have two pairs of hypotheses, one for treatments and one for the blocks.

Framing Hypotheses

Null Hypotheses

H_{01}: There is no significant difference among the population means of different groups (Treatments)

H_{02}: There is no significant difference among the population means of different Blocks

Alternative Hypotheses

H_{11}: Atleast one pair of treatment means differs significantly

H_{12}: Atleast one pair of block means differs significantly

Step 2 : **Data** is presented in a rectangular table form as described in the previous section.

Step 3 : **Level of significance** α.

Step 4 : **Test Statistic**

$$F_{0t}(\text{treatments}) = \frac{MST}{MSE}$$

$$F_{0b}(\text{block}) = \frac{MSB}{MSE}$$

To find the test statistic we have to find the following intermediate values.

i) Correction Factor: $C.F = \dfrac{G^2}{n}$ where $G = \sum_{j=1}^{m}\sum_{i=1}^{k} x_{ij}$

ii) Total Sum of Squares: $TSS = \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}^2 - C.F$

iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^{k} \dfrac{x_{i.}^2}{m} - C.F$

iv) Sum of squares between blocks: $SSB = \sum_{j=1}^{m} \frac{x_{.j}^2}{k} - C.F$

v) Sum of Squares due to Error: $SSE = TSS\text{-}SST\text{-}SSB$

vi) Degrees of freedom

Degrees of freedom (d.f.)	d.f.
Total Sum of Squares	$n-1$
Treatment Sum of Squares	$k-1$
Block Sum of Squares	$m-1$
Error of Sum Squares	$(m-1)(k-1)$

vii) Mean Sum of Squares

Mean sum of Squares due to Treatments: $MST = \frac{SST}{k-1}$

Mean sum of Squares due to Blocks: $MSB = \frac{SSB}{m-1}$

Mean sum of Squares due to Error: $MSE = \frac{SSE}{(k-1)(m-1)}$

Step 5 : Calculation of the Test Statistic

ANOVA Table (two-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	*F*-ratio
Treatments	SST	k-1	MST	$F_{0t} = \frac{MST}{MSE}$
Blocks	SSB	m-1	MSB	$F_{0b} = \frac{MSB}{MSE}$
Error	SSE	(k-1)(m-1)	MSE	
Total	TSS	n-1		

Step 6 : Critical values

Critical value for treatments $= f_{(k\text{-}1,(m\text{-}1)(k\text{-}1)),\alpha}$

Critical value for blocks $= f_{(m\text{-}1,\,(m\text{-}1)(k\text{-}1)),\alpha}$

Step 7 : Decision

For Treatments: If the calculated F_{0t} value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the treatment means, in atleast one pair.

For Blocks: If the calculated F_{0b} value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the block means, in at least one pair.

3.4.2 Merits and Demerits of two-way ANOVA

Merits

- Any number of blocks and treatments can be used.
- Number of units in each block should be equal.
- It is the most used design in view of the smaller total sample size since we are studying two variable at a time.

Demerits

- If the number of treatments is large enough, then it becomes difficult to maintain the homogeneity of the blocks.
- If there is a missing value, it cannot be ignored. It has to be replaced with some function of the existing values and certain adjustments have to be made in the analysis. This makes the analysis slightly complex.

Comparison between one-way ANOVA and two-way ANOVA

Basis of comparison	ANOVA	
	One-way	Two-way
Independent variable	One	Two
Compares	Three or more levels of one factor	Three or more levels of two factors, simultaneously
Number of observations	Need not be same in each treatment group	Need to be equal in each treatment group

Example 3.6

A reputed marketing agency in India has three different training programs for its salesmen. The three programs are Method – A, B, C. To assess the success of the programs, 4 salesmen from each of the programs were sent to the field. Their performances in terms of sales are given in the following table.

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Test whether there is significant difference among methods and among salesmen.

Solution:

Step 1 : Hypotheses

Null Hypotheses: $H_{01} : \mu_{M_1} = \mu_{M_2} = \mu_{M_3}$ (for treatments)

That is, there is no significant difference among the three programs in their mean sales.

$$H_{02} : \mu_{S_1} = \mu_{S_2} = \mu_{S_3} = \mu_{S_4} \text{ (for blocks)}$$

Alternative Hypotheses:

H_{11} : At least one average is different from the other, among the three programs.

H_{12} : At least one average is different from the other, among the four salesmen.

Step 2 : Data

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

$$F_{0t}(\text{treatment}) = \frac{MST}{MSE}$$

$$F_{0b}(\text{block}) = \frac{MSB}{MSE}$$

Step-5 : Calculation of the Test Statistic

	Methods			Total $x_{i.}$	$x_{i.}{}^2$
	A	B	C		
1	4	6	2	12	144
2	6	10	6	22	484
3	5	7	4	16	256
4	7	5	4	16	256
x_i	22	28	16	66	1140
$x_{i.}{}^2$	484	784	256	1524	

Squares

16	36	4
36	100	36
25	49	16
49	25	16
		$\sum\sum x_{ij}^2 = 408$

Correction Factor: $CF = \frac{G^2}{n} = \frac{(66)^2}{12} = \frac{4356}{12} = 363$

Total Sum of Squares: $TSS = \sum\sum x_{ij}^2 - C.F$

$= 408 - 363 = 45$

Sum of Squares due to Treatments: $SST = \frac{\sum_{i=1}^{k} x_{.j}^2}{k} - C.F$

$= \frac{1140}{3} - 363$

$= 380 - 363 = 17$

Sum of Squares due to Blocks: $SSB = \frac{\sum_{i=1}^{k} x_{.j}^2}{k} - C.F$

$= \frac{1524}{4} - 363$

$= 381 - 363$

$= 18$

Sum of Squares due to Error: $SSE = TSS - SST - SSB$

$= 45 - 17 - 18 = 10$

Mean sum of squares: $MST = \frac{SST}{k-1} = \frac{17}{2} = 8.5$

$MSB = \frac{SSB}{m-1} = \frac{18}{3} = 6$

$MSE = \frac{SSE}{(k-1)(m-1)} = \frac{10}{6} = 1.67$

ANOVA Table (two-way)

Sources of variation	Sum of squares	Degrees of freedom	Mean sum of squares	*F*-ratio
Between treatments (Programs)	17	2	8.5	$F_{ot} = \frac{8.5}{1.67} = 5.09$
Between blocks (Salesmen)	18	3	6	$F_{ob} = \frac{6}{1.67} = 3.59$
Error	10	6	1.67	
Total		11		

Step 6 : Critical values

(i) $f_{(2,\,6),0.05} = 5.1456$ (for treatments)

(ii) $f_{(3,\,6),0.05} = 4.7571$ (for blocks)

Step 7 : Decision

i) Calculated $F_{0t} = 5.09 < f_{(2,\ 6),0.05} = 5.1456$, the null hypothesis is not rejected and we conclude that there is significant difference in the mean sales among the three programs.

ii) Calculate $F_{0b} = 3.59 > f_{(3,\ 6),0.05} = 4.7571$, the null hypothesis is rejected and conclude that there does not exist significant difference in the mean sales among the four salesmen.

Example 3.7

The illness caused by a virus in a city concerning some restaurant inspectors is not consistent with their evaluations of cleanliness of restaurants. In order to investigate this possibility, the director has five restaurant inspectors to grade the cleanliness of three restaurants. The results are shown below.

Inspectors	Restaurants		
	I	II	III
1	71	55	84
2	65	57	86
3	70	65	77
4	72	69	70
5	76	64	85

Carry out two-way ANOVA at 5% level of significance.

Solution:

Step 1 :

Null hypotheses

$H_{0I}: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (For inspectors - Treatments)

That is, there is no significant difference among the five inspectors over their mean cleanliness scores

$H_{0R}: \mu_{\text{I}} = \mu_{\text{II}} = \mu_{\text{III}}$ (For restaurants - Blocks)

That is, there is no significant difference among the three restaurants over their mean cleanliness scores

Alternative Hypotheses

H_{1I}: At least one mean is different from the other among the Inspectors

H_{1R}: At least one mean is different from the other among the Restaurants.

Step 2 : Data

Inspectors	Restaurants		
	I	II	III
1	71	55	84
2	65	57	86
3	70	65	77
4	72	69	70
5	76	64	85

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

For inspectors: F_{0a} (treatments) = $= \dfrac{MST}{MSE}$

For restaurants: F_{0b} (blocks) $) = \dfrac{MSB}{MSE}$

Step-5 : Calculation of the Test Statistic

Inspectors	Restaurants			Total $x_{i.}$	$x_{i.}^2$
	I	II	III		
1	71	55	84	210	44100
2	65	57	86	208	43264
3	70	65	77	212	44944
4	72	69	70	211	44521
5	76	64	85	225	50625
$x_{.j}$	354	310	402	1066	
$x_{.j}^2$	125316	96100	161604		

Squares

5041	3025	7056
4225	3249	7396
4900	4225	5929
5184	4761	4900
5776	4096	7225
		$\sum\sum x_{ij}^2 = 76988$

Correction Factor: $$CF = \frac{G^2}{n} = \frac{(1066)^2}{15} = \frac{1136356}{15} = 75757.07$$

Total Sum of Squares: $TSS = \sum\sum x_{ij}^2 - C.F$

$= 76988 - 75757.07 = 1230.93$

Sum of Squares due to Treatments: $SST = \dfrac{\sum_{j=1}^{l} x_{i.}^2}{l} - C.F$

$= \dfrac{227454}{3} - 75757.07$

$= 75818 - 75757.07$

$= 60.93$

Sum of Squares due to Blocks: $SSB = \dfrac{\sum_{i=1}^{k} x_{.j}^2}{k} - C.F$

$= \dfrac{383020}{5} - 75757.07$

$= 76604 - 75757.07$

$= 846.93$

Sum of squares due to error: SSE $= TSS - SST - SSB$

$= 1230.93 - 60.93 - 846.93$

$= 323.07$

ANOVA Table (two-way)

Sources of variation	Sum of squares	Degrees of freedom	Mean sum of squares	*F*-ratio
Between inspectors	60.93	4	15.23	$F_{0I} = \dfrac{15.23}{40.38} = 0.377$
Between restaurants	846.93	2	423.47	$F_{0R} = \dfrac{423.47}{40.38} = 10.49$
Error	323.07	8	40.38	
Total	1230.93	14		

Step 6 : **Critical values**

(i) $f_{(4,\,8),0.05} = 3.838$ (for inspectors)

(ii) $f_{(2,\,8),0.05} = 4.459$ (for restaurants)

Step 7 : **Decision**

i) As $F_{0I} = 0.377 < f_{(4,\,8),0.05} = 3.838$, the null hypothesis is not rejected and we conclude that there is no significant difference among the mean cleanliness scores of inspectors.

ii) As $F_{0R} = 10.49 > f_{(2,\,8),0.05} = 4.459$, the null hypothesis is rejected and we conclude that there exists significant difference in atleast one pair of restaurants over their mean cleanliness scores.

POINTS TO REMEMBER

- ❖ *F*-statistic is the ratio of two independent sample variances
- ❖ If *X* and *Y* are two independent χ^2 variates with *m* and *n* degrees of freedom respectively, then $F = \frac{X/m}{Y/n}$ is said to follow *F* distribution with (*m*, *n*) degrees of freedom.
- ❖ Two independent random samples of size *m* and *n* are taken from Normal populations. Then the statistic $F = \frac{s_X^2}{s_Y^2}$ is a random variable following the *F*-distribution with *m*–1 and *n*–1 degrees of freedom.
- ❖ According to R.A. Fisher, ANOVA is the "Separation of variance, ascribable to one group of causes from the variance ascribable to other groups".
- ❖ One-way ANOVA is used to compare means in more than two groups.
- ❖ Two-way ANOVA is used to compare means in more than two groups, controlling another variable.
- ❖ Assumptions required for ANOVA are:
  - The observations follow normal distribution.
  - Experimental units assigned to treatments are random.
  - The sample observations are independent.
  - The population variances of the groups are unknown but are assumed to be equal.

EXERCISE 3

I. Choose the best Answer.

1. ANOVA was developed by

(a) S.D. Poisson (b) Karl Pearson

(c) R.A. Fisher (d) W.S. Gosset

2. ANOVA technique originated in the field of

(a) Industry (b) Agriculture

(c) Medicine (d) Genetics

3. One of the assumptions of ANOVA is that the population from which the samples are drawn is

(a) Binomial (b) Poisson (c) Chi-square (d) Normal

4. In one-way classification the total variation can be split into

(a) Two components (b) Three components

(c) Four components (d) Only one components

5. The null hypothesis in the ANOVA is that all the population means are

   (a) Equal (b) Variable

   (c) Unequal (d) none of the above

6. In one-way classification with 30 observation and 5 treatments the degrees of freedom for error is

   (a) 29 (b) 4 (c) 25 (d) 150

7. In two-way classification the total variation *TSS* is

   (a) *SST+SSB+SSE* (b) *SST-SSB+SSE*

   (c) *SST+SSB-SSE* (d) *SST+SSB*

8. In two-way classification if *TSS* = 210, *SST* = 32, *SSB* = 42 then *SSE* =

   (a) 126 (b) 74 (c) 136 (d) 178

9. In two-way classification with 5 treatments and 4 blocks the degrees of freedom due to error is

   (a) 12 (b) 19 (c) 16 (d) 15

10. The formula for comparing three or more means in one-way analysis of variance is

   (a) $F = \frac{MST}{MSE}$ (b) $F = \frac{TSS}{SST}$

   (c) $F = \frac{MSB}{MST}$ (d) $F = \frac{MST}{MSB}$

11. _______ test is used to compare three or more means.

   (a) *t* (b) χ^2 (c) *F* (d) *Z*

12. When there is no significant difference among three or more means the value of *F* will be close to

   (a) 0 (b) -1 (c) 1 (d) ∞

13. F-test is also called as

   (a) mean ratio test (b) variance ratio test

   (c) variance test (d) standard deviation ratio test

14. The Analysis of Variance procedure is appropriate for testing the equivalence of three or more population

   (a) variances (b) proportions

   (c) means (d) observations

15. In two-way classification with 'm' treatments and 'n' blocks the degrees of freedom due to error is

(a) mn-1 (b) m-1 (c) n-1 (d) $(m$-$1)(n$-$1)$

16. If the calculated value of F is greater than the critical value at the given level of significance then the H_0 is

(a) Rejected (b) Not rejected

(c) Always true (d) Sometimes true

17. ______ and ______ causes are present in Analysis of Variance techniques

(a) Chance, error (b) Fixed, block

(c) Assignable, chance (d) Assignable, fixed

18. In ANOVA, the sample observations are

(a) dependent (b) independent

(c) equal (d) unequal

19. The correction factor is ______ in ANOVA (with the usual notations).

(a) $\dfrac{\sum T_{ij}^2}{n}$ (b) $\dfrac{\sum T_{i.}^2}{n}$

(c) $\dfrac{G^2}{n}$ (d) $\dfrac{\sum T_{i.}}{n}$

20. Mean Sum of Squares is the ratio of Sum of Squarea to its

(a) number of blocks (b) number of treatments

(c) degrees of freedom (d) total sum of squares

II. Give very short answers to the following questions.

21. What is Analysis of Variance?

22. Write the applications of F-statistic.

23. What are the assumptions of ANOVA?

24. Define: Between group variance and within group variance.

25. State the hypotheses used in one-way ANOVA.

26. What are the components in a two-way ANOVA?

27. Name the causes of variation?

III. Give short answer to the following questions.

28. What are the merits and demerits of one-way classification?

29. Write the model ANOVA table for one-way classification.

30. What are the values to be found for finding the test statistic in one-way classification?

31. What are the merits and demerits of two-way classification?

32. Write the model ANOVA table for two-way classification.

33. What are the components in two-way ANOVA?

34. What are the values to be found for finding the test statistic in two way classification?

35. Compare one-way and two-way ANOVA.

IV. Give detailed answer to the following questions.

36. In a sample of 8 observations, the sum of the squares of deviations of the sample values from its sample mean was 84.4. In another sample of 10 observations it was 102.6. Test whether the two population variances are equal at 5% level.

37. Two random samples gave the following results

Sample	Size	Sample mean	Sum of squares of deviations from the mean
I	10	15	90
II	12	14	108

Test whether the populations have same variances at 5% level of significance.

38. The following data refer to the yield of wheat in quintals on plots of equal area in two agricultural blocks A and B.

	Number of plots	Mean yield	Sample variance
Block A	8	60	50
Block B	6	51	40

Is the variance of yield for block A is greater than that of block B at 5% level of significance.

39. The calories contained in 1/2 cup servings of ice-creams selected randomly from two national brands are listed here. At 5% level of significance, is there sufficient evidence to conclude that the variance of calories is less for brand A than brand B?

Brand A	330	310	300	310	300	350	380	300	300
Brand B	300	300	270	290	310	370	300	310	250

40. The carbohydrates contained in servings of some randomly selected chocolate and non-chocolate candies are listed below. Is there sufficient evidence to conclude that the variance in carbohydrates varies between chocolate and non-chocolate candies? Use = 2%.

Chocolate	29	25	18	40	41	25	32	30	38	34	25	28
Non-chocolate	39	39	37	29	30	38	39	10	29	55	29	

41. A home gardener wishes to determine the effects of four fertilizers on the average number of tomatoes produced. Test at 5% level of significance the hypothesis that the fertilizers A, B, C and D have equal average yields.

A	14	10	12	16	17
B	9	11	12	8	10
C	16	15	14	10	18
D	10	11	11	13	8

42. Three processes *X*, *Y* and *Z* are tested to see whether their outputs are equivalent. The following observations on outputs were made.

X	10	13	12	11	10	14	15	13
Y	9	11	10	12	13			
Z	11	10	15	14	12	13		

Carry out the one-way analysis of variance and state your conclusion.

43. A test was given to five students taken at random from XII class of three schools of a town. The individual scores are

School I	9	7	6	5	8
School II	7	4	5	4	5
School III	6	5	6	7	6

Carry out the one-way ANOVA.

44. A farmer applies three types of fertilizers on four separate plots. The figures on yield per acre are tabulated below.

Fertilizer	Plots			
	A	B	C	D
Nitrogen	6	4	8	6
Potash	7	6	6	9
Phosphate	8	5	10	9

Test whether there is any significant difference among mean yields of different plots and among different fertilizers.

45. Operators are tested for their efficiency in terms of number of units produced per day by five different types of machines. Test at 5% level of significance whether the operators and machines differ in terms of their efficiency?

Operators	Machine types				
	A	B	C	D	E
I	8	10	7	12	6
II	12	13	8	9	12
III	7	8	6	8	8
IV	5	5	3	5	14

ANSWERS

I.

1. (c)	2. (b)	3. (d)	4. (a)	5. (a)
6. (c)	7. (a)	8. (c)	9. (a)	10. (a)
11. (c)	12. (c)	13. (b)	14. (c)	15. (d)
16. (a)	17. (c)	18. (b)	19. (c)	20. (c)

III. 36. $F_0 = 1.06$, H_0 is not rejected

37. $F_0 = 1.02$, H_0 is not rejected

38. $F_0 = 1.25$, H_0 is not rejected

39. $F_0 = 1.34$, H_0 is not rejected

40. $F_0 = 2.52$, H_0 is not rejected

41. $F_0 = 4.59$, H_0 is rejected

42. $F_0 = 1.097$, H_0 is not rejected

43. $F_0 = 3.33$, H_0 is not rejected

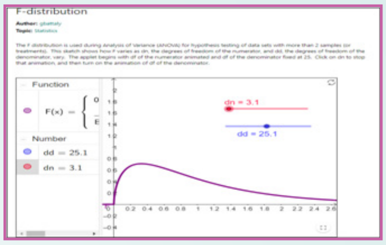
44. $F_0 =2.39$, H_0 is not rejected $F_0 = 3.59$, H_0 is not rejected

45. $F_0 = 2.53$, H_0 is not rejected $F_0 = 1.24$, H_0 is not rejected

ICT CORNER

TESTS BASED ON SAMPLING DISTRIBUTIONS – II

This activity helps to understand about *F*-DISTRIBUTION



F-distribution

Author: gbattaly

Topic: Statistics

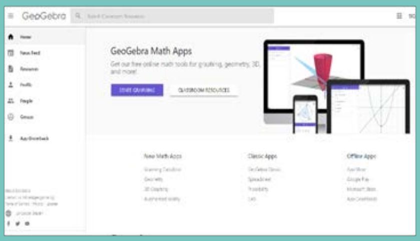
The F distribution is used during Analysis of Variance (ANOVA) for hypothesis testing of data sets with more than 2 samples (or treatments). This sketch shows how F varies as dn, the degrees of freedom of the numerator, and dd, the degrees of freedom of the denominator, vary. The applet begins with df of the numerator animated and df of the denominator fixed at 25. Click on dn to stop that animation, and then turn on the animation of df of the denominator.

Steps:

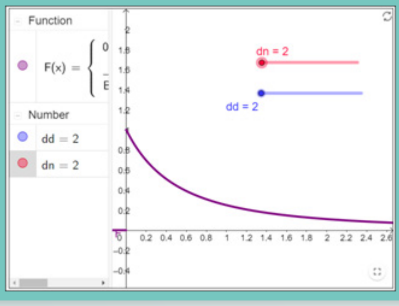
- Open the browser and type the URL given (or) scan the QR code. GeoGebra work book called "***F*-Distribution**" will appear.
- In this several work sheets for statisticsare given, open the worksheet named "*F*-Distribution"
- Drag and move the Red colour and Blue colour button or type the values in the left side box for result

Step-1

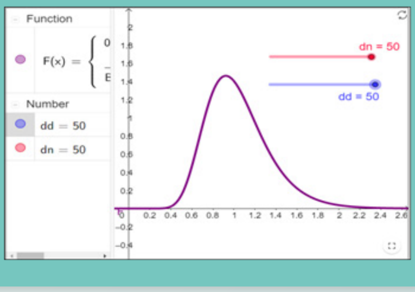
Step-2



Step-3



Step-4



Pictures are indicatives only*

URL:

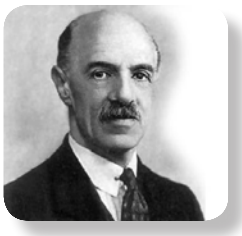
https://www.geogebra.org/m/A45YdMfJ

CHAPTER 4

CORRELATION ANALYSIS

Karl Pearson

Karl Pearson (1857-1936) was a English Mathematician and Biostatistician. He founded the world's first university statistics department at University College, London in 1911. The linear correlation coefficient is also called Pearson product moment correlation coefficent. It was developed by Karl Pearson with a related idea by Francis Galton (see Regression analysis - for Galton's contribution). It is the first of the correlation measures developed and commonly used.



Charles Spearman

Charles Edward Spearman (1863-1945) was an English psychologist and ,after serving 15 years in Army he joined to study PhD in Experimental Psychology and obtained his degree in 1906. Spearman was strongly influenced by the work of Galton and developed rank correlation in 1904.He also pioneered factor analysis in statistics.

> "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering the existence of relation and measuring the intensity of relationship is known as correlation"
>
> —*CROXTON AND COWDEN*

LEARNING OBJECTIVES

The student will be able to

- learn the meaning, definition and the uses of correlation.
- identify the types of correlation.
- understand correlation coefficient for different types of measurement scales.
- differentiate different types of correlation using scatter diagram.
- calculate Karl Pearson's coefficient of correlation, Spearman's rank correlation coefficient and Yule's coefficient of association.
- interpret the given data with the help of coefficient of correlation.

2BGIT

Introduction

"Figure as far as you can, then add judgment"

The statistical techniques discussed so far are for ***only one variable***. In many research situations one has to consider two variables simultaneously to know whether these ***two variables*** are related linearly. If so, what type of relationship that exists between them. This leads to bivariate (two variables) data analysis namely correlation analysis. If two quantities vary in such a way that movements (upward or downward) in one are accompanied by the movements(upward or downward) in the other, these quantities are said to be co-related or correlated.

The correlation concept will help to answer the following types of questions.

- Whether study time in hours is related with marks scored in the examination?
- Is it worth spending on advertisement for the promotion of sales?
- Whether a woman's age and her systolic blood pressure are related?
- Is age of husband and age of wife related?
- Whether price of a commodity and demand related?
- Is there any relationship between rainfall and production of rice?

4.1 DEFINITION OF CORRELATION

Correlation is a statistical measure which helps in analyzing the interdependence of two or more variables. In this chapter the dependence between only two variables are considered.

1. **A.M. Tuttle** defines correlation as:

   "An analysis of the co-variation of two or more variables is usually called correlation"

2. **Ya-kun-chou** defines correlation as:

   "The attempts to determine the degree of relationship between variables".

Correlation analysis is the process of studying the strength of the relationship between two related variables. High correlation means that variables have a strong linear relationship with each other while a low correlation means that the variables are hardly related. The type and intensity of correlation is measured through the correlation analysis. The measure of correlation is the correlation coefficient or correlation index. It is an absolute measure.

Uses of correlation

- Investigates the type and strength of the relationship that exists between the two variables.
- Progressive development in the methods of science and philosophy has been characterized by the rich knowledge of relationship.

4.2 TYPES OF CORRELATION

1. ***Simple (Linear) correlation*** (2 variables only) : The correlation between the given two variables. It is denoted by r_{xy}
2. ***Partial correlation (more than 2 variables):*** The correlation between any two variables while removing the effect of other variables. It is denoted by $r_{xy.z\,...}$

3. ***Multiple correlation (more than 2 variables) :*** The correlation between a group of variables and a variable which is not included in that group. It is denoted by $R_{y.(xz...)}$

In this chapter, we study simple correlation only, multiple correlation and partial correlation involving three or more variables will be studied in higher classes .

4.2.1 Simple correlation or Linear correlation

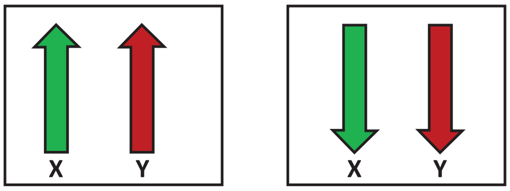
Here, we are dealing with data involving two related variables and generally we assign a symbol 'x' to scores of one variable and symbol 'y' to scores of the other variable. There are five types in simple correlation. They are

1. Positive correlation (Direct correlation)
2. Negative correlation (Inverse correlation)
3. Uncorrelated
4. Perfect positive correlation
5. Perfect negative correlation

1) Positive correlation: (Direct correlation)

The variables are said to be positively correlated if larger values of x are associated with larger values of y and smaller values of x are associated with smaller values of y. In other words, if both the variables are varying in the same direction then the correlation is said to be positive. In other words, if one variable increases, the other variable (on an average) also increases or if one variable decreases, the other (on an average)variable also decreases.

Positive or Direct Correlation

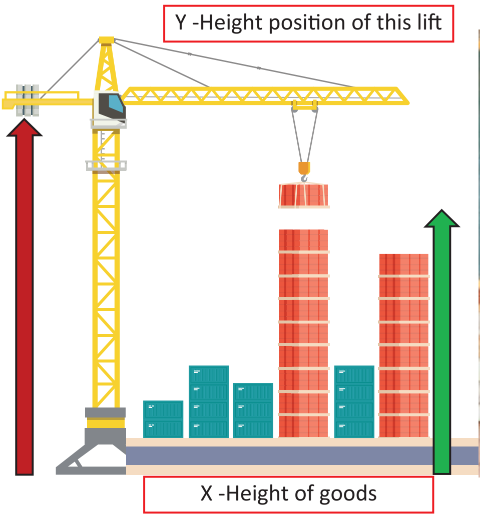


Things move in the same direction

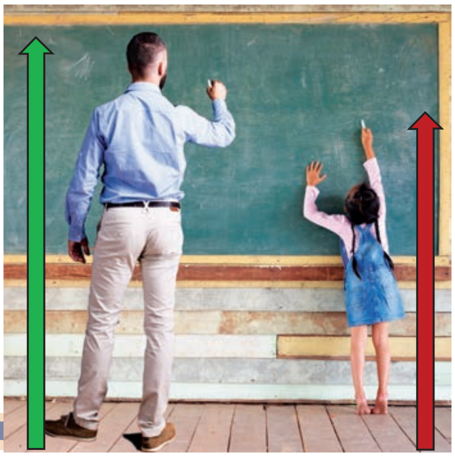
For example,

i) Income and savings

ii) Marks in Mathematics and Marks in Statistics. (*i.e.*,Direct relationship pattern exists).



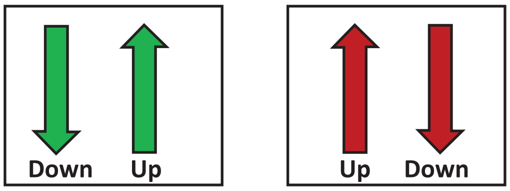
Height of the Lift increases / decreases according to the Height of goods increases / decreases.



The starting position of writing depends on the height of the writer.

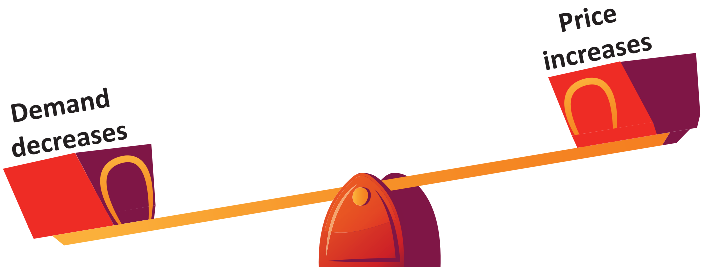
2) Negative correlation: (Inverse correlation)

Negative or Inverse relationship



Things move in opposite direction

The variables are said to be negatively correlated if smaller values of *x* are associated with larger values of *y* or larger values *x* are associated with smaller values of *y*. That is the variables varying in the ***opposite directions*** is said to be negatively correlated. In other words, if one variable increases the other variable decreases and vice versa.



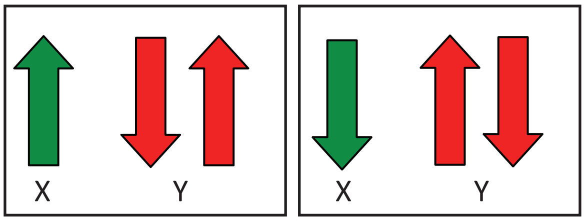
For example,

i) Price and demand

ii) Unemployment and purchasing power

3) Uncorrelated:

The variables are said to be uncorrelated if smaller values of *x* are associated with smaller or larger values of *y* and larger values of *x* are associated with larger or smaller values of *y*. If the two variables do not associate linearly, they are said to be uncorrelated. Here $r = 0$.



Important note: Uncorrelated does not imply independence. This means “do not interpret as the two variables are independent instead interpret as there is no specific linear pattern exists but there may be non linear relationship”.

4) Perfect Positive Correlation

If the values of *x* and *y* increase or decrease ***proportionately*** then they are said to have perfect positive correlation.

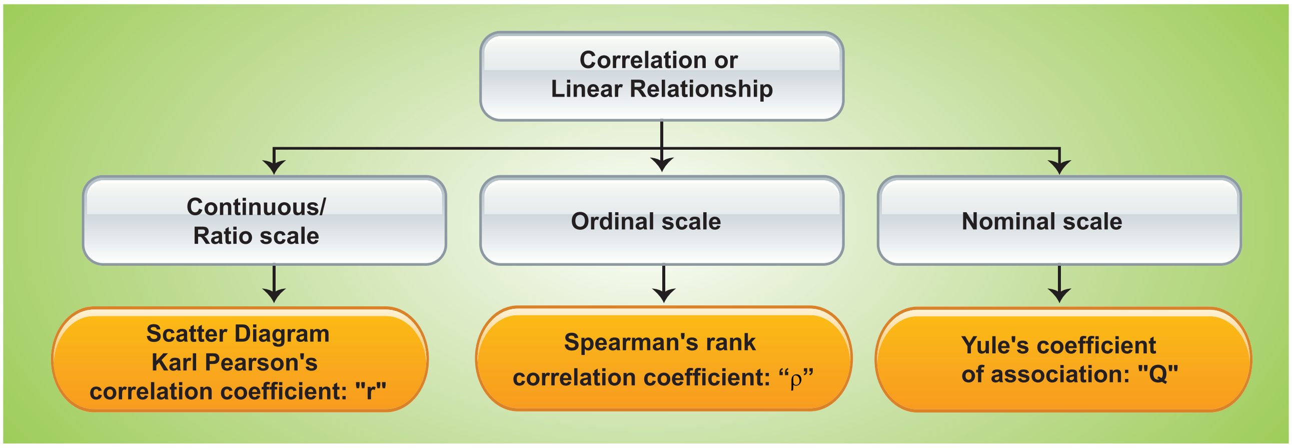
5) Perfect Negative Correlation

If *x* increases and *y* decreases ***proportionately*** or if *x* decreases and *y* increases ***proportionately***, then they are said to have perfect negative correlation.

Correlation Analysis

The purpose of correlation analysis is to find the existence of linear relationship between the variables. However, the method of calculating correlation coefficient depends on the types of measurement scale, namely, ratio scale or ordinal scale or nominal scale.

Statistical tool selection



Methods to find correlation

1. Scatter diagram
2. Karl Pearson's product moment correlation coefficient : '*r*'
3. Spearman's Rank correlation coefficient: 'ρ'
4. Yule's coefficient of Association: '*Q*'

NOTE

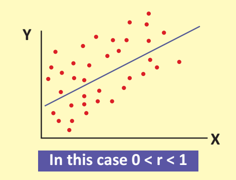
For higher order dimension of nominal or categorical variables in a contingency table, use chi-square test for independence of attributes. (Refer Chapter 2)

4.3 SCATTER DIAGRAM

A scatter diagram is the simplest way of the diagrammatic representation of bivariate data. One variable is represented along the *X*-axis and the other variable is represented along the *Y*-axis. The pair of points are plotted on the two dimensional graph. The diagram of points so obtained is known as scatter diagram. The direction of flow of points shows the type of correlation that exists between the two given variables.

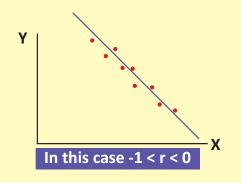
1) Positive correlation

If the plotted points in the plane form a band and they show the rising trend from the lower left hand corner to the upper right hand corner, the two variables are positively correlated.



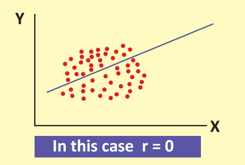
2) Negative correlation

If the plotted points in the plane form a band and they show the falling trend from the upper left hand corner to the lower right hand corner, the two variables are negatively correlated.



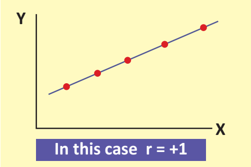
3) Uncorrelated

If the plotted points spread over in the plane then the two variables are uncorrelated.

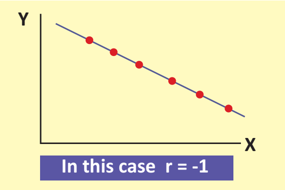


4) Perfect positive correlation

If all the plotted points lie on a straight line from lower left hand corner to the upper right hand corner then the two variables have perfect positive correlation.



5) Perfect Negative correlation



In this case r = -1

If all the plotted points lie on a straight line falling from upper left hand corner to lower right hand corner, the two variables have perfect negative correlation.

4.3.1 Merits and Demerits of scatter diagram

Merits

- It is a simple and non-mathematical method of studying correlation between the variables.
- It is not influenced by the extreme items
- It is the first step in investigating the relationship between two variables.
- It gives a rough idea at a glance whether there is a positive correlation, negative correlation or uncorrelated.

Demerits

- We get an idea about the direction of correlation but we cannot establish the exact strength of correlation between the variables.
- No mathematical formula is involved.

4.4 KARL PEARSON'S CORRELATION COEFFICIENT

When there exists some relationship between two measurable variables, we compute the degree of relationship using the correlation coefficient.

Co-variance

Let (X,Y) be a bivariable normal random variable where $V(X)$ and $V(Y)$ exists. Then, covariance between X and Y is defined as

$$\text{cov}(X,Y) = E[(X\text{-}E(X))(Y\text{-}E(Y))] = E(XY) - E(X)E(Y)$$

If (x_i,y_i), i=1,2, ..., n is a set of n realisations of (X,Y), then the sample covariance between X and Y can be calculated from

$$\text{cov}\left(X,Y\right)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}\,\overline{y}$$

4.4.1 Karl Pearson's coefficient of correlation

When X and Y are linearly related and (X,Y) has a bivariate normal distribution, the co-efficient of correlation between X and Y is defined as

$$r\left(X,Y\right)=\frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}}$$

This is also called as product moment correlation co-efficient which was defined by Karl Pearson.

Based on a given set of n paired observations (x_i,y_i), i=1,2, ... n the sample correlation co-efficient between X and Y can be calculated from

$$r\left(X,Y\right)=\frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}\;\overline{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2-\overline{y}^2}}$$

or, equivalently

$$r\left(X,Y\right)=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

4.4.2 Properties

1. The correlation coefficient between X and Y is same as the correlation coefficient between Y and X (*i.e*, $r_{xy} = r_{yx}$).
2. The correlation coefficient is free from the units of measurements of X and Y
3. The correlation coefficient is unaffected by change of scale and origin.

   Thus, if $u_i = \frac{x_i - A}{c}$ and $v_i = \frac{y_i - B}{d}$ with $c \neq 0$ and $d \neq 0$ i=1,2, ..., n

$$r=\frac{n\sum_{i=1}^{n}u_iv_i-\sum_{i=1}^{n}u_i\sum_{i=1}^{n}v_i}{\sqrt{n\sum_{i=1}^{n}u_i^2-\left(\sum_{i=1}^{n}u_i\right)^2}\sqrt{n\sum_{i=1}^{n}v_i^2-\left(\sum_{i=1}^{n}v_i\right)^2}}$$

   where A and B are arbitrary values.

Remark 1: If the widths between the values of the variabls are not equal then take c = 1 and d = 1.

Interpretation

The correlation coefficient lies between -1 and +1. *i.e.* $-1 \leq r \leq 1$

- A positive value of 'r' indicates positive correlation.
- A negative value of 'r' indicates negative correlation
- If r = +1, then the correlation is perfect positive
- If r = −1, then the correlation is perfect negative.
- If r = 0, then the variables are uncorrelated.
- If $|r| \geq 0.7$ then the correlation will be of higher degree. In interpretation we use the adjective 'highly'
- If X and Y are independent, then r_{xy} = 0. However the converse need not be true.

Example 4.1

The following data gives the heights(in inches) of father and his eldest son. Compute the correlation coefficient between the heights of fathers and sons using Karl Pearson's method.

Height of father	65	66	67	67	68	69	70	72
Height of son	67	68	65	68	72	72	69	71

Solution:

Let x denote height of father and y denote height of son. The data is on the ratio scale. We use Karl Pearson's method.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Calculation

x_i	y_i	x_i^2	y_i^2	x_iy_i
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
544	552	37028	38132	37560

$$r = \frac{8 \times 37560 - 544 \times 552}{\sqrt{8 \times 37028 - (544)^2}\ \sqrt{8 \times 38132 - (552)^2}} = 0.603$$

Heights of father and son are positively correlated. It means that on the average , if fathers are tall then sons will probably tall and if fathers are short, probably sons may be short.

Short-cut method

Let $A = 68$, $B = 69$, $c = 1$ and $d = 1$

x_i	y_i	$u_i = (xi - A)/c$ $= xi - 68$	$v_i = (y_i - B)/d$ $= y_i - 69$	u_i^2	v_i^2	u_iv_i
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total		0	0	36	44	24

$$r = \frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{\sqrt{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2}\sqrt{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2}}$$

$$r = \frac{8\times 24 - 0\times 0}{\sqrt{8\times 36-(0)^2}\ \sqrt{8\times 44-(0)^2}}$$

$$r = \frac{8\times 24}{\sqrt{8\times 36}\ \sqrt{8\times 44}}$$

$$= 0.603$$

Note: The correlation coefficient computed by using direct method and short-cut method is the same.

Example 4.2

The following are the marks scored by 7 students in two tests in a subject. Calculate coefficient of correlation from the following data and interpret.

Marks in test-1	12	9	8	10	11	13	7
Marks in test-2	14	8	6	9	11	12	3

Solution:

Let x denote marks in test-1 and y denote marks in test-2.

	x_i	y_i	x_i^2	y_i^2	x_iy_i
	12	14	144	196	168
	9	8	81	64	72
	8	6	64	36	48
	10	9	100	81	90
	11	11	121	121	121
	1	12	169	144	156
	7	3	49	9	21
Total	70	63	728	651	676

$$r = \frac{n\sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$\sum_{i=1}^{n} x_i = 70 \quad \sum_{i=1}^{n} x_i^2 = 728 \quad \sum_{i=1}^{n} x_iy_i = 676$$

$$\sum_{i=1}^{n} y_i = 63 \quad \sum_{i=1}^{n} y_i^2 = 651 \ n = 7$$

$$r = \frac{7\times 676 - 70\times 63}{\sqrt{\left[7\times 728 - 70^2\right]}\times\sqrt{\left[7\times 651 - 63^2\right]}}$$

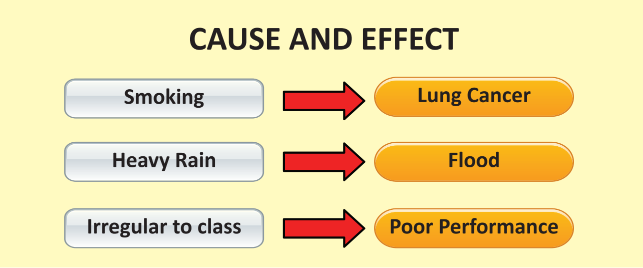
$$= \frac{4732 - 4410}{\sqrt{[5096-4900]}\times\sqrt{[7\times 651 - 3969]}}$$

$$= \frac{322}{\sqrt{196}\times\sqrt{588}} = \frac{322}{14\times 24.25} = \frac{322}{339.5} = 0.95$$

There is a high positive correlation between test-1 and test-2. That is those who perform well in test-1 will also perform well in test-2 and those who perform poor in test-1 will perform poor in test- 2.

The students can also verify the results by using shortcut method.

4.4.3 Limitations of Correlation


CAUSE AND EFFECT

Smoking → Lung Cancer

Heavy Rain → Flood

Irregular to class → Poor Performance

Although correlation is a powerful tool, there are some limitations in using it:

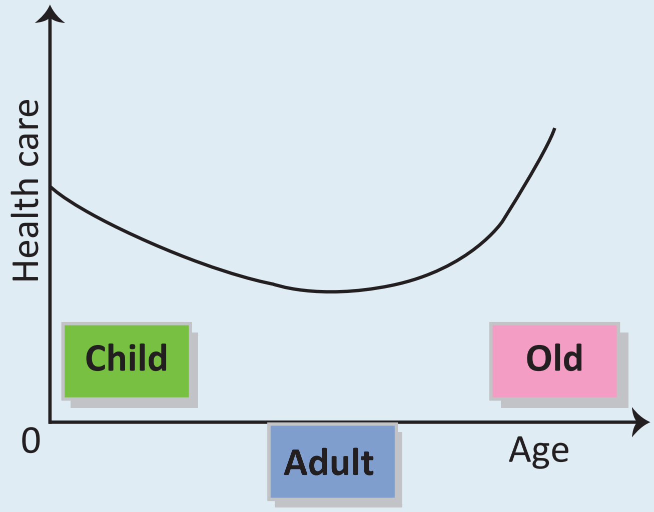
1. Outliers (extreme observations) strongly influence the correlation coefficient. If we see outliers in our data, we should be careful about the conclusions we draw from the value of r. The outliers may be dropped before the calculation for meaningful conclusion.

2. Correlation does not imply causal relationship. That a change in one variable causes a change in another.

NOTE

1. **Uncorrelated** : Uncorrelated (r = 0) implies no 'linear relationship'. But there may exist non-linear relationship (curvilinear relationship).

   Example: Age and health care are related. Children and elderly people need much more health care than middle aged persons as seen from the following graph.



However, if we compute the linear correlation r for such data, it may be zero implying age and health care are uncorrelated, but non-linear correlation is present.

2. **Spurious Correlation** : The word '**spurious**' from Latin means **'false'** or 'illegitimate'. *Spurious correlation means an association extracted from correlation coefficient that may not exist in reality.*

4.5 SPEARMAN'S RANK CORRELATION COEFFICIENT

If the data are in ordinal scale then Spearman's rank correlation coefficient is used. It is denoted by the Greek letter ρ (**rho**).

Spearman's correlation can be calculated for the subjectivity data also, like competition scores. The data can be ranked from low to high or high to low by assigning ranks.

Spearman's rank correlation coefficient is given by the formula

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

where $D_i = R_{1i} - R_{2i}$

R_{1i} = rank of i in the first set of data

R_{2i} = rank of i in the second set of data and

n = number of pairs of observations

Interpretation

Spearman's rank correlation coefficient is a statistical measure of the strength of a monotonic (increasing/decreasing) relationship between paired data. Its interpretation is similar to that of Pearson's. That is, the closer to the ±1 means the stronger the monotonic relationship.

Positive Range	Negative Range
0.01 to 0.19: "Very Weak Agreement"	(-0.01) to (-0.19): "Very Weak Disagreement"
0.20 to 0.39:"Weak Agreement"	(-0.20) to (-0.39): "Weak Disagreement"
0.40 to 0.59: "Moderate Agreement"	(-0.40) to (-0.59): "Moderate Disagreement"
0.60 to 0.79: "Strong Agreement"	(-0.60) to (-0.79): "Strong Disagreement"
0.80 to 1.0: "Very Strong Agreement"	(-0.80) to (-1.0): "Very Strong Disagreement"

Example 4.3

Two referees in a flower beauty competition rank the 10 types of flowers as follows:

Referee A	1	6	5	10	3	2	4	9	7	8
Referee B	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient and find out what degree of agreement is between the referees.

Solution:

Rank by 1st referee R_{1i}	Rank by 2nd referee R_{2i}	$D_i = R_{1i} - R_{2i}$	D_i^2
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
			$\sum_{i=1}^{n} D_i^2 = 60$

Here $n = 10$ and $\sum_{i=1}^{n} D_i^2 = 60$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 60}{10(10^2-1)} = 1 - \frac{360}{10(99)} = 1 - \frac{360}{990} = 0.636$$

Interpretation: Degree of agreement between the referees 'A' and 'B' is 0.636 and they have "strong agreement" in evaluating the competitors.

Example 4.4

Calculate the Spearman's rank correlation coefficient for the following data.

Candidates	1	2	3	4	5
Marks in Tamil	75	40	52	65	60
Marks in English	25	42	35	29	33

Solution:

Tamil		English		$D_i = R_{1i} - R_{2i}$	D_i^2
Marks	Rank (R_{1i})	Marks	Rank (R_{2i})		
75	1	25	5	-4	16
40	5	42	1	4	16
52	4	35	2	2	4
65	2	20	4	-2	4
60	3	33	3	0	0
					40

$\sum_{i=1}^{n} D_i^2 = 40$ and $n = 5$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2-1\right)}$$

$$= 1 - \frac{6 \times 40}{5\left(5^2-1\right)} = 1 - \frac{240}{5\left(24\right)} = -1$$

Interpretation: This perfect negative rank correlation (-1) indicates that scorings in the subjects, totally disagree. Student who is best in Tamil is weakest in English subject and vice-versa.

Example 4.5

Quotations of index numbers of equity share prices of a certain joint stock company and the prices of preference shares are given below.

Years	2013	2014	2015	2016	2017	2008	2009
Equity shares	97.5	99.4	98.6	96.2	95.1	98.4	97.1
Reference shares	75.1	75.9	77.1	78.2	79	74.6	76.2

Using the method of rank correlation determine the relationship between equity shares and preference shares prices.

Solution:

Equity shares	Preference share	R_{1i}	R_{2i}	$D_i = R_{1i} - R_{2i}$	D_i^2
97.5	75.1	4	6	-2	4
99.4	75.9	1	5	-4	16
98.6	77.1	2	3	-1	1
96.2	78.2	6	2	4	16
95.1	79.0	7	1	6	36
98.4	74.6	3	7	-4	16
97.1	76.2	5	4	1	1
					$\sum_{i=1}^{n} D_i^2 = 90$

$$\sum_{i=1}^{n} D_i^2 = 90 \text{ and } n = 7.$$

Rank correlation coefficient is

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2-1\right)}$$

$$= 1 - \frac{6\times 90}{7\left(7^2-1\right)} = 1 - \frac{540}{7\times 48} = 1 - \frac{540}{336} = 1 - 1.6071 = -0.6071$$

Interpretation: There is a negative correlation between equity shares and preference share prices. There is a strong disagreement between equity shares and preference share prices.

4.5.1 Repeated ranks

When two or more items have equal values (i.e., a tie) it is difficult to give ranks to them. In such cases the items are given the average of the ranks they would have received. For example, if two individuals are placed in the 8th place, they are given the rank $\frac{8+9}{2} = 8.5$ each, which is common rank to be assigned and the next will be 10; and if three ranked equal at the 8th place, they are given the rank $\frac{8+9+10}{3} = 9$ which is the common rank to be assigned to each; and the next rank will be 11.

In this case, a different formula is used when there is more than one item having the same value.

$$\rho = 1 - 6\left[\frac{\sum D_i^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{n\left(n^2-1\right)}\right]$$

where m_i is the number of repetitions of i^{th} rank

Example 4.6

Compute the rank correlation coefficient for the following data of the marks obtained by 8 students in the Commerce and Mathematics.

Marks in Commerce	15	20	28	12	40	60	20	80
Marks in Mathematics	40	30	50	30	20	10	30	60

Solution:

Marks in Commerce (X)	Rank (R_{1i})	Marks in Mathematics (Y)	Rank (R_{2i})	$D_i = R_{1i} - R_{2i}$	D_i^2
15	2	40	6	-4	16
20	3.5	30	4	-0.5	0.25
28	5	50	7	-2	4
12	1	30	4	-3	9
40	6	20	2	4	16
60	7	10	1	6	36
20	3.5	30	4	-0.5	0.25
80	8	60	8	0	0
				Total	$\sum D^2 = 81.5$

$$\rho = 1 - 6\left[\frac{\sum D_i^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{n\left(n^2 - 1\right)}\right]$$

Repetitions of ranks

In Commerce (X), 20 is repeated two times corresponding to ranks 3 and 4. Therefore, 3.5 is assigned for rank 2 and 3 with m_1=2.

In Mathematics (Y), 30 is repeated three times corresponding to ranks 3, 4 and 5. Therefore, 4 is assigned for ranks 3,4 and 5 with m_2=3.

Therefore,

$$\rho = 1 - 6\left[\frac{81.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)}{8\left(8^2 - 1\right)}\right]$$

$$= 1 - 6\frac{[81.5 + 0.5 + 2]}{504} = 1 - \frac{504}{504} = 0$$

Interpretation: Marks in Commerce and Mathematics are uncorrelated

4.6 YULE'S COEFFICIENT OF ASSOCIATION

This measure is used to know the existence of relationship between the two attributes A and B (binary complementary variables). Examples of attributes are drinking, smoking, blindness, honesty, etc.

Udny yule

Udny Yule (1871 – 1951), was a British statistician. He was educated at Winchester College and at University College London. After a year dong research in experimental physics, he returned to University College in 1893 to work as a demonstrator for Karl Pearson. Pearson was beginning to work in statistics and Yule followed him into this new field. Yule was a prolific writer, and was active in Royal Statistical Society and received its Guy Medal in Gold in 1911, and served as its President in 1924–26.The concept of Association is due to him.

Coefficient of Association

Yule's Coefficient of Association measures the strength and direction of association. "Association" means that the attributes have some degree of agreement.

2×2 Contingency Table

Attribute A	Attribute B		Total
⇩	Yes B	No β	
Yes A	(AB)	$(A\beta)$	(A)
No α	(αB)	$(\alpha\beta)$	(α)
Total	(B)	(β)	N

Yule's coefficient: $Q = \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

Note 1: The usage of the symbol α is not to be confused with level of significance.

Note 2: (AB): Number with attributes AB etc.

This coefficient ranges from −1 to +1. The values between −1 and 0 indicate inverse relationship (association) between the attributes. The values between 0 and +1 indicate direct relationship (association) between the attributes.

Example 4.7

Out of 1800 candidates appeared for a competitive examination 625 were successful; 300 had attended a coaching class and of these 180 came out successful. Test for the association of attributes attending the coaching class and success in the examination.

Solution:

$N = 1800$

A: Success in examination $\quad\quad$ α: No success in examination

B: Attended the coaching class $\quad\quad$ β: Not attended the coaching class

$$(A) = 625,\ \ (B) = 300,\ (AB) = 180$$

	B	β	Total
A	180	445	625
α	120	1055	1175
Total	300	1500	$N = 1800$

Yule's coefficient: $Q = \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$$
\begin{aligned}
&= \frac{180\times1055-445\times120}{180\times1055+445\times120} \\
&= \frac{189900-53400}{189900+53400} \\
&= \frac{136500}{243300} \\
&= 0.561 > 0
\end{aligned}
$$

Interpretation: There is a positive association between success in examination and attending coaching classes. Coaching class is useful for success in examination.

Remark: Consistency in the data using contingency table may be found as under.

Construct a 2 × 2 contingency table for the given information. If at least one of the cell frequencies is negative then there is inconsistency in the given data.

Example 4.8

Verify whether the given data: $N = 100$, (A) **= 75,** $(B) = 60$ and $(AB) = 15$ is consistent.

Solution:

The given information is presented in the following contingency table.

	B	β	Total
A	15	60	75
α	45	$\boxed{-20}$	25
Total	60	40	$N = 100$

Notice that $(\alpha\beta) = -20$

Interpretation: Since one of the cell frequencies is negative, the given data is "Inconsistent".

POINTS TO REMEMBER

- Correlation study is about finding the linear relationship between two variables. Correlation is not causation. Sometimes the correlation may be spurious.
- Correlation coefficient lies between −1 and +1.
- Pearson's correlation coefficient provides the type of relationship and intensity of relationship, for the data in ratio scale measure.
- Spearman's correlation measures the relationship between the two ordinal variables.
- Yule's coefficient of Association measures the association between two dichotomous attributes.

EXERCISE 4

I. Choose the best answer.

1. The statistical device which helps in analyzing the co-variation of two or more variables is

   (a) variance (b) probability

   (c) correlation coefficient (d) coefficient of skewness

2. "The attempts to determine the degree of relationship between variables is correlation" is the definition given by

   (a) A.M. Tuttle (b) Ya-Kun-Chou

   (c) A.L. Bowley (d) Croxton and Cowden

3. If the two variables do not have linear relationship between them then they are said to have

   (a) positive correlation (b) negative correlation

   (c) uncorrelated (d) spurious correlation

4. If all the plotted points lie on a straight line falling from upper left hand corner to lower right hand corner then it is called

   (a) perfect positive correlation (b) perfect negative correlation

   (c) positive correlation (d) negative correlation

5. If $r = +1$, then the correlation is called

   (a) perfect positive correlation (b) perfect negative correlation

   (c) positive correlation (d) negative correlation

6. The correlation coefficient lies in the interval

   (a) $-1 \le r \le 0$ (b) $-1 < r < 1$ (c) $0 \le r \le 1$ (d) $-1 \le r \le 1$

7. Rank correlation coefficient is given by

   (a) $1+\frac{6\sum_{i=1}^{n} D_i^2}{n^3-n}$ (b) $1-\frac{6\sum_{i=1}^{n} D_i^2}{n^3-n}$ (c) $1-\frac{6\sum_{i=1}^{n} D_i^2}{n^3+n}$ (d) $1-\frac{6\sum_{i=1}^{n} D_i^3}{n\left(n^2-1\right)}$

8. If $\sum D^2 = 0$, rank correlation is

   (a) 0 (b) 1 (c) 0.5 (d) −1

9. Rank correlation was developed by

   (a) Pearson (b) Spearman (c) Yule (d) Fisher

10. Product moment coefficient of correlation is

   (a) $r = \frac{\sigma_x \ \sigma_y}{\text{cov}\,(x, y)}$ (b) $r = \sqrt{\sigma_x \ \sigma_y}$ (c) $r = \frac{\text{cov}\,(x,\ y)}{\sigma_x \ \sigma_y}$ (d) $r = \frac{\text{cov}\,(x,\ y)}{\sigma_{xy}}$

11. The purpose of the study of __________ is to identify the factors of influence and try to control them for better performance.

    (a) mean (b) correlation (c) standard deviation (d) skewness

12. The height and weight of a group of persons will have ________ correlation.

    (a) positive (b) negative

    (c) zero (d) both positive and negative

13. _______ correlation studies the association of two variables with ordinal scale.

    (a) A.M. Tuttle rank (b) Croxton and Cowdon rank

    (c) Karl Pearson's rank (d) Spearman's rank.

14. _______ presents a graphic description of quantitative relation between two series of facts.

    (a) scatter diagram (b) bar diagram (c) pareto diagram (d) pie diagram

15. _______ measures the degree of relationship between two variables.

    (a) standard deviation (b) correlation coefficient

    (c) moment (d) median

16. The correlation coefficient of x and y is symmetric. Hence

    (a) $r_{xy} = r_{yx}$ (b) $r_{xy} > r_{yx}$ (c) $r_{xy} < r_{yx}$ (d) $r_{xy} \neq r_{yx}$

17. If cov $(x, y) = 0$ then its interpretation is

    (a) x and y are positively correlated (b) x and y are negatively correlated

    (c) x and y are uncorrelated (d) x and y are independent

18. Rank correlation is useful to study data in _______ scale.

    (a) ratio (b) ordinal (c) nominal (d) ratio and nominal

19. If $r = 0$ then cov(x, y) is

    (a) 0 (b) +1 (c) -1 (d) α

20. If cov$(x, y) = \sigma_x, \sigma_y$ then

    (a) $r = 0$ (b) $r = -1$ (c) $r = +1$ (d) $r = \alpha$

II. Give very short answer to the following questions.

21. What is correlation?
22. Write the definition of correlation by A.M. Tuttle.
23. What are the different types of correlation?
24. What are the types of simple correlation?
25. What do you mean by uncorrelated?
26. What you understand by spurious correlation?

27. What is scatter diagram?

28. Define co-variance.

29. Define rank correlation.

30. If $\sum D^2 = 0$ what is your conclusion regarding Spearman's rank correlation coefficient?

31. Give an example for (i) positive correlation

(ii) negative correlation (iii) no correlation

32. What is the value of 'r' when two variables are uncorrelated?

33. When the correlation coefficient is +1, state your interpretation.

III. Give short answer to the following questions.

34. Write any three uses of correlation.

35. Define Karl Pearson's coefficient of correlation.

36. How do you interpret the coefficient of correlation which lies between 0 and +1?

37. Write down any 3 properties of correlation?

38. If rank correlation coefficient $r = 0.8$, $\sum D^2 = 3$ then find n?

39. Write any three merits of scatter diagram.

40. Given that cov(x, y) = 18.6, variance of x = 20.2, variance of y = 23.7. Find r.

41. Test the consistency of the following data with the symbols having their usual meaning. $N = 1000$, $(A) = 600$, $(B) = 500$, $(AB) = 50$.

IV. Give detailed answer to the following questions.

42. Explain different types of correlation.

43. Explain scatter diagram.

44. Calculate the Karl Pearson's coefficient of correlation for the following data and interpret.

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

45. Find the Karl Pearson's coefficient of correlation for the following data.

Wages	100	101	102	102	100	99	97	98	96	95
Cost of living	98	99	99	97	95	92	95	94	90	91

How are the wages and cost of living correlated?

46. Calculate the Karl Pearson's correlation coefficient between the marks (out of 10) in statistics and mathematics of 6 students.

Student	1	2	3	4	5	6
Statistics	7	4	6	9	3	8
Mathematics	8	5	4	8	3	6

47. In a marketing survey the prices of tea and prices of coffee in a town based on quality was found as shown below. Find the rank correlation between prices of tea and prices of coffee.

Price of tea	88	90	95	70	60	75	50
Price of coffee	120	134	150	115	110	140	100

48. Calculate the Spearman's rank correlation coefficient between price and supply from the following data.

Price	4	6	8	10	12	14	16	18
Supply	10	15	20	25	30	35	40	45

49. A random sample of 5 college students is selected and their marks in Tamil and English are found to be:

Tamil	85	60	73	40	90
English	93	75	65	50	80

Calculate Spearman's rank correlation coefficient.

50. Calculate Spearman's coefficient of rank correlation for the following data.

x	53	98	95	81	75	71	59	55
y	47	25	32	37	30	40	39	45

51. Calculate the coefficient of correlation for the following data using ranks.

Mark in Tamil	29	24	25	27	30	31
Mark in English	29	19	30	33	37	36

52. From the following data calculate the rank correlation coefficient.

x	49	34	41	10	17	17	66	25	17	58
y	14	14	25	7	16	5	21	10	7	20

Yule's coefficient

53. Can vaccination be regarded as a preventive measure of Hepatitis B from the data given below. Of 1500 person in a locality, 400 were attacked by Hepatitis B. 750 has been vaccinated. Among them only 75 were attacked.

ANSWERS

I 1. (c) 2. (b) 3. (c) 4. (b) 5. (a)

6. (d) 7. (b) 8. (b) 9. (b) 10. (b)

11. (b) 12. (a) 13. (d) 14. (a) 15. (b)

16. (a) 17. (c) 18. (b) 19. (a) 20. (c)

II 30. $r = 1$

III 38. $n = 10$

40. $r = 0.85$

41. $(\alpha\beta) = -50$, The given data is inconsistent

IV 44. $r = 0.95$ it is highly positively correlated

45. $r = 0.847$ wages and cost of living are highly positively correlated.

46. $r = 0.8081$. Statistics and mathematics marks are highly positively correlated.

47. $\rho = 0.8929$ price of tea and coffee are highly positively correlated.

48. $\rho = 1$ (perfect positive correlation)

49. $\rho = 0.8$

50. $\rho = -0.905$ x and y are highly negatively

51. $\rho = -0.78$ marks in Tamil and English are negatively correlated.

52. $\rho = +0.733$

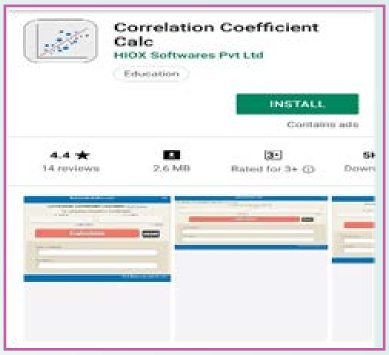
53. There is a negative association between attacked and vaccinated. There is a positive association between not attacked and not vaccinated. Hence vaccination can be regarded as a preventive measure of Hepatitis B.

ICT CORNER

CORRELATION ANALYSIS

STATS IN YOUR PALM

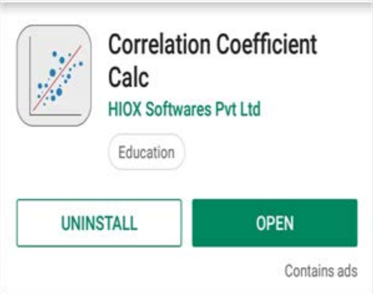
This activity is to calculate Correlation Coefficient


Correlation Coefficient Calc
HIOX Softwares Pvt Ltd
Education

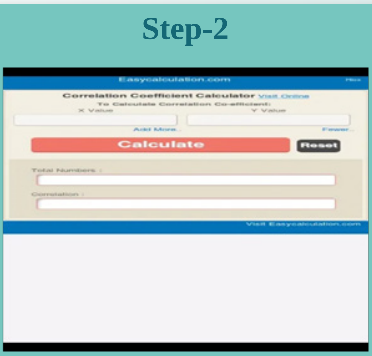
Steps:

- This is an android app activity. Open the browser and type the URL given (or) scan the QR code. (Or) search for "**Correlation Coefficient**" in google play store.
- (i) Install the app and open the app, (ii) To calculate Correlation Co-efficient in put the the values of *X* and *Y* in the given box (iii) Then click "**CALCULATE**" we will get the result.

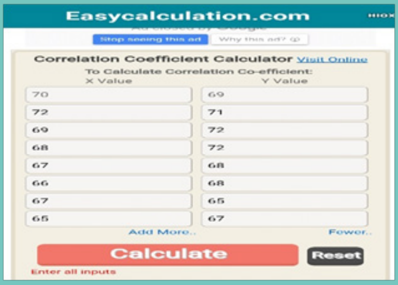
Step-1


Correlation Coefficient Calc
HIOX Softwares Pvt Ltd
Education

Step-2


Easycalculation.com
Correlation Coefficient Calculator Visit Online
To Calculate Correlation Co-efficient:
X Value
Y Value

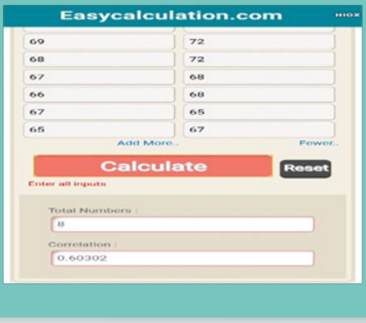
Step-3


Easycalculation.com
Correlation Coefficient Calculator Visit Online
To Calculate Correlation Co-efficient:

X Value	Y Value
70	69
72	71
69	72
68	72
67	68
66	68
67	65
65	67

Enter all inputs

Result


Easycalculation.com

69	72
68	72
67	68
66	68
67	65
65	67

Enter all inputs

Total Numbers : 8

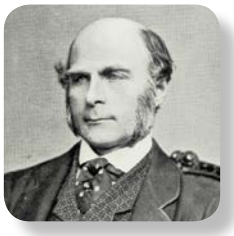
Correlation : 0.60302

Pictures are indicatives only*

URL:

https://play.google.com/store/apps/details?id=com.hiox.CreliCoefficientCalcul

CHAPTER 5 REGRESSION ANALYSIS



Francis Galton

Francis Galton (1822-1911) was born in a wealthy family. The youngest of nine children, he appeared as an intelligent child. Galton's progress in education was not smooth. He dabbled in medicine and then studied Mathematics at Cambridge. In fact he subsequently freely acknowledged his weakness in formal Mathematics, but this weakness was compensated by an exceptional ability to understand the meaning of data. Many statistical terms, which are in current usage were coined by Galton. For example, correlation is due to him, as is regression, and he was the originator of terms and concepts such as quartile, decile and percentile, and of the use of median as the midpoint of a distribution.

The concept of regression comes from genetics and was popularized by Sir Francis Galton during the late 19th century with the publication of regression towards mediocrity in hereditary stature. Galton observed that extreme characteristics (e.g., height) in parents are not passed on completely to their offspring. An examination of publications of Sir Francis Galton and Karl Pearson revealed that Galton's work on inherited characteristics of sweet peas led to the initial conceptualization of linear regression. Subsequent efforts by Galton and Pearson brought many techniques of multiple regression and the product-moment correlation coefficient.

LEARNING OBJECTIVES

The student will be able to

❖ know the concept of regression, its types and their uses.

❖ fit best line of regression by applying the method of least squares.

❖ calculate the regression coefficient and interpret the same.

❖ know the uses of regression coefficients.

❖ distinguish between correlation analysis and regression analysis.

2PUID

Introduction

The correlation coefficient is an useful *statistical tool for describing the type (positive or negative or uncorrelated) and intensity of linear relationship* (such as moderately or highly) between two variables. But it fails to give a *mathematical functional* relationship for prediction purposes. Regression analysis is a vital statistical method for obtaining functional relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one to understand how the typical value of the dependent variable (or 'response variable') changes when any one of the independent variables (regressor(s) or predictor(s)) is varied, while the other independent variables are held fixed. It helps to determine the impact of changes in the value(s) of the the independent variable(s) upon changes in the value of the dependent variable. Regression analysis is widely used for prediction.

5.1 DEFINITION

Regression analysis is a statistical method of determining the mathematical functional relationship connecting independent variable(s) and a dependent variable.

Types of 'Regression'

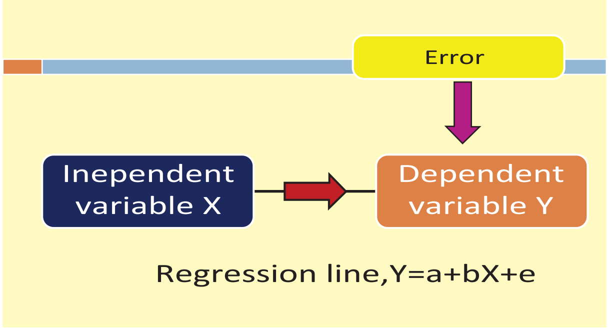
Based on the kind of relationship between the dependent variable and the set of independent variable(s), there arises two broad categories of regression *viz.*, linear regression and non-linear regression.

If the relationship is linear and there is only one independent variable, then the regression is called as simple linear regression. On the other hand, if the relationship is linear and the number of independent variables is two or more, then the regression is called as multiple linear regression. If the relationship between the dependent variable and the independent variable(s) is not linear, then the regression is called as non-linear regression.

5.1.1 Simple Linear Regression

It is one of the most widely known modeling techniques. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete and nature of relationship is linear. This relationship can be expressed using a straight line equation (linear regression) that best approximates all the individual data points.

Simple linear regression establishes a relationship between a **dependent variable (Y)** and one **independent variable (X)** using a **best fitted straight line** (also known as regression line).



NOTE

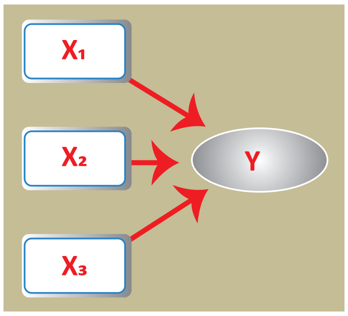
There are many reasons for the presence of the error term in the linear regression model. It is also known as measurement error. In some situations, it indicates the presence of several variables other than the present set of regressors.

The general form of the simple linear regression equation is $Y = a + bX + e$, where 'X' is independent variable, 'Y' is dependent variable, a' is intercept, 'b' is slope of the line and ' e' is error term. This equation can be used to estimate the value of response variable (Y) based on the given values of the predictor variable (X) within its domain.

5.1.2 Multiple Linear Regression

In the case of several independent variables, regression analysis also allows us to compare the effects of independent variables measured on different scales, such as the effect of price changes and the number of promotional activities.

Multiple linear regression uses two or more independent variables to estimate the value(s) of the response variable (Y).



The general form of the multiple linear regression equation is $Y = a + b_1X_1{}^{+}\, b_2X_2 + b_3X_3 + \ldots + b_tX_t + e$

Here, Y represents the dependent (response) variable, X_i represents the i^{th} independent variable (regressor), a and b_i are the regression coefficients and e is the error term.

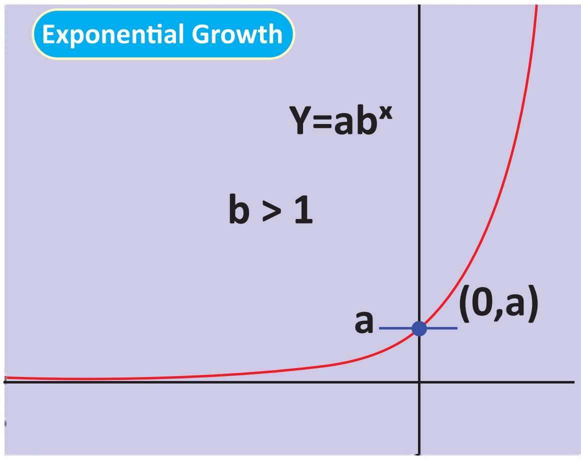
Suppose that price of a product (Y) depends mainly upon three promotional activities such as discount (X_1), instalment scheme (X_2) and free installation (X_3). If the price of the product has linear relationship with each promotional activity, then the relationship among Y and X_1, X_2 and X_3 may be expressed using the above general form as

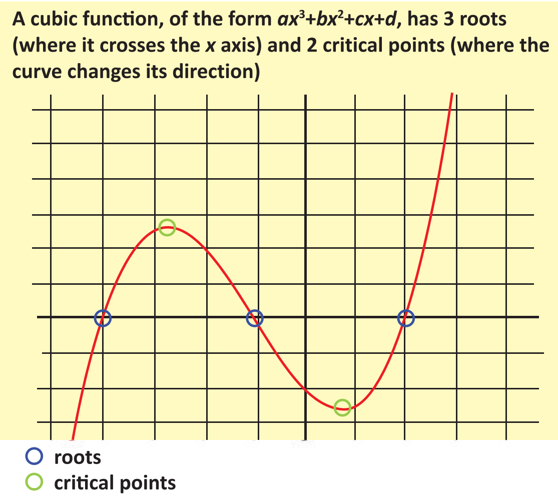
$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e\,.$$

These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building regression models for predictive purposes.

5.1.3 Non-Linear Regression

If the regression is not linear and is in some other form, then the regression is said to be non-linear regression. Some of the non-linear relationships are displayed below.




A cubic function, of the form ax^3+bx^2+cx+d, has 3 roots (where it crosses the x axis) and 2 critical points (where the curve changes its direction)

5.2 USES OF REGRESSION

Benefits of using regression analysis are as follows:

1. It indicates the **significant mathematical** relationship between independent variable (X) and dependent variable(Y). *(i.e)* Model construction
2. It indicates the **strength of impact (b)** of independent variable on a dependent variable.
3. It is used to estimate (interpolate) the value of the response variable for different values of the independent variable from its range in the given data. It means that extrapolation of the dependent variable is not generally permissible.

NOTE

Multiple linear regression and Curvilinear relationships (non-linear regression) are out of the syllabus. Basic information about them are given here, for enhancing the knowledge.

4. In the case of several independent variables, regression analysis is a way of mathematically sorting out which of those variables indeed have an impact (It answers the questions: Which independent variable matters most? Which can we ignore? How do those independent variables interact with each other?

5.3 WHY ARE THERE TWO REGRESSION LINES?

There may exist two regression lines in certain circumstances. When the variables X and Y are interchangeable with related to causal effects, one can consider X as independent variable and Y as dependent variable (or) Y as independent variable and X as dependent variable. As the result, we have (1) **the regression line of Y on X** and (2) **the regression line of X on Y**.

Both are valid regression lines. But we must judicially select the one regression equation which is suitable to the given environment.

Note: If, X only causes Y, then there is only one regression line, of Y on X.

5.3.1 Simple Linear Regression

In the general form of the simple linear regression equation of Y on X

$$Y= a + bX + e$$

the constants 'a' and 'b' are generally called as the regression coefficients.

The coefficient 'b' represents the rate of change in the value of the mean of Y due to every unit change in the value of X. When the range of X includes '0', then the intercept 'a' is $E(Y|X = 0)$. If the range of X does not include '0', then 'a' does not have practical interpretation.

If (x_i, y_i), $i = 1, 2, ..., n$ is a set of n-pairs of observations made on (X, Y), then fitting of the above regression equation means finding the estimates '$\hat{a}$' and '$\hat{b}$' for 'a' and 'b' respectively. These estimates are determined based on the following general assumptions:

i) the relationship between Y and X is linear (approximately).

ii) the error term 'e' is a random variable with mean zero.

iii) the error term 'e' has constant variance.

There are other assumptions on 'e', which are not required at this level of study.

Before going for further study, the following points are to be kept in mind.

- Both the independent and dependent variables must be measured at the interval scale.
- There must be **linear relationship** between independent and dependent variables.
- Linear Regression is very sensitive to **Outliers** (extreme observations). It can affect the regression line extremely and eventually the estimated values of Y too.

Meaning of line of "best fit"

Based on the assumption (ii), the response variable Y is also a random variable with mean

$$E(Y|X=x) = a + bx$$

In regression analysis, the main objective is finding the line of best fit, which provides the fitted equation of *Y* on *X*.

The line of 'best fit' is the line (straight line equation) which minimizes the error in the estimation of the dependent variable *Y*, for any specified value of the independent variable *X* from its range.

The regression equation $E(Y|X=x) = a + bx$ represents a family of straight lines for different values of the coefficients '*a*' and '*b*'. The problem is to determine the estimates of '*a*' and '*b*' by minimizing the error in the estimation of *Y* so that the line is a best fit. This necessitates to find the suitable values of the estimates of '*a*' and '*b*'.

5.4 METHOD OF LEAST SQUARES

In most of the cases, the data points do not fall on a straight line (not highly correlated), thus leading to a possibility of depicting the relationship between the two variables using several different lines. Selection of each line may lead to a situation where the line will be closer to some points and farther from other points. We cannot decide which line can provide best fit to the data.

Method of least squares can be used to determine the line of best fit in such cases. It determines the line of best fit for given observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line.

5.4.1 Method of Least Squares

To obtain the estimates of the coefficients '*a*' and '*b*', the least squares method minimizes the sum of squares of residuals. The residual for the i^{th} data point e_i is defined as the difference between the observed value of the response variable, y_i, and the estimate of the response variable, $\hat{y}_i$, and is identified as the error associated with the data. *i.e.*, $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$.

The method of least squares helps us to find the values of unknowns '*a*' and '*b*' in such a way that the following two conditions are satisfied:

- Sum of the residuals is zero. That is $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$.
- Sum of the squares of the residuals $E(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the least.

5.4.2 Fitting of Simple Linear Regression Equation

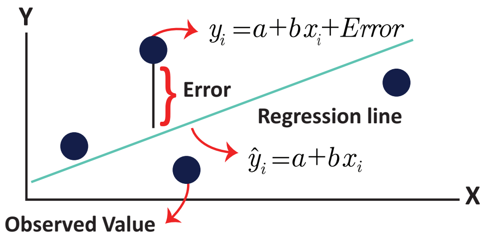
The method of least squares can be applied to determine the estimates of '*a*' and '*b*' in the simple linear regression equation using the given data $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$ by minimizing

$$E(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$i.e., \; E(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2 .$$

Here, $\hat{y}_i = a + bx_i$ is the expected (estimated) value of the response variable for given x_i.

Simple Linear Regression Model



It is obvious that if the expected value ($\hat{y}_i$) is close to the observed value (y_i), the residual will be small. Since the magnitude of the residual is determined by the values of '*a*' and '*b*', estimates of these coefficients are obtained by minimizing the sum of the squared residuals, *E(a,b)*.

Differentiation of *E(a,b)* with respect to '*a*' and '*b*' and equating them to zero constitute a set of two equations as described below:

$$\frac{\partial E(a,b)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\frac{\partial E(a,b)}{\partial b} = -2\sum_{i=1}^{n}x_i(y_i - a - bx_i) = 0$$

These give

$$na + b\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i$$

$$a\sum_{i=1}^{n}x_i + b\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_iy_i$$

These equations are popularly known as **normal equations.** Solving these equations for '*a*' and '*b*' yield the estimates $\hat{a}$ and $\hat{b}$.

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

and

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}$$

It may be seen that in the estimate of '*b*', the numerator and denominator are respectively the sample covariance between *X* and *Y*, and the sample variance of *X*. Hence, the estimate of '*b*' may be expressed as

$$\hat{b} = \frac{Cov(X,Y)}{V(X)}$$

Further, it may be noted that for notational convenience the denominator of $\hat{b}$ above is mentioned as variance of *X*. But, the definition of sample variance remains valid as defined in Chapter I, that is, $\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}^2\right)$.

From Chapter 4, the above estimate can be expressed using, r_{XY}, Pearson's coefficient of the simple correlation between *X* and *Y*, as

$$\hat{b} = r_{XY}\frac{SD(Y)}{SD(X)}.$$

Important Considerations in the Use of Regression Equation:

1. Regression equation exhibits only the relationship between the respective two variables. Cause and effect study shall not be carried out using regression analysis.
2. The regression equation is fitted to the given values of the independent variable. Hence, the fitted equation can be used for prediction purpose corresponding to the values of the regressor within its range. Interpolation of values of the response variable may be done corresponding to the values of the regressor from its range only. The results obtained from extrapolation work could not be interpreted.

Example 5.1

Construct the simple linear regression equation of *Y* on *X* if $n = 7$, $\sum_{i=1}^{n} x_i = 113$, $\sum_{i=1}^{n} x_i^2 = 1983$, $\sum_{i=1}^{n} y_i = 182$ and $\sum_{i=1}^{n} x_i y_i = 3186$.

Solution:

The simple linear regression equation of *Y* on *X* to be fitted for given data is of the form

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

The values of '*a*' and '*b*' have to be estimated from the sample data solving the following normal equations.

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (2)$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (3)$$

Substituting the given sample information in (2) and (3), the above equations can be expressed as

$$7\,a + 113\,b = 182 \qquad (4)$$

$$113\,a + 1983\,b = 3186 \qquad (5)$$

$$(4) \times 113 \Rightarrow 791\,a + 12769\,b = 20566$$

$$(5) \times 7 \Rightarrow 791\,a + 13881\,b = 22302$$

$$(-) \quad (-) \quad (-)$$

$$-1112\,b = -1736$$

$$\Rightarrow b = \frac{1736}{1112} = 1.56$$

$$b = 1.56$$

Substituting this in (4) it follows that,

$$7\,a + 113 \times 1.56 = 182$$

$$7\,a + 176.28 = 182$$

$$7\,a = 182 - 176.28$$

$$= 5.72$$

Hence, $a = 0.82$

Example 5.2

Number of man-hours and the corresponding productivity (in units) are furnished below. Fit a simple linear regression equation $\hat{Y} = a + bx$ applying the method of least squares.

Man-hours	3.6	4.8	7.2	6.9	10.7	6.1	7.9	9.5	5.4
Productivity (in units)	9.3	10.2	11.5	12	18.6	13.2	10.8	22.7	12.7

Solution:

The simple linear regression equation to be fitted for the given data is

$$\hat{Y} = a + bx$$

Here, the estimates of a and b can be calculated using their least squares estimates

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

i.e.,

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{b}\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\overline{x} \times \overline{y})}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2}$$

or equivalently $$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

From the given data, the following calculations are made with n=9

Man-hours x_i	Productivity y_i	x_i^2	$x_i y_i$
3.6	9.3	12.96	33.48
4.8	10.2	23.04	48.96
7.2	11.5	51.84	82.8
6.9	12	47.61	82.8
10.7	18.6	114.49	199.02
6.1	13.2	37.21	80.52
7.9	10.8	62.41	85.32
9.5	22.7	90.25	215.65
5.4	12.7	29.16	66.42
$\sum_{i=1}^{9} x_i = 62.1$	$\sum_{i=1}^{9} y_i = 121$	$\sum_{i=1}^{9} x_i^2 = 468.97$	$\sum_{i=1}^{9} x_i y_i = 894.97$

Substituting the column totals in the respective places in the of the estimates $\hat{a}$ and $\hat{b}$, their values can be calculated as follows:

$$\hat{b} = \frac{(9\times 894.97)-(62.1\times 121)}{(9\times 468.97)-(62.1)^2}$$

$$= \frac{8054.73-7514}{4220.73-3856.41}$$

$$= \frac{540.73}{364.32}$$

Thus, $\hat{b} = 1.48$.

Now $\hat{a}$ can be calculated using $\hat{b}$ as

$$\hat{a} = \frac{121}{9} - \left(1.48\times\frac{62.1}{9}\right)$$

$= 13.40 - 10.21$

Hence, $\hat{a} = 3.19$

Therefore, the required simple linear regression equation fitted to the given data is

$\hat{Y} = 3.19 + 1.48x$

It should be noted that the value of Y can be estimated using the above fitted equation for the values of x in its range *i.e.*, 3.6 to 10.7.

In the estimated simple linear regression equation of Y on X

$$\hat{Y} = \hat{a} + \hat{b}x$$

we can substitute the estimate $\hat{a} = \overline{y} - \hat{b}\overline{x}$. Then, the regression equation will become as

$$\hat{Y} = \overline{y} - \hat{b}\overline{x} + \hat{b}x$$

$$\hat{Y} - \overline{y} = \hat{b}(x - \overline{x})$$

It shows that the simple linear regression equation of Y on X has the slope $\hat{b}$ and the corresponding straight line passes through the point of averages $(\overline{x}, \overline{y})$. The above representation of straight line is popularly known in the field of Coordinate Geometry as 'Slope-Point form'. The above form can be applied in fitting the regression equation for given regression coefficient $\hat{b}$ and the averages $\overline{x}$ and $\overline{y}$.

As mentioned in Section 5.3, there may be two simple linear regression equations for each X and Y. Since the regression coefficients of these regression equations are different, it is essential to distinguish the coefficients with different symbols. The regression coefficient of the simple linear regression equation of Y on X may be denoted as b_{YX} and the regression coefficient of the simple linear regression equation of X on Y may be denoted as b_{XY}.

Using the same argument for fitting the regression equation of Y on X, we have the simple linear regression equation of X on Y with best fit as

$$\hat{X} = \hat{c} + b_{XY}\, y$$

$$\text{where } \hat{c} = \overline{x} - b_{XY}\,\overline{y}$$

$$b_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2}$$

The slope-point form of this equation is

$$\hat{X} - \overline{x} = b_{XY}(y - \overline{y}).$$

Also, the relationship between the Karl Pearson's coefficient of correlation and the regression coefficient are

$$b_{XX} = r_{XY}\frac{SD(X)}{SD(Y)} \text{ and } b_{YX} = r_{XY}\frac{SD(Y)}{SD(X)}.$$

5.5 PROPERTIES OF REGRESSION COEFFICIENTS

1. Correlation coefficient is the geometric mean between the regression coefficients.

$$r_{XY} = \sqrt{b_{XY} \times b_{YX}}$$

2. It is clear from the property 1, both regression coefficients must have the same sign. *i.e.,* either they will positive or negative.
3. If one of the regression coefficients is greater than unity, the other must be less than unity.
4. The correlation coefficient will have the same sign as that of the regression coefficients.
5. Arithmetic mean of the regression coefficients is greater than the correlation coefficient.

$$\frac{b_{XY} + b_{YX}}{2} \geq r_{XY}$$

6. Regression coefficients are independent of the change of origin but not of scale.

Properties of regression equation

1. If $r = 0$, the variables are uncorrelated, the lines of regression become perpendicular to each other.
2. If $r = 1$, the two lines of regression either coincide or parallel to each other.
3. Angle between the two regression lines is $\theta = \tan^{-1}\left(\frac{m_1 - m_2}{1 + m_1 m_2}\right)$ where m_1 and m_2 are the slopes of regression lines X on Y and Y on X respectively.
4. The angle between the regression lines indicates the degree of dependence between the variable.
5. Regression equations intersect at $(\overline{X}, \overline{Y})$

Example 5.3

Calculate the regression equation of X on Y from the data given below, taking deviations from actual means of X and Y.

x	12	14	15	14	18	17
y	42	40	45	47	39	45

Estimate the likely demand when the $X = 25$.

Solution:

	x_i	$u_i = x_i - 15$	u_i^2	y_i	$v_i = y_i - 43$	v_i^2	u_iv_i
	12	-3	9	42	-1	1	3
	14	-1	1	40	-3	9	3
	15	-0	0	45	2	4	0
	14	-1	1	47	4	16	-4
	18	3	9	39	-4	16	-12
	17	2	4	45	2	4	4
Total	90	0	24	258	0	50	-6

$$\bar{x} = \sum_{i=1}^{6} x_i / 6 = \frac{90}{6} = 15$$

$$\bar{y} = \sum_{i=1}^{6} y_i / 5 = \frac{258}{6} = 43$$

The regression line of U on V is computed as under

$$\hat{b}_{UV} = \frac{n\sum_{i=1}^{n} u_iv_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2} = \frac{6(-6)}{6 \times 50} = -0.12$$

$$\hat{a} = \bar{u} - \hat{b}_{UV}\,\bar{v} = 0$$

Hence, the regression line of U on V is $U = \hat{b}_{UV}\, v + \hat{a} = -0.12v$

Thus, the regression line of X on Y is $(Y-43) = -0.25(x-15)$

When $x = 25$, $y - 43 = -0.25\ (25-15)$

$$y = 40.5$$

Important Note: If $\overline{X}$, $\overline{Y}$ are not integers then the above method is tedious and time consuming to calculate b_{YX} and b_{XY}. The following modified formulae are easy for calculation.

$$b_{YX} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}$$

Example 5.4

The following data gives the experience of machine operators and their performance ratings as given by the number of good parts turned out per 50 pieces.

Operators	1	2	3	4	5	6	7	8
Experience (X)	8	11	7	10	12	5	4	6
Ratings (Y)	11	30	25	44	38	25	20	27

Obtain the regression equations and estimate the ratings corresponding to the experience x=15.

Solution:

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
	8	11	88	64	121
	11	30	330	121	900
	7	25	175	49	625
	10	44	440	100	1936
	12	38	456	144	1444
	5	25	125	25	625
	4	20	80	16	400
	6	27	162	36	729
Total	63	220	1856	555	6780

Regression equation of Y on X,

$$\hat{Y} - \overline{y} = b_{YX}\left(x - \overline{x}\right)$$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{63}{8} = 7.875$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{220}{8} = 27.5$$

The above two means are in decimal places so for the simplicity we use this formula to compute b_{YX}.

$$b_{YX} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{8\times1856 - 63\times220}{8\times555 - 63\times63}$$

$$= \frac{14848 - 13860}{4440 - 3969}$$

$$= \frac{988}{471}$$

$$b_{YX} = 2.098$$

The regression equation of Y on X,

$$\hat{Y} - \bar{y} = b_{YX}(x - \bar{x})$$

$$\hat{Y} - 27.5 = 2.098\ (x - 7.875)$$

$$\hat{Y} - 27.5 = 2.098\ x - 16.52$$

$$\hat{Y} = 2.098x + 10.98$$

When $x = 15$,

$$\hat{Y} = 2.098 \times 15 + 10.98$$

$$\hat{Y} = 31.47 + 10.98$$

$$= 42.45$$

Regression equation of X on Y,

$$\hat{X} - \bar{x} = b_{XY}(y - \bar{y})$$

$$b_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}$$

$$= \frac{8\times1856 - 63\times220}{8\times6780 - 220\times220}$$

$$= \frac{14848 - 13860}{54240 - 48400}$$

$$= \frac{988}{5840}$$

$$b_{XY} = 0.169$$

The regression equation of X on Y,

$$\hat{X} - 7.875 = 0.169\ (y - 27.5)$$

$$\hat{X} - 7.875 = 0.169y - 0.169 \times 27.5$$

$$\hat{X} = 0.169y + 3.2275$$

Example 5.5

The random sample of 5 school students is selected and their marks in statistics and accountancy are found to be

Statistics	85	60	73	40	90
Accountancy	93	75	65	50	80

Find the two regression lines.

Solution:

The two regression lines are:

Regression equation of Y on X,

$$\hat{Y} - \overline{y} = b_{YX}\left(x - \overline{x}\right)$$

Regression equation of X on Y,

$$\hat{X} - \overline{x} = b_{XY}\left(y - \overline{y}\right)$$

	x_i	y_i	$u_i = x_i - A$ $= x_i - 60$	$v_i = x_i - B$ $= x_i - 75$	u_iv_i	u_i^2	y_i^2
	85	93	25	18	450	625	324
	60 A	75 B	0	0	0	0	0
	73	65	13	–10	–130	169	100
	40	50	–20	–25	500	400	625
	90	80	30	5	150	900	25
Total	348	363	48	12	970	2094	1074

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{348}{5} = 69.6$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{363}{5} = 72.6$$

Since the mean values are in decimals format not as integers and numbers are big, we take origins for x and y and then solve the problem.

Regression equation of Y on X,

$$\hat{Y} - \bar{y} = b_{YX}\left(x - \bar{x}\right)$$

Calculation of b_{YX}

$$b_{YX} = b_{VU}\frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2}$$

$$= \frac{5 \times 970 - 48 \times 9(-12)}{5 \times 2094 - (48)^2}$$

$$= \frac{4850 + 576}{10470 - 2304}$$

$$= \frac{5426}{8126} = 0.664$$

$$b_{YX} = b_{VU} = 0.664$$

$$\hat{Y} - 72.6 = 0.664\ (x - 69.6)$$

$$\hat{Y} - 72.6 = 0.64x - 46.214$$

$$\hat{Y} = 0.664x + 26.386$$

Regression equation of X on Y,

$$\hat{X} - \bar{x} = b_{XY}\left(y - \bar{y}\right)$$

Calculation of b_{XY}

$$b_{XY} = b_{UV}\frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2}$$

$$= \frac{5 \times 970 - 48 \times (-12)}{5 \times 1074 - (-12)^2}$$

$$= \frac{4850 + 576}{5370 - 144} = \frac{5426}{5226}$$

$$b_{UV} = 1.038$$

$$\hat{X} - 69.6 = 1.038\ (y - 72.6)$$

$$\hat{X} - 69.6 = 1.038y - 75.359$$

$$\hat{X} = 1.038y - 5.759$$

Example 5.6

Is there any mistake in the data provided about the two regression lines $Y = -1.5\,X + 7$, and $X = 0.6\,Y + 9$? Give reasons.

Solution:

The regression coefficient of Y on X is $b_{YX} = -1.5$

The regression coefficient of X on Y is $b_{XY} = 0.6$

Both the regression coefficients are of different sign, which is a contrary. So the given equations cannot be regression lines.

Example: 5.7

	mean	S.D
Yield of wheat (kg. unit area)	10	8
Annual Rainfall (inches)	8	2

Correlation coefficient: 0.5

Estimate the yield when rainfall is 9 inches

Solution:

Let us denote the dependent variable yield by Y and the independent variable rainfall by X.

Regression equation of Y on X is given by

$$Y - \bar{y} = r_{XY}\frac{SD(Y)}{SD(X)}(x - \bar{x})$$

$\bar{x} = 8,\ SD(X) = 2, \bar{y} = 10,\ SD(Y) = 8, \quad r_{XY} = 0.5$

$$Y - 10 = 0.5 \times \frac{8}{2}(x - 8)$$

$$= 2\,(x - 8)$$

When $x = 9$,

$$Y - 10 = 2\,(9 - 8)$$

$$Y = 2 + 10$$

$$= 12 \text{ kg (per unit area)}$$

Corresponding to the annual rain fall 9 inches the expected yield is 12 kg (per unit area).

Example 5.8

For 50 students of a class the regression equation of marks in Statistics (X) on marks in Accountancy (Y) is $3Y - 5X + 180 = 0$. The mean marks in of Accountancy is 50 and variance of marks in statistics is $\frac{16}{25}$ of the variance of marks in Accountancy.

Find the mean marks in statistics and the coefficient of correlation between marks in the two subjects when the variance of Y is 25.

Solution:

We are given that:

n = 50, Regression equation of X on Y as $3Y - 5X + 180 = 0$

$\bar{y} = 50$, $V(X) = \frac{16}{25}V(Y)$, and $V(Y) = 25$.

We have to find (i) $\bar{x}$ and (ii) r_{XY}

(i) Calculation for $\bar{x}$

Since $(\bar{x}, \bar{y})$ is the point of intersection of the two regression lines, they lie on the regression line $3Y - 5X + 180 = 0$

Hence, $3\bar{y} - 5\bar{x} + 180 = 0$

$$3(50) - 5\bar{x} + 180 = 0$$

$$-5\bar{x} = -180 - 150$$

$$= -330$$

$$\bar{x} = \frac{-330}{-5} = 66$$

$$\bar{x} = 66$$

(ii) Calculation for coefficient of correlation.

$$3Y - 5X + 180 = 0$$

$$-5X = -180 - 3Y$$

$$X = 36 + 0.6\,Y$$

$$b_{XY} = 0.6$$

Also $b_{XY} = r_{XY}\frac{SD(X)}{SD(Y)}$

$$0.6 = r_{XY}\frac{SD(X)}{SD(Y)}$$

$$r_{XY} = \frac{0.6 \times SD(Y)}{SD(X)}$$

$$r_{XY}^2 = 0.36 \times \frac{V(Y)}{V(X)} \qquad (1)$$

Given that:

$$V(Y) = 25$$

$$V(X) = \frac{16}{25}V(Y)$$

$$= \frac{16}{25} \times 25$$

$$V(X) = 16$$

Substituting in (1) we have

$$r_{XY}^2 = \frac{0.36 \times 25}{16}$$

$$r_{XY} = \sqrt{\frac{0.36 \times 25}{16}} = 0.75$$

Example 5.9

If two regression coefficients are $b_{YX} = \frac{5}{6}$ and $b_{XY} = \frac{9}{20}$, what would be the value of r_{XY}?

Solution:

The correlation coefficient $r_{XY} = \pm\sqrt{(b_{YX})(b_{XY})}$

$$= \pm\sqrt{\frac{5}{6} \times \frac{9}{20}} = \sqrt{0.375} = 0.6124$$

Since both the signs in b_{YX} and b_{XY} are positive, correlation coefficient between X and Y is positive.

Example 5.10

Given that $b_{YX} = -\frac{8}{7}$ and $b_{XY} = -\frac{5}{6}$. Find r?

NOTE

The sign of the corelation coefficient will be the signs of the regression coefficients.

Solution:

$$r_{XY} = \pm\sqrt{(b_{YX})(b_{XY})}$$

$$= \sqrt{-\frac{8}{7} \times -\frac{5}{6}} = \sqrt{\frac{20}{21}} = -0.9759.$$

Since both the signs in b_{YX} and b_{XY} are negative, correlation coefficient between X and Y is negative.

5.6 DIFFERENCE BETWEEN CORRELATION AND REGRESSION

Correlation	Regression
1. It indicates only the nature and extent of linear relationship	It is the study about the impact of the independent variable on the dependent variable. It is used for predictions.
2. If the linear correlation is coefficient is positive / negative , then the two variables are positively / or negatively correlated	The regression coefficient is positive, then for every unit increase in x, the corresponding average increase in y is b_{YX}. Similarly, if the regression coefficient is negative , then for every unit increase in x, the corresponding average decrease in y is b_{YX}.
3. One of the variables can be taken as x and the other one can be taken as the variable y.	Care must be taken for the choice of independent variable and dependent variable. We can not assign arbitrarily x as independent variable and y as dependent variable.
4. It is symmetric in x and y, *ie.*, $r_{XY}=r_{YX}$	It is not symmetric in x and y, that is, b_{XY} and b_{YX} have different meaning and interpretations.

POINTS TO REMEMBER

- There are several types of regression - Simple linear correlation , multiple linear correlation and non-linear correlation.
- In simple linear regression there are two linear regression lines Y on X and X on Y.
- In the linear regression line $Y = a + bX + e$, where 'X' is independent variable, 'Y' is dependent variable, a' is intercept, 'b' is slope of the line and ' e' is error term.
- The point $(\overline{X},\overline{Y})$ passes through the regression lines.
- The " Method of least squares" gives the line of best fit.
- Both the regression lines have the same sign either positive of negative.
- The sign of the regression coefficient and the sign of the correlation coefficient is same.

EXERCISE 5

I. Choose the best answer.

1. ______ is widely used for prediction

   a) regression analysis b) correlation analysis

   c) analysis of variance d) analysis of covariance

2. The linear regression analysis can be classified in to

   a) 4 types b) 3 types c) 2 types d) none of the above

3. The linear equation $Y = a + bx$ is called as regression equation of

   a) X on Y b) Y on X c) between X and Y d) 'a' on 'b'

4. In regression equation $X = a + by + e$ is

   a) correlation coefficient of Y on X b) correlation coefficient of X on Y

   c) regression coefficient of Y on X d) regression coefficient of X on Y

5. $b_{YX} =$

   a) $r_{XY}\dfrac{SD(X)}{SD(Y)}$ b) $r_{XY}\dfrac{SD(Y)}{SD(X)}$ c) $\dfrac{SD(X)}{SD(Y)}$ d) $\dfrac{SD(Y)}{SD(X)}$

6. If $b_{XY} > 1$ then b_{YX} is

   a) 1 b) 0 c) > 1 d) < 1

7. In the Regression equation $\hat{Y}-\bar{y}=r_{XY}\dfrac{SD(Y)}{SD(X)}\left(x-\bar{x}\right)$, $r_{XY}\dfrac{SD(Y)}{SD(X)}$ is

   a) b_{YX} b) b_{XY} c) r_{XY} d) $\text{cov}(X,Y)$

8. Using the regression coefficients we can calculate

   a) $\text{cov}(X, Y)$ b) $SD(X)$

   c) correlation coefficient d) coefficient of variance

9. Arithmetic mean of the regression coefficients b_{XY} and b_{YX} is

   a) $> r_{XY}$ b) $\geq r_{XY}$ c) $\leq r_{XY}$ d) $< r_{XY}$

10. Regression analysis helps in establishing a functional relationship between ______ variables.

    a) 2 or more variables b) 2 variables

    c) 3 variables d) none of these

11. _____ is the Father of mental tests

    a) R.A. Fisher b) Croxton and Cowden

    c) Francis Galton d) A.L. Bowley

12. Correlation coefficient is the _______ between the regression coefficients

a) arithmetic mean　　b) geometric mean

c) harmonic mean　　d) none of the above

13. If the two lines of regression are perpendicular to each other then $r_{XY} =$

a) 0　　b) 1　　c) –1　　d) 0.5

14. If the two regression lines are parallel then

a) $r_{XY} = 0$　　b) $r_{XY} = +1$　　c) $r_{XY} = -1$　　d) $r_{XY} = \pm 1$

15. Angle between the two regression lines is

a) $\tan^{-1}\left(\frac{m_1 + m_2}{1 - m_1 m_2}\right)$　　b) $\tan^{-1}\left(\frac{m_1 m_2}{1 + m_1 m_2}\right)$

c) $\tan^{-1}\left(\frac{m_1 - m_2}{1 + m_1\ m_2}\right)$　　d) none of the above

16. $b_{XY} =$

a) $r_{XY}\frac{SD(Y)}{SD(X)}$　　b) $r_{XY}\frac{SD(X)}{SD(Y)}$

c) $r_{XY}\, SD(X)\ SD(Y)$　　d) $\frac{1}{b_{YX}}$

17. Regression equation of X on Y is

a) $Y = a + b_{YX}x + e$　　b) $Y = b_{XY}x + a + e$

c) $X = a + b_{XY}y + e$　　d) $X = b_{YX}y + a + e$

18. For the regression equation $2\hat{Y} = 0.605x + 351.58$. The regression coefficient of Y on X is

a) $b_{XY} = 0.3025$　　b) $b_{XY} = 0.605$

c) $b_{YX} = 175.79$　　d) $b_{YX} = 351.58$

19. If $b_{XY} = 0.7$ and ‘a’ = 8 then the regression equation of X on Y is

a) $Y = 8 + 0.7\ X$　　b) $X = 8 + 0.7\ Y$

c) $Y = 0.7 + 8\ X$　　d) $X = 0.7 + 8\ Y$

20. The regression lines intersect at

a) $(\overline{X}, \overline{Y})$　　b) (X, Y)　　c) $(0, 0)$　　d) $(1, 1)$

II. Give very short answer to the following questions.

21. Define regression.

22. What are the types of regression?

23. Write the two simple linear regression equations.

24. Write the two simple linear regression coefficients.

25. If the regression coefficient of X on Y is 16 and the regression coefficient of Y on X is 4, then find the correlation coefficient.

26. Find the standard deviation of Y given that $V(X)$ is 36, $b_{XY} = 0.8$, $r_{XY} = 0.5$.

III. Give short answer to the following questions.

27. Define simple linear and multiple linear regressions

28. Distinguish between linear and non-linear regression.

29. Write the regression equation of X on Y and its normal equations.

30. Write the regression equation of Y on X and its normal equations.

31. Write any three properties of regression.

32. Write any three uses of regression.

33. Write any three differences between correlation and regression.

34. If the regression equations are $\hat{X} = 64 - 0.95y$, $\hat{Y} = 7.25 - 0.95x$ then find the correlation coefficient.

35. Given the following lines of regression.

    $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$. Find the mean values of X and Y.

36. Given $\overline{x} = 90$, $\overline{y} = 70$, $b_{XY} = 1.36$, $b_{YX} = 0.61$ when $y = 50$, Find the most probable value of X.

37. Compute the two regression equations from the following data.

x	1	2	3	4	5
y	3	4	5	6	7

If $x = 3.5$ what will be the value of $\hat{Y}$?

IV Give detailed answer to the following questions.

38. Write in detail the properties of regression.

39. Explain in detail the uses of regression analysis.

40. Distinguish between correlation and regression.

41. Interpret the result for the given information. A simple regression line is fitted for a data set and its intercept and slope respectively are 2 and 3. Construct the linear regression of the form $Y = a + bx$ and offer your interpretation for 'a' and 'b'. If X is increased from 1 to 2, what is the increase in Y value. Further if X is increased from 2 to 5 what would be the increase in Y. Demonstrate your answer mathematically.

42. Using the method of least square, calculate the regression equation of X on Y and Y on X from the following data and estimate X where Y is 16.

x	10	12	13	17	18
y	5	6	7	9	13

Also determine the value of correlation coefficient.

43. The following table shows the age (X) and systolic blood pressure (Y) of 8 persons.

Age (X)	56	42	60	50	54	49	39	45
Blood pressure (Y)	160	130	125	135	145	115	140	120

Fit a simple linear regression model, Y on X and estimate the blood pressure of a person of 60 years.

44. Find the regression equation of X on Y given that $n = 5$, $\sum x = 30$, $\sum y = 40$, $\sum xy = 214$, $\sum x^2 = 220$, $\sum y^2 = 340$.

45. Given the following data, estimate the marks in statistics obtained by a student who has scored 60 marks in English.

Mean of marks in Statistics = 80, Mean of marks in English = 50, S.D of marks in Statistics = 15, S.D of marks in English = 10 and Coefficient of correlation = 0.4.

46. Find the linear regression equation of percentage worms (Y) on size of the crop (X) based on the following seven observations.

Size of the crop (X)	16	15	11	27	39	22	20
Percentage worms (Y)	24	25	34	40	35	20	23

47. In a correlation analysis, between production (X) and price of a commodity (Y) we get the following details.

Variance of $X = 36$.

The regression equations are:

$12X - 15Y + 99 = 0$ and $60\ X - 27\ Y = 321$

Calculate (a) The average value of X and Y.

(b) Coefficient of correlation between X and Y.

ANSWERS

I.	1. a)	2. c)	3. b)	4. d)	5. b)
	6. d)	7. a)	8. c)	9. b)	10. a)
	11. c)	12. b)	13. a)	14. d)	15. c)
	16. b)	17. c)	18. a)	19. b)	20. a)

II. 25) $r_{XY} = 8$

26) $SD(Y) = 3.75$

III. 34) $r_{XY} = -0.95$

35) $\overline{X} = 13$, $\overline{Y} = 17$

36) when $Y = 50$, $\hat{X} = 62.8$

37) Regression equation X on Y: $\hat{X} = Y - 2$
Regression equation Y on X: $\hat{Y} = X + 2$
when $X = 3.5$, $\hat{Y} = 5.5$

IV. 41) (1) If X increases by 1 unit then Y increases by 3 units
(2) If X increases by 3 units then Y increases by 9 units

42) (1) Regression equation of X on Y is $X = Y + 6$; when $Y = 16$, $X = 22$
(2) Regression equation of Y on X is $Y = 0.89\ X - 2.59$
(3) $b_{xy} = 1$, $b_{yx} = 0.87$, $r = 0.93$

43) $Y = 0.45\ X + 111.53$, $Y = 138.53$ when age is 60 years.

44) $a = 16.4$, $b = -1.3$
Regression equation of X on Y is : $X = 16.4 - 1.3Y$

45) $X = 86$ when $Y = 60$

46) $Y = 0.32\ X + 21.84$

47) (a) Mean of $X = 13$ and mean of $Y = 17$. (b) $r = 0.6$

CHAPTER 6 INDEX NUMBERS

Irving Fisher

Irving Fisher (1867–1947) was an American Statistician born in New York and his father was a teacher. As a child, he had remarkable mathematical ability and a flair for invention. In 1891, Fisher received the first Ph.D in economics from Yale University. Fisher had shown particular talent and inclination for mathematics, but he found that economics offered greater scope for his ambition and social concerns. He made important contributions to economics including index numbers. He edited the *Yale Review* from 1896 to 1910 and was active in many learned societies, institutes, and welfare organizations. He was a president of the American Economic Association. He died in New York City in 1947, at the age of 80.

LEARNING OBJECTIVES

The students will able to

- understand the concept and purpose of Index Numbers.
- calculate the indices to measure price and quantity changes over period of time.
- understand different tests an ideal Index Number satisfies.
- understand consumer price Index Numbers.
- understand the limitations of the construction of Index Numbers.

ZDMCW

Introduction

Index number is a technique of measuring changes in a variable or a group of variables with respect to time, location or other characteristics. It is one of the most widely used statistical methods. Index number is a specialized average designed to measure the change in a group of related variables over a period of time. For example, the price of cotton in 2010 is studied with reference to its price in 2000. It is used to feel the pulse of the economy and it reveals the inflationary or deflationary tendencies. In reality, it is viewed as barometers of economic activity because if one wants to have an idea as to what is happening in an economy, he should check the important indicators like the index number of agricultural production, index number of industrial production, and the index number business activity *etc*., There are several types of index numbers and the students will learn them in this chapter.

6.1 DEFINITION AND USES OF INDEX NUMBERS

6.1.1 Definition

An Index Number is defined as a relative measure to compare and describe the average change in price, quantity value of an item or a group of related items with respect to time, geographic location or other characteristics accordingly.

In the words of ***Maslow*** *"An index number is a numerical value characterizing the change in complex economic phenomenon over a period of time or space"*

Spiegal *defines, "An index number is a statistical measure designed to show changes in a variable on a group of related variables with respect to time, geographical location or other characteristics".*

According to ***Croxton and Cowden*** *"Index numbers are devices for measuring differences in the magnitude of a group of related variables".*

Bowley *describes "Index Numbers as a series which reflects in its trend and fluctuations the movements of some quantity".*

6.1.2 Uses

The various uses of index numbers are:

Economic Parameters

The Index Numbers are one of the most useful devices to know the pulse of the economy. It is used as an indicator of inflanationary or deflanationary tendencies.

Measures Trends

Index numbers are widely used for measuring relative changes over successive periods of time. This enable us to determine the general tendency. For example, changes in levels of prices, population, production etc. over a period of time are analysed.

Useful for comparsion

The index numbers are given in percentages. So it is useful for comparison and easy to understand the changes between two points of time.

Help in framing suitable policies

Index numbers are more useful to frame economic and business policies. For example, consumer price index numbers are useful in fixing dearness allowance to the employees.

Useful in deflating

Price index numbers are used for connecting the original data for changes in prices. The price index are used to determine the purchasing power of monetary unit.

Compares standard of living

Cost of living index of different periods and of different places will help us to compare the standard of living of the people. This enables the government to take suitable welfare measures.

Special type of average

All the basic ideas of averages are employed for the construction of index numbers. In averages, the data are homogeneous (in the same units) but in index number, we average the variables which have different units of measurements. Hence, it is a special type of average.

6.2 TYPES OF INDEX NUMBERS

(i) **Price Index Numbers**

Price index is a 'Special type' of average which studies net relative change in the prices of commodities, expressed in different units. Here comparison is made in respect of prices. Price index numbers are wholesale price index numbers and retail price index numbers.

(i) **Quantity Index Numbers**

This number measures changes in volume of goods produced, purchased or consumed. Here, the comparison is made in respect of quantity or volume. For example, the volume of agricultural goods produced, consumed, import, export etc.

(ii) **Value Index**

Value index numbers study the changes in the total value of a certain period with the total value of the base period. For example, the indices of stock-in-made, purchase, sales profit *etc.,* are analysed here.

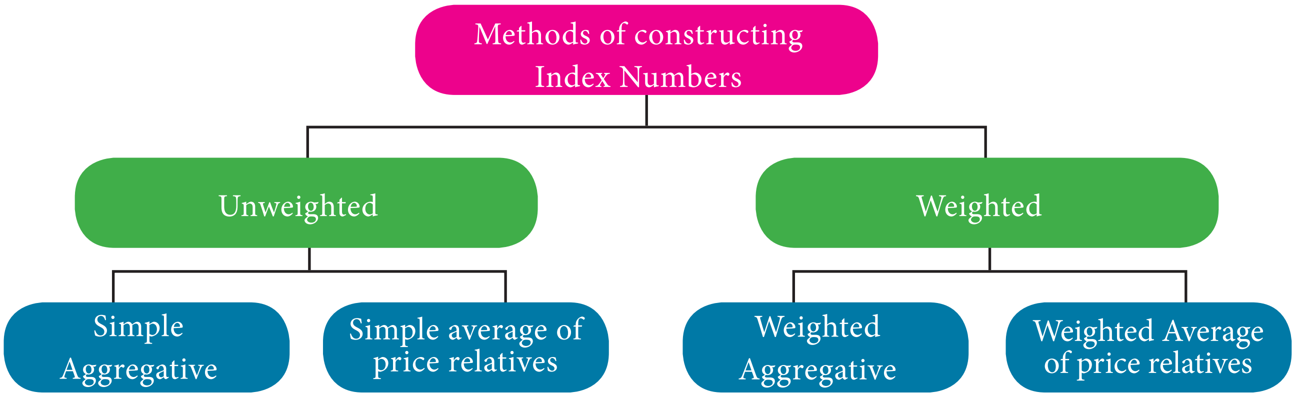
NOTE

The points and precautions that should be taken in the constructing index numbers are:

- determination of the purpose
- selection of the base period
- selection of commodities
- selection of price quotations
- selection of appropriate weight
- selection of an appropriate average
- selection of an appropriate formula

6.3 METHODS OF CONSTRUCTING INDEX NUMBERS

Different types of index number (price/quantity/value) can be classified as follows.


Methods of constructing Index Numbers
Unweighted
Weighted
Simple Aggregative
Simple average of price relatives
Weighted Aggregative
Weighted Average of price relatives

6.3.1 Unweighted Index Numbers

An unweighted price Index Number measures the percentage change in price of a single item or a group of items between two periods of time. In unweighted index numbers, all the values taken for study are of equal importance. There are two methods in this category.

(i) Simple aggregative method:

Under this method the prices of different items of current year are added and the total is divided by the sum of prices of the base year items and multiplied by 100.

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

p_1 = Current year prices for various commodities

p_0 = Base year prices for various commodities

P_{01} = Price Index number

Limitations of the simple aggregative method

(i) Relative importance of the commodities is not taken into account.

(ii) Highly priced items influence the index number

NOTE

The base period is the period against which comparison is made. Generally a year is taken as base period. The base period should be free from economic and natural disturbances.

Example 6.1

Construct the Price Index Number for the year 1997, from the following information taking 1996 as base year.

Commodities	Price in 1996 (₹)	Price in 1997 (₹)
Rice	130	115
Wheat	80	65
Sugar	75	70
Ragi	95	90
Oil	105	105
Dal	35	20

Solution:

Construction of Price Index:

Commodities	Price in 1996 (₹) (p_0)	Price in 1997 (₹) (p_1)
Rice	130	115
Wheat	80	65
Sugar	75	70
Ragi	95	90
Oil	105	105
Dal	35	20
	$\sum p_0 = 520$	$\sum p_1 = 465$

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

$$= \frac{465}{520} \times 100 = 89.42$$

Price Index in 1997, when compared to 1996 has fallen by 10.58%

Example 6.2

Calculate Price Index Number for 2016 from the following data by simple aggregate method, taking 2016 as base year.

Commodities	Price per kg	
	2015	2016
Apple	100	140
Orange	30	40
Pomegranate	120	130
Guava	40	50

Solution:

Commodities	2015 (p_0)	2016 (p_1)
Apple	100	140
Orange	30	40
Pomegranate	120	130
Guava	40	50
Total	290	360

Price index: $P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$

$$= \frac{360}{290} \times 100$$

$$= \frac{3600}{29}$$

$$P_{01} = 124.13\%$$

Price index for the year 2016 when compared to 2015 has been increased by 24.13%.

2. Simple average of price relative method

Under this method, first of all price relatives are obtained for the various items and then average of these relatives is obtained by using either arithmetic mean or geometric mean. Price relative is the price of the current year expressed as the percentage of the price of the base year. The formula for computing Index Number under this method on using Arithmetic mean and Geometric mean are given below.

If N is the member of items, p_1 is the price of the commodity with current year and p_0 is the price of the commodity in the base year then, the average Price Index Number is

(i) $P_{01} = \dfrac{\sum \dfrac{p_1}{p_0} \times 100}{N}$ (using Arithmetic mean)

(ii) P_{01} = antilog $\dfrac{\sum \log\left(\dfrac{p_1}{p_0} \times 100\right)}{N}$ (using Geometric mean)

Advantages of Average Price Index

1. It is not influenced by the extreme prices of items as equal importance is given to all items.
2. Price relatives are pure numbers; therefore the value of the average price relative index is not affected by the units of measurement of commodities included in the calculation of index numbers.

Limitations

1. Equal weights are assigned to every commodity included in the index. Each price relatives is given equal importance, but in actual practice, it is not true.
2. Arithmetic mean is very often used to calculate the average price relatives, but it has a few disadvantages. The use of geometric mean is difficult to calculate.

Example 6.3

Compute price index number by simple average of price relatives method using arithmetic mean and geometric mean.

Item	Price in 2001 (₹)	Price in 2002 (₹)
A	6	10
B	2	2
C	4	6
D	10	12
E	8	12

Solution:

Calculation of price index number by simple average of price relatives:

Item	Price in 2001 (₹) p_0	Price in 2002 (₹) p_1	$p = \frac{p_1}{p_0} \times 100$	log p
A	6	10	166.7	2.2219
B	2	2	100.0	2.0000
C	4	6	150.0	2.1761
D	10	12	120.0	2.0792
E	8	12	150.0	2.1761
			$\sum p = 686.7$	$\sum \log p = 10.6533$

(i) Price relative index number based on arithmetic mean:

$$P_{01} = \frac{\sum \frac{p_1}{p_0} \times 100}{N} = \frac{\sum p}{N} = \frac{686.7}{5} = 137.34$$

(ii) Price relative index number based on geometric mean:

$$P_{01} = \text{antilog}\left(\frac{\sum \log p}{N}\right) = \text{antilog}\left(\frac{10.6533}{5}\right)$$

$$= \text{antilog}\,(2.13066)$$

$$= 135.1$$

Hence, the price index number based on arithmetic mean and geometric mean for the year 2002 are 137.34 and 135.1 respectively.

Example 6.4

Construct simple average price relative index number using arithmetic mean for the year 2012 for the following data showing the profit from various categories sold out in departmental stores.

Profit (per week)	2010	2012
Groceries	150600	170800
Cosmetics	70000	82000
Stationery items	12000	10800
Utensils	20000	18600

Solution: Index number uning Arithmertic Mean of price relatives

	Profit in 2010 (p_0)	Profit in 2012 (p_1)	p_1/p_0 x 100
Groceries	150600	170800	$\frac{170800}{150600} \times 100 = 11341$
Cosmetics	70000	82000	$\frac{82000}{70000} \times 100 = 117.14$
Stationery items	12000	10800	$\frac{10800}{12000} \times 100 = 90.00$
Utensils	20000	18600	$\frac{18600}{20000} \times 100 = 93.00$
		Total	413.55

Simple average price relatives using A.M = P_{01} $= \dfrac{\sum \dfrac{p_1}{p_0} \times 100}{N}$

$$= \frac{413.55}{4}$$

$$=103.3875$$

$$P_{01} = 103.39$$

The average price relative index number using arithmetic mean for the year 2012 is 103.39

Example 6.5

Construct simple average price relative index number using geometric mean for the year 2015 for the data showing the expenditure in education of the children taking different courses.

Expenditure per year	2014	2015
B.Sc	24000	26000
B.Com	20000	22000
B.E	108000	12000
M.B.B.S	150000	168000

Solution:

Expenditure	Year 2014 (p_0)	Year 2015 (p_1)	$P = (p_1/p_0) \times 100$	log P
B.Sc	24000	26000	$\frac{26000}{24000} \times 100 = 108.33$	2.0346
B.Com	20000	22000	$\frac{22000}{20000} \times 100 = 110.00$	2.0414
B.E	108000	12000	$\frac{120000}{108000} \times 100 = 111.11$	2.0457
M.B.B.S	150000	168000	$\frac{168000}{150000} \times 100 = 112.00$	2.0492
				$\sum \log P = 8.1709$

$$P_{01} = \text{antilog}\left(\frac{\sum \log P}{N}\right)$$

$$= \text{antilog}\left(\frac{8.1709}{4}\right)$$

$$= \text{antilog}\ (2.04275)$$

$$= \text{antilog}\ (2.0428)$$

$$= 110.4$$

The average price relative index number using geometric mean for the year 2015 is 110.4

6.4 WEIGHTED INDEX NUMBERS

In computing weighted Index Numbers, the weights are assigned to the items to bring out their economic importance. Generally quanties consumed or value are used as weights.

Weighted index numbers are also of two types

(i) Weighted aggregative

(ii) Weighted average of price relatives

6.4.1 Weighted aggregate Index Numbers

In this method price of each commodity is weighted by the quantity sale either in the base year or in the current year. There are various methods of assigning weights and thus there are many methods of constructing index numbers. Some of the important formulae used under this methods are

a) Laspeyre's Index ($P_{01}{}^{L}$)

b) Paasche's Index ($P_{01}{}^{P}$)

c) Dorbish and Bowley's Index ($P_{01}{}^{DB}$)

d) Fisher's Ideal Index ($P_{01}{}^{F}$)

e) Marshall-Edgeworth Index ($P_{01}{}^{Em}$)

f) Kelly's Index ($P_{01}{}^{K}$)

a. Laspeyre's method

The base period quantities are taken as weights. The Index is

$$P_{01}^{L}=\frac{\sum p_1q_0}{\sum p_0q_0}\times 100$$

b. Paasche's method

The current year quantities are taken as a weight. In this method, we use continuously revised weights and thus this method is not frequently used when the number of commodities is large. The Index is

$$P_{01}^{P}=\frac{\sum p_1q_1}{\sum p_0q_1}\times 100$$

c. Dorbish and Bowley's method

In order in take into account the impact of both the base and current year, we make use of simple arithmetic mean of Laspeyre's and Paasche's formula

The Index is

$$P_{01}^{DB}=\frac{P_{01}^{L}+P_{01}^{P}}{2}$$

$$= \frac{\dfrac{\sum p_1q_0}{\sum p_0q_0} + \dfrac{\sum p_1q_1}{\sum p_0q_1}}{2} \times 100$$

d. Fisher's Ideal Index

It is the geometric mean of Laspeyre's Index and Paasche's Index, given by:

$$P_{01}^{F} = \sqrt{P_{01}^{L} \times P_{01}^{P}}$$
$$= \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$$

e. Marshall-Edgeworth method

In this method also both the current year as well as base year prices and quantities are considered.

The Index is

$$P_{01}^{ME} = \frac{\sum p_1(q_0+q_1)}{\sum p_0(q_0+q_1)} \times 100$$
$$= \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100$$

f. Kelly's method

The Kelly's Index is

$$P_{01}^{K} = \frac{\sum p_1q}{\sum p_0q} \times 100, \qquad q = \frac{q_0+q_1}{2}$$

where q refers to quantity of some period, not necessarily of the mean of the base year and current year. It is also possible to use average quantity of two or more years as weights. This method is known as fixed weight aggregative index.

Example 6.6

Construct weighted aggregate index numbers of price from the following data by applying

1. Laspeyre's method
2. Paasche's method
3. Dorbish and Bowley's method
4. Fisher's ideal method
5. Marshall-Edgeworth method

Commodity	2016		2017	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution:

Calculation of various indices

Commodity	2016		2017		p_1q_0	p_0q_0	p_1q_1	p_0q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
A	2	8	4	6	32	16	24	12
B	5	10	6	5	60	50	30	25
C	4	14	5	10	70	56	50	40
D	2	19	2	13	38	38	26	26
					$\sum p_1q_0 = 200$	$\sum p_0q_0 = 160$	$\sum p_1q_1 = 130$	$\sum p_0q_1 = 103$

(1) Laspeyre's Index:

$$P_{01}^{L} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100$$

$$= \frac{200}{160} \times 100 \;= 125$$

(2) Paasche's Index

$$P_{01}^{P} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100$$

$$= \frac{130}{103} \times 100 \;= 126.21$$

(3) Dorbish and Bowley's Index

$$P_{01}^{DB} = \frac{P_{01}^{L} + P_{01}^{P}}{2} = \frac{125 + 126.21}{2}$$

$$= 125.6$$

(4) Fisher's Ideal Index

$$P_{01}^{F} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$$

$$= \sqrt{\frac{200}{160} \times \frac{130}{103}} \times 100$$

$$= \sqrt{1.578} \times 100 = 1.2561 \times 100$$

$$= 125.61$$

(5) Marshall-Edgeworth method

$$P_{01}^{ME} = \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100$$

$$= \frac{200+130}{160+103} \times 100 = \frac{330}{263} \times 100$$

$$= 125.48$$

Example 6.7

Calculate the price indices from the following data by applying (1) Laspeyre's method (2) Paasche's method and (3) Fisher ideal number by taking 2010 as the base year.

Commodity	2010		2011	
	Prices	Quantities	Prices	Quantities
A	20	10	25	13
B	50	8	60	7
C	35	7	40	6
D	25	5	35	4

Solution: Calculations

p_0	q_0	p_1	q_1	$p_0\,q_0$	$p_0\,q_1$	$p_1\,q_0$	$p_1\,q_1$
20	10	25	13	200	260	250	325
50	8	60	7	400	350	480	420
35	7	40	6	245	210	280	240
25	5	35	4	125	100	175	140
				970	920	1185	1125

(1) Laspeyre's Index

$$P_{01}^{L} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{1185}{970} \times 100$$

$$= 122.16$$

(2)) Paasche's Index

$$P_{01}^{P} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{11254}{920} \times 100$$

$$= 122.28$$

(3) Fisher's Ideal Index

$$P_{01}^{F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{1185}{970} \times \frac{1125}{920}} \times 100$$

$$= \sqrt{1.2216 \times 1.2228} \times 100$$

$$= \sqrt{1.49377} \times 100$$

$$= 1.2222 \times 100$$

$$= 122.22$$

Example 6.8

Calculate the Dorbish and Bowley's price index number for the following data taking 2014 as base year.

Items	2014		2015	
	Prices (per kg)	Quantities (purchased)	Prices (per kg)	Quantities (purchased)
Oil	80	3	100	4
Pulses	35	2	45	3
Sugar	25	2	30	3
Rice	50	30	54	35
Cereals	35	2	40	3

Solution: Price Index by Dorbish and Bowley's Method

p_0	q_0	p_1	q_1	$p_0\,q_0$	$p_0\,q_1$	$p_1\,q_0$	$p_1\,q_1$
80	3	100	4	240	320	300	400
35	2	45	3	70	105	90	135
25	2	30	3	50	75	60	90
50	30	54	35	1500	1750	1620	1890
35	2	40	3	70	105	80	120
				1930	2355	2150	2635

$$P_{01}^{DB} = \frac{\frac{\sum p_1q_0}{\sum p_0q_0} + \frac{\sum p_1q_1}{\sum p_0q_1}}{2} \times 100$$

$$= \frac{1}{2}\left[\frac{2150}{1930} + \frac{2635}{2355}\right] \times 100$$

$$= \frac{1}{2}\left[1.1139 + 1.1188\right] \times 100$$

$$= \frac{1}{2}\left[2.2327\right] \times 100$$

$= 1.1164$ x $100 = 111.64$

Example 6.9

Compute Marshall – Edgeworth price index number for the following data by taking 2016 as base year.

Items sold out in a men's wear	2016		2017	
	Prices	Quantity	Prices	Quantity
Shirts	700	150	900	175
Pants	1000	100	1200	150
Sandals	500	70	600	100
Shoes'	1500	50	1800	60
Belts	400	100	600	150
Watches	1200	300	1500	250

Solution: Price Index by Marshall-Edgeworth Method

p_0	q_0	p_1	q_1	$p_0\,q_0$	$p_0\,q_1$	$p_1\,q_0$	$p_1\,q_1$
700	150	900	175	105000	122500	135000	157500
1000	100	1200	150	100000	150000	120000	180000
500	70	600	100	35000	50000	42000	60000
1500	50	1800	60	75000	90000	90000	108000
400	100	600	150	40000	60000	60000	90000
1200	300	1500	250	360000	300000	450000	375000
				715000	772500	897000	970500

Marshall – Edgeworth Index:

$$P_{01}^{ME} = \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100$$

$$= \frac{897000 + 970500}{715000 + 772500} \times 100$$

$$= \frac{1867500}{1487500} \times 100$$

$$= 125.55$$

Example 6.10

Calculate a suitable price index form the following data.

Commodity	Quantity	Price	
		2007	2010
X	25	3	4
Y	12	5	7
Z	10	6	5

Solution:

In this problem, the quantities for both current year and base year are same. Hence, we can conlude Kelly's Index price number.

Commodity	q	p_0	p_1	p_0q	p_1q
X	25	3	4	75	100
Y	12	5	7	60	84
Z	10	6	5	60	50
				195	234

Kelly's price Index number:

$$P_{01}^{K} = \frac{\sum p_1q}{\sum p_0q} \times 100$$

$$= \frac{234}{195} \times 100$$

$$= 120$$

6.4.2 Weighted average of price relatives

The weighted average of price relatives can be computed by introducing weights into the unweighted price relatives. Here also, we may use either arithmetic mean or the geometric mean for the purpose of averaging weighted price relatives.

The weighted average price relatives using arithmetic mean:

If $p = \frac{p_1}{p_0} \times 100$ is the price relative index and $w = p_0q_0$ is attached to the commodity, then the weighed price relative index is

$$P_{01} = \frac{\sum \left[\frac{p_1}{p_0} \times 100 \right] \times p_0 q_0}{\sum p_0 q_0} \quad = \quad P_{01} = \frac{\sum wp}{\sum w}$$

The weighted average price relatives using geometric mean:

$$P_{01} = \text{antilog}\left(\frac{\sum w \log p}{\sum w} \right)$$

Example 6.11

Compute price index for the following data by applying weighted average of price relatives method using (i) Arithmetic mean and (ii) Geometric mean.

Item	p_0	q_0	p_1
Wheat	3.0	20 kg	4.0
Flour	1.5	40 kg	1.6
Milk	1.0	10 kg	1.5

Solution:

(i) Computation for the weighted average of price relatives using arithmatic mean.

Item	p_0	q_0	p_1	w	p	log p	wp	w log p
Wheat	3.0	20	4.0	60	133.3	2.1249	7998	127.494
Flour	1.5	40	1.6	60	106.7	2.0282	6402	121.692
Milk	1.0	10	1.5	10	150.0	2.1761	1500	21.761
				$\sum w = 130$			$\sum wp = 15900$	$\sum w \log p = 270.947$

$$P_{01} = \frac{\sum wp}{\sum w} = \frac{15,900}{130} = 122.31$$

This means that there has been a 22.31 % increase in prices over the base year.

(ii) Index number using geometric mean of price relatives is:

$$P_{01} = \text{Antilog}\ \frac{\sum w \log p}{\sum w} = \text{Antilog}\ \frac{270.947}{130}$$

$$= \text{Antilog}\ (2.084) = 121.3$$

This means that there has been a 21.3 % increase in prices over the base year.

6.4.3 Quantity Index Number

The quantity index number measures the changes in the level of quantities of items consumed, or produced, or distributed during a year under study with reference to another year known as the base year.

Laspeyre's quantity index:

$$Q_{01}^{L} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

Paasche's quantity index

$$Q_{01}^{P} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

Fisher's quantity index

$$Q_{01}^{F} = \sqrt{Q_{01}^{L} \times Q_{01}^{P}}$$

$$= \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$$

These formulae represent the quantity index in which quantities of the different commodities are weighted by their prices.

Example 6.12

Compute the following quantity indices from the data given below:

(i) Laspeyre's quantity index (ii) Paasche's quantity index and (iii) Fisher's quantity index

Commodity	1970		1980	
	Price	Total value	Price	Total value
A	10	80	11	110
B	15	90	9	108
C	8	96	17	340

Solution:

Since we are given the value and the prices, the quantity figures can be obtained by dividing the value by the price for each of the commodities.

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1
A	10	8	11	10	80	88	100	110
B	15	6	9	12	90	54	180	108
C	8	12	17	20	96	204	160	340
	Total				266	342	440	558

(i) Laspeyre's quantity index

$$Q_{01}^{L} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

$$= \frac{440}{266} \times 100$$

$$= 165.4$$

(ii) Paasche's quantity index

$$Q_{01}^{P} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

$$= \frac{558}{342} \times 100$$

$$= 163.15$$

(iii) Fisher's quantity index

$$Q_{01}^{F} = \sqrt{Q_{01}^{L} \times Q_{01}^{P}}$$

$$= \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$$

$$= \sqrt{\frac{440}{266} \times \frac{558}{342}} \times 100$$

$$= 1.6428 \times 100$$

$$= 164.28$$

6.4.4 Tests for Index numbers

Fisher has given some criteria that a good index number has to satisfy. They are called (i) Time reversal test (ii) Factor reversal test (iii) Circular test. Fisher has constructed in such a way that this index number satisfies all these tests and hence it is called Fisher's Ideal Index number.

Time reversal test

Fisher has pointed out that a formula for an index number should maintain time consistency by working both forward and backward with respect to time. This is called time reversal test. Fisher describes this test as follows.

"The test is that the formula for calculating an index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as base or putting in another way the index number reckoned forward should be the reciprocal of that reckoned back ward". A good index number shoud satisfy the time reversal test.

This statement is expressed in the form of equation as $P_{01} \times P_{10} = 1$.

where

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

$$P_{10} = \sqrt{\frac{\sum p_0q_1}{\sum p_1q_1} \times \frac{\sum p_0q_0}{\sum p_1q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1} \times \frac{\sum p_0q_1}{\sum p_1q_1} \times \frac{\sum p_0q_0}{\sum p_1q_0}}$$

Hence, $P_{01} \times P_{10} = \sqrt{1} = 1$

Factor reversal test

This test is also suggested by Fisher According to the factor reversal test, the product of price index and quantity index should be equal to the corresponding value index.

In Fisher's words "Just as each formula should permit the interchange of two times without giving inconsistent results so it ought to permit interchanging the prices and quantities without giving inconsistent results. i.e, the two results multiplied together should give the true ratio".

This statement is expressed as follows:

$$P_{01} \times Q_{01} = \frac{\sum p_1q_1}{\sum p_0q_0}$$

$$\text{Now, } P_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}}$$

$$Q_{01} = \sqrt{\frac{\sum p_0q_1}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_1q_0}}$$

$$\text{Hence, } P_{01} \times Q_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1} \times \frac{\sum p_0q_1}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_1q_0}}$$

$$= \sqrt{\frac{\sum p_1q_1}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_0}}$$

$$= \frac{\sum p_1q_1}{\sum p_0q_0}$$

Circular Test

It is an extension of time reversal test. The time reversal test takes into account only two years. The current and base years. The circular test would require this property to holdgood for any two years. An index number is said to satisfy the circular test when there are three indices, P_{01}, P_{12} and P_{20}, such that $P_{01} \times P_{12} \times P_{20} = 1$.

Laspeyres, Paasche's and Fisher's ideal index numbers do not satisfy this test.

Example 6.13

The table below gives the prices of base year and current year of 5 commodities with their quantities. Use it to verify whether Fisher's ideal index satisfies time reversal test.

Commodity	Base year		Current year	
	Unit price (₹)	Quantity	Unit price (₹)	Quantity
A	4	40	5	60
B	5	50	10	70
C	8	65	12	80
D	6	20	6	90
E	7	30	10	75

Solution:

Index number by Fisher's ideal index method

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	4	40	5	60	160	240	200	300
B	5	50	10	70	250	350	500	700
C	8	65	12	80	520	640	780	960
D	6	20	6	90	120	540	120	540
E	7	30	10	75	210	525	300	750
					1260	2295	1900	3250

$$P_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

$$= \sqrt{\frac{1900}{1260} \times \frac{3250}{2295}}$$

$$P_{10} = \sqrt{\frac{\sum q_0 p_1}{\sum q_1 p_1} \times \frac{\sum q_0 p_0}{\sum q_1 p_0}}$$

$$= \sqrt{\frac{2295}{3250} \times \frac{1260}{1900}}$$

Hence, $P_{01} \times P_{10} = \sqrt{\frac{1900}{1260} \times \frac{3250}{2295} \times \frac{2295}{3250} \times \frac{1260}{1900}}$

$$= \sqrt{1} = 1$$

Fisher's Index number satisties time reveral test.

Example 6.14

Calculate the price index and quantity index for the following data by Fisher's ideal formula and verify that it statisfies the factor reversal test.

Commodity	Base year		Current year	
	Price (₹)	Quantity (`000 tonnes)	Price (₹)	Quantity (`000 tonnes)
A	40	70	40	32
B	50	84	30	80
C	60	58	25	50

Solution

Commodity	p_0	q_0	p_1	q_1	p_1q_1	p_1q_0	p_0q_0	$p_0\,q_1$
A	40	70	40	32	1280	2800	2800	1280
B	50	84	30	80	2400	2520	4200	4000
C	60	58	25	50	1250	1450	3480	3000
					5930	6770	10480	8280

Factor Reversal test: $P_{01} \times Q_{01} = \dfrac{\sum p_1q_1}{\sum p_0q_0}$

$$P_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}}$$

$$Q_{01} = \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1} \times \frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}}$$

$$= \sqrt{\frac{6770}{10480} \times \frac{5930}{8280} \times \frac{8280}{10480} \times \frac{5930}{6770}}$$

$$= \left(\sqrt{\frac{5930}{10480}}\right)^2$$

$$= \frac{5930}{10480}$$

$$= \frac{\sum p_1q_1}{\sum p_0q_0}$$

Hence, Fisher ideal index number satisfies the factor reversal test

6.5 CONSUMER PRICE INDEX NUMBERS

Consumer Price Index Numbers are computed with a view of study the effect of changes in prices on the people as consumers. These indices give the average increase in the expenses if it is designed to maintain the standard of living of base year. General index numbers fail to give an indea about the effect of the change in the general price level on the cost of living of different classes of people since a given change in the price level affects different classes of people differently.

The consumer price indices are of great significance and is given below

1. This is very useful in wage negotiations, wage contracts and dearness allowance adjustments in many countries.
2. At Government level the index numbers are used for wage policy, price policy, rent control, taxation and general economic policies.
3. Change in the purchasing power of money and real income can be measured.
4. Index numbers are also used for analyzing market price for particular kind of goods and services.

Note: *Consumer price index numbers are also called as cost of living index numbers.*

Methods of constructing consumer price Index

There are two methods of constructing consumer price index. They are:

1. Aggregate Expenditure method (or) Aggregate method
2. Family Budget method or method of weighted relative method.

1. Aggregate Expenditure method

This method is based upon the Laspeyre's method. It is widely used. The quantities of commodities consumed by a particular group in the base year are the weight.

Thus, consumer price index number $= \dfrac{\sum p_1q_0}{\sum p_0q_0} \times 100$

2. Family budget method (or) Method of weight relatives method

This method estimates an aggregate expenditure of an average family on various items and it is weighted. It is given by

consumer price index $= \dfrac{\sum wp}{\sum w}$

where

$$p = \frac{p_1}{p_0} \times 100 \text{ for each item and } w = p_0q_0$$

The family budget method is the same as "weighted average price relative method" which we have studied earlier.

Example 6.15

Calculate the consumer price index number for 2015 on the basis of 2000 from the following data by using (i) the Aggregate expenditure method (ii) the family budget (or) weighted relatives method.

Commodity	Quantity	Price	
		2000	2015
Wheat	20	15	20
Rice	8	20	24
Ghee	2	160	200
Sugar	4	40	40

Solution

(i) Calculation of cost of living index number on the basis of Aggregate expenditure method.

Commodity	q_0	p_0	p_1	p_0q_0	p_1q_0
Wheat	20	15	20	300	400
Rice	8	20	24	160	192
Ghee	2	160	200	320	400
Sugar	4	40	40	160	160
Total				940	1152

Consumer price index number for 2015

$$P_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100$$

$$= \frac{1152}{940} \times 100$$

$$\approx 112.6$$

(ii) Calculation of consumer price index number according to family budget method or weighted relative method

Commodity	q_0	p_0	p_1	$p = \frac{p_1}{p_0} \times 100$	$w = p_0q_0$	wp
Wheat	20	15	20	400/3	300	40000
Rice	8	20	24	120	160	19200
Sugar	2	160	200	125	320	40000
Ghee	4	40	40	100	160	16000
					940	115200

Consumer price index number for 2015

$$P_{01} = \frac{\sum wp}{\sum w} = \frac{115200}{940}$$

$$\approx 122.6$$

POINTS TO REMEMBER

- Index numbers are barometers of an Economy.
- It is a specilized average designed to measure the changes in a group of variables over time.
- The different types of Index Numbers are Price Index Number, Quantity Index Number and Value Index Number.
- Index numbers are classified as simple aggregative and weighted aggregative.
- The base period must be free from natural calamities.
- Laspeyeres, Paasches, Dorbish and Bowley, Fisher's ideal and Kelly's are weighed index numbers.
- Index numbers generally satisfied three tests – Time reversal, factor reversal and circular.
- Fisher's ideal index number satisfies both time and factor reversal tests.
- Many index numbers do not satisfy circular test.
- Cost of living index numbers is useful to the Government for policy making etc.

EXERCISE 6

I. Choose the best answer.

1) In simple aggregate method, the aggregate price of all items in the given year is expressed as percentage of the same in the

(a) current year (b) base year

(c) Quarterly (d) half yearly

2) If the index for 1990 to the base 1980 is 250, the index number for 1980 to the base 1990 is

(a) 4 (b) 400 (c) 40 (d) 4000

3) If Laspeyre's price index is 324 and Paasche's price index is 144, then Fisher's ideal index is

(a) 234 (b) 243 (c) 261 (d) 216

4) The index that satisfies factor reversal test is

(a) Paasche's Index (b) Laspeyre's Index

(c) Fisher's Ideal Index (d) Walsh price index

5) The Dorbish-Bowley's price index is the
   (a) geometric mean of Laspeyre's and Paasche's Price indices
   (b) arithmetic mean of Laspeyre's and Paasche's Price indices
   (c) weighted mean of Laspeyre's and Paasche's Price indices
   (d) weighted mean of Laspeyre's and Paasche's quantity indices

6) The condition for the time reversal test to hold good with usal notation is
   (a) $P_{01} \times P_{10} = 1$ (b) $P_{01} - P_{10} = 1$ (c) $P_{01} + P_{10} = 1$ (d) $P_{01}/P_{10} = 1$

7) The geometric mean of Laspeyre's and Paasche's price indices is also known as
   (a) Dorbish – Bowley's price index (b) Kelly's price index
   (c) Fisher's price index (d) Walsh price index

8) The index number for 1985 to the base 1980 is 125 and for 1980 to the base1985 is 80. The given indices satisfy
   (a) circular test (b) factor reversal test
   (c) time reversal test (d) Marshall-Edgeworth test

9) The consumer price index numbers for 1981 and 1982 to the base 1974 are 320 and 400 respectively. The consumer price index for 1981 to the base 1982 is
   (a)80 (b)128 (c)125 (d) 85

10) The consumer price index in 2000 increases by 80% as compared to the base 1990. A person I 1990 getting Rs. 60,000 per annum should now get:
   (a)Rs. 1,08,000 p.a. (b)Rs. 1,02,000 p.a.
   (c)Rs. 1,18,000 p.a. (d) Rs. 1,80,000 p.a.

II. Give very short answer to the following questions.

11. Define index number?

12. Write the uses of Index numbers.

13. Define base period.

14. State the types of Index numbers.

15. Point out the difference between weighted and unweighted index numbers.

16. Define weighted index number.

17. What is circular test?

18. State the methods of constructing consumer price index.

III. Give short answer to the following questions.

19. Give the diagrammatic representation of different types of index number.

20. Write the advantages of average price index.

21. State the methods of weighted aggregate index numbers.

22. What is the difference between the price index and quantity index numbers?

23. Write short notes on consumer price index.

24. Calculate index number from the following data by simple aggregate method taking prices of 2015 as base.

Commodity	Units	Price per kg	
		2015	2016
Wheat	Quintal	200	250
Rice	Quintal	300	400
Pulses	Quintal	400	500
Milk	Litre	2	3
Clothing	Meter	3	5

25. Compute (i) Laspeyre's (ii) Paasche's index numbers for 2010 from the following

Commodity	Price		Quantity	
	2002	2010	2002	2010
A	4	6	8	7
B	3	5	10	8
C	2	4	14	12
D	5	7	19	11

26. Calculate Fisher's ideal index method for the following data.

Commodity	2000		2001	
	Price	Quantity	Price	Quantity
A	2	7	3	5
B	5	11	6	10
C	3	14	5	11
D	4	16	4	18

IV. Give detail answer to the following questions.

27. Calculate the simple aggregate price index for the year 2013, and 2014 taking 2012 as the base year.

Categories of employees	Salary per month		
	2012	2013	2014
A	6000	6500	7200
B	12000	14000	16000
C	50000	64000	80000
D	70000	78000	84000

28 Construct the price indices from the following data by applying (1) Laspeyre's method (2) Paasche's method and (3) Fisher ideal number by taking 2010 as the base year.

Commodity	2010		2011	
	Price (₹)	Quantity	Price (₹)	Quantity
A	15	15	22	12
B	20	5	27	4
C	4	10	7	5

29. Construct (1) Laspeyre's index, (2) Paasche's index, (3) Marshall-Edgeworth index, and (4) Fisher ideal index for the following data taking 2014 as base year

Items	year 2014		year 2015	
	Price	Quantity	Price	Quantity
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	36

30. Construct Marshall–Edgeworth price index number for the following data taking 2016 as base year

Commodity	year 2016		year 2017	
	Price	Quantity	Price	Quantity
A	4	58500	6	62000
B	3.5	15630	5.5	13050
C	3	26230	5	25000
D	2.5	11360	4	10000
E	2	30000	3	31500

31. A popular consumer co-operative store located in a labour colory reported the average monthly data on prices and quantities sold of a group of selected items of mass consumption as follows.

Items	Jan 2015		Jan 2018	
	Prices ₹ Per kg	Quantity sold kg	Prices Per kg	Quantity sold kg
Veg oil	26	40	31	45
Sugar	28	90	32	100
Rice	16	120	19	20
Wheat	15	110	18	130

Compute the following indices.

(a) Laspeyre's price index for 2018 using 2015 us base year.

(b) Paache's price index for 2018 using 2015 as base year.

32. From the data given, in problem. obtain the following

(a) Laspeyre's quantity index for 2018 using 2015 as the base year.

(b) Paasche's quantity index for 2018 using 2015 as the base

(c) Compute Index number using Fisher's formula and show it satisfies time reversal test and factor reversal test

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	10	12	12	15
B	7	15	5	20
C	5	24	9	20
D	16	5	14	5

33. Construct the consumer price index number of 2015 on the from the following data using (i) the average expenditure method and (ii) the family budget method.

Commodity	Quantity consumed in 2014	Price in 2014	Price in 2015
A	6 Quintal	5	6
B	6 Quintal	6	7
C	1 Quintal	5	6
D	6 Quintal	6	7
E	4 kg	7	8
F	6 kg	8	9

34. An enquiry into the budgets of the middle class families in a city in India gave the following information.

Expenses on	Food	Rent	Clothing	Fuel	Mise
	35%	15%	20%	10%	20%
Price in 2014	450	90	225	75	120
Price in 2015	435	90	195	69	135

What change in the cost of living figures of 2015 has taken place as compared to 2014?

35. Construct the cost of living index of 2014 using family budget method.

Expenses	%	base year (2000)	year 2004
Food	40	150	174
Rent	15	50	60
Clothing	15	100	125
Fuel	10	20	25
Misc	20	60	90

36. Construct the index of 2014 from the following data for the year 2012 taking 2011 as base year as base using i) arithmetic mean and ii) geometric mean.

Item	Price (₹) in 2014	Price (₹) in 2015
A	6	10
B	2	2
C	4	6
D	10	12
E	8	12

37. Compute price index for the following data by applying weighted average of price relative method using i) arithmetic mean and ii) geometric mean.

Item	Price (₹) in 2006	Price (₹) in 2007	Quantity in 1996
A	2	2.5	40
B	3	3.25	20
C	1.5	1.75	10

ANSWERS

I. 1. (b) **2.** (c) **3.** (d) **4.** (c) **5.** (b)

6. (a) **7.** (c) **8.** (c) **9.** (a) **10.** (a)

II. 24. 127.96%

25. $P_{01}^{L} = 155.14$, $P_{01}^{P} = 158.01$

26. $P_{01}^{F} = 124.33$

III. 27. for 2013 price index = 117.75 and for 2014 price index = 135.65

28. $P_{01}^{L} = 146.5$ $P_{01}^{P} = 145.35$, $P_{01}^{F} = 145.96$

29. $P_{01}^{L} = 139.7$, $P_{01}^{P} = 139.8$, $P_{01}^{ME} = 139.8$, $P_{01}^{F} = 139.8$

30. $P_{01}^{ME} = 154.18$

31. $P_{01}^{L} = 117.53$ $P_{01}^{p} = 117.22$

32. a) $Q_{01}^{L} = 110.58$ b) Q_{01}^{p} 104.95 c) $Q_{01}^{F} = 107.73$, satisfies time reversal test and factor reversal test.

33. (i)115.84 (ii) 115.84

34. Cost of living index = 96.43% , there is a decrease of 3.57% as compared to the prices in the year 2014.

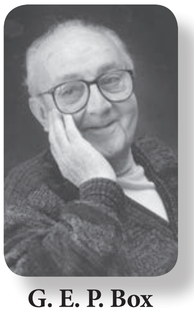
35. 122.12

36. $P_{01} = 137.34$, $P_{01} = 134.99$

37. $P_{01} = 117.74$, $P_{01} = 117.4$

CHAPTER 7

TIME SERIES AND FORECASTING



G. E. P. Box

G. E. P. Box (1919-2013) was a Britsh Statistician was "one of the great statistical minds" of the 20th century, who received his Ph.D., from the University of London, under the supervision of E. S. Pearson. He served as President of Americal Statistical Association in 1978 and of the Institute of Mathematics in 1979. His name is associated with Box-Cox transformation in addition to Box-Jenkins models in time series.

G. M. Jenkins

G. M. Jenkins (1932-1982) was a British Statistician, earned his Ph.D. degree from University College, London under the supervision of F. N. David and N. L. Johnson. He served on the Research Section Committee and Council of Royal Statistical Society in 1960's. He was elected to the Institute of Mathematical Statitics

Both Box and Jenkins contributed to Auto regressive moving average models popularly known as Box-Jenkins Models.

LEARNING OBJECTIVES

The students will be able to

❖ understand the concept of time series

❖ know the upward and downward trends

❖ calculate the trend values using semi - average and moving average methods

❖ estimate the trend values using method of least squares

❖ compute seasonal indices

❖ understand cyclical and irregular variations

❖ understand the forecasting concept

Introduction

In modern times we see data all around. The urge to evaluate the past and to peep into the future has made the need for forecasting. There are many factors which change with the passage of time. Sometimes sets of observations which vary with the passage of time and whose measurements made at equidistant points may be regarded as time series data. Statistical data which are collected, observed or recorded at successive intervals of time constitute time series data. In the study of time series, comparison of the past and the present data is made. It also compares two or more series at a time. The purpose of time series is to measure chronological variations in the observed data.

In an ever changing business and economic environment, it is necessary to have an idea about the probable future course of events. Analysis of relevant time series helps to achieve this, especially by facilitating future business forecasts. Such forecasts may serve as crucial inputs in deciding competitive strategies and planning growth initiatives.

7.1 DEFINITION

Time series refers to any group of statistical information collected at regular intervals of time. Time series analysis is used to detect the changes in patterns in these collected data.

7.1.1 Definition by Authors

According to Mooris Hamburg "A time series is a set of statistical observations arranged in chronological order".

Ya-Lun-Chou : "A time series may be defined as a collection of readings belonging to different time periods of some economic variable or composite of variables".

W.Z. Hirsch says "The main objective in analyzing time series is to understand, interpret and evaluate change in economic phenomena in the hope of more correctly anticipating the course of future events".

7.1.2 Uses of Time Series

- Time series is used to predict future values based on previously observed values.
- Time series analysis is used to identify the fluctuation in economics and business.
- It helps in the evaluation of current achievements.
- Time series is used in pattern recognition, signal processing, weather forecasting and earthquake prediction.

It can be said that time series analysis is a big tool in the hands of business executives to plan their sales, prices, policies and production.

7.2 COMPONENTS OF TIME SERIES

The factors that are responsible for bringing about changes in a time series are called the components of time series.

Components of Time Series

1. Secular trend
2. Seasonal variation
3. Cyclical variation
4. Irregular (random) variation

Approaches to time series

There are two approaches to the decomposition of time series data

(i) Additive approach

(ii) Multiplicative approach

The above two approaches are used in decomposition, depending on the nature of relationship among the four components.

The additive approach

The additive approach is used when the four components of a time series are visualized as independent of one another. Independence implies that the magnitude and pattern of movement of the components do not affect one another. Under this assumption the magnitudes of the time series are regarded as the sum of separate influences of its four components.

$$Y = T + C + S + R$$

where Y = magnitude of a time series

T = Trend,
C =Cyclical component,
S =Seasonal component, and
R = Random component

In additive approach, the unit of measurements remains the same for all the four components.

The Multiplicative approach

The multiplicative approach is used where the forces giving rise to the four types of variations are visualized as interdependent. Under this assumption, the magnitude of the time series is the product of its four components.

i.e. $Y = T \times C \times S \times R$

Difference between the two approaches

Multiplicative	Additive
(i) Four components of time series are interdependent	Four components of time series are independent
(ii) Logarithm of components are additive	Components are additive

7.3 MEASUREMENTS OF COMPONENTS

(i) Secular trend

It refers to the long term tendency of the data to move in an upward or downward direction. For example, changes in productivity, increase in the rate of capital formation, growth of population, *etc.*, follow secular trend which has upward direction, while deaths due to improved medical facilities and sanitations show downward trend. All these forces occur in slow process and influence the time series variable in a gradual manner.

Methods of Measuring Trend

Trend is measured using by the following methods:

1. Graphical method
2. Semi averages method
3. Moving averages method
4. Method of least squares

7.3.1 Graphical Method

Under this method the values of a time series are plotted on a graph paper by taking time variable on the X-axis and the values variable on the Y-axis. After this, a smooth curve is drawn with free hand through the plotted points. The trend line drawn above can be extended to forecast the values. The following points must be kept in mind in drawing the freehand smooth curve.

(i) The curve should be smooth

(ii) The number of points above the line or curve should be approximately equal to the points below it

(iii) The sum of the squares of the vertical deviation of the points above the smoothed line is equal to the sum of the squares of the vertical deviation of the points below the line.

Merits

- It is simple method of estimating trend.
- It requires no mathematical calculations.
- This method can be used even if trend is not linear.

Demerits

- It is a subjective method
- The values of trend obtained by different statisticians would be different and hence not reliable.

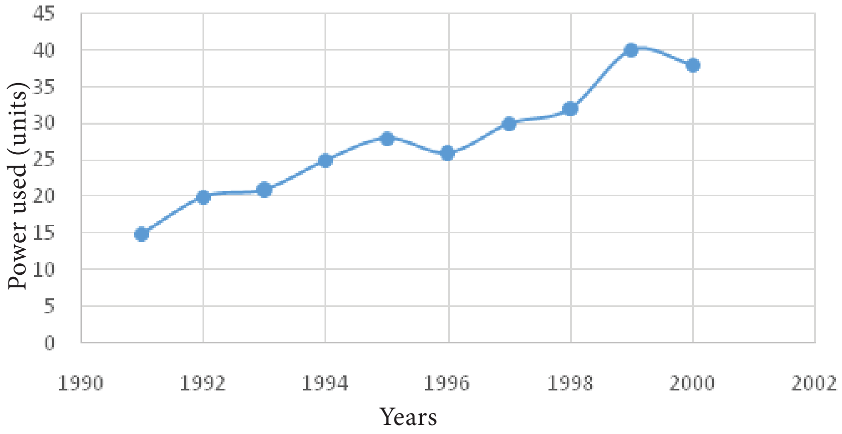
Example 7.1

Annual power consumption per household in a certain locality was reported below.

Years	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Power used (units)	15	20	21	25	28	26	30	32	40	38

Draw a free hand curve for the above data.

Solution:



7.3.2 Semi-Average Method

In this method, the series is divided into two equal parts and the average of each part is plotted at the mid-point of their time duration.

(i) In case the series consists of an even number of years, the series is divisible into two halves. Find the average of the two parts of the series and place these values in the mid-year of each of the respective durations.

(ii) In case the series consists of odd number of years, it is not possible to divide the series into two equal halves. The middle year will be omitted. After dividing the data into two parts, find the arithmetic mean of each part. Thus we get semi-averages.

(iii) The trend values for other years can be computed by successive addition or subtraction for each year ahead or behind any year.

Merits

- This method is very simple and easy to understand
- It does not require many calculations.

NOTE

In semi-average method if the difference between the semi-averages is negative then the trend values will be in decreasing order.

Demerits

- This method is used only when the trend is linear.
- It is used for calculation of averages and they are affected by extreme values.

Example 7.2

Calculate the trend values using semi-averages methods for the income from the forest department. Find the yearly increase.

Year	2008	2009	2010	2011	2012	2013
Income (in crores)	46.17	51.65	63.81	70.99	84.91	91.64

Source: The Principal Chief conservator of forests, Chennai-15. (pg. 231)

Solution:

Year	Income	3-Year semi-total	Semi-average
2008	46.17	161.63	53.877
2009	51.65		
2010	63.81		
2011	70.99	247.54	82.513
2012	84.91		
2013	91.64		

Difference between the central years = 2012 – 2009 = 3

Difference between the semi-averages = 82.513 – 53.877 = 28.636

Increase in trend value for one year = $\frac{28.636}{3} = 9.545$

Trend values for the previous and successive years of the central years can be calculated by subtracting and adding respectively, the increase in annual trend value.

Example 7.3

Population of India for 7 successive census years are given below. Find the trend values using semi-averages method.

Census Year	1951	1961	1971	1981	1991	2001	2011
Population (in lakhs)	301.2	336.9	412.0	484.1	558.6	624.1	721.4

Solution:

Trend values using semi average method

Census Year	Population (in lakhs)	3-year semi-total	3-year semi-average	Trend values
1951	301.2	1050.1	350.03	278.86
1961	336.9			350.03
1971	412.0			421.2
1981	484.1			492.37
1991	558.6	1904.1	634.7	563.54
2001	624.1			634.71
2011	721.4			705.88

Difference between the years = 2001 – 1961 = 40

Difference between the semi-averages = 634.7 – 350.03 = 284.67

Increase in trend value for 10 year = $\frac{284.67}{4} = 71.17$

For example the trend value for the year 1951 = 350.03 – 71.17 = 278.86

The value for the year 2011 = 634.7 + 71.17 = 705.87

The trend values have been calculated by successively subtracting and adding the increase in trend for previous and following years respectively.

Example 7.4

Find the trend values by semi-average method for the following data.

Year	1965	1966	1967	1968	1969	1970	1971	1972
Production of bleaching powder (in tonnes)	7.4	10.8	9.2	10.5	15.5	13.7	16.7	15

Solution:

Trend values using semi averages method

Year	Production of bleaching powder	4 year semi-total	4 year semi-average	Trend
1965	7.4	37.9	9.475	7.315
1966	10.8			8.755
1967	9.2			10.195
1968	10.5			11.635
1969	15.5	60.9	15.225	13.075
1970	13.7			14.515
1971	16.7			15.955
1972	15			17.395

Difference between the years = 1970.5 – 1966.5 = 4

Difference between the semi-averages = 15.225 – 9.475 = 5.75

Increase in trend = $\frac{5.75}{4} = 1.44$

Half yearly increase in trend = $\frac{1.44}{2} = 0.72$

The trend value for 1967 = 9.475 + 0.72 = 10.195

The trend value for 1968 = 9.475 + 3 * 0.72 = 11.635

Similarly the trend values for the other years can be calculated.

7.3.3 Moving Averages Method

Moving averages is a series of arithmetic means of variate values of a sequence. This is another way of drawing a smooth curve for a time series data.

Moving averages is more frequently used for eliminating the seasonal variations. Even when applied for estimating trend values, the moving average method helps to establish a trend line by eliminating the cyclical, seasonal and random variations present in the time series. The period of the moving average depends upon the length of the time series data.

The choice of the length of a moving average is an important decision in using this method. For a moving average, appropriate length plays a significant role in smoothening the variations. In general, if the number of years for the moving average is more then the curve becomes smooth.

Merits

- It can be easily applied
- It is useful in case of series with periodic fluctuations.
- It does not show different results when used by different persons
- It can be used to find the figures on either extremes; that is, for the past and future years.

Demerits

- In non-periodic data this method is less effective.
- Selection of proper 'period' or 'time interval' for computing moving average is difficult.
- Values for the first few years and as well as for the last few years cannot be found.

Moving averages odd number of years (3 years)

To find the trend values by the method of three yearly moving averages, the following steps have to be considered.

1. Add up the values of the first 3 years and place the yearly sum against the median year. [This sum is called moving total]
2. Leave the first year value, add up the values of the next three years and place it against its median year.
3. This process must be continued till all the values of the data are taken for calculation.
4. Each 3-yearly moving total must be divided by 3 to get the 3-year moving averages, which is our required trend values.

Example 7.5

Calculate the 3-year moving averages for the loans issued by co-operative banks for non-farm sector/small scale industries based on the values given below.

Year	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15
Loan by District Central Cooperative banks (Rupees in crores)	41.82	40.05	39.12	24.72	26.69	59.66	23.65	28.36	33.31	31.60	36.48

Solution:

The three year moving averages are shown in the last column.

Year	Loan by District Central Cooperative Banks	3-year moving total	3-year moving average
2004-05	41.82	-	-
2005-06	40.05	120.99	40.33
2006-07	39.12	103.89	34.63
2007-08	24.72	90.53	30.18
2008-09	26.69	111.07	37.02
2009-10	59.66	110	36.67
2010-11	23.65	111.67	37.22
2011-12	28.36	85.32	28.44
2012-13	33.31	93.27	31.09
2013-14	31.60	101.39	33.80
2014-15	36.48	-	-

Moving averages - even number of years (4 years)

1. Add up the values of the first 4 years and place the sum against the middle of 2nd and 3rd year. (This sum is called 4 year moving total)
2. Leave the first year value and add next 4 values from the 2nd year onward and write the sum against its middle position.
3. This process must be continued till the value of the last item is taken into account.
4. Add the first two 4-years moving total and write the sum against 3rd year.
5. Leave the first 4-year moving total and add the next two 4-year moving total and place it against 4th year.
6. This process must be continued till all the 4-yearly moving totals are summed up and centered.
7. Divide the 4-years moving total by 8 to get the moving averages which are our required trend values.

Example 7.6

Compute the trends by the method of moving averages, assuming that 4-year cycle is present in the following series.

Year	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Annual value	154.0	140.5	147.0	148.5	142.9	142.1	136.6	142.7	145.7	145.1	137.8

Solution:

The four year moving averages are shown in the last column.

Year	Annual value	4-year moving total	Centered total	4-year moving average
1998	154.0			
		-		
1999	140.5		-	
		590.0		
2000	147.0		1168.9	146.11
		578.9		
2001	148.5		1159.4	144.93
		580.5		
2002	142.9		1150.6	143.83
		570.1		
2003	142.1		1134.4	141.8
		564.3		
2004	136.6		1131.4	141.43
		567.1		
2005	142.7		1137.2	142.15
		570.1		
2006	145.7		1141.4	142.68
		571.3		
2007	145.1		-	
		-		
2008	137.8			

7.3.4 Method of least squares

Among the four components of the time series, secular trend represents the long term direction of the series. One way of finding the trend values with the help of mathematical technique is the method of least squares. This method is most widely used in practice and in this method the sum of squares of deviations of the actual and computed values is least and hence the line obtained by this method is known as the line of best fit.

It helps for forecasting the future values. It plays an important role in finding the trend values of economic and business time series data.

Computation of Trend using Method of Least squares

Method of least squares is a device for finding the equation which best fits a given set of observations.

Suppose we are given *n* pairs of observations and it is required to fit a straight line to these data. The general equation of the straight line is:

$$y = a + bx$$

where a and b are constants. Any value of a and b would give a straight line, and once these values are obtained an estimate of y can be obtained by substituting the observed values of y. In order that the equation $y = a + b\,x$ gives a good representation of the linear relationship between x and y, it is desirable that the estimated values of y_i, say $\hat{y}_i$ on the whole close enough to the observed values y_i, $i = 1, 2, \ldots, n$. According to the principle of least squares, the best fitting equation is obtained by minimizing the sum of squares of differences $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$

That is, $\sum\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - a - bx_i\right)^2$ is minimum. This leads us to two normal equations.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \qquad (7.1)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x + b\sum_{i=1}^{n} x_i^2 \qquad (7.2)$$

Solving these two equations we get the vales for a and b and the fit of the trend equation (line of best):

$$y = a + bx \qquad (7.3)$$

Substituting the observed values x_i in (7.3) we get the trend values y_i, $i = 1, 2, \ldots, n$.

Note: The time unit is usually of uniform duration and occurs in consecutive numbers. Thus, when the middle period is taken as the point of origin, it reduces the sum of the time variable x to zero $\left(\sum_{i=1}^{n} x_i = 0\right)$ and hence we get

$$a = \frac{\sum_{i=1}^{n} y_i}{n} \text{ and } b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \text{ by simplifying (7.1) and (7.2)}$$

The number of time units may be even or odd, depending upon this, we follow the method of calculating trend values using least square method.

Merits

- The method of least squares completely eliminates personal bias.
- Trend values for all the given time periods can be obtained
- This method enables us to forecast future values.

Demerits

- The calculations for this method are difficult compared to the other methods.
- Addition of new observations requires recalculations.
- It ignores cyclical, seasonal and irregular fluctuations.
- The trend can be estimated only for immediate future and not for distant future.

Steps for calculating trend values when *n* is odd:

i) Subtract the first year from all the years (x)

ii) Take the middle value (A)

iii) Find $u_i = x_i - A$

iv) Find u_i^2 and u_iy_i

Then use the normal equations:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} u_i$$

$$\sum_{i=1}^{n} u_iy_i = a\sum_{i=1}^{n} u_i + b\sum_{i=1}^{n} u_i^2$$

Find $a = \dfrac{\sum_{i=1}^{n} y_i}{n}$ and $b = \dfrac{\sum_{i=1}^{n} u_iy_i}{\sum_{i=1}^{n} u_i^2}$

Then the estimated equation of straight line is:

$$y = a + b\ u = a + b\ (x - A)$$

Example 7.7

Fit a straight line trend by the method of least squares for the following consumer price index numbers of the industrial workers.

Year	2010	2011	2012	2013	2014
Index number	166	177	198	221	225

Solution:

Year	Index Number	$X = x_i - 2010$	$u_i = X - A$ $= X - 2$	u_i^2	u_iy_i	Trend
2010	166	0	–2	4	–332	165
2011	177	1	–1	1	–177	181.2
2012	198	2	0	0	0	197.4
2013	221	3	1	1	221	213.6
2014	225	4	2	4	450	229.8
	$\sum_{i=1}^{5} y_i = 987$		$\sum_{i=1}^{5} u_i = 0$	$\sum_{i=1}^{5} u_i^2 = 10$	$\sum_{i=1}^{5} u_iy_i = 162$	

The equation of the straight line is $y = a + bx$

$= a + bu$ where $u = X - 2$

The normal equations give:

$$a = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{987}{5} = 197.4$$

$$b = \frac{\sum_{i=1}^{n} u_i y_i}{\sum_{i=1}^{n} u_i^2} = \frac{162}{10} = 16.2$$

$y = 197.4 + 16.2\ (X - 2)$

$= 197.4 + 16.2\ X - 32.4$

$= 16.2\ X + 165$

That is, $y = 165 + 16.2X$

To get the required trend values, put $X = 0, 1, 2, 3, 4$ in the estimated equation.

$X = 0, y = 165 + 0 = 165$

$X = 1, y = 165 + 16.2 = 181.2$

$X = 2, y = 165 + 32.4 = 197.4$

$X = 3, y = 165 + 48.6 = 213.6$

$X = 4, y = 165 + 64.8 = 229.8$

Hence, the trend values for 2010, 2011, 2012, 2013 and 2014 are 165, 181.2, 197.4, 213.6 and 229.8 respectively.

Steps for calculating trend values when *n* is even:

i). Subtract the first year from all the years (x)

ii). Find $u_i = 2X - (n - 1)$

iii). Find u_i^2 and $u_i y_i$

Then follow the same procedure used in previous method for odd years

Example 7.8

Tourist arrivals (Foreigners) in Tamil Nadu for 6 consecutive years are given in the following table. Calculate the trend values by using the method of least squares.

Year	2005	2006	2007	2008	2009	2010
No. of arrivals (in lakhs)	12	13	18	20	24	28

Solution:

Year x	No. of arrivals y_i	$X = x_i - 2005$	$u_i = 2X - 5$	u_i^2	$u_i y_i$
2005	12	0	–5	25	–60
2006	13	1	–3	9	–39
2007	18	2	–1	1	–18
2008	20	3	1	1	20
2009	24	4	3	9	72
2010	28	5	5	25	140
	$\sum_{i=1}^{6} y_i = 115$		$\sum_{i=1}^{6} u_i = 0$	$\sum_{i=1}^{6} u_i^2 = 70$	$\sum_{i=1}^{6} u_i y_i = 115$

The equation of the straight line is $y = a + bx$

$$= a + bu \text{ where } u = 2X - 5$$

Using the normal equation we have,

$$a = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{115}{6} = 19.17$$

$$b = \frac{\sum_{i=1}^{n} u_i y_i}{\sum_{i=1}^{n} u_i^2} = \frac{115}{70} = 1.64$$

$$\begin{aligned} y &= a + bu \\ &= 19.17 + 1.64\,(2X - 5) \\ &= 19.17 + 3.28X - 8.2 \\ &= 3.28X + 10.97 \end{aligned}$$

That is, $y = 10.97 + 3.28X$

To get the required trend values, put $X = 0, 1, 2, 3, 4, 5$ in the estimated equation. Thus,

$X = 0, y = 10.97 + 0 = 10.97$

$X = 1, y = 10.97 + 3.28 = 14.25$

$X = 2, y = 10.97 + 6.56 = 17.53$

$X = 3, y = 10.97 + 9.84 = 20.81$

$X = 4, y = 10.97 + 13.12 = 24.09$

$X = 5, y = 10.97 + 16.4 = 27.37$

Hence, the trend values for 2005, 2006, 2007, 2008, 2009 and 2010 are 10.97, 14.25, 17.53, 20.81, 24.09 and 27.37 respectively.

(ii) Seasonal variation

Seasonal variations are fluctuations within a year over different seasons.

Estimation of seasonal variations requires that the time series data are recorded at even intervals such as quarterly, monthly, weekly or daily, depending on the nature of the time series. Changes due to seasons, weather conditions and social customs are the primary causes of seasonal variations. The main objective of the measurement of seasonal variation is to study their effect and isolate them from the trend.

Methods of constructing seasonal indices

There are four methods of constructing seasonal indices.

1. Simple averages method
2. Ratio to trend method
3. Percentage moving average method
4. Link relatives method

Among these, we shall discuss the construction of seasonal index by the first method only.

7.3.5 Simple Averages Method

Under this method, the time series data for each of the 4 seasons (for quarterly data) of a particular year are expressed as percentages to the seasonal average for that year.

The percentages for different seasons are averaged over the years by using simple average. The resulting percentages for each of the 4 seasons then constitute the required seasonal indices.

Method of calculating seasonal indices

i) The data is arranged season-wise

ii) The data for all the 4 seasons are added first for all the years and the seasonal averages for each year is computed.

iii) The average of seasonal averages is calculated
(*i.e.*, Grand average = Total of seasonal averages /number of years).

iv) The seasonal average for each year is divided by the corresponding grand average and the results are expressed in percentages and these are called seasonal indices.

Example 7.9

Calculate the seasonal indices for the rain fall (in mm) data in Tamil Nadu given below by simple average method

Year	Season			
	I	II	III	IV
2001	118.4	260.0	379.4	70
2002	85.8	185.4	407.1	8.7
2003	129.8	336.5	403.1	12.0
2004	283.4	360.7	472.1	14.3
2005	231.7	308.5	828.8	15.9

Solution:

Year	Season			
	I	II	III	IV
2001	118.4	260.0	379.4	70
2002	85.8	185.4	407.1	8.7
2003	129.8	336.5	403.1	12.0
2004	283.4	360.7	472.1	14.3
2005	231.7	308.5	828.8	15.9
Seasonal total	849.1	1451.1	2490.5	120.9
Seasonal average	169.82	290.22	498.1	24.18
Seasonal index	69	118	203	10

$$\text{Grand Average} = \frac{\text{Total of seasonal averages}}{4}$$

$$= \frac{169.82 + 290.22 + 498.1 + 24.18}{4}$$

$$= \frac{982.32}{4} = 245.58$$

$$\text{Seasonal Index} = \frac{\text{Seasonal average}}{\text{Grand average}} \times 100$$

$$\text{Seasonal Index for Season I} = \frac{169.82}{245.58} \times 100 = 69.15 \approx 69$$

$$\text{Seasonal Index for Season II} = \frac{290.22}{245.58} \times 100 = 118.18 \approx 118$$

$$\text{Seasonal Index for Season III} = \frac{498.1}{245.58} \times 100 = 202.83 \approx 203$$

$$\text{Seasonal Index for Season IV} = \frac{24.18}{245.58} \times 100 = 9.85 \approx 10$$

Example 7.10

Obtain the seasonal indices for the rain fall (in mm) data in India given in the following table.

Quarter \ Year	2009	2010	2011	2012
I	38.2	38.5	55	50.5
II	166.8	250.9	277.7	197
III	612.6	773.1	717.8	706.1
IV	72.2	153.1	65.8	101.1

Solution:

Year	Quarter			
	I	II	III	IV
2009	38.2	166.8	612.6	72.2
2010	38.5	250.9	773.1	153.1
2011	55	277.7	717.8	65.8
2012	50.5	197	706.1	101.1
Seasonal total	182.2	892.4	2809.6	392.2
Seasonal average	45.55	223.1	702.4	98.05
Seasonal index	17	83	263	37

$$\text{Grand Average} = \frac{\text{Total of seasonal averages}}{4}$$

$$= \frac{45.55 + 223.1 + 702.4 + 98.05}{4}$$

$$= \frac{1069.10}{4} = 267.28$$

$$\text{Seasonal Index} = \frac{\text{Seasonal average}}{\text{Grand average}} \times 100$$

$$\text{Seasonal Index for Quarter I} = \frac{45.55}{267.28} \times 100 = 17.04 \approx 17$$

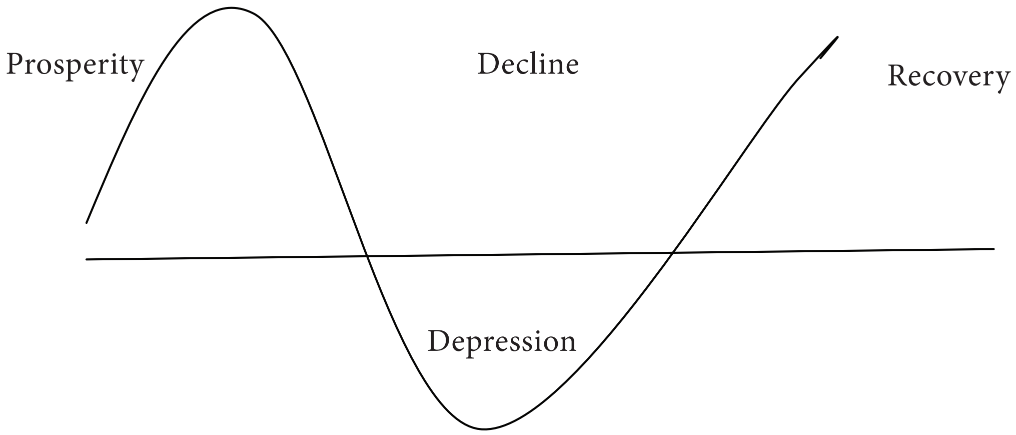
$$\text{Seasonal Index for Quarter II} = \frac{223.1}{267.28} \times 100 = 83.47 \approx 83$$

$$\text{Seasonal Index for Quarter III} = \frac{702.4}{267.28} \times 100 = 262.80 \approx 263$$

$$\text{Seasonal Index for Quarter IV} = \frac{98.05}{267.28} \times 100 = 36.69 \approx 37$$

(iii) Cyclical variation

Cyclical variations refer to periodic movements in the time series about the trend line, described by upswings and downswings. They occur in a cyclical fashion over an extended period of time (more than a year). For example, the business cycle may be described as follows.



The cyclical pattern of any time series tells about the prosperity and recession, ups and downs, booms and depression of a business. In most of the businesses there are upward trend for some time followed by a downfall, touching its lowest level. Again a rise starts which touches its peak. This process of prosperity and recession continues and may be considered as a natural phenomenon.

YOU WILL KNOW

Cyclic movements are mainly due to Trade cycle.

(iv) Irregular variation

In practice, the changes in a time series that cannot be attributed to the influence of cyclic fluctuations or seasonal variations or those of the secular trend are classified as irregular variations.

In the words of Patterson, "Irregular variation in a time series is composed of non-recurring sporadic (rare) form which is not attributed to trend, cyclical or seasonal factors".

NOTE

There is no statistical technique for measuring or isolating irregular fluctuations

Nothing can be predicted about the occurrence of irregular influences and the magnitude of such effects. Hence, no standard method has been evolved to estimate the same. It is taken as the residual left in the time series, after accounting for the trend, seasonal and cyclic variations.

YOU WILL KNOW

Irregular variation is also called erratic fluctuations.

7.4 FORECASTING

The importance of statistics lies in the extent to which it serves as the basis for making reliable forecasts, against arbitrary forecast with no statistical background.

Forecasting is a scientific process which aims at reducing the uncertainty of the future state of business and trade, not dependent merely on guess work, but with a sound scientific footing for the decision on the future course of action.

7.4.1 Definition

"Forecasting refers to the analysis of past and present conditions with a view of arriving at rough estimates about the future conditions.

NOTE

1. Long term forecasting can be found using Trend.
2. Short term forecasting can be found using seasonal variations.

According to T.S. Lewis and R.A. Fox "Forecasting is using the knowledge we have at one time to estimate what will happen at some future moment of time".

Forecasting is an important tool that serves many fields including business and industry, government, economics, environmental sciences, medicine, social science, politics and finance. Forecasting problems are often classified as short-term, medium-term, and long-term.

Short-term forecasting problems involve predicting events for a few time periods (days, weeks, months) into the future.

Medium-term forecast extends from one to two years into the future.

Long-term forecasting problems can extend beyond that by many years.

Short and medium-term forecasts are required for activities that range from operations management to budgeting and selecting new research and development projects. Long term forecasts impact issues relating to strategic planning.

POINTS TO REMEMBER

- ❖ Time series is a time oriented sequence of observations.
- ❖ Components of time series are secular trend, seasonal variations, cyclical variations and irregular (erratic) variations
- ❖ Methods of calculating trend values are graphical method, semi - averages method, moving averages method, and method of least squares.
- ❖ The line $y = a + b\,x$ found out using the method of least squares is called 'line of best fit'.
- ❖ Normal equations involved in the method of least squares are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x + b\sum_{i=1}^{n} x_i^2$$

- ❖ Seasonal indices may be found out by using simple average method.
- ❖ Forecasting is the analysis of using past and present conditions to get rough estimates of the future conditions
- ❖ Forecasting methods can be short-term, medium-term and long-term.

EXERCISE 7

2N07FE

I. Choose the best answer.

1. An overall tendency of rise or fall in a time series is called

   (a) seasonal variation (b) secular trend

   (c) cyclical variation (d) irregular variation

2. The component having primary use for short-term forecasting is

   (a) cyclical variation (b) irregular variation

   (c) seasonal variation (d) trend

3. Cyclical movements are due to

   (a) ratio to trend (b) seasonal

   (c) trend (d) trade cycle

4. Data on annual turnover of a company over a period of ten years can be represented by a

   (a) a time series (b) an index number

   (c) a parameter (d) a statistic

5. The component having primary use for long term forecasting is

   (a) cyclical variation (b) irregular variation

   (c) seasonal variation (d) trend

6. A time series is a set of data recorded

   (a) periodically (b) at equal time intervals

   (c) at successive points of time (d) all the above

7. A time series consists of

   (a) two components (b) three components

   (c) four components (d) five components

8. Irregular variation in a time series can be due to

   (a) trend variations (b) seasonal variations

   (c) cyclical variations (d) unpredictable causes

9. The terms prosperity, recession, depression and recovery are in particular attached to

   (a) secular trend (b) seasonal fluctuation

   (c) cyclical movements (d) irregular variation

10. An additive model of time series with components, T, S, C and I is

   (a) $Y = T \times S \times C \times I$ (b) $Y = T + S + C + I$

   (c) $Y = T \times S + C \times I$ (d) $Y = T \times S \times C + I$

11. A decline in the sale of ice cream during November to March is associated with

(a) seasonal variation	(b) cyclical variation
(c) irregular variation	(d) secular trend

12. Business forecasts are made on the basis of

(a) future data	(b) past data
(c) tax regulations	(d) Government policies

13. The four components of time series in a multiplicative model are

(a) independent	(b) interdependent
(c) constant	(d) additive

14. In the least square theory the sum of squares of residuals is

(a) zero	(b) minimum
(c) constant	(d) maximum

15. No statistical techniques for measuring or isolating _______is available.

(a) cyclical variation	(b) seasonal variation
(c) erratic fluctuations	(d) secular trend

II. Give very short answers to the following questions

16. What is a time series?

17. What are the components of a time series?

18. Name different methods of estimating the trend?

19. Write short notes on irregular variation.

20. Mention the methods used to find seasonal indices?

21. What are the demerits of moving averages?

22. What are the merits of method of least squares?

23. Write the normal equations used in method of least squares?

24. Define forecasting.

25. What are the three types of forecasting?

26. What is a short-term forecast?

III. Give short answers to the following questions

27. Write the uses of time series.

28. Explain semi-averages method

29. Write the merits of moving averages.

30. What is cyclical variation?

31. What is seasonal variation?

32. What are medium-term and long-term forecasts?

33. Describe the method of finding seasonal indices.

34. With what characteristic component of a time series should each of the following be associated.
    (i) An upturn in business activity
    (ii) Fire loss in a factory
    (iii) General increase in the sale of Television sets.

35. The number of units of a product exported during 1990-97 is given below. Draw the trend line using graphical method.

Year	1990	1991	1992	1993	1994	1995	1996	1997
No. of units exported (in ‘000)	12	13	13	16	19	23	21	23

36. Draw a time series graph relating to the following data and show the trend by free hand method

Year	1996	1997	1998	1999	2000	2001	2002	2003
Production in Million tonnes	40	44	42	48	51	54	50	56

37. Draw a trend line by the method semi averages

Year	1992	1993	1994	1995	1996	1997
Production of steel in Million tonnes	21	23	25	23	26	25

38. Yield of ground nut in Kharif season in India for the years 2003-04 to 2009-10 are given below. Calculate 3-year moving averages.

Year	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10
Yield (kg/hectare)	1320	909	1097	689	1386	1063	835

39. What do you understand by seasonal variations?

IV. Give detailed answers to the following questions

40. Explain the method of least squares

41. The following data states the number of ATM centers during 1995 to2001.

Year	1995	1996	1997	1998	1999	2000	2001
Number of ATM centres	50	63	75	100	109	120	135

Obtain the trend values using semi averages method

42. From the following data estimate the trend values using semi averages method

Year	2003	2004	2005	2006	2007	2008	2009	2010
Consumption of cotton (Thousands of bales)	677	696	747	755	766	777	785	836

43. Following data gives the yield of food grains in India for the years 2000-01 to 2009-10. Find the trend values using 4 year moving averages.

Year	2000-01	2001-02	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10
Yield (kg/ hectare)	1626	1734	1535	1727	1652	1715	1756	1860	1909	1798

44. Estimate the value of production for the year 1995 by using the method of least squares from the following data.

Year	1990	1991	1992	1993	1994
Production (1000s tons)	70	72	88	90	92

45. Find the following for the calculation of number of telephones for the year 2000.

(1) Fit a straight line trend by the method of least squares.

(2) Calculate the trend values.

Year	1990	1991	1992	1993	1994	1995
No. of telephones (in ^{1}00s)	20	21	23	25	27	29

46. The following data describes the export quantity of a company.

Year	1995	1996	1997	1998	1999	2000	2001
Export (in millions)	12	13	13	16	16	19	23

Fit a straight line trend and estimate the export for the year 2005.

47. Calculate seasonal indices for the rainfall data of Tamil Nadu by using simple average method.

Quarter Year	I	II	III	IV
2000-01	314.5	335.6	16.8	118.4
2001-02	260.0	379.4	70.0	85.8
2002-03	185.4	407.1	8.7	129.8
2003-04	336.5	403.1	12.0	283.4
2004-05	360.7	472.1	14.3	231.7

48. Find seasonal Indices for the rainfall data in Tamil Nadu (in mm)

Quarter Year	2009	2010	2011	2012
I	38.2	38.5	55	50.5
II	166.8	250.9	277.7	197
III	612.6	773.1	717.8	706.1
IV	72.2	153.1	65.8	101.1

49. The following table gives quarterly expenditure over a number of years. Obtain seasonal correction for the data

Year Season	2000	2001	2002	2003
I	78	84	92	100
II	62	64	70	81
III	56	61	63	72
IV	71	82	83	96

50. Find the trend values using semi averages method. The following table shows the area covered for cultivation of Ragi in Tamil Nadu (in ʻ000 hectares)

Year	2003	2004	2005	2006	2007	2008	2009	2010
Area (in ʻ000 hectares)	118	109	100	95	94	90	82	76

(Hint: Decreasing trend)

ANSWERS

I	**1.** (b)	**2.** (c)	**3.** (d)	**4.** (a)	**5.** (d)
	6. (d)	**7.** (c)	**8.** (d)	**9.** (c)	**10.** (b)
	11. (a)	**12.** (b)	**13.** (b)	**14.** (b)	**15.** (c)

III **34.** i). Cyclic variation ii). Irregular variation iii). Secular trend

37. 22.44, 23, 23.56, 24.11, 24.67, 25.22

38. –, 108.67, 898.33, 1057.33, 1046, 1094.67, –

IV **41.** 48.00, 62.67, 77.33, 92, 106.67, 121.33, 136

42. 691.66, 709.72, 727.78, 745.84, 763.91, 781.97, 800.03, 818.09

43. –, –, 1658.75, 1659.625, 1684.875, 1729.125, 1777.875, 1820.375, –, –

44. 70, 76.2, 82.4, 88.6, 94.8, 101

45. 19.52, 21.38, 23.24, 25.1, 26.95, 28.81

Number of telephones in the year 2000 is 3810

46. 10.86, 12.57, 14.29, 16, 17.71, 19.42, 21.14

Export for the year 2005 is 28 millions

47. 132, 181, 11, 77

48. 17, 83, 263, 37

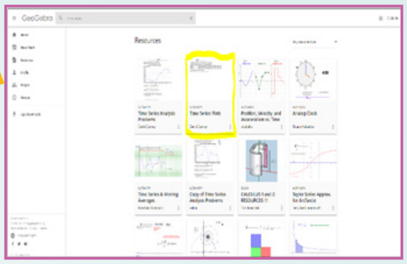
49. 117, 91, 83, 109

50. 113, 108, 103, 98, 93, 88, 83, 78

ICT CORNER

TIME SERIES AND FORECASTING

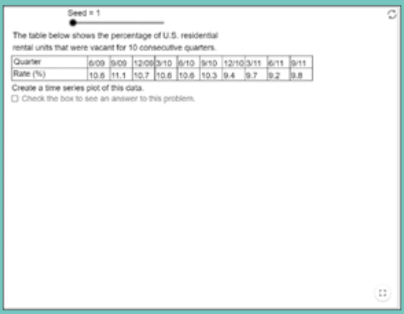
This activity helps to Understand about Times Series and Forecasting



Steps:

- Open the browser and type the URL given (or) scan the QR code.
- GeoGebra work book called "**Time Series Plot**" will open.
- Move the "seed" slider to select a new sample.
- Put the Tick Mark in the Check box to see an answer to the problem.

Step-1



Seed = 1

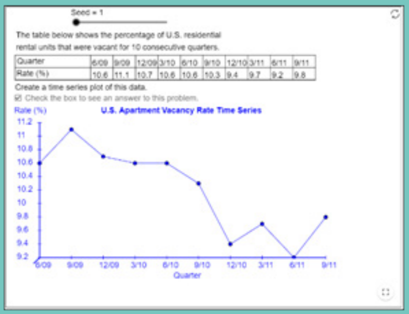
The table below shows the percentage of U.S. residential rental units that were vacant for 10 consecutive quarters.

Quarter	6/09	9/09	12/09	3/10	6/10	9/10	12/10	3/11	6/11	9/11
Rate (%)	10.6	11.1	10.7	10.6	10.6	10.3	9.4	9.7	9.2	9.8

Create a time series plot of this data.

□ Check the box to see an answer to this problem.

Step-2



Seed = 1

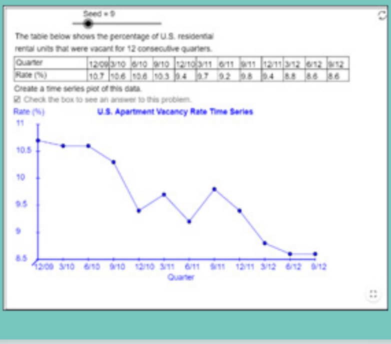
The table below shows the percentage of U.S. residential rental units that were vacant for 10 consecutive quarters.

Quarter	6/09	9/09	12/09	3/10	6/10	9/10	12/10	3/11	6/11	9/11
Rate (%)	10.6	11.1	10.7	10.6	10.6	10.3	9.4	9.7	9.2	9.8

Create a time series plot of this data.

☑ Check the box to see an answer to this problem.

Step-3



Seed = 9

The table below shows the percentage of U.S. residential rental units that were vacant for 12 consecutive quarters.

Quarter	12/09	3/10	6/10	9/10	12/10	3/11	6/11	9/11	12/11	3/12	6/12	9/12
Rate (%)	10.7	10.6	10.6	10.3	9.4	9.7	9.2	9.8	9.4	8.8	8.6	8.6

Create a time series plot of this data.

☑ Check the box to see an answer to this problem.

Pictures are indicatives only*

URL:

https://www.geogebra.org/m/Cd6dvjMV

CHAPTER

8 VITAL STATISTICS AND OFFICIAL STATISTICS

William Farr

P.C. Mahalanobis

William Farr (1807–1883) was a British Epidemiologist, regarded as one of the founders of Medical Statistics. Farr, systematically collected and analyzed Britain's Vital Statistics, is known as "the father of Modern Vital Statistics and surveillance." Farr's name features on the wall painting of the London School of Hygiene and Tropical Medicine (LSHTM) in recognition of his work. Twenty three names of public health and tropical medicine pioneers were chosen to appear on the School building in Keppel Street when it was constructed in 1926.

P.C. Mahalanobis (1893-1972) took initiatives for establishment of the Official Statistical System in India. He was the first Honorary Statistical Advisor to the Government of India. He was instrumental for establishing the present Central Statistical Office and National Sample Survey Office. He is the founder of the internationally renowned Indian Statistical Institute (ISI) at Kolkata, which has centres located at various parts of the country. ISI also concentrates on other disciplines of study which include Geology, Sociology, Computer Science and Theoretical Physics.

LEARNING OBJECTIVES

The students will be able to understand

- the importance of Vital Statistics
- various mortality rates
- different fertility rates
- different components of Life Table
- construction of Life Table for a given community
- evolvement of Statistical System in India
- the role of Official Statistics
- the establishment and roles of different divisions of CSO and NSSO.
- the present Statistical System in India

3402Y

8.1 VITAL STATISTICS

Introduction

Demography is a term, generally, concerned with human population and it is also concerned with the social implications of periodical variations taking place in the population with reference to geographical location(s).

Vital Statistics is a branch of Demography, which is the science applied to the analysis and interpretation of numerical facts regarding vital events occurring in human population such as births, deaths, marriages, divorces, migration *etc*.

The following are some of the important definitions of Vital Statistics:

> *"The whole study of mankind by heredity or environment in so far as the results of their study can be arithmetically stated"*
>
> ***–Arthur Newsholme***
>
> "Vital Statistics are conventional numerical records of marriages, births, sickness and deaths by which the health and growth of a community may be studied".
>
> ***–Benjamin***

Vital Statistics is the science of numbers applied to the life history of communities or regions.

8.1.1 Importance of Vital Statistics

Vital Statistics are quantitative measurements on live births, deaths, foetal deaths, infant deaths, fertility and so on.

- Vital Statistics are essential for conducting demographic studies on human community during specific time period.
- Vital Statistics play an important role in the development of a country especially in heath care.
- They are of great use in planning and evaluation of socio-economic and public health development of a country.
- They help to identify factors relating to fluctuations in mortality and fertility rates.
- Maternal and infant mortality are important indications of nation's health, thereby influencing the government to develop policies, funding of programs to maintain quality of health care. Timely documentation of births and deaths is essential to maintain high quality indices.
- They are also of great use for comparison of health indicators at national and international levels.
- They are useful in medical and demographical, actuarial studies and research.
- They are of great use to the government to assess the impact of various family welfare programmes implemented in a country.

- Vital Statistics reflect the changing pattern of the population of any region, community or country in terms of vital events.
- Vital Statistics help to compare two different regions or communities or countries with respect to public health based on vital events.

8.1.2 Collection of Vital Statistics

The following are the five methods normally adopted for collecting data related to various vital events:

(i) Civil Registration System
(ii) Census or Complete Enumeration method
(iii) Survey method
(iv) Sample Registration System
(v) Analytical method

(i) Civil Registration System

Civil Registration System is the most common method of collecting information on vital events. It is an administrative procedure followed by governments, to record various vital events occurring in their population.

In this method, occurrence of the vital events such as births, deaths, marriages, migration *etc*., are registered. Many countries adopt this system. Registration is done with the Authorities appointed by the respective government. In India, registration of births and deaths are made compulsory by legislation, through an act *viz*., "The Registration of Births and Deaths Act, 1969". It came into force throughout the country through a gazette notification published in 1970.

(ii) Census or Complete Enumeration Method

Census presents a comprehensive profile of the country's population. Census is conducted in most countries at intervals of ten years. The complete enumeration method normally covers data regarding age, sex, marital status, educational level, occupation, religion and other factors needed for computing Vital Statistics. However, all these information are available for the census year only.

(iii) Survey method

Ad hoc surveys are conducted in areas where the recording of births and deaths has not been done properly and periodically, particularly in those areas where registration offices have not been established. However, survey records help to provide Vital Statistics for that region only.

(iv) Sample Registration System

Vital rates are required to monitor population growth, especially for the purpose of evaluation of family planning programmes in terms of their ultimate objective of controlling fertility.

Sample Registration System is adopted at both national and state levels in India to collect the following information:

National Level

(a) Infant mortality

(b) Age specific mortality rates in rural areas

(c) Sampling variability of vital rates

State Level

(a) Differences in birth rates with respect to education, religion, parity

(b) Sex ratio

(a) Seasonality in birth and death rates

(v) Analytical Method

It is generally not possible to conduct *ad hoc* surveys to assess the population at any specific period between two consecutive census years. Population estimates of any vital event at a given time can be obtained without *ad hoc* surveys applying analytical methods which use mathematical formulae.

Calculation of Vital Rates

Generally, rate of a vital event is calculated using the formula

$$\textbf{Rate of a vital event} = \frac{Number\ of\ occurrences\ of\ the\ event\ during\ the\ specified\ period}{Size\ of\ the\ population\ exposed\ to\ the\ risk\ of\ the\ event} \times 1000$$

Rates of vital events are usually expressed '*per thousand*'.

8.1.3 Mortality and Its Measurements

Mortality refers to the deaths, which occur in the population/community/region due to sickness, accidents, etc.

Several rates are used for measuring mortality. We will discuss the following primary mortality rates.

(i) Crude Death Rate

(ii) Specific Death Rate

(iii) Infant Mortality Rate

(i) Crude Death Rate

Crude Death Rate (*CDR*) is the simplest type of death rate, which relates the number of deaths in a specific community or region to the population size of the community in a given period, preferably on yearly basis. The formula used to calculate this mortality rate for a given period is

$$CDR = \frac{D}{P} \times 1000$$

where

D: Number of deaths in the population or community during the given period, and

P: Number of persons in the population or community during the given period.

Example 8.1

There were 15,000 persons living in a village during a period and the number of persons dead during the same period was 98.

Then, the *CDR* of the village can be calculated from

$$CDR = \frac{D}{P} \times 1000$$

as

$$CDR = \frac{98}{15000} \times 1000 = 6.53 \textit{ per thousand.}$$

In some cases, information about the population may be provided in such a manner that people in the population are grouped according to their age into mutually exclusive and exhaustive age groups. The *CDR* can also be calculated based on such kind of information.

Example 8.2

People living in a town are grouped according to their age into five groups. The number of persons lived during a calendar year and the number of deaths recorded during the same period are as follows:

Age Group (in years)	0-10	10-30	30-50	50-70	70 and above
No. of Persons	5,000	10,000	15,000	10,000	2,000
No. of Deaths	125	30	30	200	1,000

Calculate crude death rate of the town.

Solution:

The total number of deaths occurred in the town, irrespective of the age, is 1385 and the population size is 42,000. Therefore, the *CDR* of the town can be calculated as

$$CDR = \frac{1385}{42000} \times 1000 = 32.98$$

Thus, the Crude Death Rate of the town is 32.98 *per thousand.*

CDR is the widely used mortality measure. It is popular, because it is very easy to compute. However, it is only a crude measure of mortality, which does not take into account of the age and sex, on the whole, of the population/community/region. The probability of death may not be same at all ages. Hence, if the age distribution of two different communities are not similar, then comparing the communities based on their *CDR* can mislead. Also, use of *CDR* may again be misleading, since the probability of death may vary over gender, though they are at the same age.

(ii) Specific Death Rate

Mortality pattern may differ in different sections/segments of the population such as age, gender, occupation *etc*.

Specific Death Rate (*SDR*) can be calculated exclusively for a section of the population. The *SDR* can be calculated for a group of persons, who are distinguished by age or gender or occupational class or marital status. The formula to calculate *SDR* is

$$SDR = \frac{D_S}{P_S} \times 1000,$$

where

D_S refers to the number of deaths in a specific section of population during the given period, and

P_S refers to the total number of persons in the specific section of population during the given period.

*SDR*s can be calculated for any age group, gender, religion, caste or community. If the death rates are calculated for different age groups, say, 0-5 years, 5-15 years, 50-60 years, they are called age specific death rates (*ASDR*s). In the age group ($x,x+n$), all the persons in the population or in its section, who have attained the age of x years and the persons with age less than $x+n$ years, are included in the age group.

If

$D(x,n)$ denotes the number of deaths in the age group ($x,x+n$) recorded in a given region during a given period, and

$P(x,n)$ denotes the number of persons in the age group ($x,x+n$) in the region during the same period, then

ASDR for the age group ($x,x+n$) for the given region during the period is given by

$$ASDR(x,n) = \frac{D(x,n)}{P(x,n)} \times 1000.$$

The death rates, calculated for persons belonging to different gender, is called as gender specific death rate. The *SDR* can also be used to compare the death rates due to different kinds of seasonal diseases such as *dengue*, *chikungunya*, *swine flu*.

The *SDR* helps to measure the death rates for different sections/segments of the population unlike *CDR*. *ASDR* and *SDR* for gender can be used to compare the death rates of the respective sections of the given population in different regions.

Example 8.3

Number of deaths recorded in various age groups in two areas, *viz.*, Area I and Area II and the population size in each age group are given in the following table.

Age Group (in years)	Area I		Area II	
	Population	No. of Deaths	Population	No. of Deaths
0-10	3000	55	7500	300
10-25	4500	30	6000	50
25-45	6000	40	8000	40
45 and over	1000	15	2000	64

Find the crude death rates and age specific death rates of Area I and Area II.

Solution:

Age Specific Death Rate can be calculated for each age group using the formula

$$ASDR(x,n) = \frac{D(x,n)}{P(x,n)} \times 1000$$

Calculation of the *ASDR* for both the areas in each age group is presented in the following table:

Age Group (x)	Area I			Area II		
	$P(x,n)$	$D(x,n)$	ASDR(x,n) (per thousand)	$P(x,n)$	$D(x,n)$	ASDR(x,n) (per thousand)
0-10	3000	55	$\frac{55}{3000}\times 1000 = 18.33$	7500	300	$\frac{300}{7500}\times 1000 = 40.00$
10-25	4500	30	$\frac{30}{4500}\times 1000 = 6.67$	6000	50	$\frac{50}{6000}\times 1000 = 8.33$
25-45	6000	40	$\frac{40}{6000}\times 1000 = 6.67$	8000	40	$\frac{40}{8000}\times 1000 = 5.00$
45 and over	1000	15	$\frac{15}{1000}\times 1000 = 15.00$	2000	64	$\frac{64}{2000}\times 1000 = 32.00$
Total	14500	140		23500	454	

$$CDR \text{ of Area I} = \frac{140}{14500}\times 1000 = 9.66 \textit{ per thousand}$$

$$CDR \text{ of Area II} = \frac{454}{23500}\times 1000 = 19.32 \textit{ per thousand}$$

Example 8.4

The following are the information about the number of persons who are affected by Diabetes and Lung Cancer and the number of persons died due to each cause of death during a calendar year in two different districts:

Cause of Death	District A		District B	
	No. of Persons		No. of Persons	
	Affected	Died	Affected	Died
Diabetes	20,000	325	22,000	400
Lung Cancer	19500	300	21,225	380

Find the Illness specific death rates for the two districts. Also, compare health conditions of both the districts with reference to these two causes of death. Assume that a person affected by Diabetes is not affected by Lung Cancer and *vice-versa*.

Solution:

The *SDR* due to the two causes of death are calculated as follows:

SDR of District A

$$SDR_{Diabetes} = \frac{D_{Diabetes}}{P_{Diabetes}} \times 1000$$

$$= \frac{325}{20000} \times 1000$$

$$SDR_{Diabetes} = 16.25 \textit{ per thousand}$$

$$SDR_{Lung\ Cancer} = \frac{D_{Lung\ Cancer}}{P_{Lung\ Cancer}} \times 1000$$

$$= \frac{300}{19500} \times 1000$$

$$SDR_{Lung\ Cancer} = 15.38 \textit{ per thousand}$$

SDR of District B

$$SDR_{Diabetes} = \frac{400}{22000} \times 1000$$

$$SDR_{Diabetes} = 18.18 \textit{ per thousand}$$

$$SDR_{Lung\ Cancer} = \frac{380}{21225} \times 1000$$

$$SDR_{Lung\ Cnacer} = 17.90 \textit{ per thousand}$$

In both the districts, death rates are more due to Diabetes in comparison with Lung Cancer. Among the two districts, District B has relatively more death rate due to both Diabetes and Lung Cancer.

The *SDR* can be calculated with respect to gender also. For example, *SDR* for males in a given region during a given period can be calculated as

$$SDR_{Male} = \frac{D_{Male}}{P_{Male}} \times 1000 \, .$$

Here, D_{Male} is the number of male deaths in the region during the specified period, and

P_{Male} is the male population size in the region during the specified period.

(iii) Infant Mortality Rate

'Infant' means a baby of age less than a year. **Infant Mortality Rate** (*IMR*) is defined as the number of infant deaths *per thousand* live births in a period or children die before they attain age of one year. The following formula is used to calculate infant mortality rate

$$IMR = \frac{D_{Infant}}{P_{Infant}} \times 1000$$

where

D_{Infant} denotes the number of deaths of infants in a population during a period, and

P_{Infant} denotes the number of live births in the population during the period.

Child Mortality Rate: Child mortality is the death of a child before the child's fifth birth day, measured as the under 5 child mortality rate (*U5MR*).

Example 8.5

The number of live births recorded and the number of infants died in a town during a given period are respectively 400 and 25. Calculate, from these information, the infant mortality rate of the town for the period.

Solution:

The *IMR* of the town is given by

$$IMR = \frac{25}{400} \times 1000$$

$$IMR = 62.50 \text{ } \textit{per thousand}.$$

8.1.4 Life Table and Its Applications

A **Life Table** is a presentation of summary of the mortality experiences of a community during a given period in the form of a table. The Life Table exhibits the number of persons living and dying at each age, on the basis of the experience of a ***cohort***. It also gives the probability of dying and living separately. The Life Table tells the life history of a ***cohort***.

Cohort is a group of individuals who born at the same time and experienced the same mortality conditions.

Uses of Life Table

- Actuaries compute rate of premiums for persons of different age groups using Life Table.
- It helps to assess the accuracy of census figures, death and birth registrations.
- It helps to evaluate the impact of family planning on population growth.
- It enables to assess the increase in the life span due to new scientific inventions, sophisticated medical treatments and improved living conditions.
- Estimates of migration can be made from Life Table.

Construction of Life Table

Construction of Life Table begins with a ***cohort*** population. The following are the standard set of components of a Life Table:

(i) Age (x)

(ii) Survivorship function

(iii) Number of deaths in the age interval (x, x+1)

(iv) Probability for a person surviving at the age x to die before x+1 years

(v) Probability for a person aged x years to survive upto x+1 years

(vi) Number of persons lived in aggregate in the age interval (x,x+1)

(vii) Number of years lived by the ***cohort*** at and above the age x years

(viii) Expectation of life.

These components are described below with their respective notations and formula required to compute each of them.

(i) x : Age, in years

(ii) $l(x)$: Number of survivors at the exact age of x years.

For instance, $l(25)$ denotes the number of persons who survive to the moment of attaining age 25 years. Hence, $l(x)$ is a decreasing function of x.

$l(0)$ is called **radix** of the Life Table or ***cohort***, which is conventionally taken as 1,00,000.

(iii) $d(x)$: Number of persons among $l(x)$ persons who die before reaching the age x years.

i.e., $d(x) = l(x) - l(x+1)$

(iv) $q(x)$: Probability for a person surviving at the age x will die before x+1 years.

$$i.e.,\ q(x) = \frac{d(x)}{l(x)}$$

It is the proportion of persons dying between the ages of x and x +1 to the number of persons surviving at the age of x years, *i.e.,* at the beginning of the corresponding age interval.

(v) $p(x)$: Probability for a person aged x years to survive up to x+1 years

$$i.e.,\ p(x) = 1 - q(x), \text{ or equivalently, } p(x) = \frac{l(x+1)}{l(x)}$$

It is the proportion of persons surviving up to (x + 1) years to the number of persons at the age of x years.

(vi) $L(x)$: Number of persons lived in aggregate in the age interval $(x,x+1)$

i.e., $L(x)=\dfrac{l(x)+l(x+1)}{2}$

or equivalently

$L(x)=l(x)-\dfrac{1}{2}d(x)$

(vii) $T(x)$: Number of years lived by the ***cohort*** at and above the age x

i.e., $T(x) = L(x)+L(x+1)+L(x+2)+ \dots$

or equivalently

$T(x+1) = T(x) - L(x)$.

Total number of years lived by the ***cohort*** after x years of age.

(viii) $e^0(x)$: Expectation of life

$e^0(x)=\dfrac{T(x)}{l(x)}$

It gives the average number of years a person of age x years is expected to survive under the existing mortality conditions.

Assumptions of Life Table

The following assumptions are made while constructing a Life Table.

(i) There is no effect of immigration and emigration on the ***cohort***. It means that the reduction in the number of the initial ***cohorts*** is merely due to deaths.

(ii) The size of ***cohort*** begins with a convenient figure, it is conventionally 1,00,000.

(iii) Deaths are uniformly distributed over each age interval.

Example 8.6

A Life Table was constructed for a ***cohort***. The following is a section of the table, wherein some of the entries are not available. Find the estimates of missing values and complete the Life Table.

Age (in years)	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
40	10, 645	-	-	-	-	1, 93, 820	-
41	10, 543	169	-	-	-	-	-

Solution:

The Life Table can be completed using the relationship among missing terms and other terms.

The number of persons who die before reaching age 40 years is calculated as

$$d(40) = l(40) - \text{l}(41)$$
$$= 10645 - 10543$$

Therefore, d(40) = 102.

Values of $q(x)$ are estimated as

$$q(40) = \frac{d(40)}{l(40)}$$

$$= \frac{102}{10645}$$

$$= 0.0095.$$

$$q(41) = \frac{d(41)}{l(41)}$$

$$= \frac{169}{10543}$$

$$= 0.0160.$$

Values of $p(x)$ are estimated from the corresponding values of $q(x)$ as

$$p(40) = 1 - q(40)$$

$$= 1 - 0.0095 = 0.9905$$

$$p(41) = 1 - q(41)$$

$$= 1 - 0.0160 = 0.9840$$

Values of $L(x)$ are estimated using its relationship with $l(x)$ and $d(x)$ as follows:

$$L(40) = \frac{l(40)+l(41)}{2}$$

$$= \frac{10645+10543}{2}$$

$$= 10{,}594$$

$$L(41) = l(41) - \frac{1}{2}d(41)$$

$$= 10{,}543 - \frac{1}{2}\times 169 = 10{,}458.5$$

$$= 10{,}459\ (Approx.)$$

The value of $T(41)$ is estimated from the given value of $T(40)$ and the estimated value of $L(40)$ from the relationship

$$T(41) = T(40) - L(40)$$

as $$T(41) = 193820 - 10594$$

$$= 1{,}83{,}226.$$

The life expectancy of the ***cohort*** at the age $x = 40$ and 41 years can be estimated using the relationship

$$e^0(x) = \frac{T(x)}{l(x)}$$

as

$$e^0(40) = \frac{1{,}93{,}820}{10{,}645} = 18.20$$

$$e^0(41) = \frac{1,83,226}{10,543} = 17.37$$

Now, the completed Life Table becomes as

x	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
40	10, 645	102	0.9905	0.0095	10,594	1, 93, 820	18.20
41	10, 543	169	0.9840	0.0160	10,459	1,83,226	17.37

Example 8.7

The following is a part of the Life Table constructed for a population, where the contents are incomplete. Evaluate the missing values using the given data and complete the Life Table.

x	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
83	3560	-	-	0.16	-	-	
84	-	508	-	0.17	-	11975	

Solution:

Values of the missing entries can be estimated from the given data applying the respective formulae as follows:

The number of persons who die before reaching age of 83 years is calculated as

$$d(83) = l(83) \times q(83)$$

$$= 3560 \times 0.16$$

$$= 569.6 = 570$$

The value of the survivorship function *l(x)* at $x = 84$ years is estimated as

$$l(84) = l(83) - d(83)$$

$$= 3560 - 570$$

$$= 2990$$

Values of $p(x)$ are estimated from the corresponding values of $q(x)$ as

$$p(83) = 1 - q(83)$$

$$= 1 - 0.16 = 0.84$$

$$p(84) = 1 - q(84)$$

$$= 1 - 0.17 = 0.83$$

Values of *L(x)* are estimated using its relationship with *l(x)* and *d(x)* as follows:

$$L(83)=\frac{l(83)+l(84)}{2}$$

$$=\frac{3560+2990}{2}$$

$$= 3{,}275$$

$$L(84)=l(84)-\frac{1}{2}d(84)$$

$$2990-\frac{508}{2}$$

$L(84) = 2736.$

The value of $T(83)$ can be estimated from the given value of $T(84)$ and the estimated value of $L(83)$ from the relationship

$T(84) = T(83) – L(83)$

$T(83) = L(83) + T(84)$

as

$T(83) = 3{,}275 + 11975 = 15{,}250$

The life expectancy of the ***cohort*** at the age x = 83 and 84 years is estimated using the relationship

$$e^0(x)=\frac{T(x)}{l(x)}$$

as

$$e^0(83)=\frac{15250}{3560}=4.28$$

$$e^0(84)=\frac{11975}{2990}=4.01$$

Now, the completed Life Table is as follows:

x	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
83	3560	570	0.84	0.16	3275	15250	4.28
84	2990	508	0.83	0.17	2736	11975	4.01

Example 8.8

A part of the Life Table of a population is given hereunder with incomplete information. Find those information from the given data and complete the Life Table.

Age (in years)	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
72	4412	-	-	-	-	-	-
73	3724	-	-	-	-	-	-
74	3201	642	-	-	-	26567	-

Solution:

Values of the missing entries can be calculated from the given data applying the respective formulae as follows:

The number of persons who die before reaching age $x = 72$ and 73 years can be calculated as

$$d(72) = l(72) - l(73)$$
$$= 4412 - 3724$$
$$= 688$$

$$d(73) = l(73) - l(74)$$
$$= 3724 - 3201$$
$$d(73) = 523.$$

Values of $q(x)$ are estimated as

$$q(72) = \frac{d(72)}{l(72)}$$
$$= \frac{688}{4412}$$
$$= 0.1559$$

$$q(73) = \frac{d(73)}{l(73)}$$
$$= \frac{523}{3724}$$
$$q(73) = 0.1404$$

$$q(74) = \frac{d(74)}{l(74)}$$
$$= \frac{642}{3201}$$
$$= 0.2006.$$

Values of $p(x)$ are estimated from the corresponding values of $q(x)$ as

$$p(72) = 1 - q(72)$$
$$= 1 - 0.1559 = 0.8441$$

$p(73) = 1 - q(73)$

$= 1 - 0.1404 = 0.8596$

$p(74) = 1 - q(74)$

$= 1 - 0.2006 = 0.7994.$

Values of $L(x)$ are estimated using its relationship with $l(x)$ and $d(x)$ as follows:

$$L(72) = \frac{l(72)+l(73)}{2}$$

$$= \frac{4412+3724}{2}$$

$$= 4{,}068$$

$$L(73) = \frac{l(73)+l(74)}{2}$$

$$= \frac{3724+3201}{2}$$

$$L(74) = l(74) - \frac{d(74)}{2}$$

$$= 3201 - \frac{642}{2}$$

$$= 2880.$$

The value of $T(x)$ is estimated for x = 72 and 73 from the given value of $T(74)$ and the estimated values of $L(72)$ and $L(73)$ as

$T(73) = L(73) + T(74)$

$= 3463 + 26567 = 30{,}030.$

$T(72) = L(72) + T(73)$

$= 4068 + 30030 = 34{,}098.$

The life expectancy of the ***cohort*** at the age x = 72, 73 and 74 years is estimated using the relationship

$$e^0(x) = \frac{T(x)}{l(x)}$$

as

$$e^0(72) = \frac{34098}{4412} = 7.73$$

$$e^0(73) = \frac{30030}{3724} = 8.06$$

$$e^0(74) = \frac{26567}{3201} = 8.30$$

The completed Life Table is as follows:

x	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
72	4412	688	0.8441	0.1559	4,068	34,098	7.73
73	3724	523	0.8596	0.1404	3,463	30,030	8.06
74	3201	642	0.7994	0.2006	2,880	26,567	8.30

Example 8.9

Find the missing values in the following Life Table:

Age (in years)	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
4	95,000	500	-	-	-	48,50,300	-
5	-	400	-	-	-	-	-

Solution:

Value of the survivorship function $l(x)$ at $x = 5$ years can be estimated as

$$l(5) = l(4) - d(4)$$

$$= 95000 - 500$$

$$= 94500$$

Values of $q(x)$ are estimated as

$$q(4) = \frac{d(4)}{l(4)}$$

$$= \frac{500}{95000}$$

$$= 0.005$$

$$q(5) = \frac{d(5)}{l(5)}$$

$$= \frac{400}{94500}$$

$$= 0.004.$$

Values of $p(x)$ are estimated from the corresponding values of $q(x)$ as

$$p(4) = 1 - q(4)$$

$$= 1 - 0.005 = 0.995$$

$$p(5) = 1 - q(5)$$

$$= 1 - 0.004 = 0.996.$$

Values of *L(x)* are estimated using its relationship with *l(x)* and *d(x)* as follows:

$$L(4) = \frac{l(4)+l(5)}{2}$$

$$= \frac{95000+94500}{2}$$

$$= 94,750$$

$$L(5) = l(5) - \frac{d(5)}{2}$$

$$= 94500 - \frac{400}{2}$$

$$= 94,300.$$

The value of *T(5)* is estimated from the given value of *T(4)* and the estimate of *L(4)* as

$T(5) = T(4) – L(4)$

$T(5) = 4850300 – 94750 = 47,55,550.$

The life expectancy of the ***cohort*** at the age $x = 4$ and 5 years is estimated using the relationship

$$e^0(x) = \frac{T(x)}{l(x)}$$

as

$$e^0(4) = \frac{4850300}{95000} = 51.06$$

$$e^0(5) = \frac{4755550}{94500} = 50.32$$

The completed Life Table is

x	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
4	95,000	500	0.995	0.005	94,750	48,50,300	51.06
5	94,500	400	0.996	0.004	94,300	47,55,550	50.32

8.1.5 Fertility and its Measurements

Fertility refers to births occurring to the women who are at child bearing age. A woman at child bearing age is defined as the age of the female who can give birth to a child. In other words, it is the reproductive age of woman.

Fertility rates are the quantitative characteristics, which are used to measure the rate of growth of the population due to births during a specified period, usually a year. The fertility rates are expressed *per thousand* women who are at child bearing age.

As like mortality rates, there are several fertility rates. Among them, the following are the basic fertility rates discussed here

(i) Crude Birth Rate

(ii) Specific Fertility Rate

(iii) General Fertility Rate

(i) Crude Birth Rate

Crude Birth Rate (*CBR*) of a region or a community relates the number of live births to size of the population of the region or the community. This quantity can be computed using the formula

$$CBR = \frac{B_t}{P_t} \times 1000$$

where

B_t denotes the number of live births occurred in a given region/community during the period t, and

P_t denotes the population size of the given region/community during the period t.

Example 8.10

The number of children born in a city during a period was 15,628 and the total population of the city in that period was 80,00,000. Find the crude birth rate of the city.

Solution:

The Crude Birth Rate can be calculated using the formula

$$CBR = \frac{B_t}{P_t} \times 1000$$

The *CBR* of the city is

$$CBR = \frac{15628}{8000000} \times 1000$$

$$= 1.95 \text{ per thousand.}$$

Example 8.11

People living in a town are grouped according to their age into nine groups. Details about the number of live birts are also grouped according to the age group of women. These information are presented in the following table:

Age Group (in years)	Less than 15	15-20	20-25	25-30	30-35	35-40	40-45	45-49	49 and above
No. of Persons	20,000	15,000	19,000	21,000	25,000	20,000	18,000	16,000	35,000
No. of Live Births	0	30	200	1,000	1500	800	500	100	0

Calculate crude birth rate of the town.

Solution:

The total number of persons in the town during the specified period can be calculated from the given information as

$$P_t = 1{,}89{,}000$$

and the total number of live births as

B_t = 4,130.

Hence, the Crude Birth Rate of the town is

$$CBR = \frac{4130}{189000} \times 1000$$

= 21.85 per thousand.

CBR is the simplest fertility rate, which provides a comprehensive idea about the population growth of any region/community. It is easy to compute for any given region.

However, *CBR* does not take into account of the age and gender distribution of the population or specifically the number of women at the child bearing age. This is a crude measure, since it includes the sections of the population who are not in the reproductive age group. It means that P_t includes the sections of the population who are not exposed to the risk of producing children, in particular, male and also the female beyond the reproductive age. *CBR* also assumes that all the women at child bearing age have the same reproductive capacity irrespective of their age, which is totally unrealistic.

(ii) General Fertility Rate

General Fertility Rate (*GFR*) of a region or a community relates the number of live births to the number of women in the reproductive age.

i.e.,

$$GFR = \frac{Number\ of\ \ live\ births}{Number\ of\ women\ at\ child\ bearing\ age} \times 1000$$

This quantity can be computed using the formula

$$GFR = \frac{B_t}{\sum_{i=a_1}^{a_2} P_t^i} \times 1000$$

where

P_t^i denotes the number of women in the reproductive age *i* years in the given region/community during the period *t*, $i = a_1$ to a_2.

In India, generally, a_1 = 15 years and a_2 = 49 years.

GFR overcomes the disadvantage of *CBR* considering only the women population at the child bearing age group, since the denominator in the above formula represents the entire women population at the reproductive age group. *GFR* expresses the increase in the women population at the child bearing age through live births.

However, *GFR* does not express the age composition of women population at the reproductive age group. Hence, two different regions/communities cannot be compared with respect to age of women using *GFR*.

Example 8.12

Women, at child bearing age, of a district are grouped into seven age groups. The number of women lived during a calendar year in the district and the number of live births recorded during the same period are as follows:

Age Group (in years)	15-20	20-25	25-30	30-35	35-40	40-45	45-49
No. of Persons	20,000	22,000	28,000	32,000	29,000	24,000	8,000
No. of Live Births	50	1500	1700	2,000	1800	500	80

Calculate the general fertility rate of the district.

Solution:

The total number of women, at child bearing age in the district during the study period can be calculated from the given information as

$$\sum_{i=15}^{49} P_t^i = 1,63,000$$

and the total number of live births as

$B_t = 7,630.$

Hence, the general fertility rate of the district is

$$GFR = \frac{7630}{163000} \times 1000$$

$$= 46.81 \textit{ per thousand.}$$

(iii) Specific Fertility Rate (SFR)

It is a well known fact that fertility is affected by several factors such as age, marriage, migration, region *etc*. But, both *CBR* and *GFR* do not take into account of this fact. In this respect, **Specific Fertility Rate** (*SFR*) is defined as

$$SFR = \frac{\text{Number of live births to the women population in the reproductive age groups of specific section in a given period}}{\text{Total number of women in the reproductive age groups of the specific section in the given period}} \times 1000$$

SFR can be calculated separately for various age groups of females who are at child bearing age such as 15-20, 20-25, and so on. The *SFR* computed with respect to different reproductive age of women is known as the **Age Specific Fertility Rate** (*ASFR*), which can be calculated using the formula

$$ASFR(x, x+n) = \frac{B_t(x, x+n)}{P_t(x, x+n)} \times 1000$$

where

$B_t(x, x+n)$ denotes the number of live births to the women in the reproductive age group (*x,x+n*) during the period *t* in the given region, and

$P_t(x, x+n)$ denotes the number of women in the reproductive age group (*x,x+n*) during the period *t* in the given region.

Grouping of women with respect to their age is essential, since capacity of women to give birth to child varies over the age of women. Thus, *ASFR* enables to compare the fertility rates of two or more different regions with respect to specific age groups. Moreover, *ASFR* can be considered as a probability value.

Example 8.13

The female population, at reproductive age, of a country is grouped into six age groups. The number of women in each group and the number of live births given by them are given the following table.

Age Group (in years)	No. of Women	No. of Live Births
15-20	1,16,610	10,668
20-25	1,13,810	17,183
25-30	1,03,130	12,722
30-35	93,500	7,283
35-40	74,120	3,656
40-45	62,900	1,340

Calculate the general fertility rate and the age specific fertility rates of the country.

Solution:

Total number of women in the country at reproductive age during the period of '*t*' is

$$\sum_{i=15}^{44} P_t^i = 5,64,070$$

and the total number of live births of the country during the same period is

$$B_t = 52,852$$

Therefore, the general fertility rate of the country is

$$GFR = \frac{52852}{564070} \times 1000$$

$$= 93.69 \textit{ per thousand.}$$

The age specific fertility rates of the country can be calculated for each age group using the formula

$$ASFR(x, x+n) = \frac{B_t(x, x+n)}{P_t(x, x+n)} \times 1000$$

where

$B_t(x, x+n)$ denotes the number of live births to the women in the reproductive age group $(x,x+n)$ during the period t in the given region, and

$P_t(x, x+n)$ denotes the number of women in the reproductive age group $(x,x+n)$ during the period t in the given region.

The *ASFRs* are calculated from the given information and are presented in the following table:

Age Group	*ASFR*
15-20	$\frac{10668}{116610}\times 1000 = 91.48$
20-25	$\frac{17183}{113810}\times 1000 = 150.98$
25-30	$\frac{12722}{103130}\times 1000 = 123.36$
30-35	$\frac{7283}{93500}\times 1000 = 77.89$
35-40	$\frac{3656}{74120}\times 1000 = 49.33$
40-45	$\frac{1340}{62900}\times 1000 = 21.30$

It is can be observed, with respect to *ASFRs*, that the women in the country falling in the age group of 20-25 years have given relatively more live births. The women at the age of 40-45 years have reproduced less number of live births.

Example 8.14

The following is the data regarding the size of female population in a country at reproductive age and the live births during a period.

Age Group (yrs)	Female Population	No. of Live Births
15–20	2,16,410	20,468
20–25	2,13,610	26,983
25–30	2,02,930	22,522
30–35	1,93,300	17,083
35–40	1,73,920	13,456
40–45	1,62,870	11,140

Calculate the general fertility rate and the age specific fertility rates of the country.

Solution:

Size of the women population of the country at reproductive age during the period ‘*t*’, is

$$\sum_{i=15}^{44} P_t^i = 11,63,040$$

and the total number of live births occurred during the same period in the country is

$$B_t = 1,11,652.$$

Hence, the general fertility rate of the country during the period is

$$GFR = \frac{111652}{1163040} \times 1000 = 96.00$$

The age specific fertility rate of the country during the same period is calculated for each reproductive age group and the rates are presented in the following table:

Age Group (in years)	*ASFR*
15 – 20	$\frac{20468}{216410} \times 1000 = 94.58$
20 – 25	$\frac{26983}{213610} \times 1000 = 126.32$
25 – 30	$\frac{22522}{202930} \times 1000 = 110.98$
30 – 35	$\frac{17083}{193300} \times 1000 = 88.38$
35 – 40	$\frac{13456}{173920} \times 1000 = 77.37$
40 – 45	$\frac{11140}{162870} \times 1000 = 68.40$

It can be observed, with respect to *ASFR*, that the women population of the country falling in the age group of 20-25 years have given relatively more live births. The women at the age of 40-45 years have reproduced less number of live births.

8.1.6 Measurement of Population Growth

Every human population, normally, may have a change in its size over a period of time. Each change in the population size may be an increase or decrease in magnitude. Sometimes, the population size may remain without any change. Such a population is known as stable population. The tendency to increase in the population size may be called as **population growth**. Every government requires information about the rate of growth of its population.

If many new born babies in a population are female, the corresponding population size may increase. If gender of the infant deaths is female, change in the population size may be downward. Hence, fertility and mortality rates, individually, do not provide knowledge on population growth.

Quantitative ideas about the growth of a population can be obtained from several measurements. Among them,

(i) Crude Rate of Natural Increase, and

(ii) Pearl’s Vital Index

may be considered as the basic indicators of population growth. These two measures can be calculated using the following formulae

$$\textbf{\textit{Crude Rate of Natural Increase}} = CBR - CDR$$

$$Pearl's\ Vital\ Index = \frac{CBR}{CDR} \times 100$$

Positive values of Crude Rate of Natural Increase indicate the net increase in the population. Similarly, negative values of Crude Rate of Natural Increase indicate the net decrease in the population.

If Pearl's Vital Index is greater than 100, then it can be regarded as the population is growing. On the other hand, if this index is less than 100, it can be regarded as the population is not growing. The above formula shows that the Vital Index can also provide knowledge on birth-death ratio of the population.

These two measures are simple and easy to calculate. They indicate whether the number of births exceeds the number of deaths. However, these two measures suffer from the limitations of *CBR* and *CDR*. They cannot be used for comparing two different populations. Also, information regarding whether the population has a tendency to increase or decrease cannot be obtained from these two measures.

8.2 OFFICIAL STATISTICS

Official statistics are the statistical information published by government agencies or other public bodies, which are collected and compiled on various aspects for administrative purposes. Official Statistics are collected in a systematic manner through a well-established Statistical System. These information include quantitative and qualitative information on all major areas of citizens' lives, such as economic and social development, living conditions, health, education and environment. Official statistics should be objective and easily accessible, produced on a continual basis so that measurement of change is possible. The following may be considered as the main functions of the Statistical System:

(i) collection, validation, compilation of data

(ii) publication/dissemination of the statistical information

(iii) maintenance of statistical standards such as definitions, classification, statistical methodology, comparability *etc.*

(iv) coordination of statistical activities

(v) training statistical personnel

(vi) independence and integrity of its functioning

(vii) international coordination

8.2.1 Early History of Statistical System in India

Statistical data collection and compilation began in India during 321-298 *BC* and are documented in Kautilya's *Arthasastra*. Later, during the Moghul's period, the details of Official Statistics can be found in *Ain-i-Akbari* written by Abul Fazal in Emperor Akbar's rule during 1590 *AD*. It contains Official Statistics of various characteristics including land classification, crop yields, measurement system, revenue *etc.*

In British India, a statistical survey was conducted, in the year 1807 by Dr.Francis Buchanan, Governor-in-Council of *East India Company*. Information were collected regarding topographical account of each district, conditions of the inhabitants, their religion and customs, details of fisheries, mines and forests, farm sizes, vegetables grown, commerce, list of useful plants and seeds. An **Official Statistical System** was established in India by *Col*. Sykes during 1847 with a Department of Statistics in *India House*. The **first Census Report of India** was published in 1848. The second Census Report was published in 1881 and since then Census was conducted every 10 years.

8.2.2 Post-Independence Indian Official Statistical System

After Independence, the need for a statistical system for monitoring socio-economic development of the country was felt by the Government of India. In 1949, Shri.P.C.Mahalanobis was appointed by the Government of India as the Honorary Statistical Advisor to the government. In the same year, he established the Central Statistical Unit. This Unit was renamed, in 1951, as Central Statistical Organization, which coordinated various statistical activities in the country. It also defined and maintained statistical standards in the country.

During the same period, National Income Committee was established in 1949 to estimate the National Income of the country. The Committee recommended the use of sampling methods for collecting information in order to fill the large gaps in the statistical information required for estimation of the National Income. Sample surveys were conducted at national level for this purpose. The first round of National Sample Survey was conducted in October 1950. Later, a separate organization under the government set-up for conducting sample surveys was formed in the name of National Sample Survey Organization.

The Central Statistical Organization and National Sample Survey Organization are now called respectively as Central Statistics Office (CSO) and National Sample Survey Office (NSSO).

The Ministry of Statistics and Programme Implementation (MoS&PI), a ministry with independent charge in the Government of India, was formed on October 15, 1999 with two wings, *viz*., National Statistical Office (NSO) and Programme Implementation.

The Government of India set up a Commission in the year 2000 under the headship of Shri.C.Rangarajan to address the growing statistical needs of the country. Based on the recommendations of the Commission, a permanent and statutory apex body, called **National Statistical Commission** (NSC), was set up in NSO on July 12, 2006. The NSC was formed to evolve policies, priorities and to maintain quality standards in statistical matters.

The NSC is constituted with an eminent statistician or a social-scientist as its Chairperson and four members - one each from the areas of Economic Statistics; Social and Environmental Statistics; Census Operations, Surveys and Statistical Information System; and National Accounts. The Chief Statistician of India is the Secretary of the Commission and the Chief Executive Officer of NITI Aayog of Planning Commission of India is an *ex-officio* member of NSC. The Chief Statistician of India is the Head of National Statistical Office and Secretary of the MoS&PI. NSC, in addition to the above responsibilities, also performs the functions of Governing Council of the NSSO since August 30, 2006.

Presently, the main sections of NSO are NSC, CSO, NSSO and a Computer Center.

8.2.2.1 Central Statistics Office

The **Central Statistics Office** is responsible for coordination of statistical activities in the country, and evolving and maintaining statistical standards. CSO is headed by a Director General, who is assisted by five Additional Director Generals. The CSO has five main divisions. The divisions and their responsibilities are presented below:

(i) National Accounts Division

This division is responsible for

- preparation of national accounts including Gross Domestic Product
- preparation of quarterly estimates of Gross Domestic Product
- estimation of Capital Stock and Consumption of fixed capital
- estimation of State-wise Gross Value Added and Gross Fixed Capital Formation
- preparation of Input-Output Transaction Tables, and
- preparation of comparable estimates of State Domestic Product.

(ii) Social Statistics Division

This division is responsible for

- statistical monitoring of the Millennium Development goals
- preparation and maintaining environmental economic accounting
- grant-in-aid for research, workshop/seminars/conferences in Official/Applied Statistics
- national/international awards for statisticians
- preparation of National Data Bank on socio-religious categories
- basic statistics for Local Level Development Pilot scheme
- conduct of time-use surveys and release of regular and *ad hoc* publications.

National Statistics Day and World Statistics Day

The Government of India declared 29th June, the birthday of Prof. P.C. Mahalanobis, as the National Statistics Day to honour his contribution to the establishment of Official Statistical System in India. As a part of the celebration of this day, essay writing competitions are conducted nationwide among postgraduate students of Statistics. The winners are honoured with awards during the celebration at New Delhi. The first National Statistics Day was celebrated on June 29, 2007.

United Nations Statistical Commission declared October 20 as the World Statistics Day. It is celebrated every five years in almost all the countries focussing on specific theme. The first World Statistics Day was celebrated all over the world on October 20, 2010. The second World Statistics Day was celebrated on October 20, 2015 on the theme *Better Data, Better Lives* to emphasize the important role of high quality official statistical information in decision-making.

All the government departments, educational departments and others dealing with Statistics take part in celebration of National Statistics Day as well as World Statistics Day.

The year 2013 was observed as the International Year of Statistics.

The State and Central Governments recruit professionals, who are trained in applications of statistical methods, for appointment as Statistical Investigators, Assistant Directors. Most of the statisticians in CSO, NSSO and in ministries now qualify Indian Statistical Service (ISS) examination conducted by Union Public Service Commission. They are trained by the National Statistical Systems Training Academy.

The first Economic Census was conducted in 1977.

The Second Five Year Plan of India followed the model developed by Prof. P.C. Mahalanobis, which focused on public sector development and rapid industrialization. The Government of India honoured him with one of the highest civilian awards *Padma Vibhushan*. The Government of India released a stamp on June 29, 1993 in commemoration of his 100th birthday. Recently, the Government of India released a commemorative coin on June 29, 2018 during the celebration of his 125*th* birthday.

(iii) Economic Statistics Division

This division is responsible for

- conducting Economic Census and Annual Survey of Industries
- compiling All India Index of Industrial Production
- collecting and compiling Energy Statistics and Infrastructure Statistics
- developing classifications like, National Industrial Classification and National Product Classification.

(iv) Training Division

This division is responsible for

- training manpower in theoretical and applied statistics to deal with the challenges of data collection, compilation, analysis and dissemination of information for policy making, planning, monitoring and evaluation
- looking after the National Statistical Systems Training Academy, which is a premier institute for developing human resource to deal with Official Statistics in India as well as at international level.

(v) Coordination and Publications Division

This division is responsible for

- coordinating the works related to statistical matters within CSO and the Ministries of Central Government and State/UT Governments
- organizing Conferences of Central and State Statistical Organizations
- celebration of National Statistics Day every year

- preparation of Results Framework Document, Citizens' Charter, Annual Action Plan and Outcome Budget of the MoS&PI
- implementation of Capacity Development Scheme and Support for Statistical Strengthening with an aim of improving the Capacity and Infrastructure of the State Statistical System for collection, compilation and dissemination of reliable Official Statistics for policy making
- coordinating implementation of recommendations of NSC
- administrative works related to Indian Statistical Institute.

8.2.2.2 National Sample Survey Office

National Sample Survey Office (NSSO), headed by a Director General, is responsible for conduct of national level large scale sample surveys in diverse fields. Primarily, data are collected through nation-wide household surveys on various socio-economic subjects. Besides these surveys, NSSO collects data on rural and urban prices and plays a significant role in the improvement of crop statistics through monitoring the area enumeration and crop estimation surveys of the State agencies. It also maintains a sampling frame of urban area units for conducting sample surveys in urban areas.

NSSO has four divisions. The divisions and their responsibilities are listed below:

(i) Survey Design and Research Division

This division, located at Kolkata, is responsible for

- technical planning of surveys
- formulation of concepts and definitions
- preparation of sampling design
- designing of inquiry schedules
- drawing up of tabulation plan
- analysing and presenting survey results.

(ii) Field Operations Division

The headquarters of this division is at Delhi. This division has a network of 6 Zonal Offices, 49 Regional Offices and 118 Sub-Regional Offices spread throughout the country. This division is responsible for collection of primary data for the surveys undertaken by NSSO.

(iii) Data Processing Division

The Division, with its headquarters at Kolkata and 6 Data Processing Centers at various places, is responsible for

- selection of sample subjects
- developing relevant software
- processing, validation and tabulation of the data collected through surveys.

(iv) Coordination and Publications Division

This Division, located at New Delhi, is responsible for

- coordinating all the activities of different Divisions of NSSO
- publishing the bi-annual journal of NSSO, titled "Sarvekshana"
- organizing National Seminars on the results of various socio-economic surveys undertaken by NSSO.

8.2.3 Present Statistical System in India

In addition to the role played by CSO and NSSO, most of the Central Ministries collect statistical information on the subjects related to the respective ministries. The statistical information are collected as by-products of administration of the ministries or for monitoring the progress of specific programmes implemented by the respective ministries. Some Ministries in Government of India, like Agriculture, Water Resources, Health, Finance, Commerce, Labour, and Industrial Development have separate statistical divisions, while most others have nucleus cells.

The Statistical System in the States is similar to the system at the Central Government. A Directorate of Economics and Statistics, functioning in each State under a decentralized system, is a nodal agency, which is responsible for the coordination of statistical activities in the State. The Directorates have statistical offices at the headquarters in each district. The district level offices collect statistical information related to all sections of economy of the respective district. The Directorates compile and publish such information as Statistical Hand Books every year. The Hand Books contain several information including estimates of area, production and yield of principal crops. In Tamil Nadu, the Directorate is functioning with the nomenclature "**Department of Economics and Statistics**". This department, with headquarters at Chennai, is headed by a Commissioner, who is assisted by a Director, 3 Additional Directors and 2 Joint Directors, in addition to Assistant Directors and supportive officials.

Generally, flow of statistical information in Indian Statistical System is upwards from village → block → district → State Government Departments → corresponding Ministries at the Centre.

In addition to CSO, NSSO and the Ministries, there are other public and private organizations in India, which also deal with collection of Official Statistics on various characteristics. Reserve Bank of India is one such organization, which collects, compiles and publishes, every year, the statistical information related to economy of the country as the "**Hand Book of Indian Economy**".

POINTS TO REMEMBER

- Vital Statistics are quantitative measurements on live births, deaths, foetal deaths, infant deaths, fertility and so on.
- Data on vital events are collected adopting the five methods - Civil Registration System, Census or Complete Enumeration method, Survey method, Sample Registration System and Analytical method.
- Census method normally covers data regarding age, sex, marital status, educational level, occupation, religion and other factors needed for computing vital statistics. Census is conducted in most countries at intervals of ten years.
- Rates of vital events are usually expressed '*per thousand*'.
- Crude Death Rate =

$$\frac{\text{No. of deaths in the population or community during the given period}}{\text{Total number of persons in the population or community during the given period}}\times 1000$$

- Specific Death Rate =

$$\frac{\text{No. of deaths in a specific section of the population during the given period}}{\text{Total number of persons in the specific section of the population during the given period}}\times 1000$$

- Infant Mortality Rate =

$$\frac{\text{No. of infant deaths in a population during the given period}}{\text{Number of live births in the population during the given period}}\times 1000$$

- ***Cohort*** is a group of individuals who born at the same time and experienced the same mortality conditions.
- A Life Table exhibits the number of persons living and dying at each age, on the basis of the experience of a ***cohort***, which also gives the life expectancy of the population.
- Radix of a Life Table refers to the number of survivors at the beginning of the table.
- Crude Birth Rate =

$$\frac{\text{No. of live births in the population during the given period}}{\text{Total number of persons in the population during the given period}}\times 1000$$

- General Fertility Rate =

$$\frac{\text{No. of live births occurred in the population during the given period}}{\text{Total number of women at the reproductive age in the population during the given period}}\times 1000$$

- Specific Fertility Rate = $\dfrac{\text{Number of live births to the women population in the reproductive age groups of specific section in a given period}}{\text{Total number of women in the reproductive age groups of the specific section in the given period}} \times 1000$
- Crude Rate of Natural Increase = Crude Birth Rate – Crude Death Rate
- Pearl's Vital Index= $\dfrac{\text{Crude Birth Rate}}{\text{Crude Death Rate}} \times 100$
- Official statistics are the statistical information collected and compiled on various aspects including all major areas of citizens' lives, such as economic and social development, living conditions, health, education and environment.
- An Official Statistical System was established in India by *Col*. Sykes during 1847 with a Department of Statistics in *India House*. The first Census Report of India was published in 1848.
- The Central Statistics Office is responsible for coordination of statistical activities in the country, and evolving and maintaining statistical standards, which has five main divisions.
- National Sample Survey Office (NSSO), headed by a Director General, is responsible for conduct of national level large scale sample surveys in diverse fields, which has four divisions.

EXERCISE 8

3ZYSH

I. Choose the best answer.

1. One of the branches of Demography is

(a) Economic Statistics (b) Vital Statistics
(c) Official Statistics (d) Agricultural Statistics

2. When there is no proper system of recording births and deaths, Vital Statistics are collected through

(a) Registration Method (b) Census Method
(c) Survey Method (d) Analytical Method

3. Compulsory registration of births and deaths was implemented in India, during the year

(a) 1947 (b) 1951 (c) 1969 (d) 1970

4. Rates of vital events are usually measured as

(a) *per* ten lakh (b) *per* ten thousand
(c) *per* thousand (d) *per* hundred

5. Death of a child before the child's fifth birth day is measured by

(a) crude death rate (b) specific death rate
(c) infant mortality rate (d) child mortality rate

6. Death rates due to different kinds of diseases can be calculated using

   (a) crude death rate (b) infant mortality rate

   (c) specific death rate (d) vital index

7. History of a *cohort* can be understood from

   (a) mortality rates (b) life table

   (c) fertility rates (d) population growth

8. Total number of women at child bearing age group is used to calculate

   (a) crude birth rate (b) general fertility rate

   (c) age specific fertility rate (d) population growth

9. Population growth can be measured using

   (a) crude birth rate, specific death rate

   (b) general fertility rate, infant mortality rate

   (c) specific fertility rate, specific death rate

   (d) crude birth rate, crude death rate

10. Vital Index measures

    (a) birth-death ratio (b) rate of vital event

    (c) infant mortality rate (d) general fertility rate

11. The first census Report of India was published in

    (a) 1858 (b) 1848 (c) 1948 (d) 1958

12. Department of Statistics in India House was established in India by Col. Sykes during

    (a) 1847 (b) 1947 (c) 1857 (d) 1887

13. The first Honorary Statistical Advisor to Government of India is

    (a) P.C. Mahalanobias (b) *Col.* Sykes

    (c) C. Rangarajan (d) Dr. Francis Buchanan

14. Ministry of Statistics and Programme Implementation was formed on

    (a) October 2, 1959 (b) October 15, 1999

    (c) November 13, 1969 (d) November 15, 1999

15. Secretary to MoS&PI and Head of NSO is

    (a) Chairman, Planning Commission (b) Director, NSSTA

    (c) Chief Statistician of India (d) CEO, NITI Aayog

16. National Statistical Commission was formed on

    (a) January 12, 2006 (b) April 12, 2006

    (c) June 12, 2006 (d) July 12, 2006

17. Celebration of National Statistics Day is one of the responsibilities of ______ Division of CSO.

    (a) Coordination and Publications (b) Training

    (c) Social Statistics (d) Economic Statistics

18. Collection of primary data for the surveys undertaken by NSSO is one of the responsibilities of ________ Division.
    (a) Survey Design and Research (b) Field Operations
    (c) Data Processing (d) Coordination and Publications

19. NSSO publishes the bi-annual journal
    (a) Sarvekshana (b) Sankhya (c) Pramana (d) Yojana

20. "Hand Book of Indian Economy" is published periodically by
    (a) State Bank of India (b) Reserve Bank of India
    (c) NSO (d) Ministry of Economics

II. Give very short answer to the following questions.

21. What is Registration method?
22. Define rate of vital event.
23. Mention the purpose of Analytical method in collecting Vital Statistics.
24. What is meant by mortality?
25. What is known as fertility?
26. What is *cohort*?
27. What is called radix of Life Table?
28. How will you calculate expectation of life?
29. Write down the formula to compute crude death rate.
30. What is the formula used to compute infant mortality rate?
31. Define crude birth rate.
32. What is specific fertility rate?
33. Write down the formula to compute general fertility rate.
34. Define vital index.
35. Specify the difference between crude birth rate and general fertility rate.
36. What are known as Official Statistics?
37. What were the Official Statistics collected by East India Company?
38. What were the earlier attempts made in India before British rule for collection of Official Statistics?
39. What are the wings of MoS & PI?
40. What are the divisions of CSO?
41. List the divisions of NSSO:
42. Which Ministries of Government of India have separate statistical divisions?

III. Give short answer to the following questions.

43. Write down the definitions of Vital Statistics.
44. Write a brief note on Census Method of collecting Vital Statistics.

45. What are various information collected under Sample Registration System?
46. If the number of deaths occurred is 980 in a town consisting of 1,50,000 persons during a period, quantify the death rate of the town using suitable formula.
47. Population size of a Hamlet in a hill station during a calendar year was 55,000 and the number of deaths recorded in the Hamlet during the same year was 185. What is the crude death rate of the village?
48. The number of women at reproductive age in a district is 70,000. Also, the number of live births and infant deaths registered in the district are respectively 10,000 and 70. Which mortality rate you will calculate? What is its value?
49. A Primary Health Centre located in a village has a record of 135 live births during a year. The number of deaths recorded in the village was 300. Among them, 5 are new born babies of age less than a year. Measure the infant mortality rate of the village.
50. What is specific death rate? What are its uses?
51. List various uses of Life Table.
52. What are the assumptions made in the construction of Life Table?
53. Population size of a District during a calendar year was 1,25,526. Also, the number of live births registered in the district during the same period was 987. Compute appropriate vital rate using this information.
54. The number of women at child bearing age in a tribal village during a year was 2275 and the number of new born babies of age less than a year in the same village was 23. How will you quantify the birth rate of the village? Find its value.
55. Write a brief note on specific fertility rate.

IV. Give detailed answer to the following questions.

56. What are the importance of Vital Statistics?
57. Calculate crude death rate of a population living in a town from the following data:

Age Group (in years)	0-10	10-20	20-40	40-60	60 and above
No. of Persons	6500	12,000	24,000	20,000	8,000
No. of Deaths	25	37	30	90	100

58. The number of deaths registered in a district with respect to age during a year and the population size in each age group are given below. Calculate the crude death rate of the district for the period.

Age Group (in years)	Below 15	15-25	25-40	40-65	65 and above
No. of Persons	40,000	88,000	90,000	60,800	23,000
No. of Deaths	40	62	100	78	20

59. The population of a District during a Census year was grouped into 8 age groups. The data regarding the number of deaths and the population size for each age group are given hereunder.

Age Group (in years)	0-5	5-15	15-25	25-40	40-50	50-60	60-65	65 and above
No. of Persons (in '000)	42,345	19046	93,578	30,724	28,874	62,087	28,473	37,693
No. of Deaths	900	798	512	186	174	213	475	883

Calculate the age specific death rates of the District.

60. Calculate the specific death rates for each age group of a population, whose size and the number of deaths are given in the following table:

Age Group (in years)	Below 10	10-20	20-40	40-60	60 and above
No. of Persons	36,000	28,000	62,000	52,000	18,000
No. of Deaths	682	204	576	878	725

61. What are the different components and their formulae of Life Table?

62. Find the missing entries in the following Life Table.

Age (in years)	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
25	75818						
26	75445					2722331	
27	75039			0.009			

63. There are missing entries in some of the columns in the following Life Table. Find the values of the missing entries.

Age (in years)	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
42	64711					1513333	
43	63787						
44	62821				62310		

64. The following is a section of a Life Table with some missing entries. Complete the Life Table.

Age (in years)	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
36	69818						
37	69032						
38	68212	850				1779254	

65. Calculate crude birth rate from the following data:

Age Group (in years)	15-20	20-25	25-30	30-40	40-49
No. of Perosns	16,000	18,000	14,000	15,000	28,000
No. of Live Births	25	30	38	28	14

66. The women population at reproductive age in a State are grouped and the population size in each group are given hereunder. The number of live births given by the women in each group are also presented. Find the general fertility rate of the State. Also, calculate the age specific fertility rate for each group.

Age Group (in years)	15-20	20-25	25-30	30-35	35-40	40-45	45-49
No. of Women	2,12,724	1,89,237	2,45,367	1,32,109	1,29,645	90,708	34,975
No. of Live Births	20,209	23,655	37,787	12,815	9,723	4,898	874

67. The number of live births occurred in a District during a calendar year are classified according to the age of mother. The female population size at child bearing age are also given.

Age Group (in years)	15-20	20-25	25-30	30-35	35-40	40-45	45-49
No. of Women	4,729	6,236	8,034	9,408	5,907	4,657	2,975
No. of Live Births	356	845	970	1,878	856	608	452

Calculate the general fertility rate of the District. Also, calculate the specific fertility rates for each of the reproductive age group.

68. The following are the information registered about the number of live births and the female population size in a town during a calendar year.

Age Group (in years)	15-20	20-25	25-30	30-35	35-40	40-45	45-49
No. of Women	1,276	3,253	5,628	7,345	6,901	4,253	3,957
No. of Live Births	218	361	693	1,305	1,031	634	390

Calculate from these information the general fertility rate of the town and the age specific fertility rates for the year.

ANSWERS

I. 1. (b) 2. (c) 3. (d) 4. (c) 5. (d)

6. (c) 7. (b) 8. (b) 9. (d) 10. (a)

11. (b) 12. (a) 13. (a) 14. (b) 15. (c)

16. (d) 17. (a) 18. (b) 19. (a) 20. (b)

III. 46. $CDR = 6.53$ 47. $CDR = 3.36$ 48. $IMR = 7.00$ 49. $IMR = 37.04$

53. $CBR = 7.86$ 54. $GFR = 10.11$

IV. 57. $CDR = 4.00$ 58. $CDR = 0.99$

59.

Age Group (in years)	0-5	5-15	15-25	25-40	40-50	50-60	60-65	65 and above
ASDR	21.25	41.90	5.47	6.05	6.03	3.43	16.68	23.43

60.

Age Group (in years)	Below 10	10-20	20-40	40-60	60 and above
ASDR	18.94	7.29	9.29	16.88	40.28

62.

x	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
25	75818	373	0.9951	0.0049	75632	27,97,968	36.90
26	75445	406	0.9946	0.0054	75242	27,22,331	36.08
27	75039	675	0.9910	0.0090	74702	26,47,629	35.28

63.

x	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
42	64711	924	0.9857	0.0143	64249	1513333	23.39
43	63787	966	0.9849	0.0151	63304	1449084	22.72
44	62821	1022	0.9837	0.0163	62310	1385780	22.06

64.

x	$l(x)$	$d(x)$	$p(x)$	$q(x)$	$L(x)$	$T(x)$	$e^0(x)$
36	69818	786	0.9887	0.0113	69425	1917301	27.46
37	69032	820	0.9881	0.0119	68622	1847876	26.77
38	68212	850	0.9875	0.0125	67787	1779254	26.08

65. *CBR* = 1.48

66. *GFR* = 106.27

Age Group (in years)	15-20	20-25	25-30	30-35	35-40	40-45	45-49
ASFR	95.00	125.00	154.00	97.00	75.00	54.00	24.99

67. *GFR* = 142.21

Age Group (in years)	15-20	20-25	25-30	30-35	35-40	40-45	45-49
ASFR	75.28	135.50	120.74	199.62	145.06	130.56	151.93

68. *GFR* = 142.03

Age Group (in years)	15-20	20-25	25-30	30-35	35-40	40-45	45-49
ASFR	170.85	110.97	123.13	177.67	149.40	149.07	98.56

References

Mukhopadhyay, P.(2016). *Applied Statistics*. Books and Allied (P) Ltd., Kolkata.

Goon, A.M., Gupta M. K. and Das Gupta B. (2016). Fundamentals of Statistics, Vol. 2. The World Press pvt Ltd, Kolkatta.

Rao, T.J. (2010). Official Statistics in India: The Past and the Present. *Journal of Official Statistics*, Vo. 26(2), 215-231.

Rao, T.J. (2013). National Statistical Commission and Indian Official Statistics. *Resonance*, December, 1062-1072.

CHAPTER 9

PROJECT WORK

Project Work is a student-centered pedagogy in which students learn about a subject through the experience of solving an open-ended problem. This necessitates to synthesize knowledge from various areas of learning, and critically and creatively apply it to real life situations.

LEARNING OBJECTIVES

The student will be able to

- ❖ Design a questionnaire for a given project
- ❖ Collect data using questionnaire.
- ❖ Compile and tabulate the collected data.
- ❖ Statistically analyze the data
- ❖ Write a brief report

SFI2T

Introduction

A mechanical engineer has to spend time in an Industry as an apprentice, a medical doctor with a hospital as a house surgeon, an auditor with an accountant, a budding lawyer with an established senior as a junior to get to know intricacies of the profession in his day to day affairs. *Where is the counterpart of this for a student studying Statistics?* It is in this background, the inclusion of **project work** in the curriculum gains importance. This would help to some extent to bridge the gap between the theory learned in the class room and application.

Characteristics

The Project work is an assignment to be carried out by students during the course, either individually or in a group under supervision of the teacher. A brief written report of this work is an integral part of the assignment. Thus the project work is a *complete assignment*. . It contains the levels of planning, execution, analysis and reporting the work done.

Advantages

1. Formulating a real world problem with statistical perspective, student acquires application knowledge through the completion of the project.
2. Project works provides the knowledge about systematic collection of data, organization of ideas and an ability to analyze them with in a stipulated time. The following intangible benefits will be derived on completion of the project work.
   - ❖ Development of the capacity to identify and correctly specify statistical problem.
   - ❖ Being aware of the assumptions and the steps to validate them.
   - ❖ Ability to interpret results.
   - ❖ Development of report writing skill.
3. It pave the way for interaction with the respondents, the ability to fit into a team, provides co-operation and provides interaction among the students and between the teachers.
4. It gives them a sense of involvement and commitment.

9.1 DESIGNING A PROJECT

There are Six stages while doing a project work.

- ❖ Stage 1 - Identify the topic of the project work
- ❖ Stage 2 - Define Your Goals. Formulate Objectives / Hypothesis
- ❖ Stage 3 - Develop Your project work Plan.
- ❖ Stage 4 - Collection of Data and creating a data file.
- ❖ Stage 5 - Analyze Data
- ❖ Stage 6 - Report writing

Stage 1 . Fixing the project work topic

The first and foremost stages of any project work is identifying the topic of the project work. The sources for selecting the topic include (i) individual experiences, (2) personal conversation, (4) day-to-day practical experience, (5) Social problems (6) Politics like opinion poll.

The criteria for fixing the topic depends about data availability, time period to complete a research, availability of expertise.

Stage 2. Clearly state the objectives / Formulate Hypotheses

Objectives:

The aspects, the project worker want to probe in the project be spelt out in clear terms as objectives. Research should not proceed until objectives are clearly spelt out.

Hypothesis

Hypothesis is the perception of the researcher. It is stated as a testable proposition subject to empirical verification. Some times it is called as research hypothesis.

Stage 3: Project work planning

It involves selecting the most appropriate methods of (i) selecting respondents, (ii) method of collecting the data, (iii) selecting the data gathering instrument and test for its reliability and validity reliability and (iv) techniques to solve the problem under investigation. Designing a research project can be compared to that of an architect designing a building. It is a plan or blue print for collection, analysis and interpretation of data.

Stage 4: Data collection and data file preparation

Once the project plan is ready, the student will know from whom/ where to collect the data and the data gathering instrument. Now he moves to collect the required data.

Data file preparation involves entering in a spread sheet and checking for errors if any.

Stage 5: Statistical Data Analysis

This activity is most important. Here, we select the appropriate statistical tool for data analysis. This necessitates a good conceptual clarity and application understanding of the subject. T*he selection of statistical tool primarily depends on* Research Objective. And subsequently on the variables involved, its measurement scale (nominal, ordinal and scale) and number of variables considered at a time.

Stage 6: Report writing

The important step in any project study is that of preparing the project report. The report records the purpose, the importance, the procedure, the findings, the limitations and the conclusion of the project study. This should be prepared in such a way that it is easily understood and is helpful to other research or project workers in a similar field.

9.2 PROJECT WORK PLAN

A project work plan find the answers to the following host of questions:

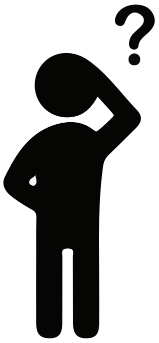
WHERE
WHY
WHEN
?
WHO
HOW
WHAT

1. What is study about?
2. Why is the study being made? (***Motivation of the study)***
3. Where will the study be carried out? (***Scope of coverage)***
4. **What type of data is required**? *(Survey data / Experimental data)*
5. What are instruments needed? (Questionnaire / Instruments)
6. From whom can be required data be found? (***Where to collect?*** ***Target group?)***
7. What periods of time will study include? (***How many times to collect***?)
8. What techniques of data collection will be used? (***Which method to follow?)***
9. How the data will be analyzed?

9.3 QUESTIONNAIRE DEVELOPMENT PROCESS

Questionnaire Format

It consists of two parts namely (1) Questions and (2) Responses.

Questions	**RESPONSES**
	
1. Name 2. Age (in completed years) 3. Sex	-- ________ Years 1. Male 2. Female

Once the researcher has decided on the specific type of questions and the response formats, the next task is the actual writing of the questions. The wording in specific questions always poses significant time investment for the researcher. The general guidelines are useful to bear in mind during the wording and sequencing of each question.

Characteristics of a questionnaire

- ❖ The wording must be clear
- ❖ Select words so as to avoid biasing the respondent
- ❖ Consider the ability of the respondent to answer the question
- ❖ Create the willingness of the respondent to answer the questions.

Evaluate the Questionnaire

Once a rough draft of the Questionnaire has been designed the researcher is obligated to take a step back and critically evaluate it. This phase may seem redundant, given all the careful thoughts that went into each question. But recall the crucial role played by the questionnaire. At this point in the questionnaire development of the following item should be considered.

- Is the question necessary?
- Is the survey too long?
- Will the question provide the answers to the survey objectives?

Pre-Test

After the questionnaire is prepared, pre-test is to be done. The process collecting information from the related respondent in small number with the framed questionnaire is called *Pre-Test*. Sometimes, the questionnaire is circulated among some competent investigators and they are asked to make suggestions for its improvement. Once this has been done and suggestion incorporated in the final form of the questionnaire is ready for the collection of data.

9.4 FEATURES OF A PROJECT REPORT

Features of a project report.

The general structure of a project report consists of three main divisions.

A. The Preliminary Section:

This includes the title page, the preface, the acknowledgement, the table of contents and the list of tables and figures.

B. The Main Body Or The Text Of The Project :

This contains the introduction to the problem, the review of previous research in a similar field, the details of the procedure, the findings, the analysis of that, and conclusions.

C. The Reference section :

This includes Foot Notes, Bibliography Appendix , Index, etc.

LANGUAGE OF THE REPORT:

Report should be written in a simple language. It should be clear, precise and simple in style, and brief. It should be written in third person or passive voice. Spelling mistakes, colloquial form of presentation should be avoided. Spelling of non-English words, if used, should be kept uniform throughout.

EXERCISE 9

I. Choose the best answer.

1. Which one is the correct sequence of activities in project work?
   a) formulating objectives, report writing, data analysis, project work plan
   b) report writing, data analysis, project work plan , formulating objectives
   c) formulating objectives, data analysis report writing, , project work plan
   d) formulating objectives, , project work plan, data analysis, report writing.

2. Qualitative data implies:
   a) Measurable b) Non measurable c) Partly measurable d) All the above

3. The measurement scale of 'taste of a coffee' is :
   a) Nominal b) Ordinal c) Interval d) Ratio

4. When the researcher uses the data of an agency, then the data is called:
   a) Quantitative data b) Qualitative data
   c) Secondary data d) Primary data

5. Opinion poll in a study is conducted:
   a) Before the process start b) After the process start
   c) Middle of the process d) At any point of time of the process

6. A study is conducted on impact of stress on blood pressure. In this study blood pressure is a
   a) Independent variable b) Dependent variable
   c) Intervening variable d) Extraneous variable

7. Which one is false in the questionnaire method?
   a) Vast coverage in less time b) This method can be adopted to any respondent
   c) Response rate may be low d) It offers greater anonymity.

8. Null hypothesis in a research is:
   a) A positive directional hypothesis b) A negative directional hypothesis
   c) Hypothesis of no difference d) None of the above.

9. When we study the effect of any new intervention is,
   a) F test b) ANOVA c) Chi Square test d) Paired *t* test

10. Thanking the people who helped as to complete the project work will come under:
    a) Reference b) Conclusion
    c) Acknowledgement d) Need not appear in the report.

II. Give very short answer to the following questions.

11. Why is a project work needed in the curriculum?
12. Define population or target group under study.

13. List the points to be noted before fixing the project topic.

14. State the situations where hypothesis testing is inevitable in a project work

15. List the components that decide the tool selection.

III. Give short answer to the following questions.

16. State the characteristics of a project

17. List the intangible benefits derived by doing a project work

18. State the stages involved in doing a project work

19. What are things to be included in the primary section of a project report?

20. What is a pretest in the questionnaire method?

IV. Give detailed answer to the following questions.

21. State the advantages of doing a project work?

22. Briefly explain the characteristics of various stages in a project work.

23. State the points kept in mind while writing the questionnaire

24. In project work planning, state the aspects to be focused?

25. Discuss the features of a project report.

Answers: *Since the answers are part and parcel of the course and none of the questions have any problem, writing answers does not arise.*

ANNEXURES

(These Annextures use for understanding purpose only and questions should not be asked in this portion)

SAMPLE PROJECT - TEMPLATE

Stage 1. Fixing the project work topic

A study on the prevalence of obesity among the students

Stage 2. Clearly state the objectives / Formulate Hypotheses

1. How the respondents are distributed in sex and community wise.

2. How obesity is prevalent among the college students?

3. Whether sex has any influence on obesity?

4. Whether diet intake and obesity are related?

5. Establish a simple linear equation (linear model) for prediction purpose between height and weight?

6. Do the male and female differs with respect to average height?

7. Whether the average weights of different diet takers can be taken as equal?

Objectives:

1. *To describe the demographic features of the respondents*
2. *To find the prevalence of obesity among the population.*
3. *To know whether sex and diet in take is associated with obesity.*
4. *To test whether the suggested linear model between height and weight is good fit to the given data.*
5. *To test whether the sex and diet intake has influence on average weight*

Hypothess:

1. *Obesity doesn't depend on sex*
2. *Diet intake is not associated with obesity.*
3. *The linear model is good fit for the data between height and weight*
4. *The average weight of male and the average height female do not differ due to sex.*
5. *Different diet intake do not affect the average height.*

Stage 3: Project work planning

Planning for Data Collection

(i). Source : a). Survey ☑ *b). Experiment*□ *c). Observational study*□

(ii) Type of data: a). Primary data ☑ *b). Secondary data* □

(iii) Population or target group: ____________

(iv) Coverage : ____________
a). Census survey □ *b). Sample survey* ☑

If you go for a sample survey, answer the following:

(v) Sampling frame : a). available ☑ *b). Not available* □

(vi) Sample size determination: _____________

(vii) Number of contacts: a). Only one time (Cross section study). ☑

b). Two times (before and after the intervention/ treatment) □

c). Several times (Longitudinal study) □

(viii) Sampling to be used? ____________________

(ix) Is it appropriate or whether will it give representative population ?

1. Yes ☑ *2. No* □

(x) The measuring instruments you intend to use? __________

(xi) In case of, questionnaire, Whether " Pre test " is conducted

1. Yes ☑ *2. No* □

(xii) Are they reliable and valid? 1. Yes □ *2. No* ☑

QUESTIONNAIRE

For this purpose the researcher has prepared a suitable questionnaire.

1. Name :

2. Sex 1.Male □ 2.Female □

3. Residence type: 1. Urban □ 2. Rural. □

4. Height in cms

5. Weight in kgs.

Stage 4: Data collection and data file preparation

Data Entry

S.No	Diet in take	Sex	Height	Weight
1	2	1	137.8	30
2	3	1	131.5	30.5
3	1	2	132.8	31.5
4	3	1	139.8	30.5
---	---	---	---	---
39	2	1	126	24
40	3	2	128.5	23.5

Stage 5. Statistical Data Analysis

The appropriate statistical tools are selected based on the four following aspects.

1. Purpose.
2. Variables involved.
3. Types of measurement scales.
4. Number of variables considered at a time.

Interpretation: While interpreting data, the researcher should give both findings/inference and conclusion. It should be remembered that finding is what you have obtained from the data and conclusion is with related to answering the research question.

Template for Statistical Tool Selection and Analysis for formula type

Research Objective:

Template for selecting suitable test for testing of hypothesis and Analysis

Identify the variables involved:

Type of measurement scales:

Number of Variables:

==

Selection of Appropriate Statistical Tool:

==

Statistical calculation:

Findings:

Conclusion:

Research Hypothesis:

1. Study design (Tick appropriate)

   One sample □ Two independent sample □

   More than two independent sample □ Related sample (before – after) □

   Repeated measure more tha 2 times (Longitudinal) □

2. Identify the appropriate parameter to be used:

   Mean □ Proportion □ Variance □ Median □

3. Number of samples considered simultaneously:

   One □ Two □ More than two □

4. Sample size: Greater than 30 (n ≥ 30) □ Less than 30 (n ≤ 29) □

5. Data set follows: Normal Distribution □ Non Normal □

==

Suitable statistical test: Parametric □ Non Parametric □

Name of the test: ______________________

TEST PROCEDURE

Test Procedure:

Null Hypothesis H_0 :

Alternative Hypothesis H_1:

Level of Significance α:

Test Statistic:

Calculation of calculated value using sample data:

Critical value from the table:

Inference:

Statistical Tool Selection based on purpose or aim

S.No.	Purpose / Aim	Stat. Tool
1	Describing or highlighting the variables	Descriptive Statistics
2	Relationship between **variables**. In terms of movements	Correlation Analysis Pearson's Correlation,Correlation with Spearman's Correlation
3	Relationship between **Attributes.**	Chi-square test for independence of attributes
4	Finding ***mathematical relationship*** for future prediction Or **impact of** independent variables on the dependent variable	**Regression Analysis** Simple Regression analysis
5	Hypothesis Testing 1. Assigning a value to the parameter 2. Comparing the efficiency of one group over other groups.	Large sample: Z tests ***Parametric Tests:*** Small Sample: , Single sample t test, Independent Sample t test, Paired Sample t test. ANOVA
6	To study the relative changes with respect to a base period in production, salary, prices etc	Index Numbers
7	In time series data to know the trend and to analyze the seasonal effect	Time series analysis Method of moving averages- method of least squares
8	To calculate the demographic details like birth rates, death rates, expected life time etc	Vital statistics -Mortality table

Report writing

RESEARCH REPORT FORMAT

Preliminaries,

Chapter I INTRODUCTION

Introduction about the research area : Broad Area

Problem Selection : Selection of the topic

Objectives of the study

Hypothesis of the study

Methodology of the study:

- ❖ Sample Design
- ❖ Source of Data - Instrument used for extracting information from the sample units
- ❖ Pre test- Reliability and validity of the data collection instrument and Pilot study
- ❖ Description of variables
- ❖ Frame work of analysis
- ❖ Significance of the study
- ❖ Period of study
- ❖ Scope of Study
- ❖ Limitations of the study
- ❖ Scheme or layout or organization of the report.

Chapter II METHODOLOGY / DATA COLLECTION

Chapter III ANALYSIS AND INTERPRETATION

Chapter IV SUMMARY AND CONCLUSION

Reference

12th standard – Statistics Practical

The syllabus for 12th standard practical are the following problems should be taken from the textbook examples or Exercises or relevant problems in real life situation. The question paper consists of two sections. Each section contains five questions. The students should answer four questions choosing two from each section.

Section A

1. Tests of Significance of a Proportion and Equality of Proportions based on *Z*-Statistic.
2. Tests of Significance of a Mean and Equality of Means based on *Z*-Statistic
3. Tests of Significance of a Mean based on *t*-Statistic.
4. Tests of Significance for equality of Means of two Independent Populations. (Independent samples '*t*' test)
5. Paired *t*-Test for dependent samples.
6. Test of Significance for Equality of Population Variances based on *F*-Statistic.
7. ANOVA for One Way Classification.
8. ANOVA for Two Way Classification.
9. Chi-square Test for Independence of attributes.
10. Chi-square Test for Goodness of fit.

Section B

1. Computation of Pearson's Correlation Coefficient.
2. Computation of Spearman's Rank Correlation.
3. Computation of Yule's coefficient of association.
4. Construction of Regression Equations.
5. Construction of Index Numbers.
6. Trend by the Method of "Moving Averages" of a Time series data.
7. Trend by the Method of "Least Squares" of a Time series data.
8. Seasonal Indices by the Method of "Simple Averages" of a Time series data.
9. Computation of CBR, ASBR, CDR,ASDR.
10. Construction of Life Table for Vital Statistics.

The Outline of the each of the problems is as follows.

1. Aim or purpose:
2. Selection of the suitable statistical tool
3. The following procedure is to be followed for the *SECTION A* and *SECTION B*.

SECTION B: Problem solving type

- Formula
- Substitution of data in the formula
- Calculation
- Result

SECTION A: Hypothesis testing

Test procedure

- Null Hypothesis H_0
- Alternative Hypothesis H_1
- Level of Significance α
- Test Statistic
- Calculation of test statistic value using sample data
- Critical value from the table
- Decision

Include graphs / diagrams wherever needed

GLOSSARY	
Acceptance region	H_0 ஐ ஏற்கும் பகுதி
Action-Reaction Theory	செயல்-எதிர்செயல் கோட்பாடு
Alternative Hypothesis	மாற்று கருதுகோள்
ANOVA	மாறுபாட்டு பகுப்பாய்வு
Association of attributes	பண்புகளின் இணை தொடர்பு
Assumption	அனுமானங்கள்
Attributes	பண்புகள்
Best fit straight line	சிறந்த பொருத்தமான நேர்கோடு
Blocks	தொகுதிகள்
Cell frequency	கட்ட நிகழ்வெண்
Central Statistics Office (CSO)	தேசிய புள்ளியியல் அலுவலகம்
Cohort	ஆய்வுக்கு எடுத்துக்கொள்ளும் பெருங்குழு
Contingency table	இணைப்புப் பட்டியல்
Correction factor	திருத்தக்காரணி
Correlation	ஒட்டுறவு
Critical region	தீர்மானிக்கும் பகுதி அல்லது H_0 ஐ மறுக்கும் பகுதி
Critical value	தீர்மானிக்கும் எல்லை மதிப்பு அல்லது தீர்மானிக்கும் மதிப்பு
Crude Death Rate (CDR)	செப்பனிடா இறப்பு விகிதம்
Cyclical Variation	சுழல் மாறுபாடுகள்
Decision Rules	முடிவெடுக்கும் விதிகள்
Decision	முடிவு
Degrees of freedom	கட்டின்மை கூறுகள்
Demography	மக்கள் தொகையியல்
Depression	வீழ்ச்சி
Descriptive Statistics	விளக்கப் புள்ளியியல்
Economic Rhythm theory	பொருளாதார ஏற்ற இறக்க கோட்பாடு
Erratic Fluctuation	ஒழுங்கற்ற வேறுபாடுகள்
Error	பிழை
Expected frequency	எதிர்பார்க்கப்படும் நிகழ்வெண்
Forecast	முன்கணிப்பு
General Fertility Rate (GFR)	பொது கருவுறுதல் விகிதம்
goodness of fit	செம்மை பொருத்தம்
grade	தரம்
Hypothesis Testing	கருதுகோள் சோதனை
Hypothesis	கருதுகோள்
Independent	சார்பற்ற
Infant Mortality Rate	குழவி இறப்பு விகிதம்
Inferential Statistics	அனுமானப் புள்ளியியல்

Irregular Variation	ஒழுங்கற்ற மாறுபாடுகள்
Large Sample	பெருங்கூறு ($n \geq 30$)
Least Squares	மீச்சிறு வர்க்கம்
Level of significance	மிகைகாண் நிலை அல்லது மிகைகாண் மட்டம்
Life Table	வாழ்நிலை அட்டவணை
Linear Regression	நேர்கோட்டு உடன்தொடர்பு
Mean sum of squares	வர்க்கங்களின் கூடுதல் சராசரி
Mean	சராசரி
Method of Least Square	மீச்சிறு வர்க்க முறை
Mortality	இறப்பு நிலை
Multiple Correlation	பல்சார் ஒட்டுறவு
Multiple Linear Regression	பல்சார் உடன்தொடர்பு
National Sample Survey Office (NSSO)	தேசிய மாதிரிக் கணக்கெடுப்பு அலுவலகம்
Negative Correlation (Inverse Correlation)	குறை ஒட்டுறவு (எதிர் ஒட்டுறவு)
Non-Linear Regression	வளைகோட்டு உடன்தொடர்பு
Normal Equations	இயல்நிலைச் சமன்பாடுகள்
Normal population	இயல் நிலை முழுமைத் தொகுதி
Null Hypothesis	இன்மை கருதுகோள்
Observed frequency	கண்டறிந்த நிகழ்வெண்
Official Statistics	நிர்வாகப் புள்ளியியல்
One tailed (left) test	ஒரு முனை (இடது) சோதனை
One tailed (right) test	ஒரு முனை (வலது) சோதனை
Opinion polling	கருத்துக் கணிப்பு
Paired t-test	இணை t-சோதனை
Parameter	தொகுதிப்பண்பளவை
Partial Correlation	பகுதி ஒட்டுறவு
Perfect Negative Correlation	முழுமையான எதிர் ஒட்டுறவு
Perfect Positive Correlation	முழுமையான நேர் ஒட்டுறவு
Pooled estimate	கூட்டப்பட்ட மதிப்பீடு
Population	முழுமைத்தொகுதி
Positive Correlation (Direct Correlation)	மிகை ஒட்டுறவு (நேர் ஒட்டுறவு)
Prediction	முன்னரே யூகித்தல்
Projection	விரிவாக்கம் செய்தல்
Proportion	விகிதசமம்
Prosperity	வளம், செழிப்பு
Radix	பெருங்குழுவில் உள்ளோர் எண்ணிக்கை
Rank Correlation	தர ஒட்டுறவு
Ratio	விகிதம்
Recession	பின்னடைவு

Recovery	மீட்சி
Regression Coefficient	உடன்தொடர்பு கெழு
Regression	உடன் தொடர்பு
Rejection Rule	H_0 ஐ மறுக்கும் விதி
Sample size	மாதிரி அளவு, கூறு அளவு
Sample Space	கூறுவெளி
Sample	மாதிரி அல்லது கூறு
Sampling distribution	மாதிரி பரவல், கூறுபரவல்
Sampling	கூறெடுத்தல்
Scatter Diagram	சிதறல் விளக்கப்படம்
Seasonal Variation	பருவ கால மாறுபாடு
Secular Trend	நீண்ட கால போக்கு
Simple Correlation	எளிய ஒட்டுறவு
Simple Linear Regression	எளிய நேர்கோட்டு உடன்தொடர்பு
skewed	சமச்சீரற்ற
Small Sample	சிறுகூறு ($n < 30$)
Specific Death Rate (SDR)	குறித்த இறப்பு விகிதம்
Specific Fertility Rate (SFR)	குறித்த கருவுறுதல் விகிதம்
Standard Deviation	திட்டவிலக்கம்
Standard Error	திட்டப்பிழை
Statistic	மாதிரிப்பண்பளவை, கூறுபண்பளவை
Statistics	புள்ளியியல்
Strength of impact	தாக்கத்தின் வலிமை
Sum of squares	வர்க்கங்களின் கூடுதல்
Symmetrical distribution	சமச்சீர் பரவல்
Symmetry	சமச்சீர்
Test of Hypothesis	கருதுகோள் சோதனை
Test statistic	மாதிரிப்பண்பளவை சோதனை, கூறுபண்பளவை சோதனை
Time Series	காலத் தொடர் வரிசை
Trade cycle	வணிகச் சுழல்
Treatment	நடத்துமுறைகள்
Two tailed test	இருமுனை சோதனை
Type I error	முதல் வகைப் பிழை
Type II error	இரண்டாம் வகைப் பிழை
Uncorrelated	ஒட்டுறவின்மை
Variance	மாறுபாட்டு அளவை
Vital Statistics	வாழ்நிலை புள்ளியியல்
Yule's Correlation	யூலின் தொடர்புக் கெழு

LOGARITHM TABLE

											Mean Difference								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
1.0	0.0000	0.0043	0.0086	0.0128	0.0170	0.0212	0.0253	0.0294	0.0334	0.0374	4	8	12	17	21	25	29	33	37
1.1	0.0414	0.0453	0.0492	0.0531	0.0569	0.0607	0.0645	0.0682	0.0719	0.0755	4	8	11	15	19	23	26	30	34
1.2	0.0792	0.0828	0.0864	0.0899	0.0934	0.0969	0.1004	0.1038	0.1072	0.1106	3	7	10	14	17	21	24	28	31
1.3	0.1139	0.1173	0.1206	0.1239	0.1271	0.1303	0.1335	0.1367	0.1399	0.1430	3	6	10	13	16	19	23	26	29
1.4	0.1461	0.1492	0.1523	0.1553	0.1584	0.1614	0.1644	0.1673	0.1703	0.1732	3	6	9	12	15	18	21	24	27
1.5	0.1761	0.1790	0.1818	0.1847	0.1875	0.1903	0.1931	0.1959	0.1987	0.2014	3	6	8	11	14	17	20	22	25
1.6	0.2041	0.2068	0.2095	0.2122	0.2148	0.2175	0.2201	0.2227	0.2253	0.2279	3	5	8	11	13	16	18	21	24
1.7	0.2304	0.2330	0.2355	0.2380	0.2405	0.2430	0.2455	0.2480	0.2504	0.2529	2	5	7	10	12	15	17	20	22
1.8	0.2553	0.2577	0.2601	0.2625	0.2648	0.2672	0.2695	0.2718	0.2742	0.2765	2	5	7	9	12	14	16	19	21
1.9	0.2788	0.2810	0.2833	0.2856	0.2878	0.2900	0.2923	0.2945	0.2967	0.2989	2	4	7	9	11	13	16	18	20
2.0	0.3010	0.3032	0.3054	0.3075	0.3096	0.3118	0.3139	0.3160	0.3181	0.3201	2	4	6	8	11	13	15	17	19
2.1	0.3222	0.3243	0.3263	0.3284	0.3304	0.3324	0.3345	0.3365	0.3385	0.3404	2	4	6	8	10	12	14	16	18
2.2	0.3424	0.3444	0.3464	0.3483	0.3502	0.3522	0.3541	0.3560	0.3579	0.3598	2	4	6	8	10	12	14	15	17
2.3	0.3617	0.3636	0.3655	0.3674	0.3692	0.3711	0.3729	0.3747	0.3766	0.3784	2	4	6	7	9	11	13	15	17
2.4	0.3802	0.3820	0.3838	0.3856	0.3874	0.3892	0.3909	0.3927	0.3945	0.3962	2	4	5	7	9	11	12	14	16
2.5	0.3979	0.3997	0.4014	0.4031	0.4048	0.4065	0.4082	0.4099	0.4116	0.4133	2	3	5	7	9	10	12	14	15
2.6	0.4150	0.4166	0.4183	0.4200	0.4216	0.4232	0.4249	0.4265	0.4281	0.4298	2	3	5	7	8	10	11	13	15
2.7	0.4314	0.4330	0.4346	0.4362	0.4378	0.4393	0.4409	0.4425	0.4440	0.4456	2	3	5	6	8	9	11	13	14
2.8	0.4472	0.4487	0.4502	0.4518	0.4533	0.4548	0.4564	0.4579	0.4594	0.4609	2	3	5	6	8	9	11	12	14
2.9	0.4624	0.4639	0.4654	0.4669	0.4683	0.4698	0.4713	0.4728	0.4742	0.4757	1	3	4	6	7	9	10	12	13
3.0	0.4771	0.4786	0.4800	0.4814	0.4829	0.4843	0.4857	0.4871	0.4886	0.4900	1	3	4	6	7	9	10	11	13
3.1	0.4914	0.4928	0.4942	0.4955	0.4969	0.4983	0.4997	0.5011	0.5024	0.5038	1	3	4	6	7	8	10	11	12
3.2	0.5051	0.5065	0.5079	0.5092	0.5105	0.5119	0.5132	0.5145	0.5159	0.5172	1	3	4	5	7	8	9	11	12
3.3	0.5185	0.5198	0.5211	0.5224	0.5237	0.5250	0.5263	0.5276	0.5289	0.5302	1	3	4	5	6	8	9	10	12
3.4	0.5315	0.5328	0.5340	0.5353	0.5366	0.5378	0.5391	0.5403	0.5416	0.5428	1	3	4	5	6	8	9	10	11
3.5	0.5441	0.5453	0.5465	0.5478	0.5490	0.5502	0.5514	0.5527	0.5539	0.5551	1	2	4	5	6	7	9	10	11
3.6	0.5563	0.5575	0.5587	0.5599	0.5611	0.5623	0.5635	0.5647	0.5658	0.5670	1	2	4	5	6	7	8	10	11
3.7	0.5682	0.5694	0.5705	0.5717	0.5729	0.5740	0.5752	0.5763	0.5775	0.5786	1	2	3	5	6	7	8	9	10
3.8	0.5798	0.5809	0.5821	0.5832	0.5843	0.5855	0.5866	0.5877	0.5888	0.5899	1	2	3	5	6	7	8	9	10
3.9	0.5911	0.5922	0.5933	0.5944	0.5955	0.5966	0.5977	0.5988	0.5999	0.6010	1	2	3	4	5	7	8	9	10
4.0	0.6021	0.6031	0.6042	0.6053	0.6064	0.6075	0.6085	0.6096	0.6107	0.6117	1	2	3	4	5	6	8	9	10
4.1	0.6128	0.6138	0.6149	0.6160	0.6170	0.6180	0.6191	0.6201	0.6212	0.6222	1	2	3	4	5	6	7	8	9
4.2	0.6232	0.6243	0.6253	0.6263	0.6274	0.6284	0.6294	0.6304	0.6314	0.6325	1	2	3	4	5	6	7	8	9
4.3	0.6335	0.6345	0.6355	0.6365	0.6375	0.6385	0.6395	0.6405	0.6415	0.6425	1	2	3	4	5	6	7	8	9
4.4	0.6435	0.6444	0.6454	0.6464	0.6474	0.6484	0.6493	0.6503	0.6513	0.6522	1	2	3	4	5	6	7	8	9
4.5	0.6532	0.6542	0.6551	0.6561	0.6571	0.6580	0.6590	0.6599	0.6609	0.6618	1	2	3	4	5	6	7	8	9
4.6	0.6628	0.6637	0.6646	0.6656	0.6665	0.6675	0.6684	0.6693	0.6702	0.6712	1	2	3	4	5	6	7	7	8
4.7	0.6721	0.6730	0.6739	0.6749	0.6758	0.6767	0.6776	0.6785	0.6794	0.6803	1	2	3	4	5	5	6	7	8
4.8	0.6812	0.6821	0.6830	0.6839	0.6848	0.6857	0.6866	0.6875	0.6884	0.6893	1	2	3	4	4	5	6	7	8
4.9	0.6902	0.6911	0.6920	0.6928	0.6937	0.6946	0.6955	0.6964	0.6972	0.6981	1	2	3	4	4	5	6	7	8
5.0	0.6990	0.6998	0.7007	0.7016	0.7024	0.7033	0.7042	0.7050	0.7059	0.7067	1	2	3	3	4	5	6	7	8
5.1	0.7076	0.7084	0.7093	0.7101	0.7110	0.7118	0.7126	0.7135	0.7143	0.7152	1	2	3	3	4	5	6	7	8
5.2	0.7160	0.7168	0.7177	0.7185	0.7193	0.7202	0.7210	0.7218	0.7226	0.7235	1	2	2	3	4	5	6	7	7
5.3	0.7243	0.7251	0.7259	0.7267	0.7275	0.7284	0.7292	0.7300	0.7308	0.7316	1	2	2	3	4	5	6	6	7
5.4	0.7324	0.7332	0.7340	0.7348	0.7356	0.7364	0.7372	0.7380	0.7388	0.7396	1	2	2	3	4	5	6	6	7

LOGARITHM TABLE

											Mean Difference								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
5.5	0.7404	0.7412	0.7419	0.7427	0.7435	0.7443	0.7451	0.7459	0.7466	0.7474	1	2	2	3	4	5	5	6	7
5.6	0.7482	0.7490	0.7497	0.7505	0.7513	0.7520	0.7528	0.7536	0.7543	0.7551	1	2	2	3	4	5	5	6	7
5.7	0.7559	0.7566	0.7574	0.7582	0.7589	0.7597	0.7604	0.7612	0.7619	0.7627	1	2	2	3	4	5	5	6	7
5.8	0.7634	0.7642	0.7649	0.7657	0.7664	0.7672	0.7679	0.7686	0.7694	0.7701	1	1	2	3	4	4	5	6	7
5.9	0.7709	0.7716	0.7723	0.7731	0.7738	0.7745	0.7752	0.7760	0.7767	0.7774	1	1	2	3	4	4	5	6	7
6.0	0.7782	0.7789	0.7796	0.7803	0.7810	0.7818	0.7825	0.7832	0.7839	0.7846	1	1	2	3	4	4	5	6	6
6.1	0.7853	0.7860	0.7868	0.7875	0.7882	0.7889	0.7896	0.7903	0.7910	0.7917	1	1	2	3	4	4	5	6	6
6.2	0.7924	0.7931	0.7938	0.7945	0.7952	0.7959	0.7966	0.7973	0.7980	0.7987	1	1	2	3	3	4	5	6	6
6.3	0.7993	0.8000	0.8007	0.8014	0.8021	0.8028	0.8035	0.8041	0.8048	0.8055	1	1	2	3	3	4	5	5	6
6.4	0.8062	0.8069	0.8075	0.8082	0.8089	0.8096	0.8102	0.8109	0.8116	0.8122	1	1	2	3	3	4	5	5	6
6.5	0.8129	0.8136	0.8142	0.8149	0.8156	0.8162	0.8169	0.8176	0.8182	0.8189	1	1	2	3	3	4	5	5	6
6.6	0.8195	0.8202	0.8209	0.8215	0.8222	0.8228	0.8235	0.8241	0.8248	0.8254	1	1	2	3	3	4	5	5	6
6.7	0.8261	0.8267	0.8274	0.8280	0.8287	0.8293	0.8299	0.8306	0.8312	0.8319	1	1	2	3	3	4	5	5	6
6.8	0.8325	0.8331	0.8338	0.8344	0.8351	0.8357	0.8363	0.8370	0.8376	0.8382	1	1	2	3	3	4	4	5	6
6.9	0.8388	0.8395	0.8401	0.8407	0.8414	0.8420	0.8426	0.8432	0.8439	0.8445	1	1	2	2	3	4	4	5	6
7.0	0.8451	0.8457	0.8463	0.8470	0.8476	0.8482	0.8488	0.8494	0.8500	0.8506	1	1	2	2	3	4	4	5	6
7.1	0.8513	0.8519	0.8525	0.8531	0.8537	0.8543	0.8549	0.8555	0.8561	0.8567	1	1	2	2	3	4	4	5	5
7.2	0.8573	0.8579	0.8585	0.8591	0.8597	0.8603	0.8609	0.8615	0.8621	0.8627	1	1	2	2	3	4	4	5	5
7.3	0.8633	0.8639	0.8645	0.8651	0.8657	0.8663	0.8669	0.8675	0.8681	0.8686	1	1	2	2	3	4	4	5	5
7.4	0.8692	0.8698	0.8704	0.8710	0.8716	0.8722	0.8727	0.8733	0.8739	0.8745	1	1	2	2	3	4	4	5	5
7.5	0.8751	0.8756	0.8762	0.8768	0.8774	0.8779	0.8785	0.8791	0.8797	0.8802	1	1	2	2	3	3	4	5	5
7.6	0.8808	0.8814	0.8820	0.8825	0.8831	0.8837	0.8842	0.8848	0.8854	0.8859	1	1	2	2	3	3	4	5	5
7.7	0.8865	0.8871	0.8876	0.8882	0.8887	0.8893	0.8899	0.8904	0.8910	0.8915	1	1	2	2	3	3	4	4	5
7.8	0.8921	0.8927	0.8932	0.8938	0.8943	0.8949	0.8954	0.8960	0.8965	0.8971	1	1	2	2	3	3	4	4	5
7.9	0.8976	0.8982	0.8987	0.8993	0.8998	0.9004	0.9009	0.9015	0.9020	0.9025	1	1	2	2	3	3	4	4	5
8.0	0.9031	0.9036	0.9042	0.9047	0.9053	0.9058	0.9063	0.9069	0.9074	0.9079	1	1	2	2	3	3	4	4	5
8.1	0.9085	0.9090	0.9096	0.9101	0.9106	0.9112	0.9117	0.9122	0.9128	0.9133	1	1	2	2	3	3	4	4	5
8.2	0.9138	0.9143	0.9149	0.9154	0.9159	0.9165	0.9170	0.9175	0.9180	0.9186	1	1	2	2	3	3	4	4	5
8.3	0.9191	0.9196	0.9201	0.9206	0.9212	0.9217	0.9222	0.9227	0.9232	0.9238	1	1	2	2	3	3	4	4	5
8.4	0.9243	0.9248	0.9253	0.9258	0.9263	0.9269	0.9274	0.9279	0.9284	0.9289	1	1	2	2	3	3	4	4	5
8.5	0.9294	0.9299	0.9304	0.9309	0.9315	0.9320	0.9325	0.9330	0.9335	0.9340	1	1	2	2	3	3	4	4	5
8.6	0.9345	0.9350	0.9355	0.9360	0.9365	0.9370	0.9375	0.9380	0.9385	0.9390	1	1	2	2	3	3	4	4	5
8.7	0.9395	0.9400	0.9405	0.9410	0.9415	0.9420	0.9425	0.9430	0.9435	0.9440	0	1	1	2	2	3	3	4	4
8.8	0.9445	0.9450	0.9455	0.9460	0.9465	0.9469	0.9474	0.9479	0.9484	0.9489	0	1	1	2	2	3	3	4	4
8.9	0.9494	0.9499	0.9504	0.9509	0.9513	0.9518	0.9523	0.9528	0.9533	0.9538	0	1	1	2	2	3	3	4	4
9.0	0.9542	0.9547	0.9552	0.9557	0.9562	0.9566	0.9571	0.9576	0.9581	0.9586	0	1	1	2	2	3	3	4	4
9.1	0.9590	0.9595	0.9600	0.9605	0.9609	0.9614	0.9619	0.9624	0.9628	0.9633	0	1	1	2	2	3	3	4	4
9.2	0.9638	0.9643	0.9647	0.9652	0.9657	0.9661	0.9666	0.9671	0.9675	0.9680	0	1	1	2	2	3	3	4	4
9.3	0.9685	0.9689	0.9694	0.9699	0.9703	0.9708	0.9713	0.9717	0.9722	0.9727	0	1	1	2	2	3	3	4	4
9.4	0.9731	0.9736	0.9741	0.9745	0.9750	0.9754	0.9759	0.9763	0.9768	0.9773	0	1	1	2	2	3	3	4	4
9.5	0.9777	0.9782	0.9786	0.9791	0.9795	0.9800	0.9805	0.9809	0.9814	0.9818	0	1	1	2	2	3	3	4	4
9.6	0.9823	0.9827	0.9832	0.9836	0.9841	0.9845	0.9850	0.9854	0.9859	0.9863	0	1	1	2	2	3	3	4	4
9.7	0.9868	0.9872	0.9877	0.9881	0.9886	0.9890	0.9894	0.9899	0.9903	0.9908	0	1	1	2	2	3	3	4	4
9.8	0.9912	0.9917	0.9921	0.9926	0.9930	0.9934	0.9939	0.9943	0.9948	0.9952	0	1	1	2	2	3	3	4	4
9.9	0.9956	0.9961	0.9965	0.9969	0.9974	0.9978	0.9983	0.9987	0.9991	0.9996	0	1	1	2	2	3	3	3	4

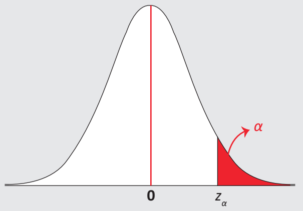
ANTI LOGARITHM TABLE

											Mean Difference								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
0.00	1.000	1.002	1.005	1.007	1.009	1.012	1.014	1.016	1.019	1.021	0	0	1	1	1	1	2	2	2
0.01	1.023	1.026	1.028	1.030	1.033	1.035	1.038	1.040	1.042	1.045	0	0	1	1	1	1	2	2	2
0.02	1.047	1.050	1.052	1.054	1.057	1.059	1.062	1.064	1.067	1.069	0	0	1	1	1	1	2	2	2
0.03	1.072	1.074	1.076	1.079	1.081	1.084	1.086	1.089	1.091	1.094	0	0	1	1	1	1	2	2	2
0.04	1.096	1.099	1.102	1.104	1.107	1.109	1.112	1.114	1.117	1.119	0	1	1	1	1	2	2	2	2
0.05	1.122	1.125	1.127	1.130	1.132	1.135	1.138	1.140	1.143	1.146	0	1	1	1	1	2	2	2	2
0.06	1.148	1.151	1.153	1.156	1.159	1.161	1.164	1.167	1.169	1.172	0	1	1	1	1	2	2	2	2
0.07	1.175	1.178	1.180	1.183	1.186	1.189	1.191	1.194	1.197	1.199	0	1	1	1	1	2	2	2	2
0.08	1.202	1.205	1.208	1.211	1.213	1.216	1.219	1.222	1.225	1.227	0	1	1	1	1	2	2	2	3
0.09	1.230	1.233	1.236	1.239	1.242	1.245	1.247	1.250	1.253	1.256	0	1	1	1	1	2	2	2	3
0.10	1.259	1.262	1.265	1.268	1.271	1.274	1.276	1.279	1.282	1.285	0	1	1	1	1	2	2	2	3
0.11	1.288	1.291	1.294	1.297	1.300	1.303	1.306	1.309	1.312	1.315	0	1	1	1	2	2	2	2	3
0.12	1.318	1.321	1.324	1.327	1.330	1.334	1.337	1.340	1.343	1.346	0	1	1	1	2	2	2	2	3
0.13	1.349	1.352	1.355	1.358	1.361	1.365	1.368	1.371	1.374	1.377	0	1	1	1	2	2	2	3	3
0.14	1.380	1.384	1.387	1.390	1.393	1.396	1.400	1.403	1.406	1.409	0	1	1	1	2	2	2	3	3
0.15	1.413	1.416	1.419	1.422	1.426	1.429	1.432	1.435	1.439	1.442	0	1	1	1	2	2	2	3	3
0.16	1.445	1.449	1.452	1.455	1.459	1.462	1.466	1.469	1.472	1.476	0	1	1	1	2	2	2	3	3
0.17	1.479	1.483	1.486	1.489	1.493	1.496	1.500	1.503	1.507	1.510	0	1	1	1	2	2	2	3	3
0.18	1.514	1.517	1.521	1.524	1.528	1.531	1.535	1.538	1.542	1.545	0	1	1	1	2	2	2	3	3
0.19	1.549	1.552	1.556	1.560	1.563	1.567	1.570	1.574	1.578	1.581	0	1	1	1	2	2	3	3	3
0.20	1.585	1.589	1.592	1.596	1.600	1.603	1.607	1.611	1.614	1.618	0	1	1	1	2	2	3	3	3
0.21	1.622	1.626	1.629	1.633	1.637	1.641	1.644	1.648	1.652	1.656	0	1	1	2	2	2	3	3	3
0.22	1.660	1.663	1.667	1.671	1.675	1.679	1.683	1.687	1.690	1.694	0	1	1	2	2	2	3	3	3
0.23	1.698	1.702	1.706	1.710	1.714	1.718	1.722	1.726	1.730	1.734	0	1	1	2	2	2	3	3	4
0.24	1.738	1.742	1.746	1.750	1.754	1.758	1.762	1.766	1.770	1.774	0	1	1	2	2	2	3	3	4
0.25	1.778	1.782	1.786	1.791	1.795	1.799	1.803	1.807	1.811	1.816	0	1	1	2	2	2	3	3	4
0.26	1.820	1.824	1.828	1.832	1.837	1.841	1.845	1.849	1.854	1.858	0	1	1	2	2	3	3	3	4
0.27	1.862	1.866	1.871	1.875	1.879	1.884	1.888	1.892	1.897	1.901	0	1	1	2	2	3	3	3	4
0.28	1.905	1.910	1.914	1.919	1.923	1.928	1.932	1.936	1.941	1.945	0	1	1	2	2	3	3	4	4
0.29	1.950	1.954	1.959	1.963	1.968	1.972	1.977	1.982	1.986	1.991	0	1	1	2	2	3	3	4	4
0.30	1.995	2.000	2.004	2.009	2.014	2.018	2.023	2.028	2.032	2.037	0	1	1	2	2	3	3	4	4
0.31	2.042	2.046	2.051	2.056	2.061	2.065	2.070	2.075	2.080	2.084	0	1	1	2	2	3	3	4	4
0.32	2.089	2.094	2.099	2.104	2.109	2.113	2.118	2.123	2.128	2.133	0	1	1	2	2	3	3	4	4
0.33	2.138	2.143	2.148	2.153	2.158	2.163	2.168	2.173	2.178	2.183	0	1	1	2	2	3	3	4	4
0.34	2.188	2.193	2.198	2.203	2.208	2.213	2.218	2.223	2.228	2.234	1	1	2	2	3	3	4	4	5
0.35	2.239	2.244	2.249	2.254	2.259	2.265	2.270	2.275	2.280	2.286	1	1	2	2	3	3	4	4	5
0.36	2.291	2.296	2.301	2.307	2.312	2.317	2.323	2.328	2.333	2.339	1	1	2	2	3	3	4	4	5
0.37	2.344	2.350	2.355	2.360	2.366	2.371	2.377	2.382	2.388	2.393	1	1	2	2	3	3	4	4	5
0.38	2.399	2.404	2.410	2.415	2.421	2.427	2.432	2.438	2.443	2.449	1	1	2	2	3	3	4	4	5
0.39	2.455	2.460	2.466	2.472	2.477	2.483	2.489	2.495	2.500	2.506	1	1	2	2	3	3	4	5	5
0.40	2.512	2.518	2.523	2.529	2.535	2.541	2.547	2.553	2.559	2.564	1	1	2	2	3	4	4	5	5
0.41	2.570	2.576	2.582	2.588	2.594	2.600	2.606	2.612	2.618	2.624	1	1	2	2	3	4	4	5	5
0.42	2.630	2.636	2.642	2.649	2.655	2.661	2.667	2.673	2.679	2.685	1	1	2	2	3	4	4	5	6
0.43	2.692	2.698	2.704	2.710	2.716	2.723	2.729	2.735	2.742	2.748	1	1	2	3	3	4	4	5	6
0.44	2.754	2.761	2.767	2.773	2.780	2.786	2.793	2.799	2.805	2.812	1	1	2	3	3	4	4	5	6
0.45	2.818	2.825	2.831	2.838	2.844	2.851	2.858	2.864	2.871	2.877	1	1	2	3	3	4	5	5	6
0.46	2.884	2.891	2.897	2.904	2.911	2.917	2.924	2.931	2.938	2.944	1	1	2	3	3	4	5	5	6
0.47	2.951	2.958	2.965	2.972	2.979	2.985	2.992	2.999	3.006	3.013	1	1	2	3	3	4	5	5	6
0.48	3.020	3.027	3.034	3.041	3.048	3.055	3.062	3.069	3.076	3.083	1	1	2	3	4	4	5	6	6
0.49	3.090	3.097	3.105	3.112	3.119	3.126	3.133	3.141	3.148	3.155	1	1	2	3	4	4	5	6	6

ANTI LOGARITHM TABLE

											Mean Difference								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
0.50	3.162	3.170	3.177	3.184	3.192	3.199	3.206	3.214	3.221	3.228	1	1	2	3	4	4	5	6	7
0.51	3.236	3.243	3.251	3.258	3.266	3.273	3.281	3.289	3.296	3.304	1	2	2	3	4	5	5	6	7
0.52	3.311	3.319	3.327	3.334	3.342	3.350	3.357	3.365	3.373	3.381	1	2	2	3	4	5	5	6	7
0.53	3.388	3.396	3.404	3.412	3.420	3.428	3.436	3.443	3.451	3.459	1	2	2	3	4	5	6	6	7
0.54	3.467	3.475	3.483	3.491	3.499	3.508	3.516	3.524	3.532	3.540	1	2	2	3	4	5	6	6	7
0.55	3.548	3.556	3.565	3.573	3.581	3.589	3.597	3.606	3.614	3.622	1	2	2	3	4	5	6	7	7
0.56	3.631	3.639	3.648	3.656	3.664	3.673	3.681	3.690	3.698	3.707	1	2	3	3	4	5	6	7	8
0.57	3.715	3.724	3.733	3.741	3.750	3.758	3.767	3.776	3.784	3.793	1	2	3	3	4	5	6	7	8
0.58	3.802	3.811	3.819	3.828	3.837	3.846	3.855	3.864	3.873	3.882	1	2	3	4	4	5	6	7	8
0.59	3.890	3.899	3.908	3.917	3.926	3.936	3.945	3.954	3.963	3.972	1	2	3	4	5	5	6	7	8
0.60	3.981	3.990	3.999	4.009	4.018	4.027	4.036	4.046	4.055	4.064	1	2	3	4	5	6	6	7	8
0.61	4.074	4.083	4.093	4.102	4.111	4.121	4.130	4.140	4.150	4.159	1	2	3	4	5	6	7	8	9
0.62	4.169	4.178	4.188	4.198	4.207	4.217	4.227	4.236	4.246	4.256	1	2	3	4	5	6	7	8	9
0.63	4.266	4.276	4.285	4.295	4.305	4.315	4.325	4.335	4.345	4.355	1	2	3	4	5	6	7	8	9
0.64	4.365	4.375	4.385	4.395	4.406	4.416	4.426	4.436	4.446	4.457	1	2	3	4	5	6	7	8	9
0.65	4.467	4.477	4.487	4.498	4.508	4.519	4.529	4.539	4.550	4.560	1	2	3	4	5	6	7	8	9
0.66	4.571	4.581	4.592	4.603	4.613	4.624	4.634	4.645	4.656	4.667	1	2	3	4	5	6	7	9	10
0.67	4.677	4.688	4.699	4.710	4.721	4.732	4.742	4.753	4.764	4.775	1	2	3	4	5	7	8	9	10
0.68	4.786	4.797	4.808	4.819	4.831	4.842	4.853	4.864	4.875	4.887	1	2	3	4	6	7	8	9	10
0.69	4.898	4.909	4.920	4.932	4.943	4.955	4.966	4.977	4.989	5.000	1	2	3	5	6	7	8	9	10
0.70	5.012	5.023	5.035	5.047	5.058	5.070	5.082	5.093	5.105	5.117	1	2	4	5	6	7	8	9	11
0.71	5.129	5.140	5.152	5.164	5.176	5.188	5.200	5.212	5.224	5.236	1	2	4	5	6	7	8	10	11
0.72	5.248	5.260	5.272	5.284	5.297	5.309	5.321	5.333	5.346	5.358	1	2	4	5	6	7	9	10	11
0.73	5.370	5.383	5.395	5.408	5.420	5.433	5.445	5.458	5.470	5.483	1	3	4	5	6	8	9	10	11
0.74	5.495	5.508	5.521	5.534	5.546	5.559	5.572	5.585	5.598	5.610	1	3	4	5	6	8	9	10	12
0.75	5.623	5.636	5.649	5.662	5.675	5.689	5.702	5.715	5.728	5.741	1	3	4	5	7	8	9	10	12
0.76	5.754	5.768	5.781	5.794	5.808	5.821	5.834	5.848	5.861	5.875	1	3	4	5	7	8	9	11	12
0.77	5.888	5.902	5.916	5.929	5.943	5.957	5.970	5.984	5.998	6.012	1	3	4	5	7	8	10	11	12
0.78	6.026	6.039	6.053	6.067	6.081	6.095	6.109	6.124	6.138	6.152	1	3	4	6	7	8	10	11	13
0.79	6.166	6.180	6.194	6.209	6.223	6.237	6.252	6.266	6.281	6.295	1	3	4	6	7	9	10	11	13
0.80	6.310	6.324	6.339	6.353	6.368	6.383	6.397	6.412	6.427	6.442	1	3	4	6	7	9	10	12	13
0.81	6.457	6.471	6.486	6.501	6.516	6.531	6.546	6.561	6.577	6.592	2	3	5	6	8	9	11	12	14
0.82	6.607	6.622	6.637	6.653	6.668	6.683	6.699	6.714	6.730	6.745	2	3	5	6	8	9	11	12	14
0.83	6.761	6.776	6.792	6.808	6.823	6.839	6.855	6.871	6.887	6.902	2	3	5	6	8	9	11	13	14
0.84	6.918	6.934	6.950	6.966	6.982	6.998	7.015	7.031	7.047	7.063	2	3	5	6	8	10	11	13	15
0.85	7.079	7.096	7.112	7.129	7.145	7.161	7.178	7.194	7.211	7.228	2	3	5	7	8	10	12	13	15
0.86	7.244	7.261	7.278	7.295	7.311	7.328	7.345	7.362	7.379	7.396	2	3	5	7	8	10	12	13	15
0.87	7.413	7.430	7.447	7.464	7.482	7.499	7.516	7.534	7.551	7.568	2	3	5	7	9	10	12	14	16
0.88	7.586	7.603	7.621	7.638	7.656	7.674	7.691	7.709	7.727	7.745	2	4	5	7	9	11	12	14	16
0.89	7.762	7.780	7.798	7.816	7.834	7.852	7.870	7.889	7.907	7.925	2	4	5	7	9	11	13	14	16
0.90	7.943	7.962	7.980	7.998	8.017	8.035	8.054	8.072	8.091	8.110	2	4	6	7	9	11	13	15	17
0.91	8.128	8.147	8.166	8.185	8.204	8.222	8.241	8.260	8.279	8.299	2	4	6	8	9	11	13	15	17
0.92	8.318	8.337	8.356	8.375	8.395	8.414	8.433	8.453	8.472	8.492	2	4	6	8	10	12	14	15	17
0.93	8.511	8.531	8.551	8.570	8.590	8.610	8.630	8.650	8.670	8.690	2	4	6	8	10	12	14	16	18
0.94	8.710	8.730	8.750	8.770	8.790	8.810	8.831	8.851	8.872	8.892	2	4	6	8	10	12	14	16	18
0.95	8.913	8.933	8.954	8.974	8.995	9.016	9.036	9.057	9.078	9.099	2	4	6	8	10	12	15	17	19
0.96	9.120	9.141	9.162	9.183	9.204	9.226	9.247	9.268	9.290	9.311	2	4	6	8	11	13	15	17	19
0.97	9.333	9.354	9.376	9.397	9.419	9.441	9.462	9.484	9.506	9.528	2	4	7	9	11	13	15	17	20
0.98	9.550	9.572	9.594	9.616	9.638	9.661	9.683	9.705	9.727	9.750	2	4	7	9	11	13	16	18	20
0.99	9.772	9.795	9.817	9.840	9.863	9.886	9.908	9.931	9.954	9.977	2	5	7	9	11	14	16	18	20

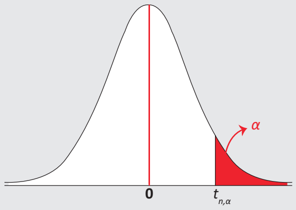
Standard Normal Distribution Values - Critical Values



$P(z > z_\alpha) = \alpha$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.0	0.5000	0.4960	0.4920	0.4880	0.4841	0.4801	0.4761	0.4721	0.4681	0.4641	0.0
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247	0.1
0.2	0.4207	0.4168	0.4129	0.4091	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859	0.2
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483	0.3
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121	0.4
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776	0.5
0.6	0.2743	0.2709	0.2676	0.2644	0.2611	0.2579	0.2546	0.2514	0.2483	0.2451	0.6
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2207	0.2177	0.2148	0.7
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867	0.8
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611	0.9
1.0	0.1587	0.1563	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379	1.0
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170	1.1
1.2	0.1151	0.1131	0.1112	0.1094	0.1075	0.1057	0.1038	0.1020	0.1003	0.0985	1.2
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823	1.3
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681	1.4
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559	1.5
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455	1.6
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367	1.7
1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294	1.8
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233	1.9
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183	2.0
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143	2.1
2.2	0.0139	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	2.2
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084	2.3
2.4	0.0082	0.0080	0.0078	0.0076	0.0073	0.0071	0.0070	0.0068	0.0066	0.0064	2.4
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048	2.5
2.6	0.0047	0.0045	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	2.6
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	2.7
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019	2.8
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	2.9
3.0	0.0014	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010	3.0
3.1	0.0010	0.0009	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0007	0.0007	3.1
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	3.2
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	3.3
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	3.4
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	3.5

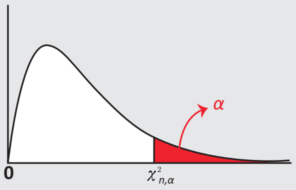
t Distribution : Critical *t* Values: $t_{n,\alpha}$



$P(T > t_{n,\alpha}) = \alpha$

df	Area in One Tails				
	0.005	0.01	0.025	0.05	0.10
	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319
24	2.797	2.492	2.064	1.711	1.318
25	2.787	2.485	2.060	1.708	1.316
26	2.779	2.479	2.056	1.706	1.315
27	2.771	2.473	2.052	1.703	1.314
28	2.763	2.467	2.048	1.701	1.313
29	2.756	2.462	2.045	1.699	1.311
30	2.750	2.457	2.042	1.697	1.310
40	2.704	2.423	2.021	1.684	1.303
50	2.678	2.403	2.009	1.676	1.299
60	2.660	2.390	2.000	1.671	1.296
100	2.626	2.364	1.984	1.660	1.290
120	2.617	2.358	1.980	1.658	1.289
Infinity	2.576	2.326	1.960	1.645	1.282

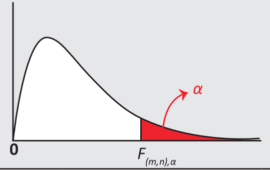
Critical Values of χ^2 Statistic



$P(\chi^2_n > \chi^2_{n,\alpha}) = \alpha$

	α										
df	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005	*df*
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750	5
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156	18
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997	20
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401	21
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796	22
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181	23
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559	24
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928	25
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290	26
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645	27
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993	28
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336	29
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672	30
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766	40
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490	50
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952	60
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215	70
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321	80
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299	90
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169	100

F-Distribution - Critical Values $\alpha = 0.01$



n	*m*																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847	6106.321	6157.285	6208.730	6234.631	6260.649	6286.782	6313.030	6339.391
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.251	4.099	4.021	3.941	3.860	3.776	3.690
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	4.010	3.858	3.780	3.701	3.619	3.535	3.449
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.815	3.665	3.587	3.507	3.425	3.341	3.255
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.656	3.505	3.427	3.348	3.266	3.181	3.094
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.455	3.312	3.162	3.084	3.003	2.920	2.835	2.746
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.660
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	3.030	2.880	2.801	2.720	2.636	2.548	2.457
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.931	2.781	2.702	2.620	2.535	2.447	2.354
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.310
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.850	2.699	2.620	2.538	2.453	2.364	2.270
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.815	2.664	2.585	2.503	2.417	2.327	2.233
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.783	2.632	2.552	2.470	2.384	2.294	2.198
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.896	2.753	2.602	2.522	2.440	2.354	2.263	2.167
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.868	2.726	2.574	2.495	2.412	2.325	2.234	2.138
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.700	2.549	2.469	2.386	2.299	2.208	2.111
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726
120	6.851	4.787	3.949	3.48	3.174	2.956	2.792	2.663	2.559	2.472	2.336	2.192	2.035	1.95	1.86	1.763	1.656	1.533

n	*m*																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882	243.906	245.950	248.013	249.052	250.095	251.143	252.196	253.253
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.413	19.429	19.446	19.454	19.462	19.471	19.479	19.487
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.703	8.660	8.639	8.617	8.594	8.572	8.549
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.858	5.803	5.774	5.746	5.717	5.688	5.658
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.619	4.558	4.527	4.496	4.464	4.431	4.398
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.938	3.874	3.841	3.808	3.774	3.740	3.705
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.511	3.445	3.410	3.376	3.340	3.304	3.267
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.284	3.218	3.150	3.115	3.079	3.043	3.005	2.967
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	3.006	2.936	2.900	2.864	2.826	2.787	2.748
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.845	2.774	2.737	2.700	2.661	2.621	2.580
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.719	2.646	2.609	2.570	2.531	2.490	2.448
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.617	2.544	2.505	2.466	2.426	2.384	2.341
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.533	2.459	2.420	2.380	2.339	2.297	2.252
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.463	2.388	2.349	2.308	2.266	2.223	2.178
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.403	2.328	2.288	2.247	2.204	2.160	2.114
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.352	2.276	2.235	2.194	2.151	2.106	2.059
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.308	2.230	2.190	2.148	2.104	2.058	2.011
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.269	2.191	2.150	2.107	2.063	2.017	1.968
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.234	2.155	2.114	2.071	2.026	1.980	1.930
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.203	2.124	2.082	2.039	1.994	1.946	1.896
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.250	2.176	2.096	2.054	2.010	1.965	1.916	1.866
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.226	2.151	2.071	2.028	1.984	1.938	1.889	1.838
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.204	2.128	2.048	2.005	1.961	1.914	1.865	1.813
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.183	2.108	2.027	1.984	1.939	1.892	1.842	1.790
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.165	2.089	2.007	1.964	1.919	1.872	1.822	1.768
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220	2.148	2.072	1.990	1.946	1.901	1.853	1.803	1.749
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204	2.132	2.056	1.974	1.930	1.884	1.836	1.785	1.731
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.118	2.041	1.959	1.915	1.869	1.820	1.769	1.714
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177	2.104	2.027	1.945	1.901	1.854	1.806	1.754	1.698
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.092	2.015	1.932	1.887	1.841	1.792	1.740	1.683
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.003	1.924	1.839	1.793	1.744	1.693	1.637	1.577
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.917	1.836	1.748	1.700	1.649	1.594	1.534	1.467
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.834	1.750	1.659	1.608	1.554	1.495	1.429	1.352

Exponential Function Table (Values of e^{-m})

m	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	1.0000	0.9900	0.9802	0.9704	0.9608	0.9512	0.9418	0.9324	0.9231	0.9139
0.1	0.9048	0.8958	0.8869	0.8781	0.8694	0.8607	0.8521	0.8437	0.8353	0.8270
0.2	0.8187	0.8106	0.8025	0.7945	0.7866	0.7788	0.7711	0.7634	0.7558	0.7483
0.3	0.7408	0.7334	0.7261	0.7189	0.7118	0.7047	0.6977	0.6907	0.6839	0.6771
0.4	0.6703	0.6637	0.6570	0.6505	0.6440	0.6376	0.6313	0.6250	0.6188	0.6126
0.5	0.6065	0.6005	0.5945	0.5886	0.5827	0.5769	0.5712	0.5655	0.5599	0.5543
0.6	0.5488	0.5434	0.5379	0.5326	0.5273	0.5220	0.5169	0.5117	0.5066	0.5016
0.7	0.4966	0.4916	0.4868	0.4819	0.4771	0.4724	0.4677	0.4630	0.4584	0.4538
0.8	0.4493	0.4449	0.4404	0.4360	0.4317	0.4274	0.4232	0.4190	0.4148	0.4107
0.9	0.4066	0.4025	0.3985	0.3946	0.3906	0.3867	0.3829	0.3791	0.3753	0.3716

Statistics – Class XII
List of Authors and Reviewers

Domain Experts

Dr. G. Gopal
Professor & Head (Retd.), Dept. of Statistics
University of Madras, Chennai

Dr. G. Stephen Vincent
Associate Professor & Head (Retd.), Dept. of Statistics
St.Joseph's College, Trichy

Dr. R. Ravanan
Principal, Presidency College, Chennai.

Dr. K. Senthamarai Kannan
Professor, Dept. of Statistics, Manonmaniam
Sundaranar University, Tirunelveli.

Dr. A. Loganathan
Professor, Dept. of Statistics, Manonmaniam
Sundaranar University, Tirunelveli.

Dr. R. Kannan
Professor, Dept. of Statistics,
Annamalai University, Chidambaram.

Dr. N. Viswanathan
Associate Professor, Dept. of Statistics,
Presidency College, Chennai.

Dr. R.K. Radha
Assistant Professor, Dept. of Statistics,
Presidency College, Chennai.

Art and Design Team

Layout

Yogesh B,
Yesurathinam
R Mathan Raj

Illustrations & Image Editing

Muthukumar R.

In-House

QC - Rajesh Thangapppan
Kamatchi Balan Arumugam
Arun Kamaraj
Jerald Wilson

Wrapper- Kathir Arumugam

Co-ordination

Ramesh Munisamy

Typist

G. Beula Lancy

Reviewers

Dr. M.R. Srinivasan
Professor & Head, Dept. of Statistics,
University of Madras, Chennai.

Dr. P. Dhanavanthan
Professor and Dean, Dept. of Statistics,
Pondicherry University, Pondicherry.

Content Writers

G. Gnanasundaram
HM (Retd.), SSV HSS, Parktown, Chennai.

P. Rengarajan
PG Asst., (Retd.), Thiyagarajar HSS, Madurai.

AL.Nagammai
PG Asst., Sevasangam GHSS, Trichy.

M. Rama Lakshmi
PG Asst., Suguni Bai Sanathana Dharma GHSS, Chennai.

Maala Bhaskaran
PG Asst., GGHSS, Nandhivaram, Kanchipuram.

M. Boobalan
PG Asst., Zamindar HSS, Thuraiyur, Trichy.

R. Avoodaiappan
PG Asst., GGHSS, Ashok Pillar, Chennai.

K. Chitra
PG Asst., Tarapore and Loganathan GHSS, Chennai.

Academic Coordinator

N. Gnanasekaran
B.T. Asst.,
Govt. Girls Hr. Sec. School, Thiruporur, Kanchipuram Dt.

ICT Coordinator

D. Vasuraj
BT Asst., PUMS, Kosapur, Puzhal Block,
Thiruvallur Dt.

QR Coordinators

R. Jaganathan , SGT,
PUMS - Ganesapuram, Polur , Thiruvannamalai.

V. Padmavathi, B.T,
GHS, Vetriyur, Ariyalur.

M. Murugesan, B.T,
PUMS. Pethavelankottagam, Muttupettai, Thiruvarur.

**This book has been printed on 80 G.S.M.
Elegant Maplitho paper.**

Printed by offset at: