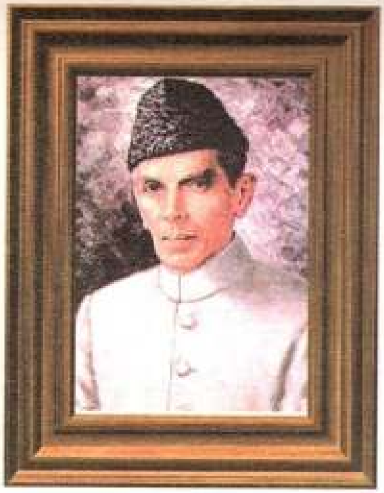
STATISTICS 11

2017-18

PUNJAB CURRICULUM AND
TEXTBOOK BOARD, LAHORE



"Education is a matter of life and death for Pakistan. The world is progressing so rapidly that without requisite advance in education, not only shall we be left behind others but may be wiped out altogether."

(September 26, 1947, Karachi)

Quaid-e-Azam
Muhammad Ali Jinnah
Founder of Pakistan

قومی ترانہ

پاک سرزمین شاد باد   کشورِ حسین شاد باد
تُو نشانِ عزمِ عالی شان   ارضِ پاکستان
مرکزِ یقین شاد باد

پاک سرزمین کا نظام   قوّتِ اخوّتِ عوام
قوم ، مُلک ، سلطنت   پایندہ تابندہ باد
شاد باد منزلِ مراد

پرچمِ ستارہ و ہلال   رہبرِ ترقّی و کمال
ترجمانِ ماضی، شانِ حال   جانِ استقبال
سایۂ خدائے ذُوالجلال

All textbooks of the Punjab Curriculum and Textbook Board carry a square shaped security sticker on the title page. The sticker exhibits a unique colour shift due to change of angle (orange to green) in the logo of Punjab Curriculum and Textbook Board, Lahore. If you scratch protective coating on the white strip above QR Code with a coin, it will reveal the registered trade name "PCTB". Moreover, if QR Code is checked with Android Application "News Reader" from your android mobile, serial number on sticker will be seen inside QR Code. This is a hallmark of the authenticity of original textbook of the Punjab Curriculum and Textbook Board. Look for the security sticker while purchasing the textbooks. If there is no sticker on the title page or it has been tampered with, do not buy such a book.

بِسۡمِ اللّٰہِ الرَّحۡمٰنِ الرَّحِیۡمِ ۝

A Textbook of

STATISTICS

For Class

11

PUNJAB CURRICULUM AND
TEXTBOOK BOARD, LAHORE

Prepared by: Punjab Curriculum & Textbook Board, Lahore.
Approved by: Federal Ministry of Education, Curriculum Wing, Islamabad.
Vide letter No F.2-1/99-Maths (Stat)

CONTENTS

Authors:

Dr. Faqir Muhammad
Professor & Controller Examinations,
Allama Iqbal Open University, Islamabad.

Mr. Amjad Mehmood
Lecturer, Punjab College,
Lahore.

Editor:
Mr. Mazhar Hayat
Subject Specialist,

Director (Manuscripts): Dr. Mobeen Akhtar
Deputy Director (Graphics): Raza-ur-Rehman

Supervised by:
Madiha Mehmood
Subject Specialist

ARTIST Aisha Waheed

Published by: Idara Farogh Urdu, Lahore
Printed by: Qudrat Ullah Printers, Lahore

Date of Printing	Edition	Impression	No of Copies	Price
May 2017	10th	10th	12,000	96.00

1 Introduction to Statistics

1.1 Introduction

The word statistics was first used by a German scholar Gotifried Achenwall about the middle of the 18th century as the science of Statecraft concerning the collection and use of data by the state. According to another pioneer statistician Yule, the word statistics occurred at the earliest in the book "The elements of universal erudition" by Baron (1770). It was used again with rather a wider definition in 1787 by E.A.W. Zimmermann in "A political survey of the present state of Europe". It appeared in Encyclopedia Britannica in 1797 and was used by Sir John Sinclair in Britain in a series of volumes published between 1791 and 1799 giving a statistical account of Scotland. In the 19th century, the word statistics acquired a wider meaning, covering numerical data of almost any subject whatever and also interpretation of data through appropriate analysis. The word **data** is used for numerical facts and a single numerical fact is **datum.**

Since early 1920's with the growth in the experimental sciences there was a need for reliable scientific methods for analyzing the results of experiments and surveys. The modern subject of statistics evolves with many of the early developers of statistical methodology being experimenters themselves with the incentive of their own practical problems.

Although the gathering and presenting of information is still an important part of statistics, the modern statistics is quite different from the early days. Statistics now-a-days includes probability theory and applied mathematics. These days computers have made to perform statistical analysis routinely which could not have been contemplated in the past. Computers do not form a part of statistical theory but may be useful in applying statistical theory to solve a practical problem.

Statistics has been defined as **the mathematical science of making decisions and drawing conclusions from data in situations of uncertainty**. It includes designing of experiments, collection, organization, summarization, analysis and interpretation of numerical data.

In the above definition of statistics, we have only considered its scientifi: meaning. In day to day usage, the word statistics refers to *numbers* or *facts*, such as statistics of births, statistics of deaths and statistics of road accidents etc. In some other situations, it has a symbolic meaning such as "do not become a statistics on the next weekend". The word statistics is also the plural of 'statistic' which is a statistical term and is a quantity calculated from the sample values.

Before proceeding further, some statistical terms and notations need to be defined and discussed. We explain each of the terms through the following example:

Example 1.1: It was observed that out of 500 college students surveyed, 300 were females. Is there evidence that more students in this college are females.

1.1.1. Definitions

Population: The total group under discussion or the group to which the results will be generalized is called population. In the example 1.1, the set of all students in the college is our population.

Sample: Sometimes the measurement of interest can not be made on the whole population, then we choose a subset of population to draw inference about the population. If the inferences from sample to population are going to be meaningful, it is imperative that the sample should be representative of the population. In the example 1.1, the set of 500 students being a subset of all college student, is our sample.

Ratio: The ratio of A to B is the fraction A/B. In the example 1.1, there are 300 females and the remaining 200 are males. So, the male female ratio is 200/300.

Proportion: A proportion is a special ratio of a part to its total. In the example 1.1, the proportion of females and males are 300/500 and 200/500 respectively. Proportion becomes percentage when multiplied by 100.

The female percentage in the example 1.1 is (300/500) × 100 i.e., 60%. 60% means 60 out of 100. The symbol % is abbreviation for **percent**.

Parameter: It is a quantity computed from a population when the entire population is available. Parameters are fixed or constant quantities and are not usually known. In the example1.1, the proportion of female students in the population is our parameter.

Statistic: It is a quantity computed from a sample. In the example 1.1, the proportion of the female students in the sample of students is a statistic. Statistic is variable because it varies from sample to sample.

Sampling Variability: As sample is a representative part of a population; there may be more than one samples in a population. Therefore, all samples from the same population may not be identical e.g. with reference to the example 1.1, another sample of 500 students surveyed may not contain 300 female students.

Experiment: Any study in which the scientist can control the allocation of treatments to the experimental units is called an experiment. Every unit must be capable of receiving every treatment, and the decision as to which unit receives which treatment is determined by an allocation mechanism. This mechanism could be that the scientist observes the unit and then decides which treatment to apply, which would be an unsatisfactory allocation mechanism because of the possibility of the subjective bias, or it could be to assign the treatments according to a rule which would be scientifically acceptable. So, all allocation mechanisms do not lead to acceptable inferences.

Sample Surveys: In a sample survey, there are no treatments. The units in the population under study are listed in a frame and a sample of units is selected from the frame using a selection mechanism. So, the distinguishing feature of a survey is the control over the selection of units.

The features which distinguishexperiments andsample surveys are control over allocation and control over selection.

Constant: Quantities which don't vary from individual to individual are called constants e.g.

$$\pi \approx 3.14159,\ e \approx 2.71828,\ 4,\ 25 \text{ etc.}$$

Order Statistic: The order statistic (OS) of data $Y_1, Y_2, Y_3, \ldots\ldots Y_n'$ is just the arrangement of data in order of magnitude. It is denoted by $Y_{(1)}, Y_{(2)}, Y_{(3)}, \ldots, Y_{(n)}$

$Y_{(1)}$ is the minimum of $Y_1, Y_2, Y_3, \ldots, Y_n$ and $Y_{(n)}$ is the maximum of $Y_1, Y_2, Y_3, \ldots, Y_n$

$Y_{(1)}$ is minimum order statistic and $Y_{(n)}$ is maximum order statistic. If the data are arranged in increasing order of magnitude, the data are said to be arranged in ***ascending order*** and if the data are arranged in decreasing order of magnitude, the data are said to be arranged in ***descending order***.

Model: It is a mathematical statement used in studying the results of an experiment or predicting the behaviour of future repetitions of the experiment.

The models involve probability distributions which describe the variability in the characteristic of interest in the population and therefore, the variability we might expect to get in different samples.

In the example 1.1, binomial model is appropriate where parameter is the proportion of students who are females in the population.

A common model for describing the makeup of an observation states that it consists of a mean plus an error so an observation can be described by means of simplest model as:

$$Y_i = \mu + \epsilon_i \qquad (1.1)$$

where, Y_i represents any individual observation, μ (a Greek letter read as meu) represents average or mean of the population and ϵ_i (a Greek letter read as epsilon) represents a random error.

Random Error (C_i): Random error is the chance variation in an observational process. It should not be confused with its synonym in common usage mistake, which means human error.

Equation (1.1) can be written as:

$$\epsilon_i = Y_i - \mu \qquad (1.2)$$

ϵ_i*'s* are usually assumed to be from a population having zero mean. As the random error sum to zero, so, there would be approximately equal number of positive and negative deviations.

The term, $(Y_i - \mu)$ is known as *deviation* of an observation Y_i from mean μ.

Equation (1.1) represents the simplest form of the linear additive model. The μ may be a single mean.

1.1.2 Notations

Sigma (Σ): It is a Greek letter and is used as a short-hand notation for sum. For example.

i.e., if we have to add five numbers Y_1, Y_2, Y_3, Y_4, and Y_5 , then

$$Y_1 + Y_2 + Y_3 + Y_4 + Y_5 = \sum_{i=1}^{5} Y_i$$

This tells us to sum all Y_i values starting at $i = 1$ (the lower limit) and stopping at $i = 5$ (the upper limit). It is always assumed that consecutive integer values are to be summed unless otherwise specified. Other examples are

i) $\Sigma 2^Y : Y \in D$ where $D = \{2, 3,4,7\}$

$$\Sigma 2^Y = 2^2 + 2^3 + 2^4 + 2^7$$

where $\in$ is read as belongs to.

ii) If $P(Y)$ denotes the probability then

$$\Sigma P(Y) = P(2) + P(3) + P(4) + P(7)$$

$$Y \in D$$

Product (π): It is a Greek letter (read as pi) and is used here as a short-hand notation for product. For example,

$$\underset{i=1}{\overset{n}{\pi}} Y_i = Y_1 \times Y_2 \times Y_3 \times .... \times Y_n$$

If we have only three observations

$$Y_1 = 2, Y_2 = 3, Y_3 = 5,$$

then $$\underset{i=1}{\overset{3}{\pi}} Y_i = Y_1 \times Y_2 \times Y_3 = 2 \times 3 \times 5 = 30$$

***n*!**: Read as n factorial and is defined as:

$$n! = n.(n\text{-}1).(n\text{-}2)....(3).\ (2).\ (1)$$

$$\Rightarrow \quad 4! = 4.3.2.1$$

$$= 24$$

and $1! = 1$

Note that $0! = 1$

1.2 Importance Of Statistics In Various Disciplines

Since the information collected in the form of data (observations) from any field will almost always involve some sort of variability (uncertainty), so, this subject has applications in almost all fields of research. The researchers use statistics in the analysis, interpretation and communication of their research findings. Some examples of the questions, which statistics might help to answer with appropriate data are:

i) how much better a yield of wheat do we get if we use a new fertilizer as opposed to a commonly used fertilizer?

ii) are a company's sales figures likely to increase in the next quarter?

iii) what dose of an insecticide be used successfully to monitor an insect population?

iv) what is the likely weather in the coming season?

It is obvious that statistics has its application in almost every field where research is carried out and findings are reported.

When Statistics is applied in Economics, it is called ***Econometrics***, When it is applied in biological sciences, it is called ***Biometry***. Similary, ***Psychometry*** and others. We give a brief account of its application in differnt fields as follows:

1.2.1 Social Sciences

In social sciences, one of the major objectives is to establish relationship that exists between certain variables. This end is achieved through postulating hypothesis and their testing by using different statistical techniques.

Most of the areas of our economy can be studied by econometric models because they help in forecasting and forecasts are important for future planning.

1.2.2 Plant Sciences

The most important aspect of statistics in plant sciences research is its role for efficient planning of experiments and drawing valid conclusions. A technique in statistics known as 'Designs of Experiments' helps introducing new varieties. Optimum plot sizes can be worked out for different crops like wheat, cotton, sugarcane and others under different environmental conditions using statistical techniques.

1.2.3 Physical Sciences

The application of Statistics in physical sciences is widely accepted. The researchers use these methods in the analysis, interpretation and communication of their research findings, linear and nonlinear regression models are used to establish cause and effect relationship between different variables and also these days computers have facilitated the experimentation and it is possible to simulate the process rather than experimentation.

1.2.4 Medical Sciences

The interest may be in the effectiveness of new drugs, effect of environmental factors, heritability, standardization of various records and other related problems. Statistics comes to rescue. It helps to plan the next investigation in order to get trustworthy information from the limited resources. It also helps to analyze the current data and integrate the information with that previously existing.

1.3 Variables

A characteristic that varies from individual to individual in a population is called a variable. The nature abhors constancy, so, the natural phenomena show variability. For example, plant height, number of plants per plot, eye colour (black, blue, green) etc.

Let Y represents the variable and Y_i (read as Y subscript i) represents the ith observation. The variables, $Y_1, Y_2, \ldots, Y_n$ form a set of n observations on the variable Y. For example, we measure the height of 5 wheat plants and observe that the height of first plant is 87 cm, the height of second plant is 90 cm, the height of third plan is 92 cm, the height of fourth plant is 89 cm and the height of fifth plant is 95 cm.
Here, $Y_1 = 87$ cm, and Y_5 is 95 cm.

A variable may be **fixed** or ***mathematical*** when its value can be determined before hand i.e., amount of fertilizer to be applied to a plot, amount of insecticide applied to control insect pests. A variable may be **random** when its value cannot be exactly determined i.e., yield from a plot, the response to an insecticide. Variables are usually of two types:

i) Quantitative variable **ii)** Qualitative variable

1.3.1 Quantitative variable

A quantitative variable is one which is capable of assuming a numerical value. For example, height of plants, weight of grains or number of students in a class.

Quantitative variables can further be placed into two types depending upon the type of measurement possible.

i) Continuous variables **ii)** Discrete variables

A continuous variable is one that can take all possible values in an interval on the number line. For example, atmospheric pressure, plant height, student height and temperature.

A discrete variable is also known as discontinuous variable. It is one that can take only isolated points on the number line. Usually these values are positive integers as these arise from counting. For example, number of students in a class, number of plants per plot, number of insects in a unit area, number of grains per plant.

1.3.2 Qualitative or Categorical variable

A qualitative variable also known as categorical variable is one which is not capable of taking numerical measurements. An observation is made when an individual is allocated to one of several mutually exclusive categories. Observations falling in each class can only be counted. For example, sex (either male or female), general knowledge (poor, moderate, good), colour (blue, green, red etc.).

Example 1.2: A sample of 5 students from a class was selected and each one of them was asked which brand of soap they use. Their responses were as follows:

Lux, Rexona, Lux, Capry, Rexona

Identify the type of variable?

Solution: As we can only categorize these observations into different brands of soap, so, the observations arise from a categorical variable.

Example 1.3: There are 7 sections of class XI in an intermediate college. The number of students in each section is as follows:

41, 45, 47, 37, 35, 45, 42

Identify the type of variable?

Solution: This data set arises from a quantitative variable because the observations are numerical values. Again the variable is discrete because the observations are isolated points on a number line.

1.4 Descriptive And Inferential Statistics

Generally, rough and crude form of the data is obtained from experiments and surveys that needs to be organized and summarized in order to describe its sense. This is where the ***descriptive statistics*** comes in to help us. It provides procedures for:

i) organizing the data collected from the sample,

ii) summarizing the data. It includes graphic representation and calculation of summary values like measures of central value and measures of variability that we call statistic (mean, proportion, variance).

iii) presenting the summaries in an understandable form for the others.

One may be interested to generalize the results of the data. For example, based upon the descriptive statistics, one might be willing to estimate the value of his measure of a central value, had he gathered data about all the subjects possessing the given character rather than a sample. This procedure of inferring about the characteristics of the population based upon the characteristics of its sample is called ***inferential statistics***.

The discipline strengthening inferential statistics is probability theory. So, our generalizations of results always involve some risks. Thus, we always make probabilistic statements and we really don't prove anything. For example, we say that the probability is high that an experimental variable affected the dependent variable.

The whole issue of descriptive and inferential statistics can be described with the help of what we call ***Statistical Problem***.

The distinguishing feature of a statistical problem is that we are trying to say something about a population based on a sample from the population only, taking into account the variability within the samples. Before we do this, we must make some assumptions about the manner in which the data was produced (these are embodied in a statistical model). Based on the model and data, statistical methods are designed which allow us to work back and make statements about the underlying population. The main aspects of the statistical problem are:

i. a clear definition of the population of interest and objectives,

ii. the design of experiment or the sampling procedure,

iii. the collection, organization and analysis of data,

iv. the selection of suitable model and the process of making statements about the population based on sample information.

Figure 1.1 illustrates clearly the main aspects of statistics. Often the underlying population will be clearly defined, in other situations the population may only be hypothetical corresponding to what might have happened in an infinite series of repetition of the experiment concerned.

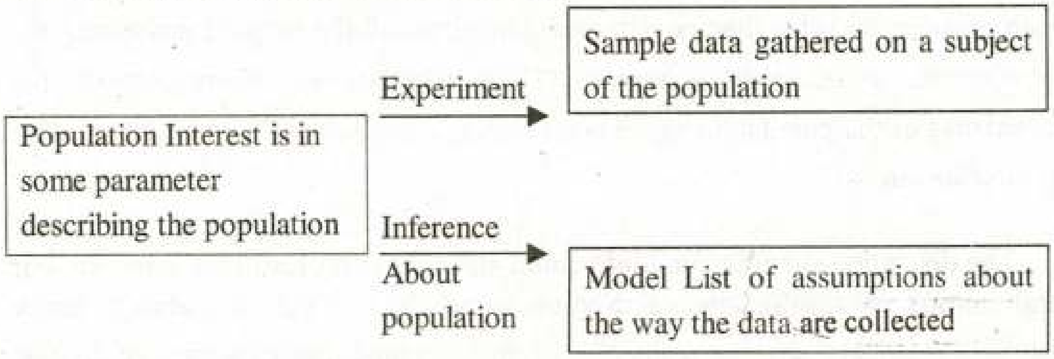

Sample data gathered on a subject of the population
Experiment
Population Interest is in some parameter describing the population
Inference About population
Model List of assumptions about the way the data are collected

Figure 1.1: Statistical problem – Real world situation

1.5 Sources of Data

Keeping in view the objectives of the study, data are collected by individual research workers or by organizations through sample surveys or experiments. The data collected may be

i) primary data   ii) secondary data

By primary data we mean the raw data, which has just been collected from the source and has not gone through any kind of statistical treatment like sorting and tabulation.

By secondary data we mean the data, which has already been collected by someone, that has undergone a statistical treatment like sorting and tabulation etc.

1.5.1 Sources of primary data

The sources of primary data are primary units like basic experimental units, individuals, households and the following methods are usually used to collect data from primary units. The method used depends very much on the nature of the primary unit.

i: Personal Investigation: The researcher conducts the experiment or survey by himself and collects data from it. The data collected is generally accurate and reliable. This is only feasible in case of small scale laboratory, field experiments or pilot surveys and is not practicable for large scale experiments and surveys because it will take too much time.

ii: Through Investigators: The trained investigators are employed to collect the data. In case of surveys, they contact the individuals and fill in the questionnaires after asking the required information. A ***questionnaire*** is an inquiry form having a number of questions designed to obtain information from the respondents. This method is usually employed by most of the organizations. This method gives a reasonably accurate information but it is very costly.

iii: Through questionnaire: The required information is obtained by sending questionnaire to the selected individuals by mail who fill in the questionnaire and return it to the investigator. This method is cheap but non-response rate is very high as most of the respondents don't bother to fill in the questionnaire and send it back.

iv: Through local sources: The local representatives or agents are asked to send requisite information who provide the information based upon their own experience. This method is quick but gives only rough estimates.

v: Through telephone: The information may be obtained by contacting the individuals on telephone. This method is quick and gives accurate information.

vi: Through internet: With the introduction of information technology, the people may be contacted through internet and the individuals may be asked to provide the pertinent information.

It is important to go through the primary data and locate any inconsistent observations before it is given a statistical treatment.

1.5.2 Sources of Secondary Data

The secondary data may be available from the following sources.

i: Government organizations: Federal and Provincial Bureau of Statistics, Crop Reporting Service-Agriculture Department, Census and Registration Organizations etc.

ii: Semi-government organizations i.e., Municipal committees, District Councils, commercial and financial institutions like banks etc.

iii: Teaching and research organizations.

iv: Research journals and newspapers.

v : Internet

Exercise 1

Ans on Page 250

1.1 Define the word statistics and explain its different meanings?

1.2 Define the following terms:

i) Population and sample   ii) Parameter and statistic

1.3 Distinguish between qualitative and quantitative variables?

1.4 i) Write the following using a summation sign with appropriate index?

a) $Y_3 + Y_4 + ... + Y_{15}$ b) $Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2$

c) $(Y_1 - \mu)^2 + (Y_2 - \mu)^2 + (Y_3 - \mu)^2$ d) $bY_{20} + bY_{21} + \dots + bY_{30}$

ii) Expand the following summation and product signs?

a) $\sum_{i=1}^{5} Y_i$ b) $\sum_{i=5}^{8} (Y_i - \mu)$ c) $\sum_{i=1}^{3} Y_i^2$

d) $\sum_{i=2}^{4} (Y_i - 9)$ e) $\underset{i=1}{\overset{3}{\pi}} Y_i$ f) $\underset{i=3}{\overset{6}{\pi}} (aY_i)^2$

g) $\underset{i=2}{\overset{4}{\pi}} (X_i - Y_i)$

1.5 Classify the following as categorical, discrete or continuous variable:

i) Sex of an insect.   ii) Weights of plants.

iii) Major crops of Pakistan.   iv) Level of satisfaction.

v) Teaching standards.   vi) Temperature measured in Fahrenheit.

1.6 Explain in detail the main aspects of a Statistical problem?

1.7 Define Descriptive and Inferential Statistics and differentiate between them.

1.8 Distinguish between primary and secondary data and give different sources from which these are obtained.

1.9 Fill in the blanks:

i) The purpose of the sample is to draw inference about __________.

ii) Proportion is always __________ or equal to one.

iii) The quantity computed from population is called __________.

iv) The quantity computed from sample is called __________.

v) The quantity which does not vary from individual to individual is called __________.

vi) Sum of the random errors equal to __________.

vii) A variable is called __________ when its value cannot be exactly determined.

viii) A variable that takes numerical values is called __________ variable.

ix) The procedure of inferring about the population characteristics using the sample is called __________.

x) First hand collected data is called __________.

1.10 Against each statement write T for true and F for false statement.

i) The value calculated from the population is called parameter.

ii) Statistics deals with single fact.

iii) Statistics can be placed in relation to each other.

iv) Primary data and ungrouped data are same.

v) A collection of observations is called data.

vi) A variable can assume the same value.

vii) A discrete variable can assume finite values between two given limits.

viii) Height measurements of students is quantitative variable.

ix) The word Statistics is at present used in four ways.

x) A small part of the population is called sample.

2 Representation Of Data

2.1 Introduction

Most scientific experiments are conducted in an attempt to answer some specific questions and generally result in the collection of data in the form of batches of numbers, usually referred to as sample. To extract information from the sample, there is need to organize and summarize the collected data. There are many ways of describing a sample. Commonly, we use either, a graph or a small set of numbers which summarize some properties of the sample such as its centre and spread.

2.2 Classification

The term classification is the process of arranging observations into different classes or categories according to some common characteristics.

The data may be classified by one or more characteristics at a time. When data is classified according to one characteristic, it is called one-way classification. When the data is classified by two characteristics at a time, it is called two-way classification. Similarly, the data classified by three characteristics is called three-way classification.

2.3 Tabulation

The process of making tables or arranging data into rows and columns is called tabulation. Tabulation may be simple, double, triple or complex depending upon the number of characteristics involved. Tables are the most common form of documentation used by the scientists.

2.3.1 Construction of tables

Following are the parts of table out of which first four are main part.

i) **Title**: A title is the heading at top of the table. The title should be brief and self explanatory. It describes the contents of the table.

ii) Column Captions and Box head: The headings for different columns are called column captions and this part of column captions is called Box-head. The column captions should be brief, clear and arranged in order of importance.

iii) Row Captions and Stub: The headings for different rows are called row captions and the part of the table containing row captions is called Stub. Row captions should be brief, clear and arranged in order of importance.

iv) Body of the Table and Arrangements of the data: The entries in different cells of column and rows in a table are called body of the table. It is the main part of the table. The data may be arranged qualitatively, quantitatively, chronologically, geographically or alphabetically.

v) Source Note: Source notes are given at the end of the table which indicate the compiling agency, publication, the data and page of the publication.

vi) Spacing and ruling: To enhance the effectiveness of a table, spacing and ruling is used. It is also used to separate certain items in the table. Thick or double lines or single lines are used to separate row captions and column captions. To indicate no entry in a cell of the table, dots (…) or dashes(---) are used. Zeroes are not used in a table for this purpose.

vii) Prefatory Notes and Footnotes: The prefatory note is given after the title of the table and the footnotes are given at the bottom of the table. Both are used to explain the contents of the table. The footnotes are usually indicated by* * *.

2.4 Frequency Distribution

To extract information from a data set, first and important step is to present it in a compact form. A frequency distribution is a compact form of data in a table which displays the categories of observations according to their magnitudes and frequencies such that the similar or identical numerical values are grouped together. The categories arc also known as groups, class intervals or classes. The number of values falling in a particular category is called the frequency of that category. It is usually denoted by *f*.

The relative frequency, denoted by r.f of a category is the proportion of observed frequency to the total frequency and is obtained by dividing observed frequency by the total frequency. The sum of the relative frequencies should be one (1) except for rounding error. The relative frequencies are important for making comparisons between two or more distributions. Otherwise, the different sample sizes of the data sets may distort comparisons.

The frequency distribution may be made for continuous data, discrete data and categorical data.

Following steps are taken into account while making a frequency distribution for continuous data:

i) Calculate range of the data, where

Range = maximum value in the data. – minimum value in the data

ii Decide about the number of classes. The minimum number of classes may be determined by the formula

Number of classes $c = 1 + 3.3 \log(n)$ (2.1)

or $c = \sqrt{n}$ (approximately) (2.2)

where n is the total number of observations in the data.

This gives roughly the number of classes. There are certain other formulae suggested to decide the number of classes. Classes are the groups of data values constituting the frequency table. Usually, the classes of equal width are defined by the numerical limits or boundaries. Each class has a starting point called its lower limit and its end point called its upper limit.

The class limits are the end points of the class intervals both included in class interval. It is convenient to choose the end points of the class interval so that no observation falls on them. This can be obtained by expressing the end points to one more place of decimal than the observations themselves. For this purpose, class limits are usually converted to class boundaries to achieve continuity in the grouped data. This is done by expressing the upper limit of the first class to one more decimal place without changing the width of the class and starting the second class from the same value as is the end of the first class and so on. The upper values in the classes are included in the next class so that the classes do not overlap.

The number of classes are important. Neither we should make too few wide classes in which most of the variation in the data is lost nor we should have too many narrow classes in which the real values in the data are hardly grouped.

iii) Decide about width of the class. It is usually abbreviated by h and is obtained by the following relation:

$$h = \frac{\text{range}}{\text{number of classes}}$$

$$h = \frac{R}{c} \text{ (approximately)} \qquad (2.3)$$

It should be noted that always a convenient near number is chosen and it is not necessary to follow the rules of rounding because we are only grouping the data.

iv) The decision about the starting point of the first class is arbitrary usually, it is started before the minimum value in such a way that the mid point, the average of lower and upper class limits of the first class is properly placed.

v) Now, an observation is taken and a mark of vertical bar is made for a class it belongs. A running tally is kept till the last observation. The tally count 卌 indicates five.

Example 2.1: Student - Height Data

The height (in cms) of 30 students measured at the time of registration is given by

91, 89, 88, 87, 89, 91, 87, 92, 90, 98, 95, 97, 96, 100, 101, 96, 98, 99, 98, 100, 102, 99, 101, 105, 103, 107, 105, 106, 107, 112.

Make a suitable frequency distribution.

Solution: To construct a frequency distribution proceedas to follows:

i) Range = Maximum value minimum value (2.4)

In this data maximum value is 112 and minimum value is 87.

So, Range = 112 87 = 25

ii) Approximate number of classes or class intervals are number of classes c is given by

= 1 + 3.3 log(30)

= 1 + 3.3 (1.4771)

= 5:87443

= 6 (approximately)

iii) Width of the class interval (h) = range / number of classes

$$= \frac{25}{6}$$

= 4.167

= 5 (approximately)

5 is chosen for convenience, one may take 4 if he / she wishes so.

iv) Minimum value is 87, we start the first class from 86 with width of the class as 5, so, our first class is 86-90 with mid point 88, the average of lower and upper class limits i.e., (86+90)/2=88. Similarly, other classes are 91-95, 96-100, ..., 111-115. It is clear that maximum value 112 is included in the last class.

It is convenient to choose the end points of the class interval so that no observation falls on them. This can be obtained by expressing the end points to one more place of decimal than the observations themselves. Therefore, suitable class boundaries for this data would be 85.5 – 90.5, 90.5 – 95.5, . . ., 110.5 – 115.5. In the class boundaries, the upper values in the classes are included in the next class so that the classes are mutually exclusive i.e., 90.5 is the upper value of the first class and is lower value of the second class. In counting this would be included in the second class interval.

The class centres Y_i's are the middles of the classes. The class centres are also known as mid or middle points and are obtained either by averaging class limits or class boundaries i.e Y_i is the middle of the first class

Y_1 = (85.5 + 90.5)/2 = 88

The other mid points are 93,98, ... , 113 respectively.

v) Starting from first observation, all the 30 observations are assigned to the classes they belong. The first observation 87 falls in the first class 86-90, a tally mark is made in the tally column against this class. The second observation 90 belongs to the first class 86–90, a tally mark is made in tally column against this class and so on, the last observation 112 belongs to the last class 111–115. The number of tally marks in the tally column against each class gives the frequency of that class. The frequency distribution is given in Table 2.1

Table 2.1: Tally count and frequency distribution for the example 2.1.

Class limits	Class boundaries	y_i	Tally	Frequency
86-90	85.5-90.5	88	卌 |	6
91-95	90.5-95.5	93	||||	4
96-100	95.5-100.5	98	卌 卌	10
101-105	100.5-105.5	103	卌 |	6
106-110	105.5-110.5	108	|||	3
111-115	110.5-115.5	113	|	1
			Total	30

It is clear from the frequency table that 6 students have height between 85.5 and 90.5 cms, 4 students have height between 90.5 and 95.5 cms and so on and I student has height between 110.5 and 115.5 cms.

The relative frequency for a class can be computed by dividing its frequency by the total frequency. The frequency distribution with relative frequencies is given in Table 2.2.

Table 2.2: Frequency distribution with relative frequencies

Class boundaries	f	$r.f$
85.5-90.5	6	6/30 = 0.200
90.5-95.5	4	4/30 = 0.133
95.5-100.5	10	10/30 = 0.333
100.5-105.5	6	6/30 = 0.200
105.5-110.5	3	3/30 = 0.100
110.5-115.5	1	1/30 = 0.033
Total	30	1.000

It should be noted that the sum of the relative frequencies is one except for rounding error.

For discrete data: In case of discrete data, each observation is a whole number. So, while making a frequency distribution, the possible values are written in a column and a tally count of each value is made for the data. The number of tally count for each value is its frequency. The corresponding relative frequency is obtained by dividing each frequency by the total number of observations. The sum of the relative frequencies should be 1 except for rounding error.

For categorical data: In case of categorical data, the categories are placed in a column and a tally count is made for each category going through the data set which gives the frequency of each category.

Example 2.2: The observations about the number of rotten potatoes from twenty equal sized samples taken from a store are available as follows:

1, 2, 4, 3, 0, 1, 2, 3, 1, 1, 0, 2, 1, 0, 2, 3, 0, 0, 1, 3

Make a frequency table

Solution: The tally count and frequency table is made by going through each observation of the data and for each observation making a mark, vertical bar | against the appropriate value of the variable. In this data, the values of the variable vary from 0 to 4. These are written in a column and a tally count is kept going through the whole data. The resulting frequency distribution is given in Table 2.3.

Table 2.3: Tally count and frequency distribution for the example 2.2.

Number of rotten potatoes	**Tally**	***f***	**r.f**
0	卌	5	5/20 = 0.25
1	卌 |	6	6/20 = 0.30
2	||||	4	4/50 = 0.20
3	||||	4	4/20 = 0.20
4	|	1	1/20 = 0.05
Total		$\Sigma f = 20$	1.00

If the range of observations in the data is large, the same method is adopted as has been explained for the continuous data.

Open-end classes

In connection with the frequency tables, the term **open end classes** is sometimes used. It means that in a frequency table, either the lower limit of the Ist class or the upper limit of the last class is not a fixed number. It may happen that both of these are not fixed numbers. The frequency tables with open end classes are formed in some practical situations. The frequency table about the age of people in a certain locality is given in the adjacent table:

Age group	Frequency
Below 5	20
5 –14	37
15 – 24	67
25 – 34	90
35 – 44	87
45 – 54	60
55 – 64	55
65 – 74	45
75 and above	20

2.5 Cumulative Frequency Distribution

A cumulative frequency distribution is a table that displays class intervals and the corresponding cumulative frequencies. The cumulative frequency is denoted by c.f and for a class interval it is obtained by adding the frequencies of all the preceding classes including that class. It indicates the total number of values less than or equal to the upper limit of that class.

The relative frequencies, cumulative frequencies and cumulative relative frequencies for data for the example 2.1 are given in Table 2.4.

Table 2.4: Cumulative distribution for the example 2.1.

Class boundaries	*f*	*r.f*	*c.f*	c.r.f
85.5-90.5	6	6/30	6	6/30 = 0.200
90.5-95.5	4	4/30	6+4 =10	10/30 = 0.333
95.5-100.5	10	10/30	10+10=20	20/30 = 0.667
100.5-105.5	6	6/30	20+6 =26	26/30 = 0.867
105.5-110.5	3	3/30	26+3 =29	29/30 = 0.967
110.5-115.5	1	1/30	29+1 =30	30/30 = 1.000

As the cumulative frequency of a class indicates the total number of values less than or equal to the upper limit of that class, so, the cumulative frequency of 20 for a class 95.5 – 100.5 means that 20 values are less than 100.5 and similarly, the cumulative frequency of the last class 110.5 – 115.5 is 30 indicating that 30 values are less than 115.5.

If we want to compare two or more distributions, we compute relative cumulative frequencies or percentage cumulative frequencies because these would be comparable. Otherwise, the differences in sample sizes will distort comparisons.

The cumulative relative frequencies which are the proportions of the cumulative frequency, denoted by *c.r.f* are obtained by dividing the cumulative frequency by the total frequency. The *c.r.f* of a class can also be obtained by adding the relative frequencies of the preceding classes including that class. As cumulative relative frequencies are proportions, the multiplication by 100 gives corresponding percentage cumulative frequencies.

The relative cumulative frequencies are obtained by dividing the cumulative frequency by the total frequency i.e., for the first class interval it is 6/30 = 0.2, for the second class interval it is 10/30 = 0.33 and so on. The percentage cumulative frequency for each class can be obtained by multiplying its cumulative relative frequency by 100. The percentage cumulative frequency for 0.200 is (0.200) (100) = 20. The percentage cumulative frequency for 0.0333 is (0.333) (100) = 33.3 and so on, the percentage cumulative frequency for 1.000 is (1.000) (100) = 100.

2.5.1 Cumulative frequency distribution for discrete data

The cumulative frequency distribution for the discrete data is obtained in the same way as for the continuous data i.e., the cumulative frequency of a class is obtained simply by adding the preceding frequencies including the frequency for that class. The relative frequencies and the cumulative frequencies for the data of example 2.2. are given below in Table 2.5.

Table 2.5: Cumulative frequency distribution of the example 2.2.

Number of rotten potatoes	*f*	*r.f*	*c.f*	c.r.f
0	5	5/20	5	5/20
1	6	6/20	11	11/20
2	4	4/20	15	15/20
3	4	4/20	19	19/20
4	1	1/20	20	20/20
Total	20	1.00		

Example 2.4: Find out the relative frequency distribution for the following data. Where *x* denotes the number of hours worked in a day by a person in a locality of 265 people.

x	6	7	8	9	10	11	12	13
f	24	66	80	48	28	14	4	1

Solution:

x	frequency	Relative frequency
6	24	24/265 = 0.09
7	66	66/265 = 0.25
8	80	80/265 = 0.30
9	48	48/265 = 0.18
10	28	28/265 = 0.11
11	14	14/265 = 0.05
12	4	4/265 = 0.02
13	1	1/265 = 0.00
Total	265	1.00

Example 2.5: Find out the relative cumulative frequency distribution from the following data of example 2.4.

x	6	7	8	9	10	11	12	13
f	24	66	80	48	28	14	4	1

Solution:

x	f	c.f	c.r.f.
6	24	24	24/265 = 0.09
7	66	24 + 66 = 90	90/265 = 0.25
8	80	90 + 80 = 90	170/265 = 0.30
9	48	170 + 48 = 90	218/265 = 0.18
10	28	218 + 28 = 90	246/265 = 0.11
11	14	246 + 14 = 90	260/265 = 0.05
12	4	260 + 4 = 90	264/265 = 0.02
13	1	264 + 1 = 90	265/265 = 0.00
Total	265	---------	---------

2.6 Graphic representation of Data

There are reasons for drawing graphs. The most compelling being that one simple graph says more than twenty pages of prose. Many graphs just represent a summary of data that has been collected to support a particular theory. It is usually suggested that the graphic representation of the data should be looked at before proceeding for the format statistical analysis.

Common uses of graphs

i) Graphs are useful for checking assumptions made about the data, i.e., the probability distribution assumed.

ii) The graphs provide a useful subjective impression as to what the results of the format analysis should be. This serves as a check on calculations and statistical methodology used. Always believe your common sense before arithmetic calculations because for some problems calculations will be obvious from a graph.

iii) Graphs often suggest the form of a statistical analysis to be carried out, particularly, the graph of model fitted to the data.

iv) Graphs give a visual representation of the data or the results of statistical analysis to the reader which are usually easily understandable and more attractive.

v) Some graphs are useful for checking the variability in the observations and outliers can be easily detected.

Outliers are the data values which are highly inconsistent with the main body of the data. These may arise because of mistakes in copying, coding or may be some values that are different from the rest of the data just their own.

Important points for drawing graphs

There are a number of points worth keeping in mind when drawing graphs. The most important of these are:

i) clearly label axis with the names of the variables and units of measurement.

ii) keep the units along each axis uniform regardless of the scales chosen for axis.

iii) keep the diagram simple. Avoid any unnecessary details.

iv) a clear and concise title should be chosen to make the graph meaningful

v) if the data on different graphs are to be measured, always use identical scales.

vi) in the scatter plots, don't join up the dots. This makes it likely that you will see apparent patterns in any random scatter of points.

The general approach, which should be used to analyze the data, is as follows:

i) construct an appropriate diagram and summary of the data and come to an initial impression concerning the question posed. This is known as exploratory data analysis.

ii) follow this up with an appropriate formal analysis of data.

iii) compare the results of the formal analysis with your initial impression, and worry if they differ greatly.

The methods described here are appropriate for data on a **single variable**. Usually, the data will be measured on a continuous scale or at least if this is not the case, then the set of possible values will be reasonably large. The types of graphs commonly used are given below:

2.6.1 Simple Bar Diagram

To get an impression of the distribution of a discrete or categorical data set, it is usual to represent it by a bar diagram. To construct a bar diagram, the values of the variable or categories are taken along *x*-axis and a bar with height equal to its frequency is drawn on each category.

The first step is to make a tally count of the data to help us to make a frequency distribution. The procedure is explained on the example 2.2. The frequency distribution is given in table 2.5 (a).

Table 2.5 (a): Frequency distribution for the data 2.2.

Number of rotten potatoes	Tally	Frequency
0	𝍸	5
1	𝍸 𝍷	6
2	𝍷𝍷𝍷𝍷	4
3	𝍷𝍷𝍷𝍷	4
4	𝍷	1
	Total	20

To construct a bar diagram, the number of rotten potatoes are taken along x-axis. The rotten potatoes vary from 0 to 4, so we mark the x-axis with 0,1,2,3 and 4. The value 0 has frequency 5, so a bar of height 5 is drawn along y-axis at point 0 on x-axis. Similarly a bar of height 6 is drawn along y-axis at point 1 on x-axis; a bar of height 4 is drawn along y-axis at point 2 and 3 on x-axis and finally a bar of height 1 is drawn on point 4. It is shown in figure 2.1.

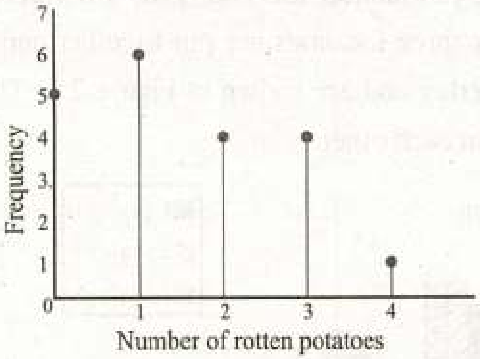


Figure 2.1: Bar diagram of rotten potatoes.

The gaps between the bars in the bar chart emphasize the gaps between the values that the discrete variable can take.

2.6.2 Multiple Bar Diagram

It is an extension of the simple bar diagram and is used to represent two or more related sets of data in the form of groups of simple bars. Its main purpose is to compare same characteristics of a variable.

Example 2.6: Following data is about the production of wheat in different localities of the Punjab for years 1987 to 1989.

Production in Kg. (thousands)			
Year	**1987**	**1988**	**1989**
Locality I	500	600	200
Locality II	600	700	400
Locality III	800	700	500

Draw an appropriate diagram for this data?

Solution: The appropriate diagram seems to be a multiple bar diagram because 3 bars, one of each locality, for each year will make the comparison between the production of three localities overtime.

To draw multiple bar diagram, the years are taken along x-axis and for each year three bars are drawn along y-axis, one for each locality to indicate the production. The bars showing production for year 1987 for three localities are put together, the bars for 1988 for three localities are put together and the bars for 1989 for three localities are put together and are shown in Figure 2.2. The bars are shaded individually to differentiate from each other.

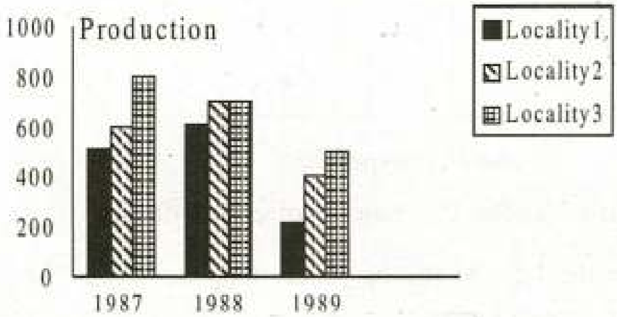


Figure 2.2: Multiple bar diagram of different localities.

2.6.3 Sub-divided Bar Diagram (Component Bar Diagram)

There are certain situations where the simple bar diagram represents the totals and it is possible to divide it further into different segments. For example, if the simple bar denotes the total population of insects caught in a field then it is possible to sub-divide it into male and female proportions.

Example 2.7: There were 500 people of blood group A (kind 1). 300 of blood group B (kind 2) and 400 of blood group O (kind 3). After classification, it was observed that for kind 1 there were 200 females, for kind 2 there were 100 females and for kind 3 there were 200 females.

Draw an appropriate diagram for this data?

Solution: The sub-divided bar diagram is useful in this situation to represent the number of males and females in each category. First construct simple bars and then divide it according to the number of males and females in each blood group category.

The simple bar and sub-divided bar diagrams are shown in Figure 2.3(a) and 2.3 (b) respectively.

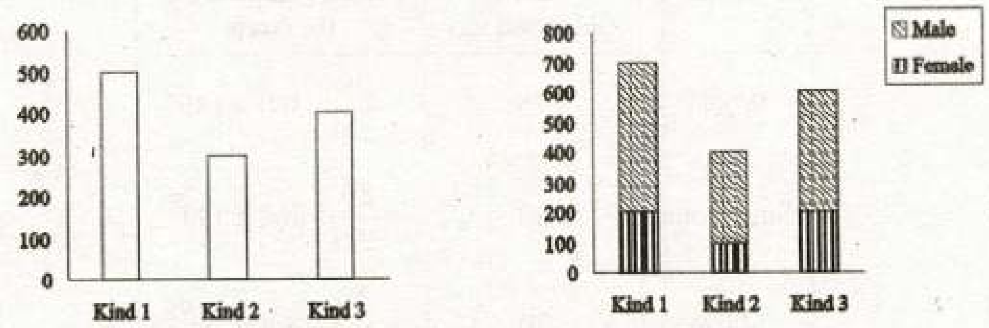


Figure 2.3(a)

Figure 2.3(b)

Figure 2.3: Simple and sub-divided bar diagrams

2.6.4 Pie Diagram

The Pie chart or Pie diagram is a division of a circular region into different sectors. It is constructed by dividing the total angle of a circle of 360 degrees into different components. The angle Q for each sector is obtained by the relation:

$$Q = \frac{\text{Component Part}}{\text{Total}} \times 360$$

Each sector is shaded with different colours or marks so that they look separate from each other.

It is a useful way of displaying the data where division of a whole into component parts needs to be presented. It can also be used to compare such divisions at different times.

Example 2.8: The data are available regarding total production of urea fertilizer and its use on different crops. Total production of urea is 200 thousand (kg) and its consumption for different crops wheat, sugarcane, maize and lentils is 75,80, 30 and 15 thousands (kg) respectively. Make an appropriate diagram to represent these data?

Solution: The appropriate diagram seems to be a pie chart because we have to present a whole into 4 component parts. To construct a pie chart, we calculate the proportionate arc of circle, i.e.,

$$\frac{75}{200} \times 360 = 135,\ \frac{80}{200} \times 360 = 144,\ \frac{30}{200} \times 360 = 54,\ \frac{15}{200} \times 360 = 27$$

Proportionate are of a circle (in degrees) for different crops are in Table 2.6.

Table 2.6: Proportionate arc of a circle for crops.

Crops	**Fertilizer (thousand kg)**	**Proportionate arc of the circle**
Wheat	75	$\frac{75}{200}\times 360^{\circ}=135^{\circ}$
Sugarcane	80	$\frac{80}{200}\times 360^{\circ}=144^{\circ}$
Maize	30	$\frac{30}{200}\times 360^{\circ}=54^{\circ}$
Lentils	15	$\frac{15}{200}\times 360^{\circ}=27^{\circ}$
Total	200	360°

Draw a circle of an appropriate radius, make the angles clockwise or anti-clockwise with the help of protractor or any other device i.e., for wheat make an angle of 135 degrees, for sugarcane an angle of 144 degrees, for maize an angle of 54 degrees and for lentils an angle of 27 degrees and hence circular region is divided into 4 sectors. Shade each sector with different colours or marks so that they look separate from each other. The pie diagram is given in figure 2.4.

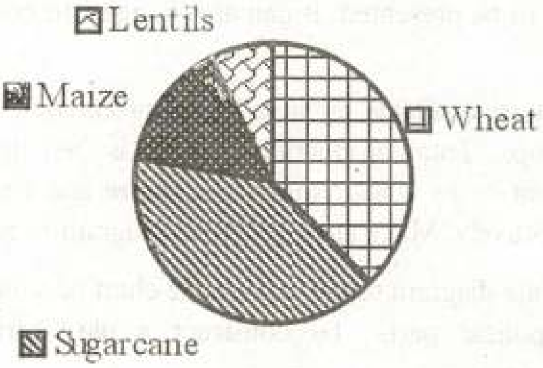


Figure 2.4

2.6.5 Histogram

A histogram is a useful graphic representation of data to get a visual impression about its distribution. It is constructed from the grouped data by taking class boundaries along x-axis and the corresponding frequencies along y-axis. If the data are in the ungrouped form, then first step is to arrange the data in the form of the grouped frequency distribution before making a histogram. The histogram may be constructed for the following two types of qualitative data.

i) Continuous grouped data ii) Discrete grouped data

Histogram: For Continuous grouped data

For the continuous grouped data the frequency distribution may be with equal class width or with unequal class width depending upon the nature of the data. To draw a histogram from the continuous grouped frequency distribution, the following steps are taken. The first two steps are common for equal / unequal class width but the third step is different.

i) Mark class boundaries of the classes along x-axis.

ii) Mark frequencies along y-axis.

iii) Draw a rectangle for each class such that the height of each rectangle is proportional to the frequency corresponding to that class. This is the case when classes are of equal width as they often are.

iv) If the classes are of unequal width, then the area instead of height of each rectangle is proportional to the frequency corresponding to that class and the height of each rectangle is obtained by dividing the frequency of the class by the width of that class.

Histogram: For Equal class Interval (data of table 2.1)

To construct a histogram, take following steps:

i) Mark the class boundaries 85.5 – 90.5, 90.5 – 95.5, ….., 110.5 – 115.5 on x-axis.

ii) Maximum frequency is 10, so label y-axis from 0 to 10.

iii) The frequency of the Ist class is 6, so the rectangle is raised uptill 6, the rectangle of the second class is raised uptill 4 and so on, the last rectangle is raised to height 1. The histogram is shown in Figure 2.5.

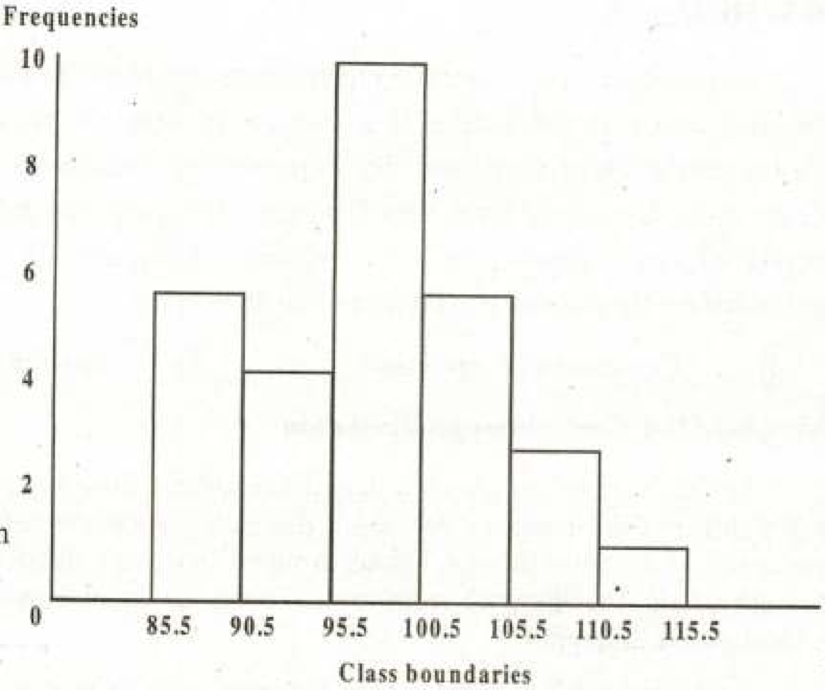


Figure 2.5: Histogram of student height data.

It may be noted that the area under a histogram can be calculated by adding up the areas of all the rectangles that constitute the histogram. The area of one rectangle is obtained by the multiplication of width of the class by the corresponding frequency i.e.,

Area of a single rectangle = width of the class × frequency of the class.

The area of above histogram is

$5(6) + 5(4) + 5(10) + 5(6) + 5(3) + 5(1) = 150$

Histogram: For Unequal Class Intervals

The principal reason for making histogram with equal class intervals is that the frequencies from class to class are directly comparable. However, there may be situations where unequal class intervals are appropriate. Firstly, in highly skewed distribution and secondly, the grouping of similar cases. In such situations while constructing a histogram, the width of the classes should be taken into account because the area of each rectangle is proportional to the frequency. This can be achieved by adjusting the heights of the rectangles. The height of each rectangle is obtained by dividing the frequency of the class by the width of that class.

Example 2.9: The frequency distribution of ages (in years) of 51 members of a locality is available adjacent table. Draw a histogram for this data?

Classes	Frequency
2 – 4	5
4 – 8	10
8 – 12	12
12 – 16	14
16 – 22	6
22 – 30	4

Solution: A look at this data, indicates that the width of the class intervals is not equal as first class has width 2; second, third, fourth classes have width 4, fifth has 6 and the last class has width 8 so, there is need to adjust the heights of the rectangles i.e., for the first class we have 2 as width of class and 5 as frequency, so height of the first class is 5/2=2.5. Similarly, for the others 10/4=2.5, 12/4=3, 14/4=3.5, 6/6=1.0, 4/8=0.5. These heights are also called adjusted frequencies. The width of the class and corresponding height of rectangles are in table 2.7

Table 2.7: Frequency distribution of the example 2.6 with adjusted heights.

Classes	Frequency	Width of class	Height of rectangle (adjusted frequency)
2 – 4	5	2	5/2 = 2.5
4 – 8	10	4	10/4 = 2.5
8 – 12	12	4	12/4 = 3.0
12 – 16	14	4	14/4 = 3.5
16 – 22	6	6	6/6 = 1.0
22 – 30	4	8	4/8 = 0.5

Taking class boundaries along *x*-axis and corresponding adjusted frequencies along *y*-axis, rectangles are drawn and the histogram is given in Figure 2.6.

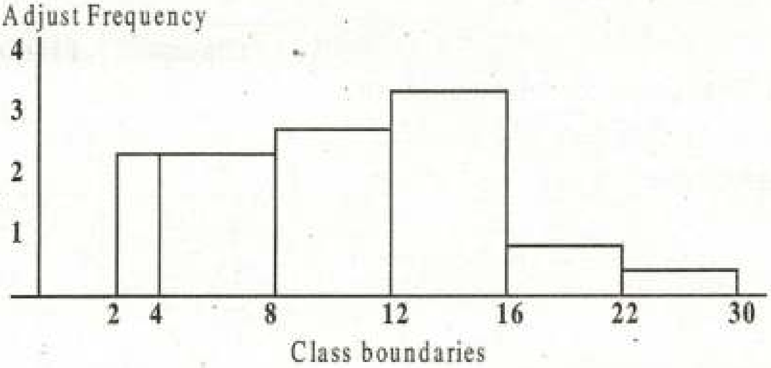


Figure 2.6: Histogram for unequal class intervals.

Histogram: discrete data

It should be noted that bar graphs are usually drawn for discrete and categorical data but there are some situations where there is need to make approximations, the histogram may be constructed.

To construct a histogram for discrete grouped data, following steps are taken:

i) mark possible values along *x*-axis.

ii) mark frequencies along *y*-axis.

iii) draw a rectangle centered on each value with equal width on each side possibly 0.5 to either side of the value.

The procedure is explained for the example 2.2.

The rotten potatoes vary from 0 to 4 so, *x*-axis is marked 0, 1, 2, 3, 4. The maximum frequency is 6 so, *y*-axis is marked from 0 to 6. A rectangle is drawn centered on each value whose height is equal to the corresponding frequency. The resulting diagram for the data is given in Figure 2.7.

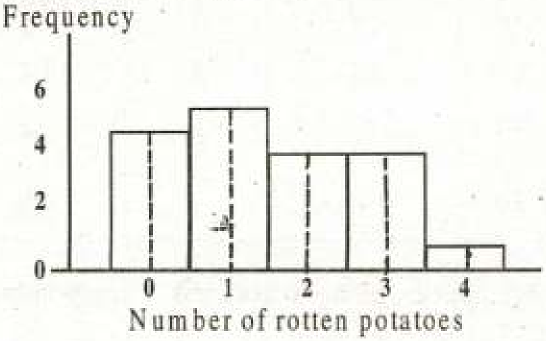


Figure 2.7: Histogram for rotten potatoes.

Advantages: The advantages of the histogram as compared to the unprocessed data are:

i) it gives range of the data.

ii) it gives location of the data.

iii) it gives clue about the skewness of the data.

iv) it gives information about the out of control situation.

2.6.6 Frequency polygon and frequency curve

A frequency polygon is a closed geometric figure used to display a frequency distribution graphically. Following steps are taken to make a frequency polygon from a frequency distribution.

i) Calculate mid values of the class boundaries.

ii) Mark these mid values along x-axis.

iii) Mark the frequencies along y-axis.

iv) Mark corresponding frequencies against each mid point, join them and extend it to x-axis.

It can also be obtained by joining the upper mid points of the rectangles of a histogram and extending ends to the x-axis. The distance from the x-axis to the plotted point corresponds to the frequency of the class. The frequency polygon smoothed is called frequency curve, which is useful to have a visual impression about the data i.e., it may help to know about the symmetry or skewness of the data. If we are interested to compare two distributions number of observations less than this is zero. For the grouped data of the example 2.1, it is shown in figure 2.8. It is clear that cumulative frequency polygon is an increasing function which starts from the lower class boundary of the first class at zero height and ends at the upper boundary of the last class with height equal to total frequency.

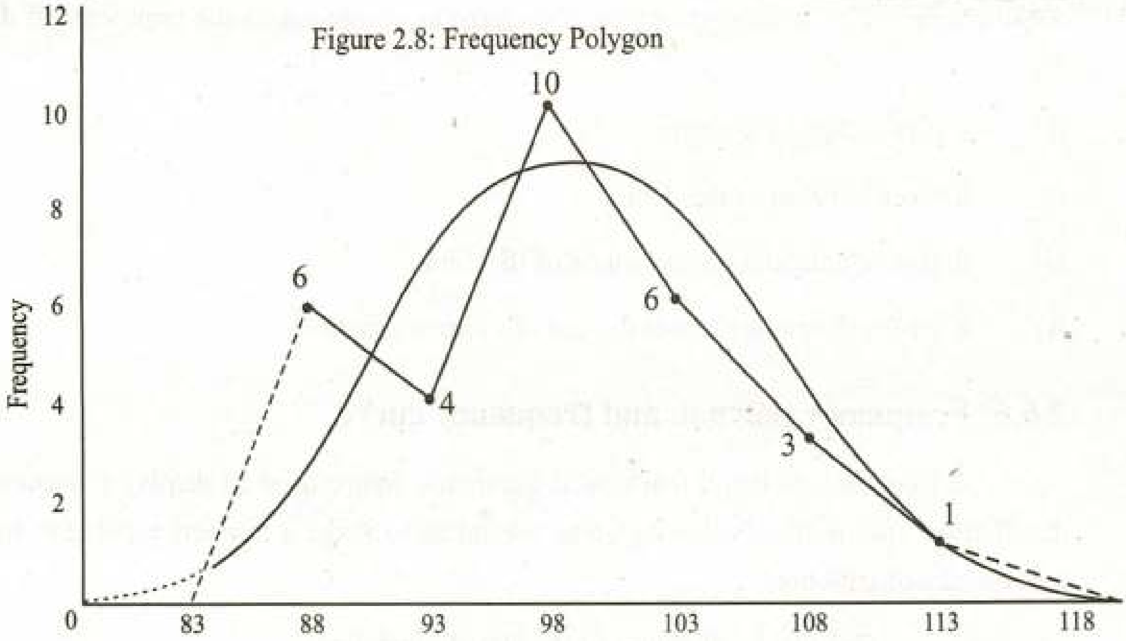


Figure 2.8: Frequency Polygon

This graph may be drawn using upper class boundaries and cumultative relative frequencies in which case it is called cumulative frequency function or polygon and can be used to locate certain values. It can be used to locate the quartiles or percentiles of the data. The figure 2.9 indicates the observation corresponding to the c.r.f. of 0.25

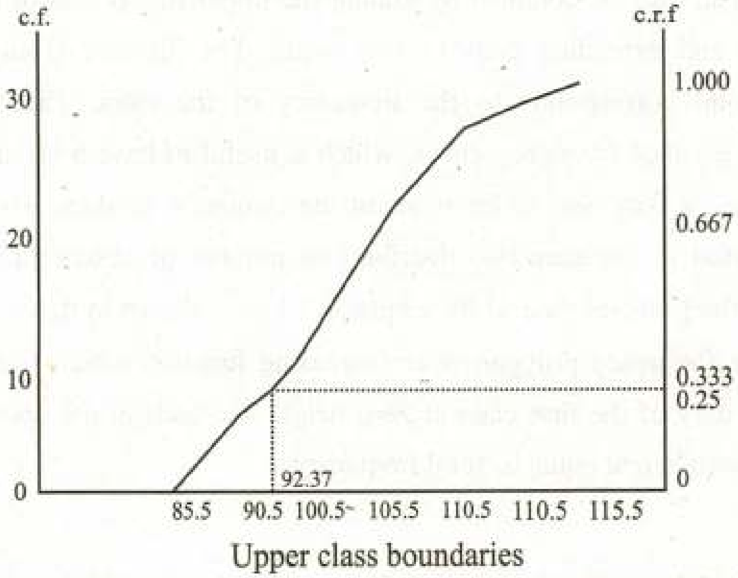


Figure 2.9: Cumulative frequency polygon of students height data.

We multiply the cumulative relative frequencies by hundred to get corresponding percentage cumulative frequencies. The c.r.f. polygon becomes the percentages along y-axis instead of c.r.f. The graph on the right hand side of figure 2.9 becomes percentage cumulative frequency polygon if we replace 0.200 by 20, the percentage cumulative frequency for 0.2000 as (0.200) (100) = 20: 0.333 by 33.3, the percentage cumulative frequency for 0.333 as (0.333) (100) = 33.3 and so on 1.00 by 100, the percentage cumulative frequency for 1.

Consider the following steps to draw a cumulative frequency polygon for discrete variable.

i) Choose horizontal axis on a graph paper and mark the data points from the smallest to the largest.

ii) Mark the vertical axis from zero to total frequency.

iii) Make a vertical jump of height equal to its frequency at the first point. Move horizontally from the top of this point until you are exactly above the second data point and make a jump equal to its frequency at the second point. Repeat this for all the data values. The cumulative frequency polygon for data of the example 2.2 is shown in Figure 2.10.

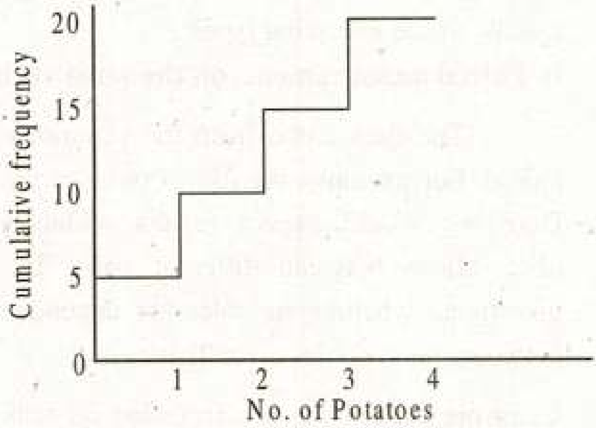


Figure 2.10: Cumulative frequency polygon.

It is clear from Figure 2.10 that there is a jump at each data value whose height is equal to its frequency. The cumulative frequency polygon is flat and horizontal between the data values. It starts from a height of zero on the left and goes to a height of total frequency at the right, being increasing function between smallest and largest values of the data set.

This graph may be drawn by taking data values along x-axis and the corresponding cumulative relative frequencies or percentage cumulative frequencies along y-axis, the relative frequencies being as heights at each point in such case it is called discrete cumulative frequency function or percentage cumulative frequency function or polygon.

2.6.7 Scatter Plots

Very often, many variables are measured on each individual. For example, we may consider two variables, height and weight of each individual in a class. Now, the resulting data set consists of n pairs of observation such as (x_i, y_i), $i = 1, 2, \ldots, n$; where each x_i denotes height and each y_i denotes weight. This is called a ***bivariate*** data set. A plot of two variables useful in such situations is scatter plot. It is obtained by taking one variable on x-axis and the other on y-axis. Each pair of values (x_i, y_i), $i = 1, 2, \ldots, n$; in the data set will contribute as a point in this bivariate plot and we usually put a cross (×) or dot (.) at the intersection of values.

A scatter plot is the best way of studying bivariate problems. The bivariate data are usually of the following types:

i) Paired measurements on the same variable

The data come from the situations where experimental units are deliberately paired. For example, the use of twins in the biological and psychological experiments. Here we would expect results within a pair of twins to be more alike than observations between different pairs. In such situations, the main interest is to investigate whether variables are dependent and if so what form of the relationship between the variables actually is.

Example 2.10: Data are recorded on milk yield of cows in the morning and in the evening.

Morning values:	4.5, 6.0, 5.5, 3.5, 4.5, 6.5, 7.0, 5.0, 4.5, 6.5
Evening values:	5.5, 6.5, 6.0, 5.5, 7.0, 5.5, 8.0, 6.0, 8.5, 7.0

The interest is whether the characteristic measured varies in any systematic pattern over the day.

Solution: As both the measurements are on same variable, the interest is therefore, not just in relationship between morning and evening measurements but also in comparing them. The line of equality is a useful visual aid for this type of data. For the scatter plot we take the morning values along x-axis and evening values along y-axis.

At the intersection of 4.5 and 5.5, we put a dot (.), similarly for 6.0, 6.5 and so on. The scatter plot is shown in Figure 2.11.

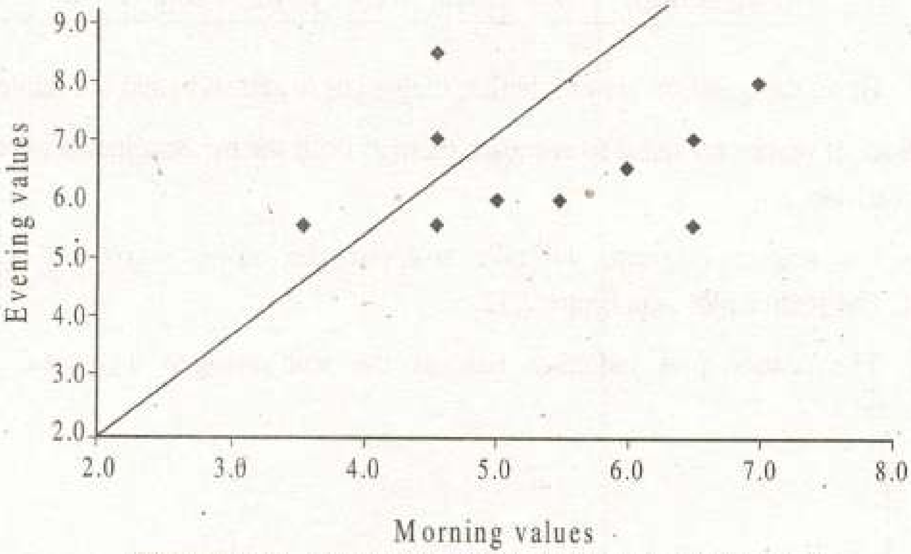


Figure: 2.11 Scatter plot of morning and evening values

Line of equality is at an angle of 45° as indicated in Figure 2.11 alongwith scatter plot. It is clear that most of the points are above the line of equality and so the milk yield in the morning and evening is not the same and more milk is obtained in the evening as compared with the morning.

With this data we can also look at the differences between morning and evening values and treat this as a one sample problem. However, we would no longer be able to see if the change was related to the initial value.

ii) Two related measurements

The pair of values may come from two variables which are related to each other. For example, samples of soil nitrogen and yield of a variety are taken in each of seven randomly selected agriculture locations. In such situations, it doesn't make any sense to compare them as both the measurements are not on the same variable.

Scatter plots are also drawn to examine the relationship between two related measurements.

Example 2.11: Samples of soil nitrogen and yield are taken in each of seven randomly selected agriculture locations. The soil nitrogen and yield are:

Location	1	2	3	4	5	6	7
Soil nitrogen	8.5	7.0	6.5	6.0	7.0	8.0	7.5
yield (kg)	9.0	7.5	6.0	4.5	6.0	7.0	6.0

Here, the question arises whether the soil nitrogen and yield are related?

Solution: It makes no sense to compare them as both the measurements are not on the same variable.

For scatter diagram, we take soil nitrogen along *x*-axis and yield along *y*-axis. The scatter plot is in figure 2.12.

The scatter plot indicates that as the soil nitrogen increases, yield also increases.

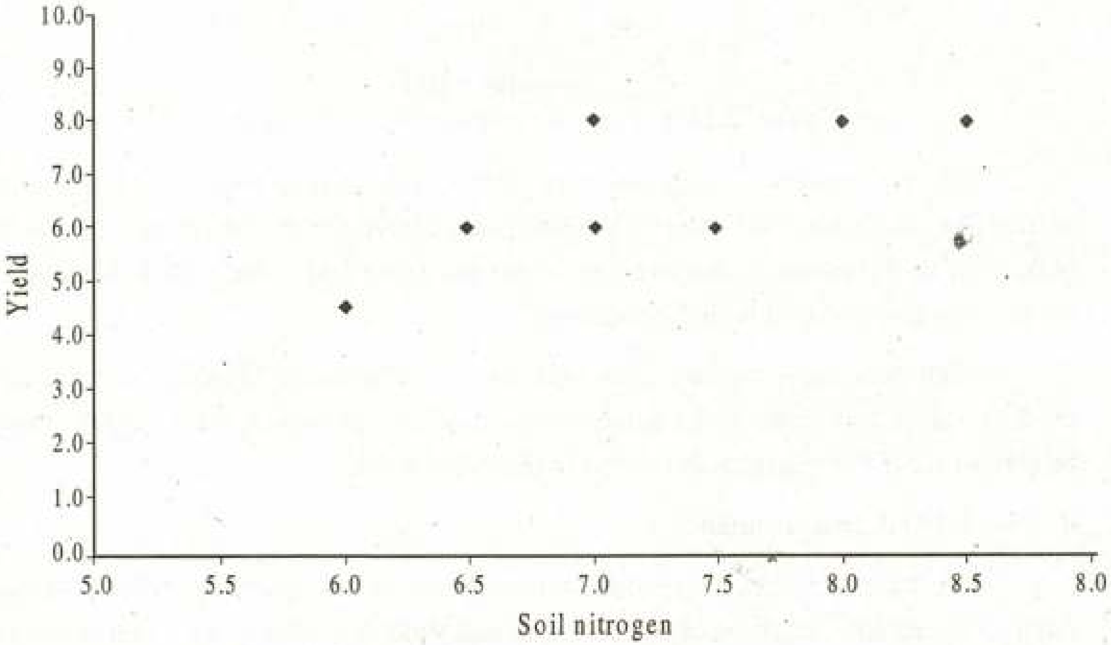


Figure 2.12: Scatter plot of soil nitrogen and yield.

2.7 Bivariate Frequency Distribution

The constructed frequency distribution, considering two variables at a time is called bivariate frequency distribution. The pairs of observations are taken into account while constructing a bivariate frequency distribution. The procedure to construct a bivariate frequency is the same except n pair of values (x_i, y_i), $i = 1, 2, \ldots, n$ are allocated at the intersection of classes of both the variables.

Example 2.12: A sample of 25 students was taken and their heights in feet (x) and weights in kilograms (y) were measured. The pairs below are (height, weight).

(5.5, 60), (5.0, 55), (4.3, 46), (5.3, 67), (4.9, 48), (5.9, 69),
(5.4, 67), (4.8, 55), (5.3, 57), (5.8, 67), (5.3, 57), (5.7, 65),
(5.8, 63), (5.9, 65), (4.8, 49), (5.3, 55), (5.1, 60), (5.7, 65),
(4.7, 50), (4.5, 50), (5.3, 60), (4.6, 53), (5.4, 62), (5.2, 59),
(4.7, 55).

Solution: The minimum and maximum height is 4.3 and 5.9 feet respectively. The minimum and maximum weight is 46 and 69 kilogram respectively.

For height, we take 4 classes with an interval of 0.5. So, the class limits for height are 4.0-4.4, 4.5-4.9, 5.0-5.4, 5.5-5.9. The corresponding class boundaries are 3.95-4.45, 4.45-4.95, 4.95-5.45, 5.45-5.95. As usual, the class boundaries have been obtained by averaging the upper class limit of one class and the lower class limit of the next class i.e., the first class boundary is

$$(4.4+4.5)/2 = 4.45 \text{ and so on.}$$

For weight, we take 5 classes with an interval of 5.0. So, the class limits are 44-49, 50-54, 55-59, 60-64, 65-69. The corresponding class boundaries are 44.5-49.5, 49.5-54.5, 54.5-59.5, 59.5-64.5, 64.5-69.5.

Starting from first pair, all the 25 pairs are assigned to the classes they belong. The first pair (5.5,60) falls in the class with height (5.45-5.95) and weight (59.5-64.5), a tally mark is made in the table against their interaction. The second pair (5.0, 55) belongs to the class with height (4.95-5.45) and weight (54.5-59.5), a tally mark is made in table against their interaction and so on, the last pair (4.7, 55) belongs to the class with the height (4.45-4.95) and weight (54.5-59.5). The number of tally marks in each cell gives the frequency of the class with certain height and weight boundaries. The bivariate frequency distribution is given in Table 2.7. (a) and 2.7 (b).

Table 2.8 (a): Bivariate frequency table

	Height			
Weight	3.95-4.45	4.45-4.95	4.95-5.45	5.45-5.95
44.5-49.5	𝍷	𝍷𝍷		
49.5-54.5		𝍷𝍷𝍷		
54.5-59.5		𝍷𝍷	𝍸	
59.5-64.5			𝍷𝍷𝍷	𝍷𝍷
64.5-69.5			𝍷𝍷	𝍸

Table 2.8 (a) is the bivariate frequency distribution (or table) after making a tally count.

Table 2.8 (a)

	Height			
Weight	3.95-4.45	4.45-4.95	4.95-5.45	5.45-5.95
44.5-49.5	1	2		
49.5-54.5		3		
54.5-59.5		2	5	
59.5-64.5			3	2
64.5-69.5			2	5

It is clear from the bivariate frequency table that there is one individual with weight between 44.5-49.5 kg and height between 3.95-4.45 feet. There are two individuals with weight between 44.5-49.5 kg and height between 4.45-4.95 feet and so on, there are 5 individuals with weight between 64.5-69.5 kg and height between 5.45-5.95 feet.

Exercise 2

Ans on Page 250

2.1 What are different methods of representation of statistical data.

2.2 Define the Histogram, the frequency polygon and the frequency curve.

2.3 What do you understand by classification and tabulation? Discuss their importance in a statistical analysis.

2.4 Distinguish between one-way and two-way tables. Illustrate your answers with examples. Also explain the following:

i) Classification according to attributes.

ii) Class limits.

iii) Length of class interval.

iv) Class frequency.

2.5 The following table gives the details of monthly budgets of two families. Represent these figures through a suitable diagram.

Items	Family *A*	Family *B*
Food	Rs. 600	Rs. 800
Clothing	Rs. 100	Rs. 100
House rent	Rs. 400	Rs. 500
Fuel and lighting	Rs. 100	Rs. 100
Miscellaneous	Rs. 300	Rs. 500
Total	**Rs. 1500**	**Rs. 2000**

2.6 Represent the following data through pie diagram.

Items of Expenditure	Amount
Food	4000
Clothing	1000
House Rent	2500
Education	1000
Fuel and light	600
Miscellaneous	2000

2.7 Define frequency Histogram. Draw a Histogram for the following frequency distribution giving the steps involved.

Mid values (*X*)	32	37	42	47	52	57	62	67
Frequency (*f*)	3	17	28	47	54	31	14	4

2.8 i. a) Write down the important points for drawing graphs?

b) In order to estimate the mean length of leaves from a certain tree, a sample of 100 leaves was chosen and their lengths are measured in millimeter. A grouped frequency table was set up and the results were as follows:

Mid value	2.2	2.7	3.2	3.7	4.2	4.7	5.2	5.7	6.2
Frequency	3	5	8	12	18	24	20	8	2

ii. a) Display the table in the form of a frequency polygon.

b) What are the boundaries of the interval whose mid point is 3.7 cm?

2.9 In a locality, total area is 500 acres where 250 acres are under sugarcane, 125 acres are under maize, 60 acres are under wheat and the remaining 65 acres are under other crops. Make a pie-diagram to represent the distribution of acreage under different crops.

2.10 A biologist was interested to know whether male spiders are longer or female spiders. He collected random samples of female and male green lynx spiders given below. Advise him?

Length of female (in mm)				**Length of male (in mm)**			
5.7	5.2	4.7	5.8	8.2	9.9	5.9	8.4
6.1	6.3	7.0	5.7	9.6	7.0	6,6	7.8
7.5	6.4	6.5	4.7	7.5	9.8	6.3	8.3
6.2	5.4	6.2	5.8	8.0	9.1	6.3	8.4
4.8	5.9	5.2	6.8	8.7	7.4	7.9	10.2
5.6	5.5						

2.11 i) What is meant by tabulation? Explain the main steps which are generally taken in tabulation?

ii) What is a frequency distribution? How is it constructed?

2.12 The following data gives the lifetime in minutes, recorded to the nearest tenth of a minute of 50 sprayed insects.

1.2	2.2	0.7	3.9	1.7	1.9	1.4	1.8	2.0	4.3
2.5	0.9	3.4	2.8	3.7	3.5	0.4	2.8	1.1	0.2
3.9	6.3	2.5	2.1	1.3	2.1	0.3	0.4	2.4	2.1
3.5	2.9	1.2	5.3	1.7	2.7	1.8	4.8	3.2	1.6
2.6	1.8	2.3	1.3	3.1	1.5	2.6	5.9	2.0	2.3

Using 8 intervals with the lowest starting at 0.1

i) Form a frequency distribution and a cumulative frequency distribution.

ii) Also draw Histogram and frequency polygon for the frequency distribution so formed.

2.13 i) What are the advantages of diagrammatic representation?

ii) Explain the following:

a) A Bar diagram b). Subdivided bar diagram

c) Multiple bar diagram.

2.14 The following data gives the record of a company's savings over the years. Draw a bar diagram to represent it:

Year	1950	1951	1952	1953	1954	1955	1956	1957
Rs.(000)	1010	2050	3458	1980	2300	1295	1520	1070

2.15 Draw a sub-divided bar diagram to represent the male and female population of four divisions of Punjab in 1961.

Division	Male	Female	Both Sexes
Lahore	35	30	65
Multan	35	31	66
Sargodha	32	28	60
Rawalpindi	21	19	40

2.16 The following information is available about the age (in years) and their weights in kilograms (age, weight). Make a scatter plot and bivarite frequency distribution.

(20, 60), (15, 55), (14, 45), (17,60), (16, 48), (22, 70), (16, 63), (14, 55), (18, 57), (19, 67), (21, 67), (17, 65), (13, 60), (15, 60), (17, 49), (19, 65), (23, 73), (21, 65), (22, 70), (14, 50), (16, 60), (19, 59), (15, 62), (17, 59), (24, 75).

2.17 Fill in the blanks:

i) Classification is the __________ of arranging data according to some common characteristics.

ii) A table has at least __________ parts.

iii) In an open-end frequency distribution, either the __________ class limit of __________ group or upper limit of the __________ class are not given.

iv) In Histogram, with unequal class intervals, the area of each rectangle is __________ to class frequency.

v) An ogive is a __________ polygon.

vi) A frequency table can be represented graphically by a __________.

vii) A Histogram is a ______________ bar Chart with ______ space between its bars.

viii) The area of each bar is __________ to the frequency it represents.

ix) If mid-points of the tops of the consecutive bars in a Histogram are joined by straight lines, a ______________ is obtained.

2.18 Against each statement write T for true and F for false statement.

i) Grouped data and primary data are same.

ii) The class mark is also named as mid point.

iii) A table has at least three parts.

iv) The graph of a time series is called Histogram.

v) Cumulative frequencies are decreasing.

vi) The data 10, 5, 7, 6, 4 is the example of grouped data.

vii) The two fold division is also named as Dichotomy.

viii) Data can be presented by means of graph.

ix) A graph of cumulative frequency curve is called polygon.

x) In constructing a Histogram, midpoints are to be taken along x-axis.

3 Measures Of Location

3.1 Introduction

The diagrammatic representation of a set of data can give us some impressions about its distribution. Even then, there remains a need for a single quantitative measure which could be used to indicate the centre of the distribution. The measures commonly used for this purpose are **mean**, **median** and **mode**. **Geometric mean** and **harmonic mean** are also sometimes used. These measures are single values, which represent the given data and are known as averages or measures of location or measures of central tendency. The name measures of location arises as these measures give an indication where to locate a distribution; The decision as to which measure is to be used depends upon the particular situation under consideration.

Properties of a good average are: -

i) It is well defined. **ii)** It is easy to calculate.

iii) It is easy to understand. **iv)** It is based on all the values.

v) It is capable of mathematical treatment.

The important types of averages are:

i) Arithmetic mean and Weighted mean. **ii)** Geometric mean

iii) Harmonic mean **iv)** Median

v) Mode

3.2 Arithmetic Mean And Weighted Mean

Arithmetic Mean is calculated by adding up all the observations and dividing the sum by the total number of observations. The Greek letter μ (meu) is used as a symbol for the population mean.

The population mean of N observations $Y_1, Y_2, ..., Y_N$ is defined as:

$$\mu = \frac{1}{N}(Y_1 + Y_2 + ... + Y_N)$$

$$= \frac{1}{\mathrm{N}}\sum_{i=1}^{N} Y_i$$

$$= \frac{\sum Y}{N} \qquad (3.1)$$

The μ is a parameter, a fixed value and is usually unknown in practice. It is also called as arithmetic mean, abbreviated as A.M. The estimate of population mean μ is sample mean and is denoted by $\overline{Y}$. $\overline{Y}$ is a statistic and its value varies from sample to sample drawn from the population. The sample mean $\overline{Y}$ of n observations $y_1, y_2, ..., y_n$ from a population is defined as:

$$\overline{Y} = \frac{1}{n}(y_1 + y_2 + ...y_n)$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$= \frac{\sum y}{n} \qquad (3.2)$$

The sample mean is a good estimate of the population mean as it is an unbiased statistic. Unbiased simply means that if means are calculated for all possible samples drawn from the population, the mean of these sample means would be equal to the population mean. The arithmetic mean has the same units as the original observation i.e., if the original observations are in centimeters, the unit of mean would be in centimeters.

Example 3.1: Arithmetic mean for ungrouped data.

Following are the data on students heights (cms).

87, 91, 89, 88, 89, 91, 87, 92, 90, 98.

There are ten observations and their sum is 902.

										Total
Y_i	89	91	89	88	89	91	87	92	90	98	902

$$\text{Arithmetic mean } (\overline{Y}) = \frac{\sum y_i}{n}$$

$$= \frac{902}{10}$$

$$= 90.20\ cm.$$

3.2.1 Updating or correcting the mean

It may happen that an observation was overlooked while calculating the mean. In such situations, the observation may be incorporated.

Suppose that in the above example 3.1, it was found later on that the last observation is 94 instead of 98. To correct the mean, we proceed as follows:

The mean of n observation is given by:

$$\bar{Y} = \frac{\sum y_i}{n} \Rightarrow \sum y = n\,\bar{Y}$$

$$= 10(902) = 9020$$

The corrected total can be obtained by subtracting 98 and adding 94 from the total 9020 so the corrected sum is given by

Corrected $\Sigma\, Y = 9020 - 98 + 94 = 9016$

So corrected mean is given by

$$\bar{Y} = \frac{\Sigma y_i}{n}$$

$$= 9016/10$$

$$= 90.16$$

A.M for grouped data

When data is lengthy, it is usually grouped into different classes and $\bar{Y}$ is calculated by the following formula:

$$\bar{Y} = \frac{f_1\, y_1 + f_2 y_2 + ... + f_k y_k}{f_1 + f_2 + ... + f_k}$$

$$= \frac{\sum_{i=1}^{k} f_i\, y_i}{\sum_{i=1}^{k} f_i} = \frac{\sum f_i y_i}{\sum f_i} \qquad (3.3)$$

where k is the number of classes. y_i is the mid point of the ith class and f_i is the corresponding frequency. Here, it is assumed that each of the observations is equal to the mid point of the class in which it occurs. This causes the value of arithmetic mean a bit different for grouped data when the same is calculated from ungrouped data. This difference is called grouping error.

4/4

Example 3.2: Find arithmetic mean consider the grouped student height data as in the table 2.1.

Solution: The first two columns indicate frequency distribution and the columns 3 and 4 are useful to calculate mean according to the definition.

C.1	f	Mid point y_i	$f_i y_i$
86-90	6	88	528
91-95	4	93	372
96-100	10	98	980
101-105	6	103	618
106-110	3	108	324
111-115	1	113	113
Total	30	---	2935

$$\sum f_i y_i = 2935$$

$$\sum f_i = 30$$

$$\bar{y} = \text{Arithmetic mean} = \frac{\sum f_i y_i}{\sum f_i}$$

$$= \frac{2935}{30}$$

$$= 97.8333\text{cm}$$

3.2.2 Properties of arithmetic mean

i) The algebraic sum of the deviations of the observations from their mean is zero. i.e.,

$$\Sigma(y_i - \bar{y}) = 0 \qquad \ldots. (3.4)$$

It can be proved as:

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

$$= \sum y_i - n\bar{Y}$$

$$= \sum y_i - n\frac{\sum y_i}{n} = 0 \quad \therefore \bar{Y} \text{ is constant for any specific values of } Y_i$$

for grouped data

$$\sum f_i\ (y_i - \bar{y}) = \sum_{i=1}^{n} f_i y_i - n\frac{\sum_{i=1}^{n} f_i y_i}{n} \qquad as\ \bar{y} = \frac{\sum_{i=1}^{n} f_i y_i}{n}$$

$$= \sum_{i=1}^{n} f_i y_i - \sum_{i=1}^{n} f_i y_i$$

$$= 0$$

Numerically, it can be easily seen from the following sample

y_i	2	4	5	6	7	6	Total 30
$y_i - \bar{y}$	3	−1	0	1	2	1	0

$$\bar{Y} = \frac{(2+4+5+6+7+6)}{6}$$

$$= \frac{30}{6} = 5$$

We see that $\sum (y_i - \bar{y}) = 0$

ii) Sometimes, it is desirable to calculate the combined mean of two or more sample means using the individual sample means and their sample sizes. It would be denoted by $\overline{Yc}$. The combined mean is calculated as:

$$\overline{Yc} = \frac{n_1 \overline{Y}_1 + n_2 \overline{Y}_2 + ... + n_k \overline{Y}_k}{n_1 + n_1 + ... + n_k} \tag{3.5}$$

$$= \frac{\sum_{i=1}^{k} n_i \overline{Y}_i}{\sum_{i=1}^{k} n_i}$$

e.g., the combined mean for two groups is given by

$$\overline{Yc} = \frac{n_1 \overline{Y}_1 + n_2 \overline{Y}_2}{n_1 + n_2}$$

If $\overline{Y}_1 = 3$ with $n_1 = 3$ and $\overline{Y}_2 = 4$ with $n_2 = 2$, then $\overline{Yc}$ is given by

$$\bar{Y}c= \frac{3(3)+2\bar{(}4)}{3+2}$$

$$= \frac{17}{5}$$

$$= 3.40$$

Similarly, $\bar{Y}c$ can be calculated for more than two groups.

(iii) The sum of squares of the deviations of the observations f rom their mea minimum i.e.,

$$\sum_{i=1}^{n} (Y - \bar{Y})^2 \text{ is minimum} \qquad (3.6)$$

forungrouped data

$\sum f_i(Y_i - \bar{Y})^2$ is minimum for frequency distribution.

It means that when we take the sum of squares of the deviations from an value a other than $\bar{Y}$, then

$$\sum_{i=1}^{n} (Y_i - a)^2 > \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

It can be proved as:

$$\sum_{i=1}^{n} (Y_i - a)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y} + \bar{Y} - a)^2.$$

$$= \sum_{i=1}^{n} \left[(Y_i - \bar{Y}) + (\bar{Y} - a)\right]^2$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - a)^2 + 2(\bar{Y} - a)\sum_{i=1}^{n} (Y_i - \bar{Y})$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - a)^2 + 0 \qquad \text{as } \sum_{i=1}^{n} (Y_i - \bar{Y})=($$

Now $\sum_{i=1}^{n} (Y_i - a)^2$ is the sum of two terms which are both positive, i. $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$ and $n(Y - a)^2$ are positive being squares. So, $\sum_{i=1}^{n} (Y_i - a)^2$ greater than single term $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$ i.e.,

$$\sum_{i=1}^{n} (Y_i - a)^2 > \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

iv) The mean is affected by the change of origin. If a constant A is added to each of the observations $Y_1, Y_2, \ldots\ldots, Y_n$ having mean $\overline{Y}$, then the mean increases by that constant.By adding$a$ to all the observations we have$a+Y_1, a+Y_2, \ldots,a+Y_n$ and the mean would be a+ $\overline{Y}$. For $a = 10$, mean woubd be 10 + $\overline{Y}$, and similarly, if a is subtracted from each Y_i then the mean is $\overline{Y} - a$.

v) The man is affected by the change of scale. If $Y_1, Y_2, \ldots\ldots, Y_n$ have mean $\overline{Y}$ then the mean after multiplying each observation by a constant a, is the mean multiplied by that constant. The mean of $aY_i, aY_2, \ldots,aY_n$. i.e.,

$$\overline{Y} = \frac{aY_1 + aY_2 + \ldots. + aY_n}{n}$$

$$= \frac{a(Y_1 + Y_2 + \ldots. + Y_n)}{n}$$

$$= a\frac{\sum_{i=1}^{n} Y_i}{n}$$

$$= a\overline{Y}$$

If $Y_1, Y_2, \ldots\ldots, Y_n$ are multiplied by 10 then the resulting mean would be 10 $\overline{Y}$.

3.2.3 Calculation of A.M. by coding / short-cut method

The arithmetic mean may be calculated by the following formula:.

$$\overline{Y} = a + \frac{\sum_{i=1}^{n} D_i}{n} \qquad (3.7)$$

Where $D_i = (Y_i - a)$ and a is an arbitrary value called provisional mean. The relation (3.7) can be derived as follows:-

$$\overline{Y} = a + \frac{\sum_{i=1}^{n} Y_i}{n}$$

$$= \frac{\sum_{i=1}^{n} (Y - a + a)}{n}$$

$$= a + \frac{\sum_{i=1}^{n} (Y - a)}{n} = a + \frac{\sum_{i=1}^{n} D}{n} \quad \text{where, } Y_i - a = D_i$$

This is for ungrouped data, and similarly, for grouped data

$$\overline{Y} = a + \frac{\sum_{i=1}^{k} f_i D_i}{\sum_{i=1}^{k} f_i} \qquad (3.8)$$

If the class intervals are equal then the arithmetic mean may also be calculated as:

$$\overline{Y} = a + \frac{\sum_{i=1}^{k} f_i u_i}{\sum_{i=1}^{k} f_i} \times h \qquad (3.9)$$

where $u_i = \frac{Y_i - a}{h}$ and h = common width of the classes.

Example 3.3: Find A.M. for the data of the examples 3.1 and 3.2 by short cut method.

Solution: Taking 90 as an arbitrary origin.

Y_i	87	91	89	88	89	91	87	92	90	98	Total
$D_i = Y_i$-90	-3	1	-1	-2	-1	1	-3	2	0	8	2

$$\overline{Y} = a + \frac{\Sigma D_i}{n}$$

$$= 90 + \frac{2}{10}$$

$$= 90.20 \text{ cm}$$

Arithmetic mean for grouped data of the example 3.2 taking $a = 98$

C.I	f_i	Y_i	D_i=Y_i-98	$f_i D_i$
86 – 90	6	88	–10	–60
91 – 95	4	93	–5	–20
96 – 100	10	98	0	0
101 – 105	6	103	5	30
106 – 110	3	108	10	30
111 – 115	1	113	15	15
Total	**30**	--	--	**–5**

$$\overline{Y} = a + \frac{\sum_{i=1}^{k} f_i D_i}{\sum_{i=1}^{k} f_i}$$

$$= 98 + \left(\frac{-5}{30}\right)$$

$$= 98 - 0.1667$$

$$= 97.8333 \text{ cm}$$

3.2.4 Merits of arithmetic mean

i) It is rigidly defined by mathematical formula.

ii) It is easy to calculate.

iii) It is easy to understand.

iv) It is based upon all the values.

v) It is stable statistic in repeated sampling experiments.

vi) Sum of the observations can be found if mean and number of observations are known.

3.2.5 Demerits of arithmetic mean

i) It is very sensitive to any marked departure from the bell shaped distribution and hence is not suitable for skewed distributions.

ii) It gives fallacious and misleading conclusions when there is too much variation in data.

iii) It is greatly affected by extreme values.

iv) It can not be calculated for open-end classes without assuming open ends.

3.2.6 Weighted Mean

Arithmetic mean is used when all the observations are given equal importance but there are certain situations in which the different observations get different weights. In these situations, weighted mean denoted by $\overline{Y}_w$ is preferred. The weighted mean of Y_1, Y_2,....,Y_n with corresponding weights $w_1, w_2,....,w_n$ is calculated as:

$$\overline{Y}_w = \frac{w_1Y_1 + w_2Y_2 + ... + w_nY_n}{w_1 + w_2 + ...,+w_n}$$

$$= \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i} = \frac{\sum w_i Y_i}{\sum w_i} \qquad (3.10)$$

Example 3.4: The following data is about the percentage kill (Y_i) and the number of insects (w_i) used in a study, the interest is to calculate the mean of the percentage kill.

Y_i	88	85.7	52.1	33.3	12.0
w_i	44	42	24	16	6

Solution: Weighted Mean $= \overline{Y}_w = \dfrac{\sum w_i Y_i}{\sum w_i}$

Now $\Sigma w_i Y_i = 44(88) + 42\,(85.7) + 24\,(52.1) + 16\,(33.3) + 6\,(12)$

$= 9326.6$

and $\Sigma w_i = 44 + 42 + 24 + 16 + 6$

$= 132$

therefore $\overline{Y}_w = \dfrac{9326.6}{132}$

$= 70.65606\%$

3.3 Geometric Mean

Geometric mean is useful measures of central tendency for positive values. It is appropriate for averaging rates and ratios. It may be appropriately calculated only for ratio scale data.

The geometric mean is defined as the nth root of the product of n positive numbers. If we have n positive values Y_1, Y_2, ..., Y_n then geometric mean, denoted by G.M is defined by

$$\text{G.M} = \sqrt[n]{Y_1 \times Y_2 .... \times Y_n} \qquad (3.11)$$

$$= (Y_1 \times Y_2 .... \times Y_n)^{1/n}$$

Taking log of both sides, we get.

$$\text{Log G.M} = \frac{1}{n}(\log Y_1 + \log Y_2 + ... + \log Y_n)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log Y_i$$

or $$\text{G.M} = \text{antilog}\left[\frac{1}{n}\sum_{i=1}^{n} \log Y_i\right] \qquad (3.12)$$

This measure is useful when dealing with relative values such as to find the average of percentage changes, independent ratios and index numbers. The formula for grouped data is:

$$G.M = \sqrt[n]{(Y_1)^{f1}(Y_2)^{f2} + ... + (Y_3)^{fk}} \qquad (3.13)$$

Where $n = \sum_{i=1}^{k} f_i$

$$G.M. = [(Y_1)^{f_1} \times (Y_2)^{f_2} \times .... \times (Y_k)^{f_k}]^{\frac{1}{n}}$$

Taking log, we get

$$\log(G.M) = \frac{1}{n}[f_1 \log(Y_1) + f_2 \log(Y_2) + ... + f_k \log(Y_k)]$$

$$= \frac{1}{n}\sum_{i=1}^{k} f_i \log(Y_i)$$

or $$G.M = \text{antilog}\left(\frac{1}{n}\sum_{i=1}^{k} f_i \log(Y_i)\right) \qquad (3.14)$$

Example 3.5: Calculate geometric mean for the following ungrouped data of the percentage changes in the weight of eight animals.

45, 30, 35, 40, 44, 32, 42, 37

Solution: we know that

$$G.M = \text{antilog}\left[\frac{1}{n}\sum_{i=1}^{n} \log Y_i\right]$$

$$\text{Log}(G.M) = \frac{1}{n}\sum_{i=1}^{n} \log Y_i$$

$$= \frac{1}{8}[\log 45 + \log 30 + \log 35 + \log 40 + \log 44 + \log 32 + \log 42 + \log 37].$$

$$= \frac{1}{8}[1.6532 + 1.4771 + 1.5441 + 1.6021 + 1.6434 + 1.5051 + 1.6232 + 1.5682].$$

$$= \frac{12.6164}{8}$$

$$= 1.57705$$

$$\Rightarrow \quad G.M = \text{antilog}(1.57705)$$

$$= 37.7616$$

Example 3.6: Compute G.M by using the basic definition for the observations:
0.5, 10.0, 2.7, 3.48, 4.7

Solution: Geometric mean of five observations is given by

$$\text{G.M.} = (Y_1\times Y_2\times Y_3\times Y_4\times Y_5)^{\frac{1}{5}}$$

$$= [(0.5)\,(1.0)\,(2.7)\,(3.48)\,(4.71)\,]^{\frac{1}{5}}$$

$$= 1.8577$$

Example 3.7: The grouped data available on insect growth population for age and corresponding frequencies are given.

Class boundaries	f_i	Y_i	**Log** Y_i	f_i **log** Y_i
0 – 4	2	2	0.3010	0.6021
4 – 8	5	6	0.7782	3.8908
8 – 12	7	10	1.0000	7.0000
12 – 16	8	14	1.1461	9.1690
16 – 20	7	18	1.2553	8.7869
20 – 24	4	22	1.3424	5.3697
24 – 28	1	26	1.4150	1.4150
Total	34	--	--	36.2334

Find geometric mean for the above data

Solution: $\therefore \text{G.M} = \text{antilog}\left[\frac{1}{n}\sum_{i=1}^{k} f_i \log(Y_i)\right],$

Where $n = \sum f_i$

To compute G.M. we calculate column 3, 4 & 5 of table

$$\sum f_i \log Y_i = 36.2334$$

$$\sum f_i = 34$$

$$\text{thus G.M.} = \text{antilog}\left[\frac{36.2334}{34}\right]$$

$$= \text{antilog}\ (1.0657)$$

$$= 11.6329$$

Example 3.8: A man gets a rise of 10% in salary at the end of his first year of service and further rise of 20% and 25% at the end of the second and third years respectively.

The rise in each case being calculated on his salary at the beginning of the year. To what annual percentage increase is this equivalent?

Solution: Suppose the initial salary of the man = 100

Increase after first year = 10%

Salary at the end of the year = 100 + 10 = 110

Salary at the end of the second year = 100 + 20 = 120

Salary at the end of the third year = 100 + 25 = 125

$$\text{G.M.} = (110 \times 120 \times 25)^{\frac{1}{3}} = 118.16$$

Annual percentage increase = 118.16 – 100 = 18.16%

Example 3.9: The frequency distribution given below has been derived from the use of working origin. If $D = Y - 18$, find Arithmetic Mean and Geometric Mean.

D	-12	-8	-4	0	4	8	12	16
f	2	5	8	18	22	13	8	4

Solution: Here, $D = Y - 18$ or $Y = D + 18$

D	f	Y	fY	log Y	f log Y
-12	2	6	12	0.776815	1.55630
-8	5	10	50	1.00000	5.00000
-4	8	14	112	1.14613	9.16904
0	18	18	324	2.25527	22.59490
4	22	22	484	1.34242	29.53324
8	13	26	338	1.41497	18.9461
12	8	30	240	1.47712	11.81696
16	4	34	136	1.53148	6.12592
Total	80	--	1696	--	104.19097

$$\text{Arithmetic Mean} = \overline{Y} = \frac{\sum fY}{\sum f}$$

$$= \frac{1696}{80}$$

$$= 21.2$$

$$\text{and G.M} = \text{Antilog}\left(\frac{\sum f \log Y}{\sum f}\right)$$

$$= \text{Antilog}\left(\frac{104.19097}{80}\right)$$

$$= 20.06$$

3.3.1 Properties of geometric mean

i) If there are k sets, each with observations $n_1, n_2, \ldots, n_k$ and G_1, G_2, G_k as their geometric means. Then the combined geometric mean G_{com} of total observations is given by

$$G_{\text{com}} = \frac{\sum_{i=1}^{k} n_i \log G_i}{\sum_{i=1}^{k} n_i} \qquad (3.15)$$

ii) If there are two sets each consisting of n positive observations $Y_1, Y_2, \ldots, Y_n$ with geometric mean G_1 and $X_1, X_2, \ldots, X_n$, with geometric mean G_2, then the geometric mean G of the ratio $Z=Y/X$ is the ratio of their geometric means i.e.,

$$G = \frac{G_1}{G_2} \qquad (3.16)$$

This can easily be proved as under:

$$\text{G} = \left[\frac{Y_1}{X_1} \times \frac{Y_2}{X_2} \times \ldots \times \frac{Y_n}{X_n}\right]^{\frac{1}{n}}$$

$$= \left[\frac{Y_1}{X_1} \times \frac{Y_2}{X_2} \times \ldots \times \frac{Y_n}{X_n}\right]^{\frac{1}{n}}$$

$$= \frac{(Y_1 \times Y_2 \times \ldots \times Y_n)^{\frac{1}{n}}}{(X_1 \times . X_2 \times \ldots \times X_n)^{\frac{1}{n}}}$$

$$= \frac{G_1}{G_2}$$

3.3.2 Merits of geometric mean

i) It is rigidly defined by a mathematical formula.

ii) It is based on all the observations.

iii) It is capable of mathematical development.

iv) It is less affected by the extreme values as compared with the mean.

3.3.3 Demerits of geometric mean

i) It becomes zero if any of the observations is zero.

ii) It is sensible only for positive values and it becomes imaginary for negative values.

3.4 Harmonic mean

The harmonic mean is particularly useful when dealing with the averages of certain types of rates and ratios. The harmonic mean of n values $Y_1, Y_2, ... Y_n$ is defined as reciprocal of the arithmetic mean of the reciprocals of the values. The harmonic mean is denoted by H.M and is defined by:

$$\text{H.M} = \text{Reciprocal of } \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i}\right) \qquad (3.17)$$

Where $Y_i \neq 0$

$$= \frac{n}{\sum_{i=1}^{n}\frac{1}{Y_i}} \qquad (3.18)$$

Harmonic mean for the grouped data is given by

$$\text{H.M} = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k}\frac{f_i}{Y_i}} \qquad (3.19)$$

Harmonic mean deals with the rates independent on each other.

Example 3.10: A tractor is running at the rate of 10 Km / hr. during the first 60 Km; at 20 Km/ hr. during second 60 Km; 30 Km/hr. during the third 60 Km; 40 Km/ hr.

during the fourth 60 Km and 50 Km/hr. during the (last) fifth 60 Km. What would be the average speed?

Solution: Harmonic mean of the values shall give the average speed.

$$\text{H.M} = \text{Reciprocal of } \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i}\right)$$

$$= \text{Reciprocal of } \left[\frac{\frac{1}{10}+\frac{1}{20}+\frac{1}{30}+\frac{1}{40}+\frac{1}{50}}{5}\right]$$

$$= \left[\frac{5}{\frac{1}{10}+\frac{1}{20}+\frac{1}{30}+\frac{1}{40}+\frac{1}{50}}\right]$$

$$= \frac{5}{0.22833}$$

$$= 21.89813 \text{ Km/hr}$$

Example 3.11: Find Harmonic mean for grouped data of the Example 3.7

Solution: We know that $\text{H.M} = \dfrac{\Sigma f_i}{\Sigma f_i/Y_i}$

The data is given by

Class boundaries	f_i	Y_i	$\frac{1}{Y_i}$	$f_i\frac{1}{Y_i}$
0 – 4	2	2	0.5000	1.0000
4 – 8	5	6	0.1667	0.8333
8 – 12	7	10	0.1000	0.7000
12 – 16	8	14	0.0714	0.5714
16 – 20	7	18	0.0556	0.3889
20 – 24	4	22	0.0454	0.1816
24 – 28	1	26	0.0385	0.0385
Total	34	--	00	3.7137

$$\sum f_i \quad = 34$$

$$\sum \frac{f_i}{Y_i} \quad = 3.7137$$

$$\therefore \text{H.M} = \frac{34}{3.7137}$$
$$= 9.1553$$

Example 3.12: Calculate Harmonic Mean and Geometric Mean from the following data:

3, 8, 11, 0, 10

Solution: It is not possible to calculate Geometric Mean and Harmonic Mean as the data involves the value 0 because while calculating Geometric Mean, the multiplication with 0 makes product of the given values zero (0) and while calculating Harmonic Mean, division by zero (0) is undefined.

3.4.1 Properties of Harmonic Mean

i) If there are k sets each with observations $n_1, n_2, ..., n_k$ and $k_1, k_2, ..., k_k$ as their harmonic means. Then the combined harmonic mean H.M$_{\text{comb}}$ of all the observations is given by:

$$\text{H.M}_{\text{comb}} = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{H_i}} \qquad (3.20)$$

3.4.2 Merits of harmonic mean

i) It is defined by a mathematical formula.

ii) It is based on all the observations.

iii) It is capable of future mathematical development.

3.4.3 Demerits of Harmonic Mean

i) It can not be calculated if any of the observations is zero.

ii) It is not simple to calculate as compared to the arithmetic mean.

iii) It gives less weight to large values and more weight to small values.

3.4.4 General relationship between A.M. , G.M. and H.M.

If $Y_1, Y_2, ..., Y_n$ are n positive observations, then the arithmetic mean, geometric mean and harmonic mean satisfy the following relation.

$$\text{A.M} \quad \geq \text{G.M.} \geq \text{H.M}$$

The three means are equal only if all the observations are identical.

3.5 Median

It is the value which divides an arranged data in order of magnitude into two equal parts. In case of odd number of observations, median is the value of $\left(\frac{n+1}{2}\right)$ th item and in case of even number of observations, median is the mean value of $(n/2)$ th and $\left(\frac{n+2}{2}\right)$ th items of a set of values arranged in ascending or descending order of magnitude i.e., defined as the middle value if the number of values is odd and the mean of the two middle values if the number of values is even.

Example 3.13: Following are the heights (cms) of 5 students measured at the time of registration. Also find median for the data of the example 3.1.

Y_i: 88.03, 94.50, 94.90, 95.05, 84.60

Solution: The ordered observations are:

84.60, 88.03, 94.50, 94.90, 95.05

Here $n = 5$, so

$$\text{Median} = \text{value of } \left(\frac{5+1}{2}\right)\text{th observation}$$

$$= \text{value of } 3^{\text{rd}} \text{ observation}$$

$$= 94.50$$

or
$$\text{Median} = y_{\left(\frac{n+1}{2}\right)}$$

$$= y_{\left(\frac{5+1}{2}\right)}$$

$= Y_{(3)}$, the third value in the ordered observations.

$= 94.50$ cms

The data of the example 3.1 is used to calculate median for even n. The ordered observations are:

87, 87, 88, 89, 89, 90, 91, 91, 92, 98.

or $$\text{Median} = \text{value of} \left(\frac{n+1}{2}\right)^{th} \text{obs.}$$

$$= \text{value of} \left(\frac{10+1}{2}\right)^{th} \text{obs.}$$

$$= \text{value of } (5.5)^{th} \text{ obs.}$$

$$= 5^{th} \text{ obs} + 0.5\ (6^{th}\ obs - 5^{th}\ obs)$$

$$= 89 + 0.5\ (90 - 89)$$

$$= 89 + 0.5\ (1) = 89.5$$

or $$\text{Median} = \frac{1}{2}\left[y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)}\right]$$

$$y_{\left(\frac{n}{2}\right)} = y_{\left(\frac{10}{2}\right)} = y_{(5)} = 89$$

$$y_{\left(\frac{n+2}{2}\right)} = y_{(6)} = 90$$

So, median is the mean of $Y_{(5)}$ and $Y_{(6)}$

$$\text{Median} = \frac{89+90}{2} = 89.5\,\text{cm}$$

For the grouped data (given in ascending order) median is calculated by the relation:

$$\text{Median} = l + \frac{h}{f}\left(\frac{n}{2} - c\right) \qquad (2.21)$$

Where l is the lower class boundary of the class containing the median.

h is width of the class containing median.

f is the frequency of the class containing median.

$\frac{n}{2}$ is used to locate the median class i.e., where the $\left(\frac{n}{2}\right)$th observation falls and this is done by looking at the class corresponding to the cumulative frequency in which $\left(\frac{n}{2}\right)$th observation lies.

c is the cumulative frequency of the class preceding to the median class.

Example 3.14: Find the median for the following student height grouped data.

Class boundaries	y	f_i	c.f
85.5 – 90.5	87	6	6
90.5 – 95.5	93	4	10
95.5 – 100.5	98	10	20
100.5 – 105.5	103	6	26
105.5 – 110.5	108	3	29
110.5 – 115.5	113	1	30

Solution: To find median class $\left(\frac{n}{2}\right)$th observation falls is $(\frac{30}{2})$ th observation.

$$\text{i.e } \frac{n}{2}\text{th observation} = \frac{30}{2}\text{th observation}$$

$$= 15 \text{ th observation}$$

The 15th observation falls in the class 95.5 – 100.5

So, Median group = 95.5 – 100.5

$$\text{Median} = l + \frac{h}{f}\left(\frac{n}{2} - c\right)$$

$$= 95.5 + \frac{5}{10}\,(15 - 10)$$

$$= 98.0 \text{ cm}$$

In case, when data is discrete but grouped, the median is calculated by using the formal definition of median.

Example 3.15: Discrete grouped data of 26 plants of cotton are taken and the number of bolls per plant observed, the data is grouped as follows:

Number of bolls	0	1	2	4	5	6	7
Number of plants	5	6	3	6	3	2	1

Solution: As n is even so the median is the mean of $Y_{(\frac{n}{2})}$ and $Y_{(\frac{n}{2}+1)}$ *obs.*

$$Y_{(\frac{n}{2})} = Y_{(13)}\ ;\ Y_{(\frac{n}{2}+1)} = Y_{(14)}$$

To locate $Y_{(13)}$ and $Y_{(14)}$ we need to make cumulative frequency column:

Number of bolls	Number of plants	c.f
0	5	5
1	6	11
2	3	14
4	6	20
5	3	23
6	2	25
7	1	26

From the c.f column it is clear that $Y_{(13)}$ and $Y_{(14)}$ are the plants which have number of bolls equal to 2. Thus 2 is the median.

3.5.1 Properties of median

i) If a constant a is added to each of the n observations $Y_1, Y_2, \ldots, Y_n$ having median M, then the median of $a + Y_1, a + Y_2, \ldots, a + Y_n$ would be $a + M$. If a is multiplied to each of the n observations, then median of $aY_1, aY_2, \ldots, aY_n$ would be aM.

ii) The sum of absolute deviations of the observations from their median is minimum i.e.,

$$\Sigma \left|Y - \text{median}\right| \text{is minimum} \qquad (3.22)$$

The bars indicate absolute value meaning thereby we are taking all the deviations positive.

iii) For a symmetrical distribution median is equidistant from the first and third quartiles i.e.,

$$Q_3 - \text{Median} = \text{Median} - Q_1 \qquad (3.23)$$

Where, Q_1 and Q_3 are first and third quartiles respectively.

3.5.2 Merits of median

i) It is quick to find.

ii) It is not much affected by exceptionally large or small values in the data.

iii) It is suitable for skewed distributions.

3.5.3 Demerits of median

i) It is not rigidly defined.

ii) It is not readily suitable for algebraic development.

iii) It is less stable in repeated sampling experiments than the mean.

iv) It is not based on all the observations.

3.6 Quantiles

Sometimes, our interest is to know the position of an observation relative to the others in a data set. For example, in the grouped student height data of the example 2.1, we may be interested to know the percentage of students having height less than some specified value. The measures used for this purpose are called quantiles or fractiles. These are usually calculated under the following headings:

i) Quartiles and Deciles ii) Percentiles

3.6.1 Quartiles

Quartiles are the values in the order statistic that divide the data into four equal parts. These are the first quartile Q_1, second quartile Q_2 (median) and third quartile Q_3. The first quartile, also known as lower quartile, is the value of order statistic that exceeds ¼ of the observations and less than the remaining ¾ observations. The second quartile is the median and the third quartile, known as upper quartile, is the value in the order statistic that exceeds ¾ of the observations and is less than remaining ¼ observations.

In case of ungrouped data, the quartiles are calculated by splitting the order statistic at the median and calculating the median of the two halves. If n is odd, the median can be included in both halves.

Example 3.16: Find Quartiles for ungrouped data of the example 3.13

Solution: We know that median of data is the mid value of the order statistic. For finding quartiles, we split the order statistic at the median and calculate the median of two halves. Since n is odd, we can include the median in both halves.

The orders statistic is

84.60, 88.03, 94.50, 94.90, 95.05

$$Q_2 = \text{Median} = Y_{\left(\frac{n+1}{2}\right)}$$

$$= Y_{(3)}, \text{ the third observation}$$

$$= 94.50$$

$$Q_1 = \text{Median of the first three value} = Y_{\left(\frac{3+1}{2}\right)}$$

$$= Y_{(2)}, \text{ the second observation}$$

$$= 88.03$$

$$Q_3 = \text{Median of the last three values} = Y_{\left(\frac{3+5}{2}\right)}$$

$= Y_{(4)}$, the fourth observation

$= 94.90$

For the grouped data (in ascending order) the quartiles are calculated as:

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right) \tag{3.24}$$

$$Q_2 = l + \frac{h}{f}\left(\frac{2n}{4} - c\right) \tag{3.25}$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right) \tag{3.26}$$

Where l is the lower class boundry of the class containing the Q1, Q2 or Q3.
h is the width of the class containing the Q1, Q2 or Q3.
f is the frequency of the class containing Q1, Q2 or Q3.
c is the cumulative frequency of the class immediately preceeding to the class containing Q_1, Q_2 or Q_3, $\left[\frac{n}{4}, \frac{2n}{4} \text{ or } \frac{3n}{4}\right]$ are used to locate Q_1, Q_2 or Q_3 group.

Example 3.17: Find quartiles for data of the example 3.2.

Solution:

Class boundaries	Y_i	f_i	c.f
85.5 – 90.5	87	6	6
90.5 – 95.5	93	4	10
95.5 – 100.5	98	10	20
100.5 – 105.5	103	6	26
105.5 – 110.5	108	3	29
110.5 – 115.5	113	1	30
Total	-	**30**	-

-To locate the class containing Q_1,

$\frac{n}{4}$ th observation $= \frac{30}{4}$ th observation

$= 7.5$ th observation

7.5 th observation falls in the group 90.5 – 95.5.

So, Q_1 group = 90.5 – 95.5

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$$

$$= 90.5 + \frac{5}{4}(7.5 - 6)$$

$$= 92.3750 \text{ cm}$$

for Q_2,

$$\frac{2n}{4} \text{th observation} = \frac{2 \times 30}{4} \text{th observation}$$

= 15 th observation falls in the group 95.5 – 100.5

So, $$Q_2 = l + \frac{h}{f}\left(2\frac{n}{4} - c\right)$$

$$= 95.5 + \frac{5}{10}(15 - 10)$$

$$= 98 \text{ cm}$$

for Q_3,

$$\frac{3n}{4} \text{ th observation} = \frac{3 \times 30}{4} \text{th observation}$$

= 22.5 th observation

So, Q_3 group = 100.5 – 105.5

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

$$= 100.5 + \frac{5}{6}(22.5 - 20)$$

$$= 100.5 + 2.0833$$

$$= 102.5833 \text{ cm}$$

3.6.2 Deciles

Deciles are the values in the order statistic that divide the data into ten equal parts. These are denoted by D_1, D_2, D_3,….,D_9. D_1 is the value of order statistic that exceeds 1/10 of the observations and less than the remaining 9/10. The fifth decile is the median and D_9, the ninth decile is the value in the order statistic that exceeds 9/10 of the observations and is less than 1/10 remaining observations.

Deciles for Ungrouped Data

To calculate deciles for the ungrouped data the following procedure may be followed.

i) Order the observations.

ii) For the *m*th decile, determine the product $\frac{m.n}{10}$. If $\frac{m.n}{10}$ is not an integer, round it up and find the corresponding ordered value but if $\frac{m.n}{10}$ is an integer, say *k*, then calculate the mean of *k*th and (k+1) th ordered observations.

Example 3.18: The data of the example 2.1 is used to explain the procedure and D_7 and D_3 have been calculated.

i) The ordered observations are

87, 87, 88, 89, 89, 90, 91, 91, 92, 95, 96, 96, 97, 98, 98, 98, 99,99, 100, 100, 101, 101, 102, 103, 105, 105, 106, 107, 107, 112.

Here $n = 30$

ii) To calculate $D_{7,}$ the (7) (30) / 10 = 21, so, we calculate the mean of 21st and 22nd observations i.e., $D_7 = (101 + 101)/2 = 101$.

To calculate $D_{3,}$ the (3) (30) / 10 = 9, so, we calculate the mean of 9th and 10th observation i.e., $D_3 = (92 + 95)/2 = 93.5$.

Deciles for grouped data

The *m*th decile for grouped data (in ascending order) is

$$D_m = l + \frac{h}{f}\left(\frac{m.n}{10} - c\right) \qquad (3.27)$$

Like the median, $\frac{m.n}{10}$ is used to locate the *m*th decile group.

l is the lower class boundary of the class containing *m*th decile.

h is the width of the class containing D_m.

f is the frequency of the class containing D_m.

n is the total number of frequencies.

c is the cumulative frequency of the class immediately preceding to the class containing D_m.

Example 3.19: Data of the example 3.17 is used to explain the procedure for grouped data. D_1 and D_7 are calculated

Calculation for D_1

$$\frac{1.n}{10}\text{th observation} = \frac{1\times 30}{10}\text{th observation}$$

$$= 3^{rd}\text{ observation}$$

So, D_1 group $= 85.5 - 90.5$

$$\therefore \quad D_1 = l + \frac{h}{f}\left(\frac{1\times n}{10} - c\right)$$

$$= 85.5 + \frac{5}{6}\ (3-0)$$

$$= 88.00\text{ cm}$$

Calculation for D_7

$$\frac{7\times n}{10}\text{th observation} = \frac{7\times 30}{10} = 21^{st}\text{ observation}$$

$$D_7 = l + \frac{h}{f}\left(\frac{7\times n}{10} - c\right)$$

$$= 100.5 + \frac{5}{6}(21-20)$$

$$= 101.3333\text{ cm}$$

3.6.3 Percentiles

These are the measures of relative standing of an observation within a data. The *pth* percentile is the value $Y_{(p)}$ in the order statistic such that *p* percent of the values are less than the value $Y_{(p)}$ and (100-p) percent of the values are greater than $Y_{(p)}$. The 5th percentile is denoted by P_5, the 10th by P_{10} and 95th by P_{95}.

Percentiles for the ungrouped data

To calculate percentiles for the ungrouped data, the following procedure is adopted:

i) Order the observations.

The procedure is explained on the data of the example 2.1 and P_{10} and P_5 have been calculated.

ii) For the *m*th percentile, determine the product $\frac{m.n.}{100}$. If $\frac{m.n.}{100}$ is not an integer, round it up and find the corresponding ordered value and if $\frac{m.n.}{100}$ is an integer, say *k*, then calculate the mean of the *Kth* and (*k*+1) *th* ordered observations.

The ordered observations of the example 2.1 are:

87, 87, 88, 89, 89, 90, 91, 91, 92, 95, 96, 96, 97, 98, 98, 98, 99,99, 100, 100, 101, 101, 102, 103, 105, 105, 106, 107, 107, 112.

To calculate $P_{10,}$ the (10) (30) / 100 = 3, so, we calculate the mean of 3rd and 4th observations i.e., $P_{10} = (88 + 89)/2 = 88.5$.

To calculate $P_{95,}$ the (95) (30) / 100 = 28.5, so, 29th observation is our 95th percentile i.e., $P_{95} = 107$.

Percentiles for the grouped data

The m_{th} percentile for grouped data (given in ascending order) is

$$P_m = l + \frac{h}{f}\left(\frac{m.n}{100} - c\right) \tag{3.28}$$

Like the median, $\frac{m.n}{100}$ is used to locate the *m*th percentile group.

l is the lower class boundary of the class containing the *m*th percentile.

h is the width of the class containing P_m.

f is the frequency of the class containing P_m.

n is the total number of frequencies.

c is the the cumulative frequency of the class immediately preceding to the class containing P_m.

The 50th percentile is the median by definition as half of the values in the data are smaller than the median and half are larger than the median.

The 25th and 75th percentiles are the lower and upper quartiles respectively. The quartiles, deciles and percentiles are also called quantiles or fractiles.

Example 3.20: Find P_{10}, P_{25}, P_{50} and P_{95} of grouped data for the example 3.17.

Solution: $\frac{10n}{100}$ th observation = $\frac{10\times 30}{100}$ th observation

= 3rd observation

So, P_{10} group= 85.5 – 90.5

$$P_{10} = l + \frac{h}{f}\left(\frac{10n}{100} - c\right)$$

$$= 85.5 + \frac{5}{6}\ (3 - 0)$$

$$= 85.5 + 2.5 = 88.00 \text{ cm}$$

$$P_{25} = l + \frac{h}{f}\left(\frac{25n}{100} - c\right)$$

$$= l + \frac{h}{f}\left(\frac{n}{4} - c\right)$$

$$= Q_1$$

Similarly, $P_{50} = Q_2$ = Median

and $P_{75} = Q_3$, already calculated under the example 3.17

$\frac{95n}{100}$ th observation $= \frac{95 \times 30}{100}$ th observation

= 28.5 th observation

So, P_{95} group= 105.5 – 110.5

$$P_{95} = l + \frac{h}{f}\left(\frac{95n}{100} - c\right)$$

$$= 105.5 + \frac{5}{3}\ (28.5 - 26)$$

$$= 105.5 + 4.1667$$

$$= 109.6667 \text{ cm}$$

The percentiles and quartiles may be read directly from the graphs of the cumulative frequency function as in chapter 2 where, Q_1 is indicated. The Q_3 may be read corresponding to a relative cumulative frequency of 0.75.

3.7 Mode

Mode is defined as the most frequent value in a data set. In case of ungrouped data, the mode can be found by inspection of the order statistic. For example, five plants having heights in cms. 87, 82, 87, 90, 89. The order statistic for this data would be 82, 87, 87, 89, 90.

87 is the value that comes twice while others are only once. So, by definition 87 is the mode of this data. If data has only one mode, then it is called unimodal. The data may have more than one mode. It may be bimodal (having two modes) or multimodal (having more than two modes). The data is said to have no mode, if every value of the data equal number of times.

The mode for the grouped data (given in ascending order) is calculated by

$$\text{Mode} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h \qquad (3.29)$$

l is the lower class boundary of the modal class.

f_m is the frequency of the modal class.

f_1 is the frequency associated with the class preceding the modal class.

f_2 is the frequency associated with the class following the modal class.

h is the width of the model class.

model class is the class in which maximum frequency lies.

Example 3.21: Find mode for the data of the example 3.17.

Solution: The maximum frequency is 10 for the class 95.5 – 100.5, so, it is a model class.

$$\therefore \ \text{Mode } (\hat{X}) = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

Here $l = 95.5,\ h = 5,\ f_m = 10,\ f_1 = 4,\ f_2 = 6$

$$\text{Mode} = 95.5 + \frac{10-4}{(10-4)+(10-6)} \times 5$$

$$= 95.5 + 3.0$$

$$= 98.5 \text{ cm}$$

Example 3.22: The table shows the distribution of the maximum loads is shot tons supported by certain cables produced by a company.

Maximum loads	**No. of Cables**
9.3 – 9.7	2
9.8 – 10.2	5
10.3 – 10.7	12
10.8 – 11.2	17
10.3 – 11.7	14
11.8 – 12.2	6
12.3 – 12.7	3
12.8 – 13.2	1

Determine its mode.

Solution:

Maximum loads	f	C.B.
9.3 – 9.7	2	9.25 – 9.75
9.8 – 10.2	5	9.75 – 10.25
10.3 – 10.7	12	10.25 – 10.75
10.8 – 11.2	17	10.75 – 11.25
11.3– 11.7	14	11.25 – 11.75
11.8 – 12.2	6	11.75 – 12.25
12.3 – 12.7	3	12.25 – 12.75
12.8 – 13.2	1	12.75 – 13.25

$$\text{Mode }(\hat{X}) = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

$$= 10{:}75 + \frac{17-12}{(17-12)+(17+14)} \times 5$$

$$= 11.06$$

Example 3.23: Find mode for the data of the example 2.1.

y	87	88	89	90	91	92	95	96	97	98	99	100	101	102	103	105	106	107	112
f	2	1	2	1	2	1	1	2	1	3	2	2	2	1	1	2	1	2	1

Mode = The value which occurs most frequently in the data.

∴ Mode = 98 cm.

3.7.1 Properties of mode

i) If a constant a is added to each of the n observations $Y_1, Y_2, ..., Y_n$ having mode m, then the mode of $a+Y_1, a+Y_2, ..., a+Y_n$ would be $a+m$.

ii) If a is multiplied with each of the n observations $Y_1, Y_2, ..., Y_n$, having mode m then the mode of $aY_1, aY_2, ..., aY_n$ would be am.

3.7.2 Merits of mode

i) It is very quick to find.

ii) It is not affected by extreme values.

3.7.3 Demerits of mode

i) It is not rigidly defined.

ii) It is not capable of further mathematical development easily.

iii) It uses only a few members of the population, so can be misleading in small data sets.

iv) It is an unstable measure like median.

v) There may be more than one values of the mode in a data set.

vi) It may not exist in many cases.

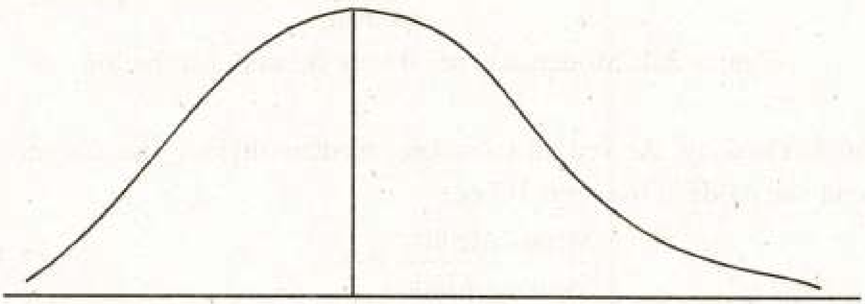
3.7.4 Empirical Relationship Between Mean, Median And Mode

The empirical relationship depends upon the shape of the distribution of the data. The distribution of a data set is called symmetrical if the frequency curve for the data is such that the left of curve to its mean is the mirror image of the portion to the right of mean. Otherwise, the distribution is called skewed. The skew may be to the right or to the left depending upon the shape of curve. The empirical relationship is described as follows:

a) In a single peaked symmetrical distributions mean, median and mode are equal i.e.,

$$\text{Mean} = \text{Median} = \text{Mode} \tag{3.30}$$

It is indicated in figure 3.1



Mean = Median = Mode

Figure 3.1: single peaked symmetrical distribution.

b) For moderately positively skewed distributions, the following empirical relation holds.

$$\text{Mean} > \text{Median} > \text{Mode} \tag{3.31}$$

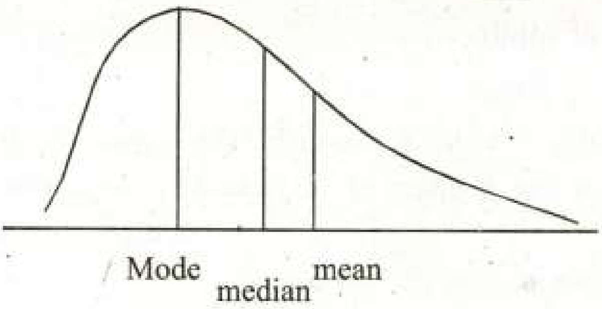


Figure 3.2: Moderately Positively skewed distribution

c) For moderately negatively skewed distributions, the following empirical relation holds.

$$\text{Mean} < \text{Median} < \text{Mode} \tag{3.32}$$

It is indicated in figure 3.3.

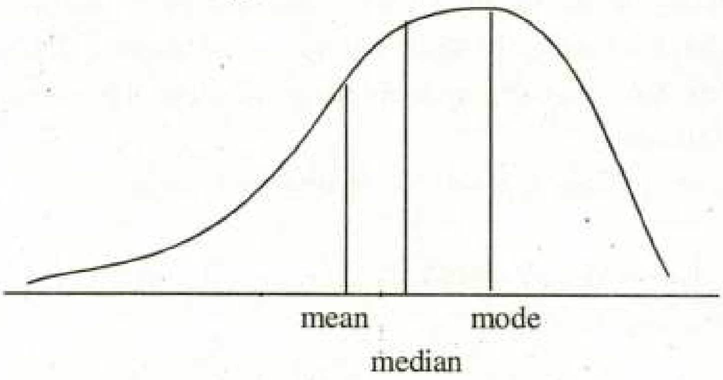


Figure 3.3: Moderately negatively skewed distribution

d) For moderately skewed distributions median divides the distance between mean and mode in the ratio 1:2 i.e.,

$$\frac{\text{Mean - Median}}{\text{Median - Mode}} = \frac{1}{2} \tag{3.33}$$

or Mode = 3 Median – 2 Mean

Example 3.24: If mode = 15 and Median = 12, find mean.

Solution: If Mode = 15, Median = 12, Mean = ?

We know that

$$\text{Mode} = 3\text{ Median} - 2\text{ Mean}$$

$$\text{Mean} = \frac{3\text{Median} - \text{Mode}}{2}$$

$$= \frac{3(12) - 15}{2} = \frac{36 - 15}{2} = 10.5$$

Example 3.25: Mean and Median of a frequency distribution are 45 and 30 respectively. Find mode of the distribution.

Solution: Mean = 45, Median = 30, Mode = ?

We know that

$$\text{Mode} = 3\text{ Median} - 2\text{ Mean}$$

$$= 3(30) - 2(45) = 90 - 90 = 0$$

3.8 Selecting a Suitable Measure of Central Tendency

To select an appropriate measure for a situation, certain factors are taken into account. This includes the type of variable, the purpose of the statistic for which it would be used and the type of distribution.

For the quantitative variables, arithmetic mean is usually appropriate. For the categorical variables, the median and mode are appropriate depending upon the type of categories. For example, if we consider the eye colour, then mode is appropriate but if the categories are income groups, then the appropriate measure is median.

The type of the distribution is an important aspect to evaluate which statistic is appropriate. If the distribution is symmetrical, all the measures i.e., mean, median and mode being equal are equally good. In the skewed distributions median is preferred as it is not affected by the extreme observations. Medians are also preferred to the means when the sample constitutes only small part of the population. Geometric and harmonic means are useful for averaging rates and ratios.

16

Exercise 3

Ans on Page 250

3.1 i) Define arithmetic mean, geometric mean and harmonic mean. Explain the situations when each of them is used perfectly?

ii) Tne relation between arithmetic mean (A.M), geometric mean (G.M) and harmonic mean (H.M) is

$$\text{A.M} \geq \text{G.M.} \geq \text{H.M.}$$

Under what situation these are equal.

3.2 Find the geometric mean of 50, 67, 39, 40, 36, 60, 54, 43.

3.3 A man traveling 100 kilometers has 5 stages at equal intervals. The speed of the man in the various stages was observed to be 10, 16, 20, 14, 15 kilometers per hour.

Find the average speed at which the man travels.

3.4 Calculate mean, median and mode for table 2.5

3.5 Calculate mean, median and mode for the grouped data of table 2.7

3.6 Calculate the following for the data of exercise 2.8 (b)

i) Calculate an estimate for the mean leaf length.

ii) Construct a cumulative frequency table and use it to estimate the sample median.

3.7 What do you understand by weighted mean? In what circumstances is it preferred to ordinary mean and why?

3.8 Define the mode of a frequency distribution. How does it compare with other types of averages?

3.9 i) Write down the empirical relation between mean, median and mode for unimodal distribution of moderate asymmetry. Illustrate graphically the relative positions of the mean, median and mode for frequency curves which are skewed to the right and to the left.

ii) For a certain frequency distribution, with the mean and median 45 and 36 respectively, find the mode approximately using the empirical relation between the three.

3.10 Bilal gets a rise of 10% in salary at the end of his first year of service and further rise of 20% and 25% at the end of the second and third year respectively. The rise in each case being calculated on his salary at the beginning of the year. To what annual percentage increase in this equivalent.

3.11 Find the Mean for the following distribution.

Classes	0–10	10–40	40–90	90–100	100–105	105–120	120–140
f	40	110	150	200	120	30	20

3.12 The frequency distribution given below has been derived from the use of working origin. If $D = X - 18$, find arithmetic mean and Geometric mean.

D	–12	–8	–4	0	4	8	12	16
f	2	5	8	18	22	13	8	4

3.13 The reciprocals of 11 values of x are given below:

0.0500, 0.0454, 0.0400, 0.0333, 0.0285, 0.0232,
0.0213, 0.02000, 0.0182, 0.0151, 0.0143.

Calculate Harmonic and Arithmetic Mean of the data.

3.14 Reciprocals of x are given below:

0.0267, 0.0235, 0.0211, 0.0191, 0.0174, 0.0160, 0.0148

Calculate Harmonic Mean of the data.

3.15 Three cities A, B, C are equidistant from each other. Fatima travels from A to B at the speed of 30 miles per hour by car. From B to C at speed of 50 miles per hour. Determine her average speed for the entire trip.

3.16 Harmonic Mean and Geometric Mean of two numbers are 3.2 and 4 respectively. Find their Arithmetic Mean and both the numbers as well.

3.17 The Arithmetic Mean and Geometric Mean of three numbers are 34 and 18 respectively. Find all the three numbers, when the Geometric Mean of the first two numbers is 9.

3.18 Find out

i) The average rate of motion in the case of a person who rides the first mile at the rate of 10 miles per hour the next mile at the rate of 8 miles per hour and the third at the rate of 6 miles per hour.

ii) Increase in population which in the first decade has increased 20% in the next 25% and in the third 4%.

3.19 The given table shows the distribution of the maximum loads in short tons supported by certain cables produced by a company. Determine Mean, Median and Mode.

Maximums Loads	No. of cables
9.3 – 9.7	2
9.8 –10.2	5
10.3 – 10.7	12
10.8 – 11.2	17
11.3 – 11.7	14
11.8 – 12.2	6
12.3 – 12.7	3
12.8 – 13.2	1

3.20 Compute Mean, Median, Mode, 6th Decile, and 74th percentile for the data given in the table:

Classes	Frequency
0.7312 – 0.7313	10
0.7314 – 0.7315	15
0.7316 – 0.7317	20
0.7318 – 0.7319	25
0.7320 – 0.7321	30
0.7322 – 0.7323	8
0.7324 – 0.7325	2

3.21 Find the value Q_3, D_5, P_5 and mode for the following data:

Groups	Frequency	Groups	Frequency
0 – 4.9	3	25–29.9	13
5 – 9.9	4	30–34.9	13
10 – 14.9	9	35–39.9	5
15 – 19.9	11	40–44.9	2
20 – 24.9	15	45–49.9	2

3.22 If for any frequency distribution the Mean is 45 and the Median is 30. Find Mode approximately, using formula connecting the three.

3.23 A bus traveling 200 miles has ten stages at equal intervals. The speed of the bus in the various stages was observed to be 10, 15, 20, 75, 20, 30, 40, 50, 30, 40 miles per hour. Find the average speed at which the bus has traveled.

3.24 The following data has been obtained from a frequency distribution of a continuous variable x after making the substitution $u = \dfrac{x-136.5}{6}$.

U	–4	–3	–2	–1	0	1	2	3
f	2	5	8	18	22	13	8	4

Find Harmonic Mean.

3.25 Salman obtained the following marks in a certain examination. Find the weighted mean if weights 4, 3, 3, 2 and 2 respectively are allotted to the subjects.

English	**Urdu**	**Math**	**Stat**	**Physics**
73	82	80	57	62

3.26 Calculate weighted mean for the following items:

Items	**Expenditures**	**Weights**
Food	290	7.5
Rent	54	2
Clothing	98	1.5
Fuel and light	75	1.0
Other Items	75	0.5

3.27 For a certain distribution, if $\Sigma\,(x-15)=5$, $\Sigma\,(x-18)=0$, $\Sigma(x-211)=-21$ What is the value of A.M and why?

3.28 Arithmetic Mean of 15 values is 20 and by adding 3 more values, the mean remains 20. Find the new three values if ratio is $a{:}b{:}c$ ::3:2:1.

3.29 State the following as true or false.

i) In a symmetrical distribution, mean, median and mode are equal.

ii) The algebraic sum of the deviations for a set of observations from their mean is always zero.

iii) Median is affected by extreme observations.

iv) Mode is not affected by extreme observations.

v) Frequency polygon is an increasing function.

vi) For highly skewed distributions, median is preferred over mean.

vii) Mean of a data set remains unchanged if a constant is added in each observation.

3.30 What type of averages would you prefer to average the following:

i) Marks obtained in an examination.

ii) Growth rate of population of different cities.

iii) Height of students.

iv) Size of agricultural holdings.

v) Increase in salaries.

vi) IQ level of students in a class.

3.31 Fill in the blanks

i) An average obtained to represent a data is called __________.

ii) A good average should not be effected by __________ values.

iii) Sum of deviations from mean is always __________.

iv) Median is a value that divides an ordered data into __________ parts.

v) For estimating an average rate of change of population __________ is a better average.

vi) The mean and median of two values is always _______.

vii) In qualitative data, the most suitable average is __________.

viii) A distribution having two modes is called _________ distribution.

ix) In symmetrical distribution, the three averages mean, median and mode are _______.

x) If extreme large or small values are changed, values of __________ are not effected.

3.32 Write T for true and F for false against each statement.

i) The sum of deviations of the values from mean is minimum.

ii) Geometric mean is possible only for negative values.

iii) The median divides the data into two halves.

iv) The third quartile is the median.

v) A distribution having only one mode is called uni-modal distribution.

vi) Arithmetic mean depends on all the values of the data.

vii) Mean, Median and Mode in a symmetrical distribution are not equal.

viii) Harmonic Mean can be calculated if any value is zero.

ix) Median is not affected by extreme values.

x) Geometric Mean cannot be calculated if any value is negative.

4 Measures of Dispersion

$0 \leq P(A) \leq 1$

4.1 Introduction

The measure of central tendency does not tell us any thing about the spread of the values in a set, because any two sets with vast difference in magnitude of their variability may have the same central tendency. Look at the following two data sets:

Data set a: 8, 7, 5, 8, 6 Data set b: 1, 4, 7, 10, 12

These two data sets have same mean 6.8 but differ in their variations from the central value. There is more variation in the date set b as compared to the data set a. This illustrates the fact that measure of central tendency is not sufficient.

To give a sensible description of data, a numerical quantity called measure of dispersion or variability that describes the spread of the values in a set of data is required.

Two types of measures of dispersion or variability are defined:

i) Absolute measures ii) Relative measures

The absolute measures are defined in such a way that they have units (meters, grams etc) same as those of the original measurements, whereas the relative measures have no units as these are ratios.

The most common measures of absolute variability are:

a) Range b) Quartile Deviation

c) Mean Deviation d) Variance

e) Standard Deviation

These are also called measures of dispersion or measures of spread.

The relative measures are discussed in article 4.2

4.1.1 Range

The range of n values $Y_1, Y_2, \ldots, Y_n$ is defined as the difference between the largest and smallest observation. If $Y_{(1)}$ is smallest in magnitude and $Y_{(n)}$ is largest in magnitude then range denoted by R, is defined by;

$$R = Y_{(n)} - Y_{(1)} \qquad (4.1)$$

This is very simple measure of variability and only takes into account two most extreme observations.

Example 4.1: Calculate range for the following observations (in cms):

84.2, 87.5, 80.7, 92.4, 91.9, 86.5, 85.4

Solution: Here $Y_{(1)} = 80.7 \text{ and } Y_{(7)} = 92.4$

so $R = 92.4 - 80.7$

$= 11.7cm$

Range for Grouped Data: In case of grouped data, it may be calculated by the following formula:

R = mid value of the highest class – mid value of the lowest class

Example 4.2: The following frequency Distribution gives the weights of 90 cotton bales.

Weights	70 – 74	75 – 79	80 – 84	85 – 89	90 – 94	95 – 99
Frequencies	1	7	17	29	20	16

Find its range.

Solution:

Weights	f	**Class boundary**
70 – 74	1	69.5 – 74.5
75 – 79	7	74.5 – 79.5
80 – 84	17	79.5 – 84.5
85 – 89	29	84.5 – 89.5
90 – 94	20	89.5 – 94.5
95 – 99	16	94.5 – 99.5
Total	90	– –

The mid value of first group is $\frac{69.5 + 74.5}{2} = 72$ and the mid value of last group is $\frac{94.5 + 99.5}{2} = 97$, so

$Y_{(n)} = 97.0 = Y_{(1)} = 72.0$

$R = Y_{(n)} - Y_{(1)} = 97.0 - 72.0 = 25$

Merits of range:

i) It is easy to calculate.

ii) It is a useful measure in small samples.

Demerits of range:

i) It is not based on all the observations.

ii) It depends only upon the extreme observations.

4.1.2 Quartile Deviation

This measure is based on quartiles Q_1 and Q_3 and is denoted by Q.*D.* It is calculated as

$$Q.D = \frac{Q_3 - Q_1}{2} \qquad (4.2)$$

It is also known as semi inter quartile range.

This measure cannot be negative because the upper quartile must be atleast as large as the lower quartile. A small value of quartile deviation indicates a small amount of variability whereas larger values indicate more variability in the data set. It measures half of the difference between the upper and lower quartiles.

Example 4.3: Calculate quartile deviation for the data of the example 3.16.

Solution: $Q_1 = 88.03$ cms, $Q_3 = 94.90$ cms

Therefore, $$Q.D = \frac{94.90 - 88.03}{2}$$

$$= 3.435 \text{ cms}$$

The formula of quartile deviation for grouped data is the same as for ungrouped data. i.e.,

$$Q.D = \frac{Q_3 - Q_1}{2}$$

Example 4.4: Calculate quartile deviation for the data of the example 3.17

Solution: $Q_1 = 92.3750$, $Q_3 = 102.5833$

$$Q.D = \frac{102.5833 - 92.3750}{2} = 5.104 \text{ cm}$$

Example 4.5: For the data given below:

1030, 1590, 1070, 1670, 1110, 1710, 1190, 1720, 1230, 1740, 1310, 1745, 1332, 1775, 1870, 1350, 1430, 1870, 1950 and 1460,

calculate Quartile Deviation and co-efficient of Quartile Deviation.

Solution: Arraying the data

1030, 1070, 1110, 1190, 1230, 1310, 1332, 1350, 1430, 1460, 1590, 1670, 1710, 1720, 1740, 1745, 1775, 1870, 1870, 1950.

Here, $n = 20$

$$Q_1 = \text{Value of}\left(\frac{n+1}{4}\right)\text{th item}$$

$$= \text{Value of}\left(\frac{20+1}{4}\right)\text{th item}$$

$$= \text{Value of (5.25) th item}$$

$$\therefore Q_1 = \text{5th value} + 0.25\,(\text{6th value} - \text{5th value})$$
$$= 1230 + 0.25(1310 - 1230)$$
$$= 1250$$

$$Q_3 = \text{Value of } 3\left(\frac{n+1}{4}\right)\text{th item}$$

$$= \text{Value of } 3\left(\frac{20+1}{4}\right)\text{th item}$$

$$= \text{Value of 15.75 th item}$$

$$\therefore Q_3 = \text{15th value } + 0.75\,(\text{16th value } - \text{ 15th value})$$
$$= 1740 + 0.75(1745 - 1740)$$
$$= 1743.75$$

$$\therefore \text{Q.D} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{1743.75 - 1250}{2} = 246.88$$

$$\text{Co-efficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{1743.75 - 1250}{1.743.75 + 1250} = 0.16$$

Example 4.6: For the following frequency distribution, find quartile deviation and co-efficient of quartile deviation.

Marks	**10 – 20**	**20 – 30**	**30 – 40**	**40 – 50**	**50 – 60**
Frequency	3	8	14	7	4

Solution:

Marks	f	Cumulative frequency (C.F.)
10 – 20	3	3
20 – 30	8	3 + 8 = 11
30 – 40	14	11 + 14 = 25
40 – 50	7	25 + 7 = 32
50 – 60	4	32 + 4 = 36
Total	36	

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$$

$n = 36,\ \frac{n}{4} = \frac{36}{4} = 9$th observation so, $l = 20,\ h=10,\ f=8,\ c=3$

$$\therefore Q_1 = 20 + \frac{10}{8}(\frac{36}{4} - 3) = 27.5,$$

$n = 36,\ \frac{3n}{4} = \frac{3\times36}{4} = 27$th observation so, $l = 40,\ h=10,\ f=7,\ c=25$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

$$\therefore Q_3 = 40 + \frac{10}{7}\left(3.\frac{36}{4} - 25\right) = 42.8$$

$$\therefore \text{ Q.D} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{42.8 - 27.5}{2} = 7.65$$

$$\text{Co - efficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{42.8 - 27.5}{42.8 + 27.5} = 0.22$$

Example 4.7: Find the Semi – Interquartile range and co-efficient of Quartile deviation for the data given below about the ages in a locality.

Ages	20	30	40	50	60	70	80
Frequency	3	61	132	158	140	51	2

Solution:

y_i	f	C.B.	c.f.
20	3	15 – 25	3
30	61	25 – 35	64
40	132	35 – 45	196
50	158	45 – 55	354
60	140	55 – 65	494
70	51	65 – 75	545
80	2	75 – 85	547
	547		

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$$

$$n = 547, \frac{n}{4} = \frac{547}{4} = 136.7 = 137\text{th observation so,}$$

$$l = 35,\ h=10,\ f=132,\ c=64$$

$$Q_1 = 35 + \frac{10}{132}\left(\frac{547}{4} - 64\right) = 40.51$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

$$n = 547, \frac{3n}{4} = \frac{3\times547}{4} = 410.1 = 410\text{th observation so,}$$

$$l = 55,\ h=10,\ f=140,\ c=354$$

$$\therefore Q_3 = 55 + \frac{10}{140}\left(3\times\frac{547}{4} - 354\right) = 59.02$$

$$\text{Semi} - \text{Inter Quartile Range} = \frac{Q_3 - Q_1}{2} = \frac{59.02 - 40.15}{2} = 9.255$$

$$\text{Co-efficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{59.02 - 40.15}{59.02 + 40.15} = \frac{18.87}{99.17} = 0.19$$

Merits of quartile deviation

i) It is easy to calculate.

ii) It is not affected by extreme observations.

Demerits of quartile deviation

i) It is not based on all the observations.

ii) Q.D. will be the same value for all the distributions having the same quartiles.

4.1.3 Mean deviation

It is defined as the mean of the absolute deviations of observations from mean, median or mode. By absolute deviations we mean that we consider all the deviations as positive. It is denoted by M.D. and is calculated as

$$M.D = \frac{\Sigma|Y-M|}{n} \quad \text{(M is mean or median or mode)} \qquad (4.3)$$

Example 4.8: For ungrouped data of the example 3.13, find mean deviation

Solution: Y_i : 88.03, 94.50, 94.90, 95.50, 84.60.

$$\Sigma Y = 88.03 + 94.50 + \ldots. + 84.60 = 457.08$$

$$\overline{Y} = \frac{\Sigma Y_i}{n}$$

$$= \frac{457.08}{5} = 91.416$$

Mean Deviation (M. D) = $\dfrac{\Sigma\left|Y_i - \overline{Y}\right|}{n}$

Where $\left|Y_i - \overline{Y}\right|$ are

$$\begin{aligned}
|88.03 - 91.416| &= 3.386 \\
|94.50 - 91.416| &= 3.084 \\
|94.90 - 91.416| &= 3.484 \\
|95.05 - 91.416| &= 3.634 \\
|64.60 - 91.416| &= 6.816 \\
\Sigma|Y_i - \overline{Y}| &= 20.404
\end{aligned}$$

$$\text{M.D.} = \frac{20.404}{5} = 4.0808\,\text{cm}$$

Dealing with the grouped data, mean deviation is calculated by multiplying the absolute deviations from mean with the corresponding frequencies and then taking the mean i.e.,

$$\text{M.D.} = \frac{\Sigma f_i \left|Y_i - \overline{Y}\right|}{\Sigma f_i}$$

Example 4.9: For the data of the example 3.2, find mean deviation for grouped data.

Solution:

$$\overline{Y} = \frac{\Sigma f_i Y_i}{\Sigma f_i} = 97.8333 \text{ cm}$$

C. I.	Y_i	f_i	$(Y_i - \overline{Y})$	$f_i \left|Y_i - \overline{Y}\right|$
86 – 90	88	6	-9.8333	58.9998
91 – 95	93	4	-4.8333	19.3332
96 – 100	98	10	0.1667	1.6670
101 – 105	103	6	5.1667	31.0002
106 – 110	108	3	10.1667	30.5001
111 – 115	113	1	15.1667	15.1667
Total		30		156.6670

$$\text{Mean deviation (M.D)} = \frac{\Sigma f_i \left|Y_i - \overline{Y}\right|}{\Sigma f_i}$$

$$= \frac{156.6670}{30}$$

$$= 5.2222 cm$$

Mean deviation from median is defined in terms of absolute deviations from median as:

$$\text{M.D.} = \frac{\Sigma \left|Y_i - \text{Median}\right|}{n} \qquad (4.4)$$

The mean deviation from median for the data of the example 3.13 is calculated as:

$\lvert Y_i - \text{Median} \rvert$
$\lvert 88.03 - 94.50 \rvert = 6.47$
$\lvert 94.50 - 94.50 \rvert = 0.0$
$\lvert 94.90 - 94.50 \rvert = 0.40$
$\lvert 95.05 - 94.50 \rvert = 0.55$
$\lvert 84.60 - 94.50 \rvert = 9.90$
$\Sigma \lvert Y_i - \text{median} \rvert = 17.32$

$$\text{So} \quad \text{M.D} = \frac{17.32}{5}$$

$$= 3.464$$

The mean deviation from median for grouped data is:

$$\text{M.D} = \frac{\Sigma f_i \lvert Y_i - \text{Median} \rvert}{\Sigma f_i}$$

The calculations for mean deviation from median taking median equals to 98 for the example 4.9 are:

Y_i	f_i	$Y_i - 98$	$\lvert Y_i - 98 \rvert$	$f_i \lvert Y_i - 98 \rvert$
88	6	-10	10	60
93	4	-5	5	20
98	10	0	0	0
103	6	-5	5	30
108	3	-10	10	30
113	1	-15	15	15
Total	30	---	---	155

$$\text{M.D} = \frac{155}{30} = 5.167 \text{ cm}$$

Properties of mean deviation

i) M .D from median is less than any other value i.e.,

$$\frac{\Sigma \lvert Y_i - \text{median} \rvert}{n} \text{ is least}$$

ii) It is always greater than or equal to zero i.e.,

$$\text{M.D} \geq 0$$

iii) For symmetrical distributions, the following relation holds

$$\text{M.D} = \frac{4}{5}\sigma \qquad (4.5)$$

Where σ is the standard deviation.

Merits of mean deviation

i) It is easy to calculate.

iii) It is based on all the observations.

Demerits of mean deviation

i) It is affected by the extreme values.

ii) It is not readily capable of mathematical development.

iii) It does not take into account the negative signs of the deviations from some average.

Example 4.10 Find mean deviation from median for the following frequency distrib ution.

Ages (Years)	5-10	10-15	15-20	20-25
Frequency	10	20	30	15

Also calculate the co-efficient of mean deviation.

Solution:

Ages	f	$c.f.$	Y_i	$\lvert Y_i - \bar{Y}\rvert$	$f_i \lvert Y_i - \bar{Y}\rvert$
5 – 10	10	10	7.5	8.75	87.5
10 – 15	20	30	12.5	3.75	75
15 – 20	30	60	17.5	1.25	37.5
20 – 25	15	75	22.5	6.25	93.75
Total:	75	-	-	-	293.75

$$\tilde{Y} = l + \frac{h}{f}\left(\frac{n}{2} - c\right)$$

$$\tilde{Y} = 15 + \frac{5(37.5 - 30)}{30} = 16.25$$

$$\text{Mean Deviation from median} = \frac{\Sigma f\left|Y - \tilde{Y}\right|}{\Sigma f} = \frac{293.75}{75} = 3.92$$

$$\text{Co-efficient of mean deviation} = \frac{\text{Mean Deviation from median}}{\text{Median}}$$

$$= \frac{3.92}{16.25} = 0.24$$

4.1.4 The Variance

Variance of the observations is defined as mean of squares of deviations of all the observations from their mean . When it is calculated from the population the variance is called population variance and is denoted by σ^2 and when it is calculated from the sample, based on n values $Y_{1,}\ Y_2 , \ldots\ldots Y_n$ is called sample variance. The

Population variance σ^2 is defined as $\sigma^2 = \dfrac{\sum(y_i - \mu)^2}{\text{N}}$

The sample variance S^2 for un-grouped data is defined as:

$$S^2 = \frac{\Sigma(y - \bar{y})^2}{n} \qquad .(4.6)$$

Short formula for variance is given by

$$S^2 = \frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2 \qquad (4.7)$$

For a frequency distribution, the sample variance S^2 is defined as:

$$S^2 = \frac{\Sigma f_i \ (y_i - \bar{y})^2}{\Sigma f_i} \qquad .(4.8)$$

Short formula for grouped data is given by

$$S^2 = \frac{\Sigma f\, y^2}{\Sigma f} - \left(\frac{\Sigma f\, y}{\Sigma f}\right)^2 \qquad (4.9)$$

If A is an arbitrary value such that $D = y - A$, S^2 for ungrouped data is given by

$$S^2 = \frac{\Sigma D^2}{n} - \left(\frac{\Sigma D}{n}\right)^2 \qquad (4.10)$$

for grouped data

$$S^2 = \frac{\Sigma fD^2}{\Sigma f} - \left(\frac{\Sigma fD}{\Sigma f}\right)^2 \qquad (4.11)$$

When data are grouped into a frequency distribution with equal class intervals of size h and $u = \frac{y - A}{h}$, then

$$S^2 = h^2\left[\frac{\Sigma fu^2}{\Sigma f} - \left(\frac{\Sigma fu}{\Sigma f}\right)^2\right] \qquad (4.12)$$

Merits and Demerits of Variance.

The variance is based on all the observations of a series. It is easy to calculate and simple to understand. It is affected by extreme values.

4.1.5 Standard Deviation

The standard deviation is defined as the positive square root of the mean of the squares of the deviations of values from their mean. In other words, standard deviation is a positive square root of variance. It is denoted by S and is given by

For ungrouped data

$$S = \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{n}} \qquad (4.13)$$

In short cut method

$$S = \sqrt{\frac{\Sigma Y^2}{n} - \left(\frac{\Sigma Y}{n}\right)^2} \qquad (4.14)$$

For frequency distribution

$$S = \sqrt{\frac{\Sigma f(Y - \bar{Y})^2}{\Sigma f}} \qquad (4.15)$$

In short cut method

$$S = \sqrt{\frac{\Sigma f Y^2}{\Sigma f} - \left(\frac{\Sigma f Y}{\Sigma f}\right)^2} \qquad (4.16)$$

If $D = Y - A$, the deviations of Y from any arbitrary value A then standard deviation is

$$S = \sqrt{\frac{\Sigma D^2}{n} - \left(\frac{\Sigma D}{n}\right)^2} \qquad (4.17)$$

For frequency distribution, the formula becomes

$$S = \sqrt{\frac{\Sigma f\, D^2}{\Sigma f} - \left(\frac{\Sigma f D}{\Sigma f}\right)^2} \qquad (4.18)$$

For coding variable $u = \frac{y - A}{h}$, the formula becomes

$$S = h\sqrt{\frac{\Sigma f\, u^2}{\Sigma f} - \left(\frac{\Sigma f\, u}{\Sigma f}\right)^2} \qquad (4.19)$$

Example 4.11: Calculate variance and standard deviation for the data: 3,6,2,1,7,5.

Solution:

Y	$Y - \overline{Y}$	$(Y - \overline{Y})^2$
3	-1	1
6	2	4
2	-2	4
1	-3	9
7	3	9
5	1	1
24	0	28

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{24}{6} = 4$$

$$\text{Variance} = S^2 = \frac{\Sigma(Y - \overline{Y})^2}{n} = \frac{28}{6} = 4.67$$

$$\text{Standard deviation} = S = \sqrt{\frac{\Sigma(Y - \overline{Y})^2}{n}}$$

$$= \sqrt{4.67}$$

$$= 2.16$$

Example 4.12: Calculate variance and standard deviation from the following frequency distribution.

Wages	30-35	35-40	40-45	45-50	50-55	55-60
Frequency	12	18	29	32	16	8

Solution:

Wages	f	y	fy	$f(y-\bar{y})^2$
30 – 35	12	32.5	390	1723
35 – 40	18	37.5	675	882
40 – 45	29	42.5	1232.5	116
45 – 50	32	47.5	1520	288
55 – 60	16	52.5	840	1024
60 – 65	8	57.5	460	1352
Total:	115		5117.5	5390

$$\bar{Y} = \frac{\Sigma f Y}{\Sigma f} = \frac{5117.5}{115}$$

$$\text{variance}\,(S)^2 = \frac{\Sigma f\,(y-\bar{y})^2}{\Sigma f}$$

$$= \frac{5390}{115} = 46.87$$

$$S = \sqrt{46.87}$$

$$= 6.846$$

Example 4.13: Calculate variance and standard deviation by using any provisional mean from the data: 3,5,7,13,15,17,23,27.

Solution:

Y	$D = y - 15$	D^2
3	-12	144
5	-10	100
7	-8	64
13	-2	4
15	0	0
17	2	4
23	8	64
27	12	144
Total	-10	524

$$\text{Variance}\quad (S^2) = \frac{\sum D^2}{n} - \left(\frac{\sum D}{n}\right)^2$$

$$= \frac{524}{8} - \left(\frac{-10}{2}\right)^2$$

$$= 65.6 - 1.562$$

$$= 63.938$$

$$\text{Standard Deviation } (S) = \sqrt{\frac{\sum D^2}{n} - \left(\frac{\sum D}{n}\right)^2}$$

$$= \sqrt{63.938} = 7.99$$

PROPERTIES OF THE VARIANCE AND STANDARD DEVIATION

1. The variance and standard deviation of a constant is zero. If a is a constant, then

   var (a) = 0

   S.D. (a) = 0

2. The variance and standard deviation are independent of origin.

   var $(y + a)$ = var (y)

var $(y - a)$ = var (y)

and S.D. $(y + a)$ = S.D. (y)

S.D. $(y - a)$ = S.D. (y)

3. When all the values are multiplied with a constant, the variance of the values is multiplied by square of the constant and their standard deviation is multiplied by the constant i.e.,

var $(ay) = a^2$, var (y)

var $(y/a) = (1/a^2)$ var (y)

and S.D $(ay) = |a|$ S.D (y)

S.D $(y/a) = |1/a|$ S.D (y)

4. The variance/standard deviation of the sum or difference of two independent variables is the sum of their respective variances/standard deviation for independent variables x and y.

var $(x + y)$ = var (x) + var (y)

var $(x - y)$ = var (x) + var (y)

and S.D. $(x + y)$ = S.D. (x) + S.D. (y)

S.D. $(x - y)$ = S.D. (x) + S.D. (y)

5. If sets of data consisting $n_1, n_2, \ldots, n_k$ values having corresponding means y_1, $\bar{y}_2, \ldots, \bar{y}_k$ and variances $S_1^2, S_2^2, \ldots, S_k^2$, the variance of combined set of data is given by

$$S_c^2 = \frac{\sum n_i \left[S_i^2 + (\bar{y}_i - \bar{y}_c)^2\right]}{\sum n_i} \qquad (4.20)$$

where, $\bar{y}_c$[1] $= \dfrac{n_1\,\bar{y}_1 + n_1\,\bar{y}_2 + \ldots + n_k\,\bar{y}_k}{n_1 + n_2 + \ldots + n_k}$

4.2 Co-efficient Of Variation and Other Relative Measures

The most important of all the relative measures of dispersion is co-efficient of variation. Co-efficien of variation is a relative measure of dispersion and independent of units of measurement and expressed in percentage. It is used to compare the variability of different sets of data. The group which has lower value of

[1] $\bar{y}_c$ stands for combined mean notation

co-efficient of variation is comparatively more consistent. The co-efficient of variation is defined as:

$$\text{Co-efficient of variation} = \text{C.V.} = \frac{S}{\overline{y}} \times 100 \text{ (for sample)}$$

$$\text{C.V} = \frac{\sigma}{\mu} \times 100 \text{ (for population)}$$

As it is a ratio of the two quantities with the same units, so is a dimensionless quantity i.e., for the same data whether it is in millimeters, centimeters or meters, etc. The co-efficient of variation remains the same and has no unit.

As the co-efficient of variation expresses variability relative to the mean, it is called a measure of relative variability or relative dispersion.

The co-efficient of variability for the example 4.7 is given by

$$\text{C.V} = \frac{S}{\overline{y}} \times 100$$

$$S = 6.8837$$

$$\overline{y} = 97.8833$$

$$\text{so} \quad \text{C.V.} = \frac{6.8837}{97.8833} \times 100 = 7.04\%$$

Large value of C.V indicates that the observations have much spread relative to the size of the mean and vice versa.

This measure can be used to compare the variability of two or more populations. It will take the same value for two or more populations if in each population, the standard deviation is directly proportional to the mean. In such situation, we say that two or more populations are consistent. For example, to compare the consistency of two methods, each method was tried on 16 soil samples and the corresponding results obtained are:

	Method I	Method II
$\overline{y}$	15.0	10.5
S	1.4	1.1
CV(%)	9.3	9.5

The CVs being almost equal indicate that both the methods are equally reliable. We actually do not compare the standard deviations, since the means will apparently be widely different.

Some other relative measures of dispersion are:

a) Co-efficient of Range = $\dfrac{Y_{(n)} - Y_{(1)}}{Y_{(n)} + Y_{(1)}}$ (4.21)

b) Co-efficient of Q.D. = $\dfrac{Q_3 - Q_1}{Q_3 - Q_1}$ (4.22)

c) Mean co-efficient of dispersion = $\dfrac{\text{M.D}}{\bar{y}}$ (4.23)

d) Median co-efficient of dispersion = $\dfrac{\text{M.D}}{\text{Median}}$ (4.24)

e) Co-efficient of standard deviation = $\dfrac{S}{\bar{y}}$ (4.25)

Example 4.14: Calculate co-efficient of variation and co-efficient of standard deviation from the following frequency distribution.

y	0	1	2	3	4
f	17	9	6	5	3

Solution:

y	f	fy	$y - \bar{y}$	$f(y - \bar{y})^2$
0	17	0	1.2	24.48
1	9	9	0.2	0.36
2	6	12	0.8	3.81
3	5	15	1.8	16.2
4	3	12	2.8	23.52
Total	40	48	--	68.4

$$\bar{y} = \frac{\sum fy}{\sum f} = \frac{48}{40} = 1.2$$

$$S = \sqrt{\frac{\sum f(y - \bar{y})^2}{\sum f}}$$

$$= \sqrt{\frac{68.4}{40}} = 1.307$$

$$\text{Co-efficient of S.D} = \frac{\text{Standard deviation}}{\text{Mean}} = \frac{1.307}{1.2} = 1.089$$

$$\text{Co-efficient of variation} = \frac{\text{S.D.}}{\bar{y}} \times 100 = \frac{1.307}{1.2} \times 100 = 108.92\%$$

Example 4.15: Given the following results, find the combined co-efficient of variation.

$$n_1 = 100 \qquad S_1 = 2.4 \qquad \bar{y}_1 = 12.6$$

$$n_2 = 120 \qquad S_2 = 4.2 \qquad \bar{y}_2 = 15.8$$

$$n_3 = 150 \qquad S_3 = 3.7 \qquad \bar{y}_3 = 10.5$$

Solution: Combined mean is given by

$$\bar{y}_c = \frac{n_1\bar{y}_1 + n_2\bar{y}_2 + n_3\bar{y}_3}{n_1 + n_2 + n_3} = \frac{100(12.5) + 120(15.8) + 150(10.5)}{100+120+150}$$

$$= 12.76$$

Combined variance is given by

$$S_c^{\,2} = \frac{n_1\left[S_1^2 + (\bar{y}_1 - \bar{y}_c)^2\right] + n_2\left[S_2^2 + (\bar{y}_2 - \bar{y}_c)^2\right] + n_3\left[S_3^2 + (\bar{y}_3 - \bar{y}_c)^2\right]}{n_1 + n_2 + n_3}$$

$$= \frac{100\left[5.76 + (12.5 - 12.76)^2\right] + 120\left[17.64 + (15.8 - 12.76)^2\right] + 150\left[13.69 + (10.5 - 12.76)^2\right]}{n_1 + n_2 + n_3}$$

$$= \frac{6628.19}{370}$$

$$S^2{}_c = 17.9140\,, \quad S_c = 4.23255$$

$$\therefore \; C.Vc = \frac{S_c}{\bar{Y}_c} \times 100$$

$$= \frac{4.2325}{12.76} \times 100 = 33.17\%$$

4.3 Moments

The measures of location alongwtih measures of variability are useful to describe a data set but fail to tell anything about the shape of the distribution. For this purpose, we need to define certain other measures. Some important measures about the shape of the distribution depend upon what we call moments. These measures are discussed under skewness and kurtosis.

4.3.1 Moments about mean

The moments about mean are the mean of deviations from the mean after raising them to integer powers. The rth population moment about the mean is denoted by μ_r is defined as:

For ungrouped data

$$\mu_r = \frac{\sum_{i=1}^{N}(y_i-\mu)^r}{N} \qquad (4.26)$$

Where, $r = 1,2, \ldots.$

and the corresponding sample moments about mean $\bar{y}$, denoted by m_r is given by

$$m_r = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^r}{n} \qquad (4.27)$$

So, the first moment m_1 is given by

$$m_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^1}{n}$$

m_1 is always zero as the numerator $\sum_{i=1}^{n}(y_i-\bar{y})=0$

Second moment m_2 is given by: $m_2 = \dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}$

This is the same as variance S^2.

Third moment m_3 is given by: $m_3 = \dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^3}{n}$

and fourth moment m_4 is given by: $m_4 = \dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^4}{n}$

If the data are grouped then the *r*th sample moment about the mean $\bar{y}$ is defined as:

$$m_r = \frac{\sum_{i=1}^{n} f_i(y_i - \bar{y})^r}{n}$$

$$\text{where } n = \sum_{i=1}^{n} f_i \qquad (4.28)$$

$$\text{so } m_1 = \frac{\sum_{i=1}^{n} f_i(y_i - \bar{y})^1}{n} = \frac{\sum f(y-\bar{y})}{\sum f} = 0$$

$$m_2 = \frac{\sum_{i=1}^{n} f_i(y_i - \bar{y})^2}{n} = \frac{\sum f(y-\bar{y})^2}{\sum f}$$

$$m_3 = \frac{\sum_{i=1}^{n} f_i(y_i - \bar{y})^3}{n} = \frac{\sum f(y-\bar{y})^3}{\sum f}$$

$$m_4 = \frac{\sum_{i=1}^{n} f_i(y_i - \bar{y})^4}{n} = \frac{\sum f(y-\bar{y})^4}{\sum f}$$

Example 4.16: Calculate first four moments about the mean for the following set of marks obtained in the examination.

45, 32, 37, 46, 39, 36, 41, 48 and 36.

Solution:

y	$y-\bar{y}$	$(y-\bar{y})^2$	$(y-\bar{y})^3$	$(y-\bar{y})^4$
32	-8	64	512	4096
36	-4	16	64	256
36	-4	16	64	256
37	-3	9	27	81
39	-1	1	1	1
41	1	1	1	1
45	5	25	125	625
46	6	36	216	1296
48	8	64	512	4096
360	0	232	186	10706

$$\bar{y} = \frac{\sum y}{n} = \frac{360}{9} = 40$$

$$m_1 = \frac{\sum (y-\bar{y})^1}{n} = \frac{0}{9} = 0$$

$$m_2 = \frac{\sum (y-\bar{y})^2}{n} = \frac{232}{9} = 25.78$$

$$m_3 = \frac{\sum (y-\bar{y})^3}{n} = \frac{186}{9} = 20.67$$

$$m_4 = \frac{\sum (y-\bar{y})^4}{n} = \frac{10706}{9} = 1189.56$$

Example 4.17: Find first four moments about the mean for the following data.

y	22	27	32	37	42	47	52
f	1	4	8	11	15	9	2

Solution:

y	f	fy	$y - \bar{y}$	$f(y - \bar{y})$	$f(y - \bar{y})^2$	$f(y - \bar{y})^3$	$f(y - \bar{y})^4$
22	1	22	-17	-17	289	4913	83521
27	4	108	-12	-48	576	6912	82944
32	8	256	-7	-56	392	2744	19208
37	11	407	-2	-22	44	88	176
42	15	630	3	45	135	405	1215
47	9	423	8	72	576	4608	36864
52	2	104	13	26	338	4394	57122
Total	50	1950	–	0	2350	5250	281050

$$\bar{y} = \frac{\Sigma fy}{\Sigma f} = \frac{1950}{50} = 39$$

$$m_1 = \frac{\Sigma f(y-\bar{y})}{\Sigma f} = 0$$

$$m_2 = \frac{\Sigma f(y-\bar{y})^2}{\Sigma f} = \frac{2350}{50} = 47$$

$$m_3 = \frac{\sum f(y-\bar{y})^3}{\sum f} = \frac{-5250}{50} = -105$$

$$m_4 = \frac{\sum f(y-\bar{y})^4}{\sum f} = \frac{281050}{50} = 5621$$

4.3.2 Moment about an arbitrary value

The *r*th sample moment about any arbitrary origin *a* denoted by m'_r is defined as:

$$m'_r = \frac{\sum_{i=1}^{n}(y_i - a)^r}{n} = \frac{\sum_{i=1}^{n} D_i^r}{n} \tag{4.29}$$

where, $D_i = (y_i - a)$

so

$$m'_1 = \frac{\sum_{i=1}^{n}(y_i - a)}{n} = \frac{\sum_{i=1}^{n} D_i}{n} \tag{4.30}$$

$$m'_2 = \frac{\sum_{i=1}^{n}(y_i - a)^2}{n} = \frac{\sum_{1=1}^{n} D_1^2}{n} \tag{4.31}$$

$$m'_3 = \frac{\sum_{i=1}^{n}(y_i - a)^3}{n} = \frac{\sum_{i=1}^{n} D_i^3}{n} \tag{4.32}$$

$$m'_4 = \frac{\sum_{i=1}^{n}(y_i - a)^4}{n} = \frac{\sum_{i=1}^{n} D_i^4}{n} \tag{4.33}$$

The moments about the mean are usually called central moments and the moments about any arbitrary origin *a* are called non-central moments or raw moments.

The *r*th sample moment for grouped data about any arbitrary origin *a* denoted by m'_r is defined as:

$$m'_r = \frac{\sum_{i=1}^{n} f_i (y_i - a)^r}{\sum f} = \frac{\sum_{i=1}^{n} f_i D_i^r}{\sum f} \qquad (4.34)$$

$$m'_1 = \frac{\sum_{i=1}^{n} f_i (y_i - a)}{\sum f} = \frac{\sum_{i=1}^{n} f_i D_i}{\sum f} \qquad (4.35)$$

$$m'_2 = \frac{\sum_{i=1}^{n} f_i (y_i - a)^2}{\sum f} = \frac{\sum_{i=1}^{n} f_i D_i^2}{\sum f} \qquad (4.36)$$

$$m'_3 = \frac{\sum_{i=1}^{n} f_i (y_i - a)^3}{\sum f} = \frac{\sum_{i=1}^{n} f_i D_i^3}{\sum f} \qquad (4.37)$$

$$m'_4 = \frac{\sum_{i=1}^{n} f_i (y_i - a)^4}{\sum f} = \frac{\sum_{i=1}^{n} f_i D_i^4}{\sum f} \qquad (4.38)$$

$$m_1 = m'_{1,} - m'_1 = 0 \text{ .... } (4.38.a)$$

$$m_2 = m'_2 - (m'_1)^2 \text{ .... } (4.38.b)$$

$$m_3 = m'_3 - 3m'_2\, m'_1 + 2(m'_1)^3 \text{........................}(4.38.c)$$

$$m_4 = m'_4 - 4m'_3\, m'_1 + 6m'_2(m'_1)^2 - 3(m'_1)^4 \text{ ... } (4.38.d)$$

Example 4.18: Find first four moments about the mean for ungrouped data of the example 2.1.

Example 4.18: The moments can be calculated directly by using relation (4.27) or relation 4.29 by selecting an arbitrary origin *a*. We calculate moments about origin *a* taking $a = 98$ and then calculate moments about mean by using the relations 4.40 to 4.43.

Y_i	$D_i = Y_i = 98$	D_i^2	D_i^3	D_i^4
87	-11	121	-1331	14641
91	-7	49	-343	2401
89	-9	81	-729	6561
88	-10	100	-1000	10000
89	-9	81	-729	6561
91	-7	49	-343	2401
87	-11	121	-1331	14641

92	-6	36	-216	1296
90	-8	64	-512	4096
98	0	0	0	0
95	-3	9	-27	81
97	-1	1	-1	1
96	-2	4	-8	16
100	2	4	8	16
101	3	9	27	81
96	-2	4	-8	16
98	0	0	0	0
99	1	1	1	1
98	0	0	0	0
100	2	4	8	16
102	4	16	64	256
99	1	1	1	1
101	3	9	27	81
105	7	49	343	2401
103	5	25	125	625
107	9	81	729	6561
105	7	49	343	2401
106	8	64	512	4096
107	9	81	729	6561
112	14	196	2744	38416
12929	-11	1309	-917	124225

$$\because m'_r = \frac{\sum D_i^r}{n}$$

$$\Rightarrow \quad m'_1 = \frac{\sum D_i}{n}$$

$$m'_1 = -11/30 = -0.3667$$

$$m'_2 = \sum D_i^2/n$$

$$= 1309/30 = 43.6333$$

$$m'_3 = \sum D_i^3/n$$

$$= -917/30 = -30.5667$$

$$m'_4 = \sum D_i^4/n$$

$$= 124225/30 = 4140.8333$$

The moments about mean are given by

$$m_1 = m'_1 - m'_1$$
$$= 0$$
$$m_2 = m'_2 - (m'_1)^2$$
$$= 43.6333 - (-0.3667)^2$$
$$= 43.6333 - 0.1345$$
$$= 43.4988$$
$$m_3 = m'_3 - 3m'_2\ m'_1 + 2(m'_1)^3$$
$$= -30.5667 - 3(43.6333)(-0.3667) + 2(-0.3667)^3$$
$$= -30.5667 + 48.0010 - 0.0989$$
$$= 17.3354$$
$$m_4 = m'_4 - 4m'_3\ m'_1 + 6m'_2\ (m'_1)^2 - 3(m'_1)^4$$
$$= 4140.8333 - 4(-30.5667)(-0.3667)$$
$$+ 6(43.6333)(-0.3667)^2 - 3(-0.3667)^4$$
$$= 4139.5 - 44.8352 + 35.2039 - 0.0542$$
$$= 4131.1478$$

4.3.3 Moments for grouped data

The moments for grouped data about an arbitrary origin with equal class interval h may be written as:

$$m'_r = h^r \frac{\sum_{i=1}^{k} f_i \left(\frac{y_i - a}{h} \right)^r}{n} = h^r \frac{\sum_{i=1}^{k} f_i\, u_i^r}{n} \qquad (4.39)$$

where $u_i = \dfrac{y_i - a}{h}$

The moments about an arbitrary origin and moments about mean have the following relationships.

$$m_1 = m'_1 - m'_1 \quad = \quad 0 \qquad (4.40)$$
$$m_2 = m'_2 - (m'_1)^2 \qquad (4.41)$$

$$\text{Co-efficient of skewness} = \frac{\text{Mean - Mode}}{\text{Standard deviation}}$$

$$= \frac{67.45 - 67.35}{2.92} = 0.034$$

Example 4.21: The weight of 38 male students at a university are given in the following frequency table:

Weight	118-126	127-135	136-144	145-153	154-162	163-171
f	3	5	9	12	5	4

Calculate Bowley's co-efficient of skewness.

Solution:

Weights	f	C.B.	c.f.
118 – 126	3	117.5 – 126.5	3
127 – 135	5	126.5 – 135.5	8
136 – 144	9	135.5 – 144.5	17
145 – 153	12	144.5 – 153.5	29
154 – 162	5	153.5 – 162.5	34
163 – 171	4	162.5 – 171.5	38
Total:	38	- - -	-

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$$

$$= 135.5 + \frac{9}{12}\left(\frac{38}{4} - 8\right)$$

$$= 136.62$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

$$= 144.5 + \frac{9}{12}\left(3 \times \frac{38}{4} - 17\right)$$

$$= 153.13$$

$$\text{Median} = Q_2 = l + \frac{h}{f}\left(\frac{2n}{4} - c\right)$$

$$= 144.5 + \frac{9}{12}\left(2 \times \frac{38}{4} - 17\right)$$

$$= 146$$

$$\therefore \text{Bowly's co-efficient of skewness} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

$$= \frac{153.13 + 136.62 - 2(146)}{153.13 - 137}$$

$$= \frac{-2.25}{16.13}$$

$$= -0.139$$

4.5.1 Kurtosis

The word kurtosis is used to indicate the length of the tails and peakedness of symmetrical distributions. Symmetrical distributions may be platykurtic, mesokurtic (normal) or leptokurtic.

The mesokurtic is the usual normal distribution. Leptokurtic is more peaked and has many values around the mean and in the tails away from the mean whereas platykurtic is bit flat and has more values between the mean and tails as compared to the mesokurtic (normal) distribution. Figure 4.3 shows these shapes of distributions.

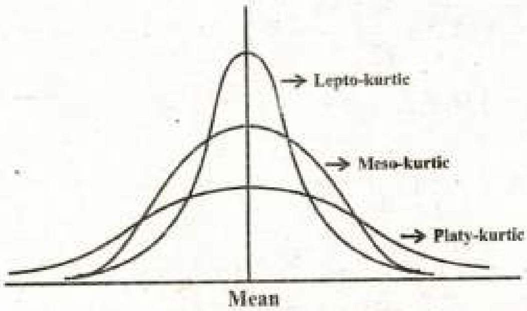


Figure 4.3: Mesokurtic, platykurtic and leptokurtic distributions

One of the numerical measures used to know about the symmetry of a distribution is $\sqrt{\beta_1}$ (β_1 is a Greek letter read as beta one) and is defined as:

$$\sqrt{\beta_1} = \frac{\mu_3}{\sqrt{(\mu_2)^3}} \qquad (4.46)$$

It is dimensionless quantity. If $\sqrt{\beta_1} = 0$, the distribution is symmetrical. If $\sqrt{\beta_1} < 0$, the distribution is negatively skewed and for $\sqrt{\beta_1} > 0$, the distribution is positively skewed. The population parameters are estimated from the corresponding sample statistic. The estimate of $\sqrt{\beta_1}$ is denoted by $\sqrt{b_1}$ and is defined as:

$$\sqrt{b_1} = \frac{m_3}{\sqrt{m_2^3}} \qquad (4.47)$$

It should be noted that these sample statistics tell us only about the particular data set under consideration and not for the whole population.

Some other measures of skewness are:

a) **Karl Pearson's first co-efficient of skewness**

$$\frac{\text{Mean - Mode}}{S} \qquad (4.48)$$

b) **Karl Pearson's second co-efficient of skewness**

$$\frac{3(\text{Mean - Median})}{S} \qquad (4.49)$$

These coefficients are pure numbers and these are zero for symmetrical distributions, negative for negatively skewed distributions and positive for positively skewed distributions.

c) **Bowley's coefficient of skewness based on quartiles**

$$S_k = \frac{\text{Q}_1 + \text{Q}_3 - 2\,\text{median}}{\text{Q}_3 - \text{Q}_1} \qquad (4.50)$$

It is a pure number and lies between –1 and +1. For symmetrical distributions its value is zero.

The skewness of the distribution of a data set can easily be seen by drawing histogram or frequency curve.

Example 4.20: The heights of 100 college students measured to nearest inch are given in the following table:

Height	60-62	63-65	66-68	69-71	72-74
f	5	18	42	27	8

Calculate co-efficient of skewness.

Solution:

Heights	f	y	$D = y - 67$	fD	fD^2	*C.B.*	*c.f.*
60 – 62	5	61	-6	-30	180	59.5 – 62.5	5
63 – 65	18	64	-3	-54	162	62.5 – 65.5	23
66 – 68	42	67	0	0	0	65.5 – 68.5	65
69 – 71	27	70	3	81	243	68.5 – 71.5	92
72 – 74	8	73	6	48	288	71.5 – 74.5	100
Total	**100**			**45**	**873**		

$$\text{Mean} = A + \frac{\sum f D}{\sum f}$$

$$= 67 + \frac{45}{100} = 67.45 \text{ inches}$$

$$\text{Standard deviation} = \sqrt{\frac{\sum f D^2}{\sum f} - \left(\frac{\sum f D}{\sum f}\right)^2}$$

$$= \sqrt{\frac{873}{100} - \left(\frac{45}{100}\right)^2} = 2.92 \text{ inches}$$

$$\text{Mode} = \hat{X} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

$$= 65.5 + \frac{(42 - 18)}{(42 - 18) + (42 - 27)} \times 3$$

$$= 67.35 \text{ inches}$$

$$m'_1 = \frac{\sum f\,y}{\sum f} \qquad m'_2 = \frac{\sum f\,y^2}{\sum f}$$

$$m'_3 = \frac{\sum f\,y^3}{\sum f} \qquad m'_4 = \frac{\sum f\,y^4}{\sum f}$$

4.4 Sheppad's Correction for Grouping Error

In case of grouped data, we proceed to calculate first four moments by replacing all the members of a class by the mid value of the respective class. The choice of class boundaries and the mid values, affects the values of our approximation to the first four moments. The first and third moments are not affected that much as the second and fourth moments because, in case of second and fourth moments all the deviations become positive and the grouping error accumulates. Sheppard has suggested the following corrections for second and fourth moments in case of grouped data where the frequency curve of the grouped data approaches to the base line gradually and slowly at each end of the distribution.

$$\text{corrected } m_2 = m_2 - \frac{h^2}{12} \qquad (4.44)$$

$$\text{corrected } m_4 = m_4 - \frac{h^2}{2} m_2 + \frac{7}{240} h^4 \qquad (4.45)$$

For the example 4.19 corrected m_2 and m_4 are:

$$\text{corrected } m_2 = 45.8055 - \frac{5^2}{12} = 43.722$$

$$\text{corrected } m_4 = 4917.3692 - \frac{5^2}{2}(45.8055) + \frac{7}{240}(5^4) = 4363.0296$$

4.5 Skewness

The word skewness means lack of symmetry of a distribution. A symmetrical distribution is one in which mean, median and mode are identical and the portion of frequency polygon to the left of the mean is the mirror image of the portion to the right of the mean. If a distribution is not symmetrical, it is called skewed or asymmetrical.

The figure 4.1 shows the three types of distributions, i.e., the symmetrical distribution in figure 4.1(a), positively skewed distribution in figure 4.1(b) and negatively skewed distribution in figure 4.1(c). A positively skewed distribution is one whose tail extends to the right hand side and a negatively skewed distribution has longer tail towards left hand side. The positions of mean, median and mode are shown in figure 4.1

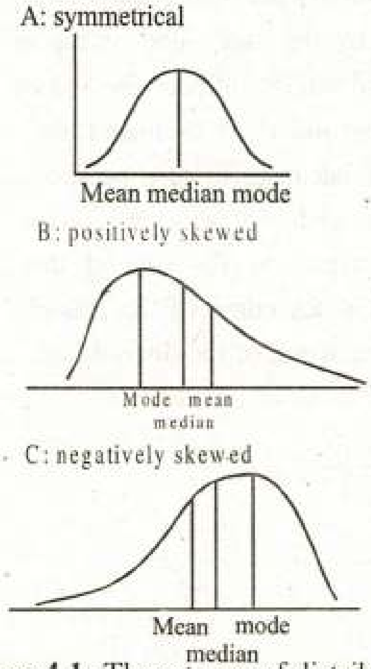


Figure 4.1: Three types of distributions.

The skewness may be very extreme and in such a case these are called J-shaped distributions. This is shown in figure 4.2.

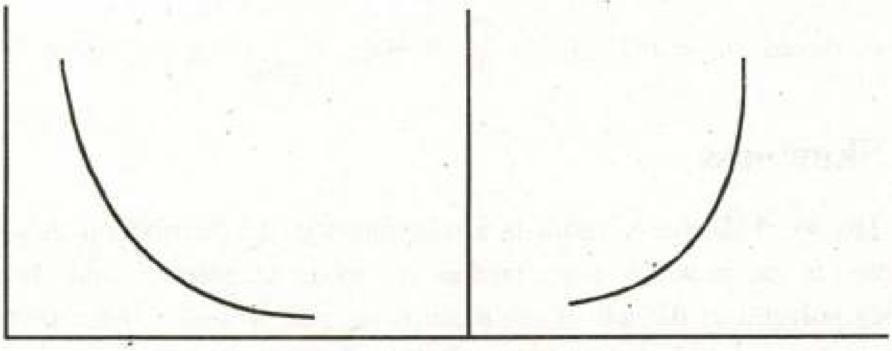


Figure 4.2: J-shaped distributions.

$$m_3 = m'_3 - 3m'_2\, m'_1 + 2(m'_1)^3 \tag{4.42}$$

$$m_4 = m'_4 - 4m'_3\, m'_1 + 6m'_2\, (m'_1)^2 - 3(m'_1)^4 \tag{4.43}$$

Moments about mean are calculated by using the above relations in which one needs to calculate moments about mean the arbitrary origin first or these can be directly calculated by using formula moment about mean.

Example 4.19: Find first four moments about an arbitrary origin $a = 98$ for grouped data of the example 3.2

Also find moments about mean by using the moments about origin.

Solution: As width of classes are equal, we use the formula involving h. First the moments about origin $a = 98$ are calculated as follows:

y_i	f_i	$u_i = (y_i\text{-}98)/5$	$f_i u_i$	$f_i u_i^2$	$f_i u_i^3$	$f_i u_i^4$
88	6	-2	-12	24	-48	96
93	4	-1	-4	4	-4	4
98	10	0	0	0	0	0
103	6	1	6	6	6	6
108	3	2	6	12	24	48
113	1	3	3	9	27	81
Total	30	---	-1	55	5	235

We know that:

$$m'_r = \frac{h^r \,\Sigma f_i\, u_i^r}{n}$$

here $h = 5$, $n = \sum f_i = 30$

$$m'_1 = h(\sum f_i\, u_i)/n$$

$$= 5(-1)/30 = -0.1667$$

$$m'_2 = h^2 \left(\sum f_i\, u_i^2\right)/n$$

$$= 5^2\,(55)/30 = 45.8333$$

$$m'_3 = h^3 \left(\sum f_i\, u_i^3\right)/n$$

$$= 5^3\,(5)/30 = 20.8333$$

$$m'_4 = h^4 \left(\sum f_i\, u_i^4\right)/n$$

$$= 5^4\,(235)/30 = 4895.8333$$

The moments about the mean are

$$m_1 = m'_1 - m'_1 = 0$$

$$m_2 = m'_2 - (m'_1)^2$$

$$= 45.8333-(-0.1667)^2$$

$$= 45.8055$$

$$m_3 = m'_3 - 3m'_2 \; m'_1 + 2(m'_1)^3$$

$$= 20.8333-3(45.8333)(-0.1667)+2(-0.1667)^3$$

$$= 20.8333+22.9167-0.0093$$

$$= 43.7407$$

$$m_4 = m'_4 - 4m'_3 \; m'_1 + 6m'_2 \; (m'_1)^2 - 3(m'_1)^4$$

$$= 4895.8333 - 4(-20.8333)(-0.1667) + 6(45.8333)(-0.1667)^2 - 3(-0.1667)^4$$

$$= 4895.8333+13.8916+7.6419 + 0.0023$$

$$= 4917.3692$$

The first four moments calculated from the same data in ungrouped form and grouped form are slightly different. This is because of the assumption that each observation in a class is equal to mid point of that class while grouping the data.

4.3.4 Moment about zero

If the variable y assumes n values $y_1, y_2, y_3, \ldots, y_n$ then rth moment about zero can be obtained by taking a = 0, so, for relation 4.29

$$m'_r = \frac{\sum y^r}{n}$$

Putting r = 1,2,3 and 4 we get

$$m'_1 = \frac{\sum y}{n} \qquad m'_2 = \frac{\sum y^2}{n}$$

$$m'_3 = \frac{\sum y^3}{n} \qquad m'_4 = \frac{\sum y^4}{n}.$$

m'_1, m'_2, m'_3 and m'_4 are the first four moments about zero.

For frequency Distribution, the raw moment about zero are given by

4.29 The daily income of employees range from Rs.0 to Rs.18. They are grouped in intervals of Rs.2 and class frequencies from the lowest to the highest class are 5, 39, 69, 41, 29, 22, 16, 7 and 5. Find the co-efficient of skewness.

4.30 First four moments of distribution about $x = 2$ are 1, 2.5, 5.5 and 16, calculate mean and co-efficient of variation.

4.31 Find moments about mean β_1 and β_2. Given the first 4 moments about $y = 20$ as -2, 15, -25 and 80 respectively.

4.32 What is meant by skewness and kurtosis. What aspects of the frequency curve are measured by them.

4.33 Second moment about mean of two distributions are 9 and 16 while fourth moment about mean are 230 and 780 respectively, which of the distribution is

i) Leptokurtic ii) Platykurtic

4.34 What can you say about skewness in each of the following cases?

i) Median is 26.01 while two Quartiles are 13.73 and 28.29.

ii) Mean = 140 and Mode = 148.7.

iii) First three moments about 16 are 0.35, 2.9 and 1.93 respectively

4.35 i) The second moment about mean of two distributions are 13.76 and 63.0 while the fourth moments about the mean are 528.06 and 9500 respectively. Which of the distributions is

a) Leptokurtic b) Mesokurtic c) Platykurtic.

ii) The fourth central moment of a symmetrical distribution is 243. What would be the value of standard deviation for which distribution is mesokurtic?

4.36 The second moment about mean of a distribution is 25, what would be the value of fourth moment about mean if the distribution is

i) LeptoKurtic ii) MesoKurtic iii) PlayKurtic.

4.37 Which of the following is correct for a negatively skewed distribution;

i) A. M. is greater than mode. ii) A. M. is less than mode.

iii) A. M. is greater than median.

4.38 What would be the shape and the name of the distribution if

i) Mean = Median = Mode ii) Mean > Median > Mode.

iii) Mean < Median < Mode iv) $\beta_1 = 0$ and $\beta_2 = 3$

v) $\beta_1 = 0$ and $\beta_2 = 5$

4.39 Against each statement, write T for true and F for false statement.

i) The sum of squares of the deviations for a data set from the median is minimum.

ii) The sum of absolute deviations for a data set from the mean is minimum.

iii) If each of the observations in a data set is multiplied by a constant, the variance of the resulting data set increases.

iv) If a constant is added in each of the observations in a data set, then the variance of the resulting data set increases.

v) Standard deviation is a positive square root of variance.

vi) Range is a measure of absolute dispersion.

vii) A relative measure is independent of unit.

viii) Mean deviation is not based on all the observations.

ix) Semi-inter quartile Range is also called inter quartile range.

x) The standard deviation is dependent upon origin.

4.15 Calculate mean deviation (about median) for the distribution given below:

Groups	Frequencies
100 – 110	56
110 – 120	59
120 – 130	61
130 – 140	68
140 – 150	77
150 – 160	59
160 – 170	51
170 – 180	42
180 – 190	36
190 – 200	25

4.16 Calculate standard deviation by using arithmetic mean and also by using any provisional mean and compare the results for the data given below:

3, 5, 7, 13, 15, 17, 23, 27

4.17 A manufacturer of television tubes produces two types *A* and *B* of tubes. The tubes have respective mean life times as $\overline{X}_A$ = 1496 hours and $\overline{X}_B$ = 1895 hours and standard deviations S_A = 280 hours and

S_B = 310 hours. Which tube has the greatest:

i) absolute dispersion ii) relative dispersion.

4.18 i) What are moments about mean and about an arbitrary value. Give the relation between them.

ii) Define the moment ratios b_1 and b_2.

4.19 Computer calculated mean and standard deviation from 20 observation as 42 and 5 respectively. It was later discovered at the time of checking that it had copied down two values as 45 and 38 whereas the correct values were 35 and 58 respectively. Find the correct value of co-efficient of variation.

4.20 A distribution consists of 3 components with frequency 100, 120 and 150 having means: 5.5, 15.8 and 10.5 and standard deviations: 2.4, 4.2, and 3.7 respectively. Find the co-efficient of variation for the combined distribution.

4.21 Calculate first four moments about mean for the following set of examination marks.

45, 32, 37, 46, 39, 36, 41, 48 and 36.

4.22 Calculate first four moments for the following distribution of wages about $y = 10$. Find moments about mean

Earning	5	6	7	8	9	10	11	12	13	14	15
f	1	2	5	10	20	51	22	11	5	3	1

4.23 First three moments of a distribution about $y = 4$ are 1, 4 and 10 respectively. Find co-efficient of variation. Is the distribution symmetrical or positively skewed or negatively skewed.

4.24 First four moments of a certain distribution about $y = 17.5$ are 0.3, 74, 45 and 12125 respectively. Find out whether the distribution is lepto kurtic or platyckurtic.

4.25 What can you say of skewness in each of the following cases:

i) $Q_2 = 26.01$, $Q_3 = 38.29$, $Q_1 = 13.73$

ii) Mean =1403 and Mode = 1487

4.26 Given that $\Sigma f = 76$, $\Sigma fy = 572$, $\Sigma fy^2 = 4848$, $\Sigma fy^3 = 44240$ and $\Sigma fy^4 = 42580$. Find first four moments about mean and test the distribution for symmetry and kurtosis.

4.27 If distribution has mean 1403 and mode 1487, what can you say about the skewness.

4.28 Lower and upper quartiles of a distribution are 142.36 and 167.73 respectively while median is 153.50. Find co-efficient of skewness.

4.6 Calculate Mean Deviation from Mean, Mean co-efficient of dispersion and variance from the data given below:

Weights (kg)	No. of students	Weights (kg)	No. of students
50 – 53	23	65 – 68	66
53 – 56	24	68 – 71	49
56 – 59	39	71 – 74	38
59 – 62	46	74 – 77	21
62 – 65	54	77 – 80	12

also calculate Range, Quartile Deviation, and co-efficient of Quartile Deviation.

4.7 Calculate quartile deviation for the data given below:

Groups	25-50	50-75	75-100	100-125	125-150	150-175
Frequency	10	12	16	17	20	18

also calculate co-efficient of Quartile Deviation.

4.8 Calculate standard deviation, variance and co-efficient of variation from the following data:

y	525	500	475	450	425	400	375
f	24	35	46	37	47	34	22

4.9 The mean of a set of 10 values is 25.2 and its standard deviation is 3.72 for another set of 15 values mean and standard deviation are 25.2 and 4.05 respectively. Find the combined standard deviation of the 25 values taken together.

4.10 For a group of 50 boys, the mean score and the standard deviation of scores on a test are 59.5 and 8.38 respectively, for a group of 40 girls, the mean and standard deviation are 54.0 and 8.23 respectively on the same test. Find the mean and standard deviation for the combined group of 90 children.

4.11 By multiplying each number 3, 6, 1, 7, 2, 5 by 2 and then adding 5, we obtained 11, 17, 7, 19, 9, 15. What is the relationship between standard deviation and means for the two sets of numbers?

4.12 The scores obtained by 5 students on a set of examination papers were 70, 50, 60, 70, 50. Their scores are changed by

i) adding 10 point to scores ii) increasing all scores by 10%.

What effect will these changes have on mean and on standard deviation?

4.13 Compute the mean wages and the co-efficient of variation for the employees working in two factories are given in the following table:

Wages	**Factories**	
	Factory *A*	**Factory *B***
30 – 35	12	4
35 – 40	18	10
40 – 45	29	31
45 – 50	32	67
50 – 55	16	35
55 – 60	8	15

4.14 Compute median and mean deviation from median for the data given below:

Daily Wages	**No. of Domestic Servants**
6	5
8	10
10	18
12	20
14	22
16	14
18	7
20	3
22	1

The leptokurtic distribution may be composite of two normal distributions with the same mean but different variances. The platykurtic distribution may be the composite of two normal populations with the same variance but different means.

The dimensionless measure of kurtosis based on the moments is β_2 and is defined as:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \tag{4.51}$$

If $\beta_2 = 3$, the distribution is mesokurtic (normal). If $\beta_2 < 3$, the distribution is platykurtic and if $\beta_2 > 3$, distribution is leptokurtic. If this measure is calculated by using sample moments then

$$b_2 = \frac{m_4}{m_2^2} \tag{4.52}$$

Here, b_2 is the estimate of β_2

Another measure of kurtosis not widely used is Percentile co-efficient of Kurtosis, and is denoted by K.

$$K = \frac{Q.D}{P_{90} - P_{10}} \tag{4.53}$$

Where, Q.D. stands for quartile deviation and P_{90} and P_{10} are the 90th and 10th percentiles respectively. K is 0.263 for a normal distribution and lies between 0 and 0.50.

Example 4.22: Calculate $\sqrt{b_1}$, b_2 for ungrouped data of the example 4.18 and for grouped data of the example 4.19.

Solution:

For ungrouped data: For ungrouped data of the example 4.18 we have,

$$m_2 = 43.4988, m_3 = 17.3354, m_4 = 4131.1478$$

$$\begin{aligned} so, \quad \sqrt{b_1} &= \frac{m_3}{\sqrt{m_2^3}} \\ &= \frac{17.3354}{\sqrt{(43.4988)^3}} \\ &= 0.0604 \end{aligned}$$

$$b_2 = \frac{m_4}{m_2^2}$$

$$= \frac{4131.1478}{(43.4988)^2}$$

$$= 2.1833$$

For grouped data: For grouped data of the example 4.19, we have

$m_2 = 43.7220,\quad m_3 = 43.7407,\quad m_4 = 4363.0296$

$$\text{so,}\quad \sqrt{b_1} = \frac{m_3}{\sqrt{m_2^3}} = \frac{43.7407}{\sqrt{(43.7220)^3}} = 0.1513$$

$$b_2 = \frac{m_4}{m_2^2} = \frac{4363.0296}{(43.7220)^3} = 2.2824$$

Exercise 4

Ans on Page 251

4.1 What do you understand by dispersion? What are the most usual methods of measuring dispersion, indicate the advantages and disadvantages of these methods?

4.2 Define mean deviation and its co-efficient. Discuss its advantages and uses.

4.3 i) What is semi-inter quartile Range.

ii) Define range and discuss its uses.

4.4 Explain the difference between absolute dispersion and relative dispersion. Describe the properties of the standard deviation.

4.5 i) Define various measures of dispersion and given their formulae.

ii) The following table gives the marks of students:

Marks	30-39	40-49	50-59	60-69	70-79
f	8	87	190	86	20

Calculate:

a) Quartile deviation

b) Co-efficient of skewness

xi) Co-efficient of variation is a Relative measure of dispersion.

xii) The first moment about mean is one.

xiii) If b_1=3,the distribution will be symmetrical.

xiv) Mean Deviation is always less than the standard deviation.

xv) The normal Distribution is also known as Mesokurtic Distribution.

4.40 Fill in the blanks.

i) A measure of dispersion is __________.

ii) A measure of dispersion expressed as a co-efficient is called ________ measures of dispersion.

iii) Sum of absolute deviations are minimum if computed from _________.

iv) The value of standard deviation does not __________ if a constant is added or subtracted from all observations.

v) Co-efficient of variation is always _________ from unit of _________.

vi) A data having least C.V. is considered more _________.

vii) The lack of symmetry is called _________.

viii) In a symmetrical distribution, the quartile deviation is _________ from the median on both sides.

ix) Shepherd correction is applicable when the frequency distribution tends to _________ in both directions.

x) A relative measure of dispersion is the _________ between absolute dispersion and the average.

5 Index Numbers

$0 \leq P(A) \leq 1$

data

5.1 Introduction

The buying power of a rupee varies from time to time as the amount of a commodity. One çould buy by 10 rupees in 1960 now costs about 60 rupees in 1995, so to make meaningful comparisons overtime it is necessary to take into account the variability in the buying power of a rupee. For example to compare the cost of 2-year college education today with its cost in 1960, it is necessary to consider the buying power of a rupee today as compared with the buying power of a rupee in 1960. Similarly one may be interested to know the average hourly wages of a labourer in 1995 as compared with their wages in 1960. The numbers known as Index numbers are computed for this purpose and measure the relative change in a variable overtime. These are usually constructed for the variables such as prices, quantities, wages, investment and cost of living to help governments, economists and business people.

An index number is a ratio or an average of ratios usually expressed as a percentage. To construct index numbers two or more time periods are considered. The values at one of the time periods are taken as a base. This ratio of the values at the other time periods to the base period when expressed as a percentage show percentage change in the value from the value of the base period. The index number is usually denoted by I_n and is calculated by the relation:

$$I_n = \frac{\text{price in current year } n}{\text{price in base year}} \times 100 \qquad \text{.... (5.1)}$$

We will use the following notations in this chapter.

p_{0i} = price of the ith commodity in the base year.

q_{0i} = quantity of the ith commodity in the base year.

p_{0n} = price of the ith commodity in the current year.

q_{0n} = quantity of the ith commodity in the current year.

p_{0n} = price Index number for current year . P_{01} means index number for the year next to base year and so on.

Q_{0n} = Quantity Index number for current year. Q_{01} means index number for the year next to base year and so on.

I_n is also used to denote Index number for current year in the literature.

The index number for the base year is always taken as 100.

Example 5.1: The data given below is available about the price of wheat for the years 1989 to 1994. The interest is to compare the price of wheat in these years taking 1989 as the base year.

Year	1989	1990	1991	1992	1993	1994
Price	85	96	112	124	130	160

Solution:

The price index for each year is calculated by the ratio.

$$\therefore I_n = \frac{\text{price in current year}}{\text{price in 1989}} \times 100$$

The index number for 1989 is the ratio of price in 1989 to the price in 1989 expressed as a percentage i.e.,

(85/85)(100) = 100

The index number for 1990 is

(96/85)(100) = 112.94 and so on. The price indices are

Year	Price	Price index (1989 base)
1989	85.00	100
1990	96.00	112.94
1991	112.00	131.76
1992	124.00	145.88
1993	130.00	152.94
1994	160.00	188.24

The price index column indicates percentages of 1989 price for each year. For example, the price in 1992 is 145.88% of the price in 1989. So, the price of wheat is 45.88 % higher in 1992 as compared with wheat price in 1989.

5.1.1 Types of index numbers

Index numbers are generally classified into the following two types.

i. Simple index numbers.

ii. Composite or aggregate index numbers.

Simple index number

An index number is called a simple index number when it measure a relative change in a single variable with respect to a base year. For example index numbers for wages of labourers, index number of wheat prices and index number for the volume of a commodity (produced, purchased, sold, consumed etc.) overtime. In example 5.1 above, simple index numbers have been calculated.

If we are calculating price index (P_{on}) then the formula is

$$P_{on} = \frac{\text{price in year } n}{\text{price in base year}} \times 100 \tag{5.2}$$

Similarly, wage index and other indices are calculated.

Example 5.2: Compare the daily wages of unskilled labourers in Lahore over the time period 1988-93 where the following data is available from the Pakistan Economic survey 1993 taking 1988 as base year. Wages are in rupees.

Year	1988	1989	1990	1991	1992	1993
Wages	46	51	58	71	71	86

Solution: The daily wage index I_n for each year is the ratio

$$W_{on} = \frac{\text{daily wages in year } n}{\text{daily wages in year 1988}} \times 100$$

Using this formula and taking 1988 wages as the base, the wage index numbers are calculated for each year and are given below in the last column.

Year	Wages (Rupees)	Wages index (1988 base)
1988	46	100.00
1989	51	110.87
1990	58	126.08
1991	71	154.35
1992	71	154.35
1993	86	186.96

The wage index of 126.08 for 1990 indicates that wages have been increased by 26.08 percent as compared to base year 1988. Similarly, the wages in 1993 have been increased 86.96 percent.

Composite or aggregate index numbers

An index number is called a composite (aggregate) index number when it measures a relative change in two or more variables with respect to a base year. For example index numbers for comparing two sets of prices from a wide variety of commodities, index numbers for comparing two sets of the quantities of the commodities from a wide variety of commodities. These are calculated in following two ways:

i. Unweighted index numbers

ii. Weighted index numbers

The unweighted and weighted indices may measure changes in price, quantity or value of a commodity. Accordingly these may be

a. Price index numbers

b. Quantity index numbers

c. Value index numbers

(a) and (b) are discussed in article 5.4 and 5.5.

The value index number denoted by V_{on} is $\frac{\Sigma p_n q_n}{\Sigma p_o q_o} \times 100$ (5.3)

where

$\Sigma p_n q_n =$ total value of all commodities in a given year.

$\Sigma p_o q_o$ = total value of all commodities in the base year.

The simple and composite indices are discussed in detail under headings 5.4 and 5.5.

5.1.2 Limitations of index numbers

Some limitations of index numbers are given below:

1. It is not possible to take into account all changes in product.
2. All index numbers are not suitable for all purposes.
3. There may be errors in the choice of base periods.
4. These are simply rough indications of the relative changes.
5. Different methods of construction of index numbers give different results.

5.1.3 Use of Index numbers

i) The price index numbers are used to measure the average price changes in a commodity or a group of commodities with the passage of time relative to the base period. This helps in comparing prices of one commodity with another.

ii) The price index numbers are used to measure the buying power of the money.

iii) The consumer price indices (CPI) are used as a factor to cancel out the effect of inflation or deflation by the governments as these measure the changes in the prices of consumer goods.

iv) The wholesale price index numbers (WPI) help in the adjustment of contract prices and payments by industrial organization as these measure changes in producer's selling prices.

v) The quantity index numbers are used to measure changes in the quantities produced, a consumed, purchased, sold, exported or imported.

vi) The index numbers are helpful for the economists and the businessmen to describe the existing conditions and help plan near future.

vii) The index numbers of import and export prices help to measure charges in terms of trade of country.

5.2 Construction of price index numbers

The construction of different index numbers involves the following main steps.

a. Purpose and scope

The first step is to define the purpose of index numbers. It should be clearly mentioned why, where and what changes are to be measured.

b. Selecting components

i. Commodities. ii. Price of commodities.

As an index number is constructed to represent a particular purpose. So, it is important to select the commodities to be included keeping in view the cost of collecting data. The items should be precisely defined in terms of quality specifications and relevant data should be readily available overtime as some items become obsolete with the introduction of new products. For practical purposes, the number of items selected should not be less than twenty. The prices of these items should be collected from different places keeping in view the quality of the items. The sampling should be carried out with care as the posted or listed prices are not the retail prices sometimes.

c. Choosing the base year

The purpose of constructing index numbers is to make comparisons. So, the base year should be chosen with care to be a year of normal prices and should not be a year too far from the current year. It is also possible to use an average of prices for several years to act as a base year e.g. it may be an average of 3 years. Usually base year is taken fixed because of comparison purposes but it may be taken as variable in case we calculate what we call "link relatives" discussed under 5.4 (iii)

d. Choosing the weights

The weights chosen should indicate the relative importance of various commodities to be included in the construction of an index. As all the commodities selected are not equally important so they should be given different weights. For example wheat should be given more weight as compared with tea.

The weights chosen for the price index may be the quantities of the base year or current year. It depends upon the situation what weight to be used. The quantities may be quantities produced or quantities marketed because in agricultural economics, a high proportion of the food produced is often consumed by the families themselves and not sold in the markets.

The weights may be formulated as follows.

Let W_{oi} denotes the weight for the ith commodity in the base year.

V_{oi} is the value of the marketed or produced commodity in base year then

Value of the commodity = price × quantity

$$V_{oi} = P_{oi} q_{oi}$$

The weight

$$W_{oi} = \frac{V_{oi}}{\Sigma V_{oi}} \tag{5.4}$$

Here, ΣV_{oi} is the total value of the k items in the base year

e. Choosing the average

An index number can be constructed as average of ratios according to the definition of index number. For example, consider k commodities then the arithmetic mean of k ratios each computed for single commodity is

$$P_{on} = \frac{\sum_{i=1}^{k} \frac{P_{ni}}{P_{oi}}}{k} \tag{5.5}$$

Other averages such as geometric mean, harmonic mean or median can also be used instead of arithmetic man. Although geometric mean is more appropriate for averaging ratios but in practice arithmetic mean is used for convenience.

5.3 Unweighted index numbers

The idea is to give equal weights to each item in the index. The unweighted index numbers are computed in the following two ways:

i. Simple aggregate Index

It is the ratio of the sum of prices (quantities) of commodities for a given year to the sum of prices (quantities) of the same commodities in the base year, expressed as a percentage.

The price index denoted by P_{on} in calculated by the formula

$$P_{on} = \frac{\sum_{i=1}^{k} P_{ni}}{\sum_{i=1}^{k} P_{oi}} \times 100 \qquad = \frac{\sum P_n}{\sum P_o} \times 100 \tag{5.6}$$

where k is the number of commodities P_{oi} and P_{oi} are the prices of the commodities in the current and the base years respectively, n denotes the current years, 0 denotes the base year and i stands for the number of commodities.

These indices suffer from drawback that changes in the measuring units may affect the value of the index which ultimately lessen their usefulness for making meaningful comparisons. Secondly these indices use equal weight for all commodities whereas all commodities are not equally important.

ii. Average Relative Index

It is the average of simple index numbers calculated individually for each commodity. For k commodities it is calculated by the relation.

$$P_{on} = \frac{1}{k} \sum_{i=1}^{k} \frac{P_{ni}}{P_{oi}} \tag{5.7}$$

The simple index numbers P_{ni} / P_{oi} are also called price relatives. Thus average index is the average of price relatives. Usually, the average used is arithmetic mean but other averages such as geometric mean or median may be used.

These indices suffer from the drawback that these don't use weights for the different commodities according to their importance but changes in the measuring units don't affect the index value.

Example 5.3: Calculate the unweighted price index for 1994 when the procurement/ support prices of agricultural commodities in rupees per 40 Kg in 1980 and 1994 are given as follows:

Commodity	Prices	
	1980	1994
Wheat	58	160
Rice	118	360
Potato	27	19
Onion	80	84

Solution:

i. **Simple aggregate index**

The simple aggregate price index for 1994 is

$$P_{on} = \frac{160+360+19+84}{58+118+27+80} \times 100 = \frac{623}{283} \times 100 = 220{\cdot}14 \quad = \frac{\sum_{i=1}^{k} P_{ni}}{\sum_{i=1}^{k} P_{oi}} \times 100$$

This indicates that the prices of the above 4 commodities in 1994 are 120.14% higher than they were in 1980.

ii. **Average relative index**

The simple price index number for wheat is given by

$$P_{on} = \frac{160}{58} = 2.7586 \text{ or } 275.86\%$$

The simple price index for rice is given by

$$P_{on} = \frac{360}{118} = 3.0508 \text{ or } 305.08\%$$

Similarly, the index numbers for potato and onion are

$$\frac{19}{27} = 0.7037 \text{ or } 70.37\% \text{ and } \frac{84}{80} = 1.05 \text{ or } 105\% \text{ respectively,}$$

So, the average relative index for commodities is given by

using arithmetic mean as average

$$P_{on} = \frac{2.7586+3.0508+0.7037+1.05}{4} = 1.8908 \; or \, 189.08\%$$

This index indicates that the prices are 89.08% higher in 1994 as compared with 1980.

using median as average

One can use the median as average, then the median value of these 4 values is obtained by arranging them in ascending order as follows:

0.7037, 1.05, 2.7586, 3.0508

As 4 is an even number so the median is the average of middle two val ues i.e.,

(1.05 + 2.7586)/2 = 1.9043, so our index number is

1.9043(100)% = 190.43%

using geometric mean as average

Y	log Y
1.05	0.0212
2.7586	0.4407
3.0508	0.4844
0.7037	0.1526
Total:	0.7937

$$\begin{aligned} \text{G.M.} &= \text{Antilog } (\Sigma \log Y / n) \\ &= \text{Antilog } (0.7937/4) \\ &= \text{Antilog } (0.1984) = 1.5791 \end{aligned}$$

So, the index is

(1.5791) (100) = 157.91

In this example, wheat and onion had equal weights but clearly wheat is produced more as compared with onion, so it is usually recommended to use weights in the index proportional to the value of the production of each item.

Example 5.4: Calculate index numbers of price, using 1962 as base

i) Mean ii) Median are used.

Years	**Commodities**			
	Firewood	**Softcake**	Kerosene oil	Match
1962	3.25	2.50	0.20	0.06
1963	3.44	2.80	0.22	0.06
1964	3.50	2.00	0.25	0.06
1965	3.75	2.50	2.50	0.06

Solution:

Years	**Price relatives**				Total	**Index numbers**	
	Fire-wood	Soft-cake	Kerosene oil	Match		Mean	Median
1962	100	100	100	100	400	100	100
1963	106	112	110	100	428	107	108
1964	108	80	125	100	413	103	104
1965	115	100	125	100	440	110	108

Method for finding the average relative Index number.

With the introduction of new commodities and quality goods in the market, the tastes and habits of the people change overtime. This brings change in the relative importance of the commodities. In such situations, it becomes necessary to change the base year and a quantity called **Link relative** is calculated instead of **price relative** for each year. The link relative is a quantity computed by taking the price of the previous year as a base. These are also expressed as percentages like price relative.

$$\text{Link relatives} = \frac{\text{Price of an item for current year}}{\text{Price of an item for previous year}} \times 100$$

$$= \frac{P_n}{P_{n-1}} \times 100$$

Link relatives are not directly comparable because they have no fixed base. To make them comparable, they are converted to what we call **Chain indices**. The chain index for the first year is taken as 100 and then the chain index for succeeding year is obtained by multiplying its link relative with the chain index of the proceeding year and dividing the results by 100. One can note that these chain indices are just the price relatives computed by taking the first year as a base.

Advantages of chain base methods

This method is rarely used because the price relatives give the same results. However, it has certain advantages.

i) Link relatives are useful to make year to year comparisons.

ii) New items can be substituted for old items provided the number of items remains the same.

iii) The weighing system can be changed according to the change in the relative importance of different items.

iv) Changes in the Geographical coverage can be accommodated.

Example 5.5: Compute link relatives and chain index for the data of the example 5.2.

Year	Wages	Link relative	Index numbers
1988	46	100	100
1989	51	$\frac{51}{46}\times100=110.87$	$\frac{110.87\times100}{100}=110.87$
1990	58	$\frac{58}{51}\times100=113.73$	$\frac{113.73\times110.87}{100}=126.09$
1991	71	$\frac{71}{58}\times100=122.41$	$\frac{122.41\times126.09}{100}=154.35$
1992	71	$\frac{71}{71}\times100=100$	$\frac{100\times154.35}{100}=154.35$
1993	86	$\frac{86}{71}\times100=121.13$	$\frac{121.13\times154.35}{100}=186.96$

Example 5.6: Find chain index numbers for the price data given below. The price of the commodities are in Rs. per 40 Kg.

Years	Commodities			
	Wheat	Rice	Potato	Onion
1980	58	118	27	80
1981	60	120	30	90
1982	75	130	30	95
1983	90	150	40	100

Solution: First we compute link relative, by the relation

$$\text{Link relative} = \frac{P_n}{P_{n-1}} \times 100$$

The link relative for wheat for the year 1980 is taken as 100, for 1981 it is 60/58 $\times 100 = 103.45$; for 1982 it is 75/60 $\times$ 100 = 125.00 and so on for all the other commodities and are given below:

Year	Link relatives					
	Wheat	Rice	Potato	Onion	Total	Mean
1980	100	100	100	100	400	100
1981	103.4	101.7	112.5	112.5	428.7	107.2
1982	125.0	108.3	105.6	105.6	438.9	109.7
1983	120.0	115.4	105.3	105.3	474.0	118.5

The chain index for 1980 is 100; The chain index for 1981 is obtained by multiplying 100 with the link relative of 1981 and dividing it by 100; i.e., (107.19)(100)/100=107.19; the chain index for 1982 is (109.72)(107.19)/ 100=117.61 and so on. These are given in the adjoining table:

Year	Chain Indices
1980	100
1981	$\frac{107.19\times100.00}{100}=107.19$
1982	$\frac{109.72\times107.19}{100}=117.61$
1983	$\frac{118.49\times117.91}{100}=139.36$

5.4 Weighted index number

In the weighted index numbers the different commodities in the combination receive their weights proportional to their importance. These are calculated by the following two types:

5.4.1. Weighted aggregate index

It is the ratio of the sum of weighted commodity prices (quantities) to the sum of weighted commodity prices (quantities) in the base year, expressed as a percentage, the weights being the corresponding quantities (prices).

The price index denoted by P_{on} (weighted) is calculated by the following formula.

$$P_{on} = \frac{\sum_{i=1}^{k} P_{ni}\, q_{oi}}{\sum_{i=1}^{k} P_{oi}\, q_{oi}} \times 100 \qquad (5.8)$$

where, p_{ni}, p_{oi} are the prices of the commodities in the current and base year and q_{ni}, q_{oi} are the corresponding quantities respectively.

Weighted index numbers are of various kinds. The most common are discussed below:

a. Laspeyre's Index Number

It was named after the name of an economist Etienne Laspeyres who introduced it. It is denoted by P_{on} and is calculated by the following formula.

Index number for prices

$$P_{on} = \frac{\sum_{i=0}^{k} P_{ni}\, q_{oi}}{\sum_{i=0}^{k} P_{oi}\, q_{oi}} \times 100 \qquad (5.9)$$

Here the weights are base year quantities and the idea of using base year quantities as weights for current year prices is that the base year quantities don't change overtime. This is true for every day consumer commodities but for others the increase in price is followed by the decrease in the quantity consumed so more weights are given to the commodities whose prices have increased.

Index number for quantities

$$Q_{on} = \frac{\sum_{i=0}^{k} q_{ni}\, p_{oi}}{\sum_{i=0}^{k} q_{oi}\, p_{oi}} (100) \qquad (5.10)$$

Here the weights are base year prices.

b. Paasche's Index Number

It is a weighted index and unlike Laspeyre's index, it uses current year quantities as weights rather than base year quantities for the price index. It is denoted by P_{on} and is calculated by the formula

$$P_{on} = \frac{\sum_{i=0}^{k} p_{ni}\, q_{ni}}{\sum_{i=0}^{k} p_{oi}\, q_{ni}} (100) \qquad (5.11)$$

Unlike Laspeyres index this index gives less weight to the commodities whose prices have increased. Similarly, the quantity index number is

$$Q_{on} = \frac{\sum_{i=1}^{k} q_{ni}\, p_{ni}}{\sum_{i=1}^{k} q_{oi}\, p_{ni}} (100) \qquad (5.12)$$

c. Fisher's ideal Index Number

This index number uses the fact that Laspeyres index gives more weight to commodities whose prices have increased and Paasche's index gives less weight to commodities whose prices have increased so, there should be an index number that should be in between these two index numbers. This aim is achieved by taking the geometric mean of these index numbers. So, the Fisher's index number P_{on} is given by

$$P_{on} = \sqrt{\text{Laspeyre} \times \text{Paasche}} \qquad (5.13)$$

$$= \sqrt{\left(\frac{\Sigma p_{ni}\, q_{oi}}{\Sigma p_{oi}\, q_{oi}}\right)\left(\frac{\Sigma p_{ni}\, q_{ni}}{\Sigma p_{oi}\, q_{ni}}\right)} \times 100 \qquad (5.14)$$

Example 5.7: Complete index numbers from the following data using 1964 as base.

i) Laspeyre's Index. ii) Paasche's Index iii) Fisher's Index

for the following data using 1964 as base.

Items	1964		1967	
	Price	Quantity	Price	Quantity
A	10	12	12	15
B	9	15	5	20
C	5	24	9	20
D	10	5	14	5

Solution:

Items	1964		1967		$p_o q_o$	$p_n q_o$	$p_n q_n$	$p_o q_n$
	p_o	q_o	p_n	q_n				
A	10	12	12	15	120	144	180	150
B	9	15	5	20	105	75	100	140
C	5	24	9	20	120	216	180	100
D	10	5	14	5	50	70	70	80
Total:	-	-	-	-	425	505	530	470

i) Laspeyre's Index $= \dfrac{\sum P_n\ q_o}{\sum P_o\ q_o} \times 100 \dfrac{505}{425} \times 100 = 118.8$

ii) Paasche's Index $= \dfrac{\sum P_n\ q_n}{\sum P_o\ q_n} \times 100$

$$= \frac{530}{470} \times 100 = 112.8$$

iii) Fisher's Index $= \sqrt{\dfrac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \dfrac{\Sigma p_n q_n}{\Sigma p_o q_n}} \times 100$

$$= \sqrt{\frac{505}{425} \times \frac{530}{470}} \times 100$$

$$= 115.8$$

5.5 Consumer Price index (CPI) and Wholesale Price Index (WPI)

Federal Bureau of Statistics (FBS) calculates the following price indices in the country to measure price changes overtime.

i) Consumer Price Index (CPI) ii) Sensitive Price Index (SPI)

iii) Wholesale Price Indices (WPI)

We shall discuss meaning and construction with reference to Pakistan.

5.5.1 Consumer Price Index (CPI)

CPI is constructed to measure the aggregate change in the cost of a fixed basket of goods and services purchased at current prices with its cost at a given period called the base, which is always taken as hundred. It is also called cost of living index.

The CPI was computed for the first time in early 1950's with base 1948-49 for industrial workers in Lahore, Karachi and Sialkot only as a measure of inflation. These days, it is being calculated for four different income groups and occupational groups in 25 big cities of the country and covers 464 items of consumption in the basket of goods and services which represent their taste, habits and customs.

To construct the CPI, the prices of the items to be included are sampled from different locations. The weights to different commodities are given keeping in view their importance to make the indices reliable. The weight assigned to each commodity is the average percentage expenditure on it to the total expenditure of a family. The present weights used are based on the results of a Family Budget Survey conducted for this purpose. Due to change in income, taste and other seasonal and geographical factors, the weights are different for different income groups, occupational categories, cities and various commodity groups.

5.5.2 Construction of CPI

The construction of consumers price index number involves the following steps:

1. Deciding the category of the people
2. Family Budget inquiry
3. Selection of items
4. Price quotations
5. Choice of weights

1. Deciding category of the people.

The first and the most important step in the construction of consumer price index numbers is the decision regarding the category of the people for whom the index numbers are going to be constructed. It should be decided before hand and whether they are for clerks, industrial, coolies.

2. Family Budget Inquiry

After deciding the category of the people an adequate number of families should be selected during a normal period. Family budget inquiry is conducted on the basis of random sampling. This inquiry would give information regarding,

i) the qualities and quantities of the items consumed by them. Index different heads such as food articles, clothing house rent, fuel lighting, education gifts, newspaper, transport etc.

ii) the retail prices of the items.

iii) amount spent by various items.

3. Selection of Items

With the help of family budget inquiry, it becomes easier to select the items which should be included in the construction of consumer's price index numbers. Only these items should be included which are largely used by that class of people and which are not subject to wide variation in quantity, supply and prices.

4. Price Quotations

The price quotations should be retail prices and not the whole rate prices. The prices should be obtained from the shops, publications of the govt. and official reporters of deciding locality.

5. Choices of Weights

All the items that enter into family budget are not of equal importance and thus weights must be assigned to the items. There are two types of weights:

ii) **Quantity Basis:** It means the quantity of the different items consumed in the base mean.

iii) **Value Basis:** It means total values of the items consumed by each group. It is calculated by multiplication of the quantities consumed and the prices, there are two methods.

a) **Aggregative Expenditure Method**

According to this method, the quantities consumed in the base year used as weights. It is the base year weighted index given by Laspeyre's

$$P_{on} = \frac{\Sigma P_n q_o}{\Sigma P_o q_o} \times 100$$

b) **Family Budget Method**

This method is the weighted average of price relatives. In this method, the family budget of a large number of families are carefully studied and the aggregate expenditure of the average family on various items is estimated. The amount of money spent by the families concerned are calculated from a family budget inquiry. The formula is

$$P_{on} = \frac{\Sigma IW}{\Sigma W} \times 100$$

Where, $I = \frac{P_n}{P_o} \times 100$

$$W = p_o q_o$$

Example 5.8: An inquiry into the budgets of the middle class families in a city of England gave the following information:

Expenses on	**Food** **35%**	**Rent** **15%**	**Clothing** **20%**	**Fuel** **10%**	**Misc.** **20%**
Price (1928)	150	30	75	25	10
Price (1929)	145	30	65	23	15

What changes in cast of living the figures of 1929 show as compared to 1928?

Solution:

Expansion	***W***	**Price (1928)** (p_o)	**Price (1929)** (p_n)	$\frac{P_n}{P_o}\times 100$	***IW***
Food	35	150	145	97	3395
Rent	15	30	30	100	1500
Clothing	20	75	65	87	1740
Fuel	10	25	23	92	920
Misc.	20	40	45	113	2260
Total	100	-	-	-	9815

cost of living for 1929

$$P_{on} = \frac{\Sigma IW}{\Sigma W} = \frac{9815}{100} = 98.15$$

Interpretation of CPI computed and circulated by FBS

The following tables are taken from the Pakistan Economic Survey (1993-1994) where, CPI has been calculated by the Federal Bureau of Statistics (FBS) for items and for different income groups.

Consumer Price Index (on annual basis)			
Index	**Weight**	**% change**	**% point Contribution (1993-94/1992-93)**
General	100.00	10.61	10.61
Food, Beverages	49.90	10.40	5.19
House rent	17.76	9.83	1.75
Fuel & lighting	5.62	14.74	0.83
Household furniture equipment	2.34	5.21	0.12
Miscellaneous	2.44	22.11	0.54

The table indicates that highest increase recorded during the period is 22.11% in the miscellaneous group followed by the fuel and lighting group which recorded an increase of 14.74%. The lowest increase was recorded in the household, furniture and equipment group.

Since weights vary across the commodity groups, the highest contribution to overall CPI increase has been made by food, tobacco, beverages and group which is calculated by computing the ratio 100/49.90 = 2.004 and then dividing the change 10.40 by 2.004 i.e., 10.40/2.004 = 5.19 and similarly, for the house rent group i.e., 100/17.76 = 5.6306, then 9.83/5.63 = 1.75 and so on.

The following table gives the consumer price index for different income groups in Pakistan.

	Income group	**% change 1994/1993**	**Ratio of individual group index to overall CPI**
I.	Upto Rs.1000	8.27	97.18
II.	Rs.1001-2500	8.56	100.59
III.	Rs.2501-4500	8.91	104.70
IV.	Above Rs.4500	8.67	101.88
V.	All groups (combined)	8.51	100.00

The table indicates that there is an increasing trend in the indices of first three income groups starting from 8.27 to 8.91, then the index for 4th income group is 8.67 whereas the combined index is 8.51. The last column is the ratio of individual group to overall for the first group it is 8.27/8.51(100)=97.18; for the second group it is 8.56/8.51(100) = 100.59, and so on.

As the CPI is used to cancel out the effect of inflation so these indices suggest that the measures should be taken to protect the group I and II as compared with the other groups.

CPI and rate of inflation

The common way used to measure inflation in Pakistan is through CPI. While calculating annual rate of inflation one should compare the current year CPI with that of last year. The average annual rate of inflation over a longer period of time can be calculated by taking the average of those years. The inflation rate during the years mentioned below is computed using the CPI with base 1980-1981. The rate of inflation for 1989-1990 is [(177.33-167.23)/ 167.23] × 100 = 6.04 and similarly others.

Period	Index	Rate of Inflation
1988-89	167.23	---
1989-90	177.33	6.04
1990-91	199.78	12.66
1991-92	218.99	9.62
1992-93	239.26	9.26
1993-94	266.00	11.18
Total:		48.76

The annual average rate of inflation for 5 years is 48.76/5 = 9.75%

Measurement of purchasing power of money

The inverse of CPI can be used to measure the purchasing power of money. Since the base of CPI is 100 and the Pakistani rupee is also convertible into 100 paisas, the purchasing power of rupee is [I/CPI] × 100. Through this approach the purchasing power of Pakistani rupee in January 1995 as compared to 1980-1981 has come down to paisas 33 only.

5.5.2 Sensitive Price Indicator (SPI)

SPI is calculated in the same way with the same formula as the CPI but the difference is that it includes only 46 essential commodities instead of 464 in the CPI.

5.5.3 Wholesale price index (WPI)

This index indicates change in producer's selling prices and is not an indicator of wholesale prices as the name indicates.

These are computed from the information collected by sampling the producer's selling prices. Weights are derived on the basis of the value of the marketable surplus of commodities available for sale. These are computed with the same formulas as for the CPI. The following table gives the wholesale price indices of selected items taken from Pakistan Economic Survey (1993-94).

Consumer Price Index
(on annual basis)

Index	Weight	% change	% point Contribution (1993-94/1992-93)
General	100.00	12.71	12.71
Food	50.63	12.42	6.29
Raw material	8.97	18.69	1.68
Fuel, lighting and lubricants	11.79	22.65	2.67
Manufacturers	24.06	6.84	1.65
Building material	4.55	11.33	0.52

The table indicates that the wholesale price index increased by 12.71 percent. The highest increase of 22.65% is recorded in the 'Lubricants' group and on the other hand, the lowest increase of 6.84% is recorded in the 'Manufacturers' group. The percentage point contribution in the last column is calculated in the

same way as for the consumer price index in the previous table. It should be noted that in these indices it is valid to compare adjacent years, such as the value of 106.30 in 1992 and the value of 110.40 in 1993. The year to year change is 110.40-106.30 = 4.10 points or

$$\frac{4.10(100)}{106.30} = 3.86\%$$

Exercise 5

Ans on Page 253

5.1 What is an index number? Give the uses of an index number.

5.2 Define an index number. Discuss the main points involved in the construction of index numbers of prices.

5.3 Define an index number and describe the different types of index number.

5.4 Discuss the important problems involved in the construction of index number of prices.

5.5 Compare the following concepts:

i. Simple and composite index.

ii. Fixed and chain base method.

5.6 What is the weighted index number?

5.7 Find the index number of price from the following data taking average price of all years as the base.

Years	1970	1971	1972	1973	1974	1975	1976	1977
Prices	15	19	21	30	37	38	40	48

5.8 Given the prices of a commodity per maund for the period 1945 to 1960 as;

Year	Prices	Year	Prices	Year	Prices
1945	12.7	1950	24.85	1955	15.65
1946	18.97	1951	20.90	1956	16.15
1947	19.70	1952	19.80	1957	20.20
1948	13.50	1953	23.65	1958	25.25
1949	15.65	1954	24.55	1959	32.40

Construct index numbers correct to 2 decimal place:

i. 1945 as base.

ii. Average price of all the year as base prices.

5.9 Find index number using

i. 1977 as base

ii. average of the price as base:

Years	Prices	Years	Prices	Years	Prices
1977	22.5	1980	30	1983	37.5
1978	25	1981	35	1984	47.5
1979	27.5	1982	32.5	1985	45

5.10 For the following data, find index numbers taking

i. 1930 as base ii. Average of 1st 3 years as base

iii. the year 1935 as base

Years	Prices	Years	Prices	Years	Prices
1930	4	1933	7	1936	9
1931	5	1934	8	1937	10
1932	6	1935	10	1938	11

5.11 The following figures show the wholesale prices of refined petroleum per gallon in UK for the year specified. On the basis of 1923=100, construct a series of price relatives.

Years	Prices	Years	Prices
1923	13	1928	11 ¼
1924	13 ½	1929	10 ½
1925	13	1930	12 ½
1926	11 ½	1931	10 ¼
1927	12 3/7	1932	10 ½

5.12 Construct index numbers of prices for the following data taking 1960 as base:

Years	Prices	Years	Prices
1960	50	1965	72
1961	52	1966	73
1962	55	1967	75
1963	57	1968	71
1964	62	1969	70

5.13 The prices in Rs. Per maund of coal sold during the year 1953-58 as given below: Compute index number of price for the year 1953 as base.

Years	Prices	Years	Prices	Years	Prices
1953	14.95	1955	15.10	1957	16.28
1954	14.95	1956	15.65	1958	16.28

5.14 From the data given below, compute, the index numbers of prices, taking 1962 as base.

Commodities (Prices in Rs.)				
Year	**Firewood**	**Soft cake**	**Kerosene oil**	**Matches**
1962	3.25	2.50	0.20	0.06
1963	3.44	2.80	0.22	0.06
1964	3.50	2.00	0.25	0.06
1965	3.75	2.50	0.25	0.06

5.15 Compute index numbers of prices from the following data taking 1981 as base and using median as average.

Year	Prices		
	A	*B*	*C*
1981	18	85	52
1982	22	76	60
1983	28	80	66
1984	31	95	80

5.16 Find chain index numbers (using geometric mean to average the relatives) for the following data of prices, taking 1970 as the base year.

Commodities	Years				
	1970	**1971**	**1972**	**1973**	**1974**
A	40	43	45	42	50
B	160	162	165	161	168
C	20	29	52	23	27
D	240	245	247	250	255

5.17 The following table gives the average whole sale prices in rupees per unit of gold, wheat, cotton during the year 1912–1917. Construct index number with 1912 as base using

i) A.M. ii) G.M.

Commodities	Average price in Rs. Per unit					
	1912	**1913**	**1914**	**1915**	**1916**	**1917**
Gold	25.3	30.8	33.4	35.5	35.4	36.0
Wheat	17.3	14.5	4.9	5.7	17.7	11.6
Cotton	7.8	5.4	6.7	5.6	7.2	10.2

5.18 Construct chain indices for the following years, taking 1940 as base.

Item	Years				
	1940	**1941**	**1942**	**1943**	**1944**
Wheat	2.80	3.40	3.60	4.00	4.20
Rice	2.95	3.60	2.90	2.75	2.75
Maize	3.10	3.50	3.40	4.50	3.70

5.19 Construct index numbers for 1963 assuming 1953 as base period by

i) Laspeyre's formula ii) Paasche's formula.

Commodities	1953		1963	
	Price	Quantity	Price	Quantity
A	2	50	10	40
B	3	10	8	5
C	4	5	4	5

5.20 Compute the weighted index numbers for 1964 from the following data with 1960 as base.

Commodities	Price		Quantity	
	1960	1964	1960	1964
Milk	3.95	4.25	97.75	104.36
Cheese	34.80	38.90	78	83
Butter	61.56	59.70	118	116

5.21 Compute Fisher's index number for the following data.

Commodities	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	7	70	5	49
B	5	27	7	28
C	10	35	9	29
D	9	50	4	42
E	3	16	10	25

5.22 Construct the following with the help of data given below.

Fisher's ideal index taking 1970 as base.

Commodities	Total Production (tons)		Harvest Price (Rs)	
	1970	1971	1970	1971
Rice	71	26	3.80	3.50
Barley	107	83	2.90	1.90
Maize	72	48	2.90	1.80

5.23 Calculate Fisher's Ideal index from the following data with 1965 as base year.

Commodities	1965		1970	
	Price	Quantity	Price	Quantity
A	4.6	102	9.50	96
B	3.7	15	7.36	28
C	10.2	17	8.42	21
D	8.9	19	9.87	13

5.24 Define weighted and unweighted index numbers and explain why weighted index numbers are preferred over unweighted index numbers.

5.25 Calculate Laspeyre's, Paasche's and Fisher's ideal index for the following data with 1992 as base.

Item	Average price (Rs)		Quantity (Units)	
	1992	1993	1992	1993
Wheat flour	4.38	4.57	20Kg	16Kg
Rice	14.15	15.58	10Kg	12Kg
Moong pulse	18.67	17.28	1Kg	1Kg
Gram pulse	10.41	16.36	1Kg	1Kg

5.26 The following figures give the average annual prices in U.K. for beef and mutton.

Years	Price		Year	Price		Year	Price	
	Beef	Mutton		Beef	Mutton		Beef	Mutton
1935	54	75	1938	62	62	1941	72	85
1936	54	73	1939	61	68	1942	76	90
1937	61	73	1940	72	85	1943	79	90

Construct index number of meat prices giving weights 2 and 1 for beef and mutton respectively. Take the year 1935 as the base year.

5.27 The following table shows the average price in rupees for wheat, rice and barley.

Year	Price		
	Wheat	Rice	Barley
1980	175.5	480.4	82.4
1981	180.3	509.7	90.6

Taking 1980 as base year, construct price index number by weighted average of relative method for the year 1981 using the weight 20 for wheat, 12 for rice and 4 for barley.

5.28 An inquiry into the budgets of the middle class families in a city of England gave the following information. What change in cost of living the figures of 1929 show as compared in 1928.

Expenses on	Food 35%	Rent 15%	Clothing 20%	Fuel 10%	Misc. 20%
Price (1928)	150	30	75	25	40
Price (1929)	145	30	65	23	45

5.29 Fill in the blanks:

i) The changes in whole sale or retail price are studied in ________.

ii) The volume or quantity of goods are compared by ________.

iii) In ________ both quantities and prices are used.

iv) Index number are used for ________ business activity and in discovering ________ fluctuation and business ________.

v) The purpose of index number may be ________ or ________.

vi) The two method of selection base periods are ________ and ________.

vii) The base period in fixed base should be ________.

viii) ________ process is must in ________ method for comparison purpose.

ix) Geometric mean is a suitable average in ________ method.

x) Un-weighted indices are classified into simple ________ indices and simple ________.

5.30 Against each statement, write T for true and F for false statement.

i) Six steps are involved in the construction of index numbers of prices.

ii) In price relative, the given year price is divided by the base year price.

iii) Laspeyre's Index number is also named as current year weighted index number.

iv) Fisher's Index number is the Geometric Mean of the Laspeyre's and Paasche's Index number.

v) Aggregate expenditure method and Family Budget method are the types of the consumer price index number.

vi) The Index numbers are calculated in percentages.

vii) Index numbers are statistical barometers.

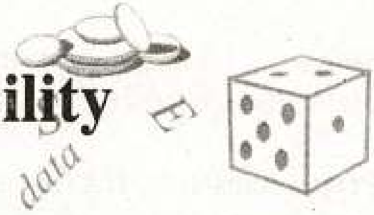
viii) In chain base method the base year is fixed.

ix) The most suitable average for Index numbers is Harmonic Mean.

x) The smaller the size of the sample, the greater would be the accuracy.

6 Probability

$0 \leq P(A) \leq 1$



6.1 Introduction

If you bought seven tickets for a raffle out of 700 tickets sold altogether. Each of 700 tickets is as likely as any other to be drawn for first prize. you would say that you had 7 chances out of 700, or a single chance out of 100 for winning the first price.

Probability gives us a measure for the **likelihood** that some thing will happen. However, probability cannot predict the number of times that an occurrance actually happens. Most of the decisions that affect our daily lives are based upon likely hood and not on absolute certainty.

In this chapter, we shall develop methods to deal with problems concerned with chance events.

Some definitions, terminologies and notations are explained below to enable the student to certain categories of situation precisely and move briefly.

Sets: A set is a well defined collection of distinct objects. The objects making up a set are called its elements. A set is usually denoted by a capital letter i.e., *A, B, C* etc. while its elements are denoted by small letters i.e., *a, b, c* etc. For example, the set *A* that consists of first five positive integers can be described as:

$$A = \{1, 2, 3, 4, 5\}$$

Here, for 3 belongs to set *A*, we write $3 \in A$; and read it as 3 belongs to A, while for 6 does not belong to set *A*, we write $6 \notin A$ and read it as 6 does not belong to *A*.

Null Set: A set that contains no element is called null set or empty set. It is denoted by { } or Φ.

Subset: If every element of a set A is also an element of a set B, A is said to be a subset of B and it is denoted by; $A \subseteq B$

Proper Subset: If A is a subset of B, and B contains at-least one element which is not an element of A, A is said to be a proper subset of B and is denoted by;

$$A \subset B$$

Finite and Infinite Sets: A set is finite, if it contains a specific number of elements, i.e., while counting the members of the set, the counting process comes to an end otherwise the set is an infinite set. For example; $A = \{1,2,3,5\}$, $B = \{x, y, z, t, u\}$ and $C = \{x | x \text{ is month of years}\}$ are finite sets.

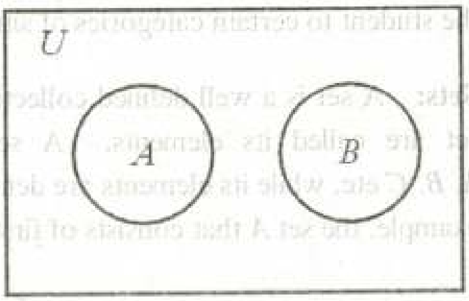
Whereas $D = \{2,4,6,8,...\}$ and $E = \{y | y \text{ is a point on a line}\}$ are infinite sets.

Universal Set: A set consisting of all the elements of the sets under consideration is called the universal set. It is denoted by U.

For example, if $A=\{1,2,3\}$, $B=\{2,4,5,6\}$, $C=\{8,10\}$, then $U=\{1,2,3,4,5,6,7,8,9,10\}$,

Disjoint Set: Two sets A and B are said to be disjoint sets, if they have no elements in common i.e., if $A \cap B = \Phi$, A and B are disjoint sets.

$A=\{6,8,10,12\}$ and $B=\{1,4,9,11\}$ are disjoint sets as $A \cap B = \Phi$



$A \cap B = \Phi$

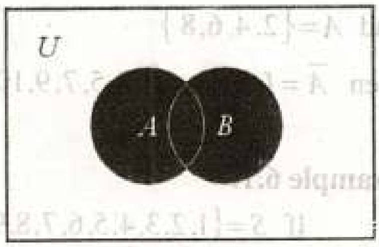
Overlapping Sets: Two sets A and B are said to be overlapping sets, if they have at-least one element in common, i.e., if $A \cap B \neq \Phi$ and none of them is the subset of the other set then A and B are overlapping sets.

For example $A=\{1,3,5,7\}$ and $B=\{1,4,8\}$ are overlapping sets, as $A \cap B = \{1\} \neq \Phi$ and none of A and B is the subset of the other.

Venn Diagram:

Venn diagram is a diagram in which universal set U is represented by a rectangle and its subset is represented by a circle. In other words Venn diagram represents the relationship between sets by means of diagram.

Union of Sets: Union of two sets A and B is a set that contains the elements either belonging to A or to B or to both. It is denoted by $A \cup B$ and read as A union B. For example, if $A=\{1,2,3,4,5\}$ and $B=\{2,4,6,8,10\}$ then $A \cup B=\{1,2,3,4,5,6,8,10\}$

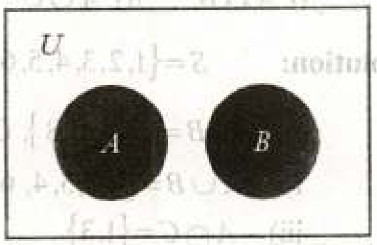


A ∪ B is shaded area

Let: $A=\{2, 4, 6\}$

$B=\{1, 3, 5\}$

then $A \cup B=\{1, 2, 3, 4,5, 6\}$

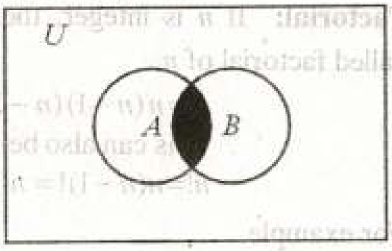


A ∪ B is shaded area

Intersection of Sets: Intersection of two sets A and B is a set that contains the elements belonging to both A and B. It is denoted by $A \cap B$ and read as A intersection B. For example, if

$A=\{1,2,3,4,5,6\}$ and $B=\{2,3,6,7\}$

then $A \cap B=\{2,3,6\}$

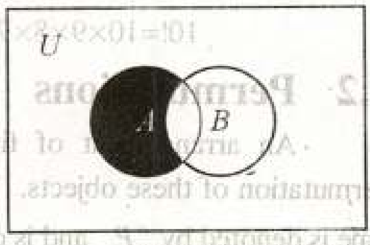


A ∩ B is shaded area

Difference of Sets: The difference of a set A and a set B is the set that contains the elements of the set A which are not contained in B. The difference of sets A and B is denoted by $A-B$. For example,

if $A=\{1,2,3,4,5,6,7,8\}$ and $B=\{2,4,6,8\}$

then $A-B=\{1,3,5,7\}$



A − B is shaded area

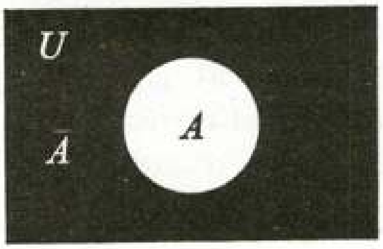
Complement of a Set: Complement of a set A denoted by $\overline{A}$ or A^c, is defined as $\overline{A} = U - A$.

For example if

$U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

and $A = \{2, 4, 6, 8\}$

then $\overline{A} = U - A = \{1, 3, 5, 7, 9, 10\}$



$\overline{\text{A}}$ is shaded area

Example 6.1:

If $S = \{1,2,3,4,5,6,7,8,9,10\}$, $A = \{1,2,3,4\}$ $B = \{2,4,6,8\}$, $C = \{1,3,5;7\}$ and $D = \{2,4\}$, then find :

i) $A \cup B$ ii) $A \cup C$ iii) $A \cap C$ iv) $C \cap B$ v) $\overline{C}$ vi) $\overline{A}$

Solution: $S = \{1,2,3,4,5,6,7,8,9,10\}$, $A = \{1,2,3,4\}$

$B = \{2,4,6,8\}$, $C = \{1,3,5,7\}$ and $D = \{2,4\}$

i) $A \cup B = \{1,2,3,4, 6, 8\}$ ii) $A \cup C = \{1,2,3,4,5,7\}$

iii) $A \cap C = \{1,3\}$ iv) $C \cap B = \Phi$

v) $\overline{C} = \{2,4,6,8,9,10\}$ vi) $\overline{A} = \{5,6,7,8,9,10\}$

Factorial: If n is integer, the product of first n positive integers denoted by $n!$ is called factorial of n.

$n! = n(n-1)(n-2)$ 3.2.1

This can also be written as :

$n! = n(n-1)! = n(n-1)(n-2)!$

For example,

$2! = 2 \times 1 = 2$

$4! = 4 \times 3 \times 2 \times 1 = 24$

$10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 3628800$

6.2 Permutations

An arrangement of finite number of objects in a definite order is called permutation of these objects. The number of ways of arranging n objects taken r at a time is denoted by ${}^nP_r$ and is defined as:

$$^nP_r = \frac{n!}{(n-r)!}$$

The number of permutation of n objects out of which n_1 are alike of one kind, n_2 are alike of second kind and so on, n_k are alike of kth kind is given by

$$\binom{n}{n_1, n_2, \ldots n_k} = \frac{n!}{n_1!\, n_2!\, n_3! \ldots .\ n_k!}$$

Example 6.2: How many distinct four–digit numbers can be formed from the following integers 1, 2, 3, 4, 5, 6 if each integer is used only once?

Solution: $\because\ ^nP_r = \frac{n!}{(n-r)!}$

Here, $n = 6$ and $r = 4$

$$\therefore {}^6P_4 = \frac{6!}{(6-4)!} = \frac{6!}{2!} = \frac{6\times5\times4\times3\times2\times1}{2\times1} = 360$$

Example 6.3: Evaluate: i) $^5P_3$ ii) $^{10}P_6$

Solution:

i) $$^5P_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{5\times4\times3\times2\times1}{2\times1} = 60$$

ii) $$^{10}P_6 = \frac{10!}{(10-6)!} = \frac{10!}{4!} = \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{4\times3\times2\times1} = 151200$$

6.3 Combinations

When a selection of objects is made without paying regard to the order of selection, it is called combination. The number of combinations of n things taken r at a time is denoted by ^nC_r or by $\binom{n}{r}$ and is defined by

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Example 6.4: Evaluate: (i) ${}^5C_3$ (ii) ${}^6C_2$ (iii) ${}^8C_5$

Solution:

i) $${}^5C_3 = \frac{5!}{3!(5-3)!} = \frac{5!}{3!\ \ 2!} = 10$$

ii) $${}^6C_2 = \frac{6!}{2!(6-2)!} = \frac{6!}{2!\ \ 4!} = 15$$

iii) $${}^8C_5 = \frac{8!}{5!(8-5)!} = \frac{8!}{5!\ \ 3!} = 56$$

Random Experiment: Random experiment is an experiment which produces different outcomes even if it is repeated a large number of times under similar conditions. A random experiment has the following properties:

i) The experiment can be repeated any number of times.

ii) A random trial consists of at least two possible outcomes.

Sample Space: A set representing all possible outcomes of a random experiment is called sample space. It is denoted by *S*. Each element in a sample space is called sample point. For example, when a coin is tossed, the sample space is given by

$$\text{S} = \{\text{H, T}\}$$

If a coin is tossed two times, the sample space is given by

$$\text{S} = \{\text{HH, HT, TH, TT}\}$$

In throwing a die, the sample space is given by

$$S = \{1,2,3,4,5,6\}$$

When two dice are thrown, the sample space is given by

$$\begin{aligned}\text{S} = \{&(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),\\ &(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),\\ &(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),\\ &(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),\\ &(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),\\ &(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}\end{aligned}$$

Event: Any subset of the sample space is called an event. In a sample space there can be two or more events consisting of sample points. For example, when a fair dice is rolled, the coming up of an even number upward is an event i.e., {2, 4, 6} is an event. Similarly, the coming up of odd numbers is an event i.e., {1, 3, 5} is an event.

Simple Event: If an event consists of one sample points, it is called simple event. For example, when two coins are tossed, the event {TT} that two tails appear is a simple event.

Compound Event: If an event consists of more than one sample points, it is called a compound event. For example, when two dice are rolled, an event B, the sum of two faces is 4 i.e., $B = \{(1, 3), (2, 2), (3, 1)\}$ is a compound event.

Independent Events: Two events A and B are said to be independent, if the occurrence of one does not affect the occurrence of the other. For example, in tossing two coins, the occurrence of a head on one coin does not affect in any way the occurrence of a head or tail on the other coin.

Dependent Events: Two events A and B are said to be dependent, if the occurrence of one event affects the occurrence of the other event.

Mutually Exclusive Events: Two events A and B are said to be mutually exclusive, if they cannot occur together i.e., $A \cap B = \Phi$

In other words, the two events are called mutually exclusive events, if they are disjoint. For example, in toss of a coin, either the head or the tail will appear, but they cannot appear together. The appearance of head and the appearance of tail are mutually exclusive.

Equally Likely Events: Two events are said to be equally likely, if they have the same chance of occurrence. For example, when a coin is tossed, it is just as likely to occur heads as to occur tails.

Exhaustive Events: When a sample space S is partitioned into some mutually exclusive events, such that their union is the sample space itself, the event are called exhaustive event. Let a die is rolled, the sample space is given by

$S = \{1, 2, 3, 4, 5, 6\}$

Let $A = \{1, 2\}$, $B = \{3, 4, 5\}$, $C = \{6\}$

A, B and C are mutually exclusive events and their union $A \cup B \cup C = S$ is the sample space, so the events A, B and C are exhaustive.

6.4 Probability

Classical or "A prior" definition: If there are n equally likely, mutually exclusive and exhaustive outcomes and m of which are favourable to the occurrence of an event A then the probability of the occurrence of the event A, denoted by $P(A)$ is defined by the ratio m/n i.e.,

$$P(A) = \frac{\text{no.of favourable outcomes}}{\text{no.of possible outcomes}} = \frac{m}{n}$$

Relative frequency or a posteriori definition: If an experiment is repeated a large number of times say n under uniform conditions and if the event A occurs m times, then the probability of the occurrence of the event A is defined by the relative frequency m/n which approaches a limits as n increases i.e.,

$$P(A) = \underset{n\to\infty}{\text{limit}} \frac{m}{n}$$

Mathematical Definition: The probability that an event A will occur, is the ratio of the number of sample points in A to the total number of sample points in S. i.e.,

$$P(A) = \frac{\text{no. of sample points in } A}{\text{no. of sample points in } S} = \frac{n(A)}{n(S)}$$

$P(A)$ must satisfy the following axioms:

i) $P(A)\geq 0$, which means that, probability of an event cannot be negative.

ii) $0\leq P(A)\leq 1$ i.e., Probability of an event lies between 0 and 1

iii) $P(S)=1$, which means that, sum of the probabilities is equal to one.

iv) If A and B are two mutually exclusive events, then

$$P(A\cup B) = P(A)+P(B)$$

Examples 6.5: A student solved 128 questions from first 200 questions of a book to be solved. What is the probability that he will solve the remaining all questions?

Solution: $n = 200$, $m' = 128$, $m = n - m' = 200 - 128 = 72$

$$\because P(A) = \frac{m}{n}$$

$$\therefore P(A) = \frac{72}{200} = 0.36$$

Example 6.6: A bag contains 4 red and 6 green balls out of which 3 balls are drawn. Find the probability of drawing

i) 2 red and 1 green balls. ii) all red balls.

iii) one green ball. iv) no red ball.

Solution:

Red	Green	Total
4	6	10

Balls to be drawn = 3

$$\text{Sample space} = \binom{10}{3} = 120$$

i) Let A be the event of drawing 2 red and one green ball

$$n(A) = \binom{4}{2}\binom{6}{1} = 36$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{36}{120} = 0.30$$

ii) Let B be the event of drawing all red balls.

$$n(B) = \binom{4}{3}\binom{6}{0} = 4$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{4}{120} = 0.033$$

iii) Let C be the event of drawing one green ball

$$n(C) = \binom{4}{2}\binom{6}{1} = 36$$

$$P(C) = \frac{n(C)}{n(S)} = \frac{36}{120} = 0.30$$

iv) Let D be the event of drawing no red ball

$$n(D) = \binom{4}{0}\binom{6}{3} = 20$$

$$\therefore P(D)=\frac{n(D)}{n(S)}=\frac{20}{120}=0.17$$

Example 6.7: If two fair dice are thrown, what is the probability of getting

i) a double six. ii) a sum of 8 or more dots.

Solution: Sample space is given by

$$\begin{aligned} S = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}\end{aligned}$$

$$\Rightarrow \quad n(S) = 36$$

i) Let A be the event that a double six occurs

$$A = \{(6,6)\} \Rightarrow n(A) = 1$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{1}{36}$$

ii) Let B be the event that a sum 8 or more dots occurs.

$$B = \begin{aligned}&\{(2,6),(3,5),(4,4),(5,3),(6,2),(3,6)(4,5)\ (4,6)\\ &(5,4),(5,5),(5,6),(6,3),(6,4),(6,5),(6,6)\}\end{aligned}$$

$$\Rightarrow n(B) = 15$$

$$\therefore \quad P(B) = \frac{n(B)}{n(S)} = \frac{15}{36} = \frac{5}{12}$$

Example 6.8: Six white balls and four black balls which are indistinguishable apart from colour, are placed in a bag. If six balls are taken from the bag, find the probability that their being three white and three black.

Solution:

White	Black	Total
6	4	10

$$\Rightarrow n(S) = \binom{10}{6} = 210$$

Let A be the event that three white and three black balls are taken

$$n(A) = \binom{6}{3}\binom{4}{3} = 80$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{80}{210} = \frac{8}{21}$$

Example 6.9: A fair die is tossed. Find the probability that the number on the uppermost face is not six.

Solution: Sample space is given by

$$S = \{1,2,3,4,5,6\} \Rightarrow n(S) = 6$$

Let A be the event that the uppermost face is 6 and $\overline{A}$ be the event that face is not 6. Then

$$A = \{6\} \Rightarrow n(A) = 1$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{1}{6} \text{ and } P(\overline{A}) = 1 - P(A) = 1 - \frac{1}{6} = \frac{5}{6}$$

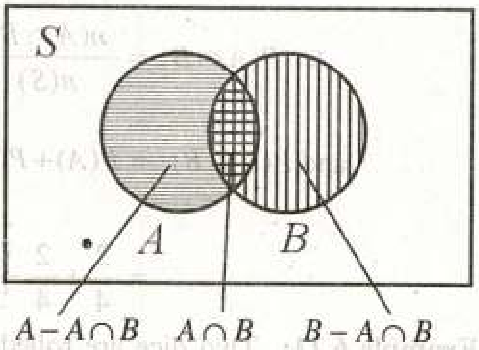
Theorem not Mutually Exclusive Events

Statement: If A and B are two not mutually exclusive events, the probability atleast one of them happens is the probability that event A occurs plus the probability that event B occurs minus the probability that both A and B occur simultaneously i.e.,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof: The event A or B can be expressed as the union of two mutually exclusive events A and $B - A \cap B$, then

$$A \cup B = A \cup (B - A \cap B)$$



By taking probability on both the sides we have,

$$P(A \cup B) = P[A \cup (B - A \cap B)]$$

$$= P(A) + P(B - A \cap B) \quad \text{........ (i)}$$

We can express B as a union of two mutually exclusive events $A \cap B$ and $B - (A \cap B)$ then

$$B = (A \cap B) \cup (B - A \cap B)$$

By taking probability on both the sides we have,

$$P(B) = P[(A \cap B) \cup (B - A \cap B)]$$

$$P(B) = P(A \cap B) + P(B - (A \cap B))$$

$$P(B) = P\,[A \cap B) + P(B - A \cap B)]$$

$$P(B) - P(A \cap B) = P(B - A \cap B)$$

By putting the value of $P(B - A \cap B)$ in equation (i) we get,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example 6.10: A coin is tossed twice, points of the sample space are HH, HT, TH, TT and each sample point with probability $\frac{1}{4}$.

If A and B are the events that head at first coin and tail on second coin respectively. Then find $P(A \cup B)$.

Solution: Sample space is given by

$$S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\} \Rightarrow n(S) = 4$$

$$\therefore \quad A = \{\text{HH, HT}\} \qquad \Rightarrow n(A) = 2$$

$$\therefore \quad P(A) = \frac{n(A)}{n(S)} = \frac{2}{4}$$

$$B = \{\text{TT, HT}\} \Rightarrow n(B) = 2$$

$$\therefore \quad P(B) = \frac{n(B)}{n(S)} = \frac{2}{4}$$

$$A \cap B = \{\text{ HT}\} \Rightarrow n(A \cap B) = 1$$

$$\therefore \quad P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{4}$$

and $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

Example 6.11: Two dice are rolled. If A and B are respectively the events that the sum of points is 8 and both dice should give odd numbers, Then find $P(A \cup B)$.

Solution: Sample space is shown in the example 6.7, where we see

$$n(S) = 36$$

$$A = \{(2,6),(3,5),(4,4),(5,3),(6,2)\} \Rightarrow n(A) = 5$$

$$\therefore \quad P(A) = \frac{n(A)}{n(S)} = \frac{5}{36}$$

$$\therefore B = \{(1,1),(1,3),(1,5),(3,1),(3,3),(3,5),(5,1),(5,3),(5,5)\} \Rightarrow n(B) = 9$$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{9}{36}$$

$$A \cap B = \{(3,5),(5,3)\} \Rightarrow n(A \cap B) = 2$$

$$\therefore P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{2}{36}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{5}{36} + \frac{9}{36} - \frac{2}{36} = \frac{1}{3}$$

Addition Theorem for Mutually Exclusive Events

If A and B are two mutually exclusive events then the probability either of them happening is the sum of their respective probabilities i.e.,

$$\boxed{P(A \cup B) = P(A) + P(B)}$$

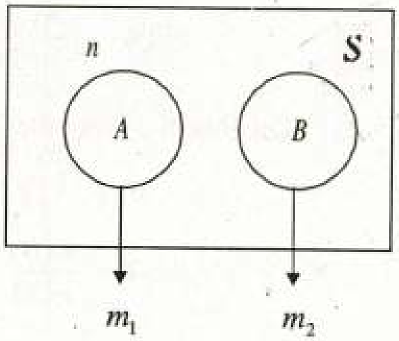
Proof: Let n be the number of sample points in a sample space S. Let A and B be the two mutually exclusive events in the sample S, such that event A contains m_1 sample points and event B contains m_2 sample points. Then $A \cup B$ will contain the sample points belonging to either A or B.

As A and B are two mutually exclusive events. i.e., $A \cap B = \Phi$

The sample points for $A \cup B$ will be $m_1 + m_2$

$$P(A \cup B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B)$$

$$\therefore P(A \cup B) = P(A) + P(B)$$



Example 6.12: A pair of dice are rolled. Find the probability that the sum of the uppermost dots is either 6 or 9.

Solution: Sample space is shown in the example 6.7, where we see

$$n(S) = 36$$

Let A be the event that the sum of the uppermost dots is 6, then

$$A = \{(1,5),(2,4),(3,3),(4,2),(5,1)\} \Rightarrow n(A) = 5$$

$$\therefore\ P(A) = \frac{n(A)}{n(S)} = \frac{5}{36}$$

Let B be the event that the sum of the uppermost dots is 9, then

$$B = \{(3,6),(4,5),(5,4),(6,3)\} \Rightarrow n(B) = 4$$

$$\therefore\ P(B) = \frac{n(B)}{n(S)} = \frac{4}{36} = \frac{1}{9}$$

Since A and B are mutually exclusive events.

$$\therefore P(A \cup B) = P(A) + P(B) = \frac{5}{36} + \frac{4}{36} = \frac{1}{4}$$

Example 6.13: A pair of dice is thrown. Find the probability of getting a total of either 5 or 11.

Solution: Sample space is shown in the example 6.7, where we see

$$n(S) = 36$$

Let A be the event that a total of 5 occurs, then

$$A = \{(1,4),(2,3),(3,2),(4,1)\} \Rightarrow n(A) = 4$$

$$\therefore\ P(A) = \frac{n(A)}{n(S)} = \frac{4}{36} = \frac{1}{9}$$

Let B be the event that a total of 11 occurs.

$$B = \{(5,6),(6,5)\} \Rightarrow n(B) = 2$$

$$\therefore\ P(B) = \frac{n(B)}{n(S)} = \frac{2}{36} = \frac{1}{18}$$

As events A and B are mutually exclusive, then

$$P(A \cup B) = P(A) + P(B) = \frac{1}{9} + \frac{1}{18} = \frac{1}{6}$$

Example 6.14: Three horses A, B and C are in a race. A is twice as likely to win as B and B is twice as likely to win as C, then

i) What are their respective chance of winning.

ii) What is the probability that B or C wins

Solution: Let $P(C) = P$

$$P(B) = 2P(C) = 2P$$
$$P(A) = 2P(B) = 2(2P) = 4P$$

i) Since A, B and C are mutually exclusive and collectively exhaustive events. Therefore, the probabilities must be equal to one i.e.,

$$P(A) + P(B) + P(C) = 1$$
$$\Rightarrow 4P + 2P + P = 1$$
$$\Rightarrow 7P = 1$$
$$\Rightarrow P = \frac{1}{7}$$
$$\therefore P(A) = \frac{4}{7},\ P(B) = \frac{2}{7} \text{ and } P(C) = \frac{1}{7}$$

ii) As B and C are mutually exclusive events, so

$$P(B \cup C) = P(B) + P(C) = \frac{2}{7} + \frac{1}{7} = \frac{3}{7}$$

6.4.1 Conditional Probability

If two events A and B are defined on a sample space S and if probability of B is not equal to zero, then the conditional probability of an event A given that B has occurred is written as $P(A/B)$ and is defined as:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \text{ where } P(B) > 0$$

If P(B)=0, the conditional probability is undefined

Multiplication Theorem For Independent Events

If A and B are two independent events, then the probability that A and B happen is the product of their respective probabilities, i.e.,

$$\boxed{P(A\cap B) \;=\; P(A)P(B)}$$

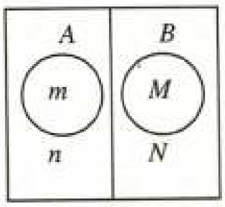
Proof: Let event A has n possible outcomes and m favourable outcomes and event B has N possible outcomes and M favourable outcomes then,

$$P(A) \;=\; \frac{m}{n} \text{ and } P(B) = \frac{M}{N}$$

As A and B are independent events so there will be nN possible outcomes for the joint events A and B. As each of n possible outcomes for A are associated with N possible outcomes for B. Similarly, each favourable outcomes for A is associated with each favourable outcomes for B. To have the favourable outcomes for compound event $A\cap B$. Then the total favourable outcomes for A and B are mM.

$$P(A\cap B) \;=\; \frac{mM}{nN} = \frac{m}{n}\times\frac{M}{N}$$

$$\text{or } P(A\cap B) \;=\; P(A)\,P(B)$$



Multiplication Theorem for not Independent Events

If A and B are two not mutually independent events, then the probability that both A and B happens is the probability of event A multiplied by the conditional probability of B given that event A has already occurred or is the probability of event B multiplied by the conditional probability of event A given that event B has already occurred i.e.,

$$\boxed{P(A\cap B) \;=\; P(A)P(B/A) \text{ or } P(A\cap B) \;=\; P(B)P(A/B)}$$

Proof: Let us have n sample points in a sample space S and A and B are not independent events such that event A has $m_{1,}$ event B has m_2 and $A\cap B$ has m_3 sample points.

Then, $P(A)=\frac{m_1}{n}$ and $P(B)=\frac{m_2}{n}$

$$P(A\cap B)=\frac{m_3}{n}$$

Now, $P(A\cap B)=\frac{m_3}{n}\times\frac{m_1}{m_1}=\frac{m_3}{m_1}\times\frac{m_1}{n}$

Here, $P(A)=\frac{m_1}{n}$ and $\frac{m_3}{m_1}=P(B/A)$

$\therefore\quad P(A\cap B)=P(B/A)\ P(A)$

or $P(A\cap B)=\frac{m_3}{n}=\frac{m_3}{n}\times\frac{m_2}{m_2}=\frac{m_3}{m_2}\times\frac{m_2}{n}$

Here, $\frac{m_2}{n}=P(B)$ and $\frac{m_3}{m_2}=P(A/B)$

$\therefore\quad P(A\cap B)=\ P(A/B)\ P(B)$

Example 6.15: Two *a*'s and two *b*'s are arranged in order, all arrangements are equally likely given that the last letters in order is *b*. Find the probability that 2 *a*'s are together.

Solution: $S=\{aabb, abba, bbaa, baab, baba, abab\}\ \Rightarrow n(S)=6$

Let *A* be the event that the two *a*'s are together

$$A\ =\{aabb, bbaa, baab\}\qquad\Rightarrow n(A)=3$$

$$P(A)\ =\frac{n(A)}{n(S)}=\frac{3}{6}=\frac{1}{2}$$

Let *B* be the event that the last letter is b

$$B\ =\{aabb, abab, baab\}\qquad\Rightarrow n(B)=3$$

$$\therefore\quad P(B)\ =\frac{n(B)}{n(S)}=\frac{3}{6}$$

$$A\cap B\ =\{aabb, baab\}\qquad\Rightarrow n(A\cap B)=2$$

$$\therefore\quad P(A\cap B)\ =\frac{n(A\cap B)}{n(S)}=\frac{2}{6}=\frac{1}{3}$$

$$P(A/B)\ =\frac{P(A\cap B)}{P(B)}=\frac{1/3}{1/2}=\frac{2}{3}$$

Example 6.16 In tossing two coins find:

i) The probability of two heads given that a head on the first coin.

ii) The probability of two heads given that atleast one head.

Solution: Sample space is

$$S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\} \Rightarrow n(S) = 4$$

Let A be the event that the head appears on the first coin.

$$A = \{\text{HH}, \text{HT}\} \Rightarrow n(A) = 2$$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{2}{4} = \frac{1}{2}$$

Let B be the event that two heads appear

$$B = \{\text{HH}\} \qquad \Rightarrow n(B) = 1$$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{1}{4}$$

$$A \cap B = \{HH\} \qquad \Rightarrow n(A \cap B) = 1$$

$$\therefore P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{4}$$

$$P\left(B/A\right) = \frac{P(A \cap B)}{P(A)} = \frac{1}{2}$$

ii) Let C be the event that atleast one head appears

$$C = \{\text{HH}, \text{HT}, \text{TH}\} \qquad \Rightarrow n(C) = 3$$

$$P(C) = \frac{n(C)}{n(S)} = \frac{3}{4}$$

$$B = \{\text{HH}\} \Rightarrow n(B) = 1$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{1}{4}$$

$$\therefore P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/4}{1/2} = \frac{1}{2}$$

$$B \cap C = \{HH\} \Rightarrow n(B \cap C) = 1$$

$$P(B \cap C) = \frac{n(B \cap C)}{n(S)} = \frac{1}{4}$$

$$\therefore P(B/C) = \frac{P(B \cap C)}{P(C)} = \frac{1/4}{3/4} = \frac{1}{3}$$

Example 6.17: A and B are two independent events. If $P(A) = 0.40$, $P(B) = 0.30$. Find the probabilities

i) $P(A \cap B)$ *ii)* $P(A/B)$ *iii)* $P(B/A)$

iv) $P(A \cup B)$ *v)* $P(\overline{A} \cap \overline{B})$ *iv)* $P(\overline{A}/\overline{B})$

Solution: We have

$$P(A) = 0.40, \quad P(B) = 0.30$$

Since A and B are independent events, therefore,

i) $P(A \cap B) = P(A)\,P(B) = (0.40)\,(0.30) = 0.12$

ii) $P(A/B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.12}{0.30} = 0.40$

iii) $P(B/A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0.12}{0.40} = 0.3$

iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$= 0.40 + 0.30 - 0.12 = 0.70 - 0.12$

$= 0.58$

v) $P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B})$

$= 1 - P(A \cup B) = 1 - 0.58 = 0.42$

vi) $P(\overline{A}/\overline{B}) = \dfrac{P(\overline{A} \cap \overline{B})}{P(\overline{B})} = \dfrac{P(\overline{A} \cap \overline{B})}{1 - P(B)}$

$P(\overline{A}/\overline{B}) = \dfrac{0.42}{1 - 0.30} = \dfrac{0.42}{0.70} = 0.6$

Example 6.18: Three missiles are fired at a target. If the probabilities of hitting the target are 0.4, 0.5 and 0.6 respectively and if the missiles are fired independently what is the probability that at least two missiles hit the target?

Solution:

$$\begin{aligned} P(A) &= 0.40 \Rightarrow P(\overline{A}) = 0.6 \\ P(B) &= 0.5 \Rightarrow P(\overline{B}) = 0.5 \\ P(C) &= 0.6 \Rightarrow P(\overline{C}) = 0.4 \end{aligned}$$

Let D be the event that at least two shots hit the targets.

$$\begin{aligned} P(D) &= P(A\cap B\cap \overline{C}) + P(A\cap \overline{B}\cap C) + P(\overline{A}\cap B\cap C) + P(A\cap B\cap C) \\ &= P(A)\,P(B)\,P(\overline{C}) + P(A)\,P(\overline{B})\,P(C) + P(\overline{A})\,P(B)\,P(C) + P(A)\,P(B)\,P(C) \\ &= 0.4\times0.5\times0.4 + 0.4\times0.5\times0.6 + 0.6\times0.5\times0.6 + 0.4\times0.5\times0.6 \\ &= 0.08+0.12+0.18+0.12 \end{aligned}$$

or $P(D) = 0.50$

Example 6.19: A purse contains 2 silver,4 copper and a second purse contains 4 silver, 3 copper coins. If a coin is selected at random from one of the purses. What is the probability that it is a

i) Silver coin ii) Copper coin.

Solution:

	Purse I	**Purse II**
Silver coin	2	4
Copper coin	4	3
Total coin	6	7
P (Purse I)	½	½

i) Let A be the event that the selected coin is silver

$$P(A) = P(\text{Purse I})\,P\left(\text{Silver coin}/_{\text{Purse I}}\right) + P(\text{Purse II})\,P\left(\text{Silver coin}/_{\text{Purse II}}\right)$$

$$= \left(\frac{1}{2}\times\frac{2}{6}\right) + \left(\frac{1}{2}\times\frac{4}{7}\right) = \frac{2}{12} + \frac{4}{14} = \frac{38}{84} = \frac{19}{42}$$

ii) Let B be the event that the coin selected is a copper, then

$$P(B) = P(\text{Purse I})\,P\left(\text{Coper coin}\big/_{\text{Purse I}}\right) + P(\text{Purse II})\,P\left(\text{Copper coin}\big/_{\text{Purse II}}\right)$$

$$= \left(\frac{1}{2}\times\frac{4}{6}\right)+\left(\frac{1}{2}\times\frac{3}{7}\right)=\frac{4}{12}+\frac{3}{14}=\frac{23}{42}$$

Exercise 6

Ans on Page 255

6.1. i) Define permutation and combination and discriminate between these two?

ii) Let $A = \{1, 3, 5, 7\}$, $B = \{2, 4, 6, 8\}$, $C = \{1,2,3,4,5\}$ and $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$

list the elements of the following:-

a) $A \cap B$ b) $C \cup A$ c) $\overline{A \cap C}$ d) $B \cap C$

6.2 Evaluate the following

i) 7! ii) 16!/8! iii) 17!/8!.4!

iv) ${}^{9}p_{6}$ v) ${}^{15}P_{6}$ vi) ${}^{5}P_{5}$

vii) ${}^{5}c_{3}$ viii) ${}^{52}c_{13}$ ix) ${}^{11}c_{4}$

6.3 Let $A = \{1, 4\}$, $B = \{2, 3\}$, $C = \{3\}$ be the subsets of the universal set $S = \{1, 2, 3, 4\}$, find

i) $A \times B$ ii) $B \times A$ iii) $A \times A$ iv) $B \times B$

6.4 What do you understand by

i) sample space ii) event iii) sample point iv) simple and compound event?

6.5 Define mutually exclusive, independent and dependent events.

6.6 Explain with examples the terms; random experiment, sample space and an event.

6.7 State and prove the multiplicative law of probability for two events A and B that are not statistically independent.

6.8 Find the probability of each of the following:

i) A head appears in tossing a fair coin.

ii) A 5 appears in rolling a six faced cubical dice.

iii) An even number appears when a perfect cubical die is rolled.

6.9 What is the probability of selecting a card of diamonds from a pack of playing cards consisting of the usual 52 cards.

6.10 Show that in a single throw with two die, the chance of throwing more than 7 is equal to that of throwing less than 7.

6.11 A bag contains 12 balls of which 3 are marked, if 5 balls are drawn out together. What is the probability that 3 of the marked balls are among these 5?

6.12 What is the probability of throwing either 7 or more than 10 with two dice?

6.13 A bag containing 2 red, 3 green, 5 blue and 2 yellow balls. Find the probability that balls of all colours, are represented in a sample if four balls are selected at random.

6.14 A bag contains 5 white and 7 black balls. If 3 balls are drawn from the bag, what is the probability that;

i) All are white. ii) Two are white and one is black.

iii) All are of the same colour?

6.15 Determine the probability for the following events:

i) the sum 8 appears in a single toss of a pair of fair dice.

ii) A sum 7 or 11 comes up in a single toss of a pair of fair dice.

iii) A ball drawn at random from a bag containing 5 red, 6 white, 4 blue and 3 orange balls is either red or blue.

6.16 Two cards are drawn at random from a well shuffled pack of 52 cards. Find the probability that:

i) One is king and other is queen. ii) Both are of same colour.

iii) Both are of different colours.

6.17 A bag contains 9 white and 12 black balls. Find the probability of drawing 5 black balls out of the bag containing 21 balls.

6.18 From a bag containing 5 white, and 3 black balls, 2 are drawn at random. Find the chance that both are of the same colour.

6.19 A set of eight cards contains one joker. *A* and *B*, are two players, choose 5 cards at random, *B* takes the remaining 3 cards, what is the probability that *A* has a joker?

6.20 From a pack of 52 cards, two cards are drawn. What is the probability that one is king and the other is queen.

6.21 In a poker hand consisting of 5 cards. What is the probability of holding?

i) 2 aces and 2 kings.

ii) 5 spades.

6.22 A bag containing 14 identical balls, out of which 4 are red, 5 black and 5 white balls, if three balls are drawn from the bag. Find the chance that

i) 3 are red. ii) At least two are white.

6.23 A marble is drawn at random from a box containing 10 red, 30 white, 20 blue and 15 orange marbles. Find the probability that it is:

i) orange or red. ii) not red or orange.

ii) not blue. iv) red, white or blue.

6.24 What is meant by conditional probability?

6.25 State and prove multiplication laws of probabilities for independent and dependent events.

6.26 *A* and *B* can solve 60% and 80% of the problems in a book respectively. What is the probability that either *A* or *B* can solve a problem chosen at random?

6.27 A class contains 10 men and 20 women out of which half men and half women have brown eyes. Find the probability that a person chosen at random is a man or has brown eyes.

6.28 A box contains 9 tickets numbered 1 to 9. If 3 tickets are drawn from the box one at a time, find the probability that they are alternately either odd even odd or even odd even.

6.29 Two drawings each of 3 balls are made from a bag containing 5 white and 8 black balls. The balls are not being replaced before the next trial. What is the probability that the first drawing will give 3 white balls and the second drawing will give three black balls.

6.30 Three balls are drawn successively from a box containing 6 red balls, 4 white balls and 5 blue balls. Find the probability that they are drawn in the order red, white and blue if each ball is

i) replaced ii) not replaced.

6.31 One bag contains 4 white and 2 black balls another bag contains 3 white and 5 black balls. If one ball is drawn from each bag, find the probability that.

i) both are white ii) both are black.

6.32 Two urns contain respectively 3 white and 7 red balls 15 black and 10 white balls, 6 red and 9 black balls. One ball is taken from each urn. What is the probability that both will be of the same colour?

6.33 The probability that a man will alive in 25 years is 3/5 and that his wife will be alive in 25 years is 2/3, find the probability that.

i) Both will alive in 25 years.

ii) Only the man will alive in 25 years.

iii) Only the wife will alive in 25 years.

iv) Atleast one will alive in 25 years.

v) None of them will alive in 25 years.

vi) At the most one will alive in 25 years.

6.34 Three cards are drawn at random from an ordinary pack of 52 cards. Find the probability that they will consists of a Jack, a queen and a king.

6.35 Two cards are drawn from a well shuffled pack of 52 cards. Find the probability that they are both aces if the first cards drawn is:

i) replaced. ii) not replaced.

6.36 Urn *A* contains 5 red balls and 3 white balls and urn *B* contains 2 red balls and 6 white balls.

i) If a ball is drawn from each urn what is the probability that both are of the same colour?

ii) If two balls are drawn from each urn what is the probability that all 4 balls are of the same colour?

6.37 Three Ghori missiles are fired at a target. If the probabilities of hitting the target are 0.4, 0.5 and 0.6 respectively and the missiles are fired independently, what is the probability that atleast 2 missiles hit the target?

6.38 Assume that *X* is a number chosen at random from the set of integers between 1 and 14 respectively. What is the probability that

i) *X* is a single digit number. ii) *X* is a multiple of 5 or 6.

6.39 What are the odds for the occurrence of an even if its probability is 4/7?

6.40 Suppose that it is 9 to 7 against a person *A* who is now 35 years of age living till he is 65 and 3 to 2 against a person *B* now 45 years of age living till he is 75. Find the probability that one of these persons will be alive 30 years hence.

6.41 A purse contains 2 silver and 4 copper and second purse contains 4 silver and 3 copper coins. If a coin is selected at random from one of the purses. What is the probability that it is a

i) Silver coin. ii) Copper coin.

6.42 One purse contains 1 sovereign and 3 shillings, a second purse contains 2 sovereign and 4 shillings and third purse contains 3 sovereign and 1 shilling. If the coin taken out of the purse is selected at random. Find the chance that it is sovereign.

6.43 The probability that a student will get a grade of *A*, *B* or *C* in statistics course are 0.09, 0.15 and 0.53 respectively, what is the probability that the student will get a grade lower than *C*.

6.44 3 coins are tossed, what is the probability of getting atleast one head?

6.45 *A* and *B* play 12 games of chess out of which 6 are won by *A* and 4 by *B* and two games end in a tie. They agree to play a tournament consisting of 3 games. Find the probability that

i) A wins all the 3 games. ii) Two games end in tie.

iii) *A* and *B* won alternately. iv) *B* wins atleast 1 game.

6.46 For two independent events *A* and *B*, $P(A) = 0.25$ and $P(B) = 0.40$. Find $P(A \cap B)$?

6.47 For 2 rolls of a balanced die, find the probability of getting 1st a five and then a number less than 4.

6.48 If two cards are drawn from an ordinary deck of 52 cards. What is the probability that both will be diamonds, if the drawing is without replacement?

6.49 *A*, *B* and *C* take turns in throwing a die for a prize to be given to one who first obtains 6. Compare their chances of success.

6.50 First bag contains 4 white balls and 3 black balls and second bag contains 3 white and 5 black balls. One ball is drawn from the first bag and placed unseen in the second bag. What is the probability that a ball now drawn from the second bag is black?

6.51 From a bag containing 4 white and 5 black balls, 2 balls are drawn at random. Find the probability that they are of same colour.

6.52 3 groups of children contains respectively 3 girls and 1 boy, 2 girls and 2 boys and 1 girl and 3 boys. One child is selected at random from each group. Find the chance that the selected group comprises of one girl and two boys.

6.53 *A* can hit a target 4 times in 5 shots, *B* can hit 2 times in 5 shots and *C* can hit 2 times in 4 shots. Find the probability that

i) 2 shots hit. ii) at least two shots hit.

6.54 Fill in the blanks:

i) A set is any _________ collection of _________ things.

ii) A set containing only one element is called _________ set.

iii) A set of subsets of set is called _________ of the set.

iv) A _________ event contains more than one element.

v) Two events are _________ if they have no common point.

vi) If $A \cup B = S$ then A and B are _________ events.

vii) If $n(A) = n(B)$, then A and B are _________ events.

viii) The number of subsets of a set containing n points are _________.

ix) The orderly arrangements of r distinct things out of n are called _________ and denoted by _________.

x) A non-orderly arrangement of things is called _________.

6.55 Against each statement, write T for true and F for false statement.

i) The null set is also named as the impossible event.

ii) A part of a set is called improper subset of the set.

iii) A subset of sample space is called sure event.

iv) An event consisting of only one sample point is called compound event.

v) If a coin is tossed four times, the number of sample points will be 22.

vi) If A and B are mutually exclusive events then, $P(A \cup B) = P(A) + P(B)$.

vii) The probability of drawing a red card from 52 cards is 13/15.

viii) The complementary events are always mutually exclusive events.

ix) When a coin is tossed, the sample space is {HH, HT}.

x) If A and B are independent, then $P(A \cup B) = P(A) + P(B)$.

7 Random Variables

$0 \leq P(A) \leq 1$

7.1. Introduction

Every random experiment results in two or more outcomes and usually the interest is in a particular aspect of the outcomes of the experiment. For example, when a pair of dice is thrown, the interest may be in the total of upturned dots on both dice. In case of this experiment, the total may be 2 (one on each die), 3 (one on the first die and two on the other) and so on. It may be 4, 5, 6, 7, 8, 9, 10, 11 or 12. In the language of probability, these values associated with outcomes are the values of a so-called random variable. An other example is the total number of children in each of the fifty randomly chosen families. If no family has more than 5 children then the values of this random variable i.e., number of children in each family, would be 0,1,2,3,4,5 i.e., no child, one child, two children, 3 children, 4 children and five children respectively.

A variable whose values depend upon the outcomes of a random experiment is called a **random variable**. We will denote the random variable by the capital letters *X*, *Y* or *Z* and their values by the corresponding small letters *x*, *y* or *z*.

Example 7.1 : Let a pair of dice be thrown and *Y* denote the random variable that is the sum of upturned values on the two dice. There are 36 outcomes and *Y* assigns to the outcome (1, 1) the real number $1 + 1 = 2$. It assigns to the outcome (2, 1), the real number $2 + 1=3$ and so on uptil the outcome (6, 6), the real number $6 + 6 = 12$ so, the values assigned are 2,3,4,5,6,7,8,9,10,11 and 12.

7.2 Random numbers and their generation

Random numbers are a sequence of digits from the set {0, 1, 2, ... , 9}. So that, at each position in the sequence, each digit has the same probability 0.1 of being selected irrespective of the actual sequence, so far constructed. The probability is 0.1

because out of ten digits {0, 1, 2, ..., 9} each digit has equal probability i.e., 1/10 or 0.1. These are also known as random digits.

The simplest ways of achieving such numbers are games of chance such as dice, coins, cards or by repeatedly drawing numbered slips out of a hat. These are usually grouped purely for convenience of reading but this would become very tedious for long runs of each digits. Fortunately tables of random digits are now widely available (see table 7.1).

For implementation on computers to provide sequences of such digits easily and quickly, the most common methods are called **Pseudo random techniques**. Here, digits will re-appear in the same order (i.e., cycle) eventually but for a good technique the cycle might be tens of thousands of digits long. Of course the **Pseudo random digits** as the title says, are not truly random. In fact, they are completely deterministic but they do exhibit most of the properties of random digits.

Generally, these methods involve the recursive formula as

$$x_{n+1} = ax_n + b (\text{mod}\, m); \; n = 0,1,2,3, ... \qquad .... \qquad (7.1)$$

Here a, b and m are suitably chosen integer constants and the seed x_1 (a starting number) is an integer. By mod m we means that if the answer is greater than m, then divide it by m and keep the remainder as a random number. Use of this formula gives rise to a sequence of integers each of which is in the range 0 to m-1. We simply run these together to give our sequence of pseudo random digits. Clearly this to be of any value; m, a and / or b should be large.

Example 7.2: Let $a = 13$, $b = 0$ and $m = 16$. Generate 4 random numbers.

Solution: According to the relation

$$x_{n+1} = ax_n + b \;\; (\text{mod } 16) \text{ for } n = 0, 1, 2, ...$$

Let a seed x_0 be 5, then for $n = 0$, we have

$$
\begin{aligned}
x_1 &= 13x_o + b \pmod{16} \\
&= 13(5) + 0 \pmod{16} \\
&= 65 \pmod{16} \\
&= 1 (\text{dividing } 65 \text{ by } 16, \text{ the remainder is } 1)
\end{aligned}
$$

For $n = 1$, we have,

$$
\begin{aligned}
x_2 &= 13(1) + 0 \pmod{16} \\
&= 13
\end{aligned}
$$

For $n = 2$, we have,

$$
\begin{aligned}
x_3 &= 13(13) + 0 \pmod{16} \\
&= 9
\end{aligned}
$$

Similarly, for $n = 3$, we have,

$$
\begin{aligned}
x_4 &= 13(9) + 0 \pmod{16} \\
&= 5
\end{aligned}
$$

So, the random numbers are 1, 13, 9, 5.

7.3 Application of random numbers

The random numbers have widely applicability in the simulation techniques (also called Monte Carlo Methods) which have been applied to many problems in the various sciences and are useful in the situations where direct experimentation is not possible, the cost of conducting an experiment is very high or the experiment takes too much time.

The random number tables are constructed from the random numbers where, the random numbers are grouped for the purposes of reading. The groups may consist of 2-digit, 3-digit, 4-digit or 5-digit sequence of random numbers (0, 1, 2, 3, 4, 5, 6, 7, 8, 9). A five digit sequence of these are given in table 7.1. We may use this table for 1-digit random numbers, 2-digit random numbers and so on uptil 5-digit random numbers depending upon the situation by using it row-wise or column-wise.

Table 7.1: Random Numbers

3	3	6	8	1	7	5	0	2	9	3	0	0	7	8	0	6	5	5	8	2	4	3	6	6
3	2	4	8	9	1	2	8	9	5	5	8	6	2	3	2	2	5	7	6	2	9	6	8	3
6	8	0	7	8	9	4	2	0	0	2	5	4	8	8	5	0	2	2	7	2	5	2	9	6
4	7	6	3	8	9	3	1	2	1	5	7	8	7	7	0	5	3	7	2	0	3	4	8	8
8	0	6	4	5	8	1	4	7	4	6	7	7	3	4	1	1	4	2	1	0	7	3	4	0
5	5	2	7	1	8	0	3	5	4	7	3	9	6	0	4	4	1	8	7	2	8	2	9	2
7	3	4	5	7	8	1	9	0	5	1	7	2	0	8	8	3	3	3	4	2	7	4	9	7
7	8	2	1	2	3	8	0	7	5	1	8	5	6	2	7	8	9	7	0	9	9	9	2	1
2	0	3	9	3	4	1	2	4	7	5	9	7	7	3	9	2	4	3	7	7	6	1	2	6
5	9	2	5	3	4	7	6	5	3	6	6	1	1	9	9	7	6	8	4	2	1	4	4	0
0	9	3	2	9	7	9	7	6	8	7	2	9	6	8	5	6	2	9	7	1	7	5	6	1
1	2	8	6	1	6	2	8	6	4	9	9	8	3	2	6	3	5	4	3	9	5	0	0	5
5	8	3	5	3	4	3	9	4	4	2	9	6	8	3	1	2	2	9	3	2	5	1	2	5
9	1	5	5	8	2	9	1	2	5	2	7	8	5	2	7	9	3	3	4	3	9	7	6	8
4	0	3	8	0	4	6	3	4	0	1	4	7	0	6	4	7	2	9	3	7	5	4	7	9
2	9	1	1	9	8	8	3	4	5	2	7	2	5	7	9	9	3	5	5	5	9	6	3	3
6	2	0	8	1	2	0	7	6	2	8	8	5	4	3	9	5	2	3	9	9	7	7	2	9
1	5	2	2	8	4	5	8	8	4	9	3	1	1	9	3	0	9	9	2	6	8	4	8	3
2	9	2	2	8	1	4	7	0	4	4	2	1	6	7	4	2	0	2	0	4	7	8	3	8
8	5	9	6	6	0	9	4	8	3	2	0	8	0	4	6	7	6	0	9	9	0	2	8	6
5	5	6	1	2	4	0	2	8	2	9	3	9	5	6	6	6	2	7	8	1	1	2	3	8
2	2	6	5	2	5	9	4	1	1	0	6	0	6	8	3	9	0	6	1	7	8	0	1	9
6	1	5	2	1	1	7	4	9	6	7	4	3	2	8	7	8	2	1	3	6	5	4	3	9
2	4	3	2	4	7	3	7	6	9	3	9	1	1	6	3	4	8	6	5	4	3	5	5	7
9	4	8	1	4	1	6	8	1	7	5	2	6	0	7	7	7	0	8	2	8	4	4	9	0
9	1	3	3	6	0	7	1	6	7	0	9	0	9	9	0	2	7	3	1	3	4	5	6	3
3	3	3	4	8	4	1	1	0	8	2	0	1	3	6	3	7	1	5	6	4	5	7	4	6
8	4	2	3	6	8	5	4	0	9	0	6	9	8	7	4	7	4	1	5	9	1	6	0	4
9	0	0	4	5	6	8	3	2	7	2	3	3	9	6	5	1	3	0	8	8	2	3	1	9
7	9	9	7	8	9	8	6	0	0	3	4	2	6	0	4	8	6	3	2	8	0	9	2	7
7	8	4	0	8	2	4	7	4	9	5	0	4	7	8	7	0	8	0	7	5	4	4	4	2
6	2	9	4	0	4	3	7	8	2	0	8	3	1	3	2	8	6	9	1	8	5	0	5	9
5	2	0	5	2	6	4	1	7	9	7	4	4	5	4	3	5	1	8	9	5	3	8	7	3
2	5	0	3	4	6	6	6	9	2	3	0	9	7	5	7	3	2	4	2	3	8	4	0	8
3	6	4	5	3	5	7	2	5	0	6	5	3	8	6	8	0	3	3	9	9	0	9	2	5
7	7	3	7	0	7	5	7	3	1	6	5	0	0	1	8	5	9	5	0	9	4	5	0	0
6	3	3	7	1	3	6	5	4	6	8	2	9	3	6	0	1	7	7	5	6	0	4	3	6
6	1	5	5	9	2	3	9	7	0	0	5	6	8	1	8	4	0	7	8	0	4	0	8	7
9	9	4	3	3	4	9	2	2	0	5	8	1	9	3	6	7	3	8	2	4	8	3	2	7
6	3	8	0	4	5	2	0	0	3	2	9	0	6	6	7	2	3	8	0	5	6	3	7	8
7	8	5	4	3	6	6	1	2	9	5	3	1	3	8	3	6	2	4	2	7	2	1	2	1
7	4	9	6	3	8	9	8	5	0	8	7	4	9	6	9	4	7	6	2	5	2	0	9	9
8	5	8	2	2	9	9	4	1	9	1	8	6	3	6	6	7	8	1	2	7	4	3	7	2
2	2	2	7	0	0	1	3	5	1	7	9	2	3	8	7	3	6	6	3	6	0	5	8	4
1	5	9	2	3	5	2	4	6	5	9	5	4	7	6	7	6	7	8	5	3	8	5	9	1
1	8	5	2	6	3	3	1	6	8	5	6	9	6	8	1	2	1	8	5	3	3	3	5	0
8	2	5	9	0	0	3	9	7	8	6	4	5	4	7	2	9	5	1	6	2	8	3	8	2
0	3	8	9	8	9	9	6	7	5	7	0	5	0	9	7	2	4	6	2	5	2	6	6	6
4	8	1	5	3	8	8	3	0	6	6	5	8	5	7	0	5	1	3	6	9	9	8	2	2
9	6	9	5	5	6	1	5	9	3	2	0	4	9	8	2	9	4	9	7	9	1	3	5	1
1	2	0	0	6	9	3	3	0	9	6	4	9	9	6	4	8	8	6	8	3	7	2	2	3
5	6	9	9	7	9	8	9	7	8	3	8	7	3	0	0	6	6	9	2	4	2	4	6	9
9	3	9	5	3	1	6	9	7	9	9	2	4	0	8	5	5	6	9	7	3	0	6	5	8
5	4	8	7	7	5	8	4	9	6	4	1	8	8	0	6	7	2	0	5	9	3	0	2	2
2	1	7	7	5	8	6	0	3	2	4	4	9	9	0	6	5	5	0	1	9	9	9	0	2

To explain the use of a random number table, consider the following example 7.3.

Example 7.3: Count the number of heads and tails when a single coin is thrown 12 times without throwing a coin.

Solution: The first step is to number both the heads and tails staying within 0 to 9. Let 0, 2, 4, 6, 8 (even numbers including zero) indicate heads and the odd numbers 1, 3, 5, 7, 9 represent tails.

The second step is to open the random number table given in table 7.1 and select arbitrarily a column, say column 1 and select arbitrarily a row say row 4 and start reading a set of 1-digit i.e., 4, 8, 5, 7, 7, 2, 5, 0, 1, 5, 9, 4. Third step is to interpret them as H, H, T, T, T, T, H, T, H, T, T, T, H as was decided head for 0, 2, 4, 6, 8 and tail for 1, 3, 5, 7, 9. So, there are 5 heads and 7 tails. These are given in the following frequency table.

Outcome	**frequency**
Head (one head)	5
Tail (0 head)	7

Similarly. to count the number of heads when two coins are thrown 20 times, the first step is to take two-digit random numbers. If both are even including zero, then both are heads; if both are odd then both are tails (0 head), and if one is even and one is odd then there is one head. The second step is to open the random number table and select arbitrarily a column say column 2 and select arbitrarily a row say row 2 and start reading two digit numbers i.e., 24, 80, 76, 06, 52, 34, 92, 03, 92, 93, 28, 83, 15, 03, 91, 20, 52, 92, 59, 56. Third step is to interpret them as HH, HH, TH, HH, TH, TH, TH, HT, TH, TT, HH, HT, TT, HT, TT, HH, TH, TH, TT, TH. So the frequency table would be

Outcome	*f*
0 head	4
1 head	11
2 head	5

Another method known as probability proportional to size (PPS) is used when we have information about the probabilities of the outcomes.

Example 7.4: If a coin is thrown thrice or three coins are tossed once) then the number of heads may be 0 (all tails) 1, 2 and 3 and the corresponding relative frequencies would be as given in adjoining table.

Find the sequence of the number of heads 20 times without throwing a coin?

Solution: Here we use 3-digits random numbers because probabilities are given to three decimals.

Number of heads	Probability
0	0.125
1	0.375
2	0.375
3	0.125

The first step is to make a cumulative probabilities (c.p) column.

The second step is to assign random numbers (*r*.no) from 0 to 999 because we have 3-digit probabilities. Note that in each class the random number assigned is one less than the number formed by the corresponding cumulative frequency. The reason is that the random numbers are from 0 to 999 not from 1 to 1000. These are shown in the table 7.2.

Table 7.2: cumulative probabilities and range.

Heads	Prob.	c.p.	r.no.
0	0.125	0.125	000-124
1	0.375	0.500	125-499
2	0.375	0.875	500-874
3	0.125	1.000	875-999

The third step is to open the random number table 7.1 and select arbitrarily a column, say column 4, and select arbitrarily a row, say row 2, the 3-digit random numbers are 441, 833, 789, 924, 976, 562, 635, 122, 793, 472, 993, 952, 309, 420, 676, 662, 390, 782, 348, 773 and the corresponding number of heads are 1,2, 2, 3, 3, 2, 2, 0, 2, 1, 3, 3, 1, 1, 2, 2, 1, 2, 1 and 2.

441 indicates 1 head because it is in the class corresponding to 1 head.

833 indicates 2 heads because it is in the class corresponding to 2 heads.

789 indicates 2 heads because it is in the class corresponding to 2 heads.

924 indicates 3 heads because it is in the class corresponding to 3 heads and so on. The corresponding frequency table is as given in table 7.3.

Table 7.3: Frequency Distribution

Heads	0	1	2	3
f	1	6	9	4

7.4 Concept of random variables and their construction from different fields.

As we are familiar by now that random variables arise from the outcomes of random experiments by associating a value to each outcome. The following examples may help explain them in detail.

Example 7.5: Consider an experiment in which three students in a class are asked to take one of the two courses Biology (B) or Computer Science (C).

Solution: Define the random variable Y by

Y = The number of students taking computer science.

The possible values are

No. student takes computer science so, $y = 0$

One student takes computer science so, $y = 1$

Two students take computer science so, $y = 2$

Three students take computer science so, $y = 3$

The possible outcomes of the experiment are:

BBB	first takes biology, second takes biology, third takes biology.
CBB	first takes computer science, second takes biology third takes biology.
BCB	first takes biology, second takes computer science, third takes biology.
BBC	first takes biology, second takes biology, third takes computer science.
CCB	first takes computer science, second takes computer science, third takes biology.
CBC	first takes computer science, second takes biology, third takes computer science.
BCC	first takes biology, second takes computer science, third takes computer science.
CCC	first takes computer science, second takes computer science, third takes computer science.

These are represented as follows:

BBB	CBB	BCB	BBC	CCB	CBC	BCC	CCC
0	1	1	1	2	2	2	3

Note that the values of the random variable Y are isolated points 0, 1, 2 and 3.

Example 7.6 : Consider an experiment of recording the time (minutes) taken by the customers to wait for its turn in a utility store while standing in a queue.

Define the random variable Y where Y is the time taken by the customers.

Solution: The first customer may take 5.0 minutes, the second may takes 6.0 minutes, the third may take 5.30 minutes, fourth may takes 12.0 minutes and so on.

It may be noted that Y may take any value in an interval on a number line.

7.5 Discrete and continuous random variables

Discrete random variable

A random variable is called discrete if the set of values it takes is a collection of isolated points on a real line i.e., the sample space S is a discrete sample space.

The outcomes of an experiment are noted and the value of the random variable is a number appropriately assigned to each outcome by some rule.

Example 7.7: Three coins are tossed and let the random variable Y denote the number of heads. Then the outcomes and the values of Y are given in adjoining table.

Outcomes	**Value of Y**
HHH	3
HHT	2
HTH	2
HTT	1
THH	2
THT	1
TTH	1
TTT	0

The values of Y are whole numbers. So, the sample space is discrete. Thus Y is a discrete random variable.

Example 7.5 above is also another example of a discrete random variable.

Continuous random variable: A random variable is called continuous if the set of values it takes is an entire interval on the number line i.e., the sample space S is continuous. The outcomes of an experiment are represented by the points on a line and the value of the random variable is a number appropriately assigned to each point by some rule.

Example 7.8: Consider the experiment of measuring the height of students in a statistics class. If the minimum height of a student is 5.0 feet and the maximum height is 5.8 feet, then the variable Y, the height of students, takes values between 5.0 and 5.8 feet i.e., in the interval 5.0 – 5.8 feet.

Example 7.6 is also an example of a continuous random variable.

Exercise 7

Ans on Page 257

7.1 i) Define random variable and give an example to explain it.

ii) What are random numbers and how these are generated. Also give an example to explain their application.

iii) Classify each of the following random variables as either discrete or continuous.

a) The number of pages in a book.

b) The number of questions asked in an oral examination.

c) The life time of a light bulb.

d) The amount of rainfall at a particular location during different months of 1996.

7.2 Generate the first 6 random digits using the pseudo-random number generator with

$m = 100,\ a = 21, b = 7, x_0 = 10$

7.3 Count the number of heads and tails when a single coin is tossed 10 times without throwing a coin.

7.4 Two coins are tossed. Let Y denotes the number of heads, then the possible number of heads and their corresponding probabilities are given in the adjoining table

Y	$p(Y)$
0	0.25
1	0.50
2	0.25

Find the sequence of number of heads 20 times without throwing a coin?

7.5 Let the digits 0, 1, 2, 3, 4 represent head and 5, 6, 7, 8, 9 represent tail, use random numbers to simulate 20 flips of a coin.

7.6 Two students in a class are asked to take one of the two courses Mathematics (M) or Biology (B). Define the random variable Y as number of students taking Biology. Write down the possible outcomes and the values assigned by the random variable Y.

7.7 Two coins are tossed and let the random variable y denote the number of heads. Write down the possible outcomes and the values assigned to the random variable.

7.8 There are three children in a family. Let the random variable denote the number of boys in a family. Write down the possible outcomes and the values assigned by the random variable assuming equal chances for boys and girls.

7.9 Four balls are drawn from a bag containing 5 white and 3 black balls. If X denotes the number of white balls drawn, then write down the possible values of the random variable X.

7.10 Differentiate between discrete and continuous random variable and give an example of each.

7.11 Fill in the blanks:

i) Random numbers are _________ by some _________ process.

ii) Random numbers are obtained in such a way that each digit has _________ probability.

iii) Random variable is also called _________ variable.

iv) A random variable assuming only a finite number of values is called_________ random variable.

v) A random variable assuming all possible values in a _________ is called _________ random variable.

vi) The sum of probabilities of events of a sample space is always _______________.

vii) A probability function is _________ function.

viii) The mean of probability distribution is called _________.

ix) $E(X) = \sum Xf(X)$ if it _________ absolutely.

x) A random variable may be _________ or _________.

8 Probability Distributions

$0 \leq P(A) \leq 1$

8.1 Introduction

Whenever we talk about random experiments, there is the need to associate a numerical value with each of their outcomes, in order to study them. As a result, two types of variables arise.

i. Discrete random variable. ii. Continuous random variable.

A discrete random variable almost always arises in connection with counting and a continuous random variable is one whose values are typically obtained by measurements.

In case of a discrete random variable, its probability distribution describes how much of the probability is placed on each of its possible values with the total of all these probabilities equal to 1. The probability distribution of a discrete random variable is usually called its probability mass function.

In case of a continuous random variable, we cannot talk about the probability on a point instead we talk about the probability on any interval of the values the random variable takes, with the total of the probabilities equal to 1. The probability distribution of a continuous random variable is called its probability density function.

The probability distribution of a discrete random variable is usually written with the help of a function, called its formula or it can be described with the help of a two column table (like frequency distribution) where one column gives the values (intervals of values in case of continuous random variables) and the other column gives the probabilities.

8.2 Probability Mass Function

As the value of a discrete random variable is determined by the outcome of a random experiment, one can associate with each possible value of the discrete random variable a probability that a random variable will take on that value. The probability mass function of a discrete random variable *Y* describes the values of *Y* and the probability associated with each value of *Y*. Usually, it is written in a two column table where one column gives the values of the random variable *Y* and the other column gives the probabilities associated with each value.

Example 8.1: A coin is tossed three times and let the random variable *Y* denote the number of upturned heads. Find the probability mass function?

Solution: There are 8 possible outcomes. The possible outcomes and the values assigned to them (according to the number of upturned heads) are given in the adjoining table:

Table A

Outcomes	Value of *Y*
H H T	2
H T H	2
H T T	1
T H H	2
T H T	1
T T H	1
T T T	0
H H H	3

It is clear that *Y* takes the values 0,1,2,3 because in three tosses of a coin (or three coins are tossed at one times) there are 8 possible columns and their detail is

No head (all tails) (T T T)
One head (H T T, T H T, T T H)
Two heads (H H T, H T H, T H H)
Three heads (H H H)

No head (all tails) can occur only once so, the probability of no head is 1/8 according to the definition of probability.

One head can occur 3 times so, the probability is 3/8. Two heads can occur 3 times so the probability is 3/8. Three heads can occur once so the probability is 1/8.

We can write P(*Y*=0), P(*Y*=1), P(*Y*=2), P(*Y*=3) read as probability of Y equals to no head, one head, two heads and three heads respectively. So, the probability function is given in the adjoining table:

Y	P(*Y*=*y*)
0	1/8
1	3/8
2	3/8
3	1/8

8.3 Probability Density Function

The probability density function of a continuous random variable Y is specified by a smooth curve such that the total area under the curve is unity. The probability that Y falls in any particular interval is the area under the curve against the interval.

Consider the weight (in kilograms) of a student in a class of 30 students taking Statistics. The weights measured to the nearest hundred of a kilogram are 60.50, 60.80, 55.40, 53.70, 50.75,….., 45.00 and 49.80. The minimum weight is 45.00 and maximum is 60.80. In this situation the probability histogram approaches a smooth curve. The area under the curve is unity and it cannot go below the horizontal scale. The probability that the weight is between 55 and 57 kg is the area under the curve and above this interval.

Let a and b be two numbers and Y is the random variable. Define the following events:

i) $a < Y < b$ is the event that the value of Y is between a and b.

ii) $Y < a$ is the event that the value of Y is less than a.

iii) $Y > b$ is the event that the value of Y is greater than b.

8.4 Simple Univariate Discrete And Continuous Distributions

The probability distributions of the discrete random variables are represented in a tabular form by the values of the random variable and the corresponding probabilities. For example, when a die is thrown then each upturned face (1,2,3,4,5 and 6) has the same probability of 1/6 of its occurrence. Thus the probability distribution in the tabular form is given in the adjoining table.

Y	$P(y)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

This probability distribution can be expressed in the form of the following formula such that the probabilities $P(Y=y)$ can be expressed by the function $f(y)$

$f(y) = 1/6$ for $y=1,2,3,\ldots,6$

$= 0$ otherwise

This is probability distribution for the number of upturned points when a die is thrown. This is known as discrete uniform distribution. It should be noted that every function defined for the values of a random variable cannot serve as a probability distribution unless it satisfies the condition given under 8.4.1.

The simplest form of the continuous distributions is the continuous uniform distribution prepared by.

$$f(y) = 1/b\,(b-a) \qquad a< y < b \tag{8.1}$$

Note that y takes the values between a and b.

If $a = 0$ and $b = 1$ then

$f(y) = 1 \qquad 0 < y < 1$

$= 0$ otherwise.

This function is shown in figure 8.1.

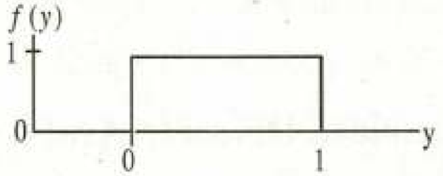


figure 8.1: - Continuous uniform distribution

The area under the probability density function is also 1. So, the width of rectangle is unit interval with height 1. To calculate the probability, area under the curve can be calculated which would be the required probability. Very often these distributions are based on the empirical evidence or prior knowledge.

It should be clear that in case of discrete distributions, the probability of an event is obtained by inserting the value of the random variable in the probability function but in case of continuous distributions, the probability is obtained by calculating the area under the curve and above the interval on which the event is defined.

It is to be noted that *P* probability of *Y* greater than *P* or equal to a and less than or equal to *b* written as $P(a \leq Y \leq b)$ is not equal to $P(a < Y < b)$ in case of discrete distributions because the probabilities at *a* and *b* are not included. In case of continuous distributions these probabilities are equal because area at a point is always zero so the probability at *a* and at *b* is always zero i.e.,

zero. So, the probability at *a* and at *b* is always zero i.e.,

8.4.1 Properties of Probability Mass Function And Probability Density Function

Probability Mass Function

Let *Y* be a discrete random variable and *P(y)* be its probability mass function. The *P(y)* must satisfy the following two conditions

i) $0 \leq P(y) \leq 1$ for each possible value of *Y*.

i.e., the probability is a number between 0 and 1.

ii) $\Sigma P(y) = 1$

i.e., the summation over probabilities for all possible values *y* of the random variable *Y* should add up to 1.

Example 8.2: A committee of size 3 is to be selected at random from 3 women and 5 men. Obtain a probability distribution for the number of women selected for the committee.

Solution:

No. of Women	No. of Men	Total
3	5	8

Number of women = 3 Number of Men = 5 so total = 8

Number of selected persons =3, So total number of sample points = $\binom{8}{3}$ = 56

Let X be the number of women in the committee, these can be 0,1,2,3 and their respective probabilities are

$$\text{P (no woman)} = \frac{\binom{3}{0}\binom{5}{3}}{\binom{8}{3}} = \frac{10}{56}, \quad P\text{ (one woman)} = \frac{\binom{3}{1}\binom{5}{2}}{\binom{8}{3}} = \frac{30}{56}$$

$$\text{P (two women)} = \frac{\binom{3}{2}\binom{5}{1}}{\binom{8}{3}} = \frac{15}{56}, \quad P\text{ (three women)} = \frac{\binom{3}{3}\binom{5}{0}}{\binom{8}{3}} = \frac{1}{56}$$

The probability distribution is given by:

x	P(x)
0	$\frac{10}{56}$
1	$\frac{30}{56}$
2	$\frac{15}{56}$
3	$\frac{1}{56}$

Example 8.3: From an urn containing 4 red and 6 white round marbles. A man draws three marbles at random without replacement. If X is a random variable which denotes the number of red marbles drawn, what is the probability distribution of X?

Solution:

Number of red marbles	Number of white marbles	total marbles = 10
= 4	= 6	

$$\text{Number of sample points} = \binom{10}{3} = 120$$

If the random variable X denotes the number of red marbles, the possible values of X are 0, 1, 2, 3 with their respective probabilities as:

$$P(X=0)=\frac{\binom{4}{0}\binom{6}{3}}{\binom{10}{3}}=\frac{5}{30},\quad P(X=1)=\frac{\binom{4}{1}\binom{6}{2}}{\binom{10}{3}}=\frac{15}{30}$$

$$P(X=2)=\frac{\binom{4}{2}\binom{6}{1}}{\binom{10}{3}}=\frac{9}{30},\quad P(x=3)=\frac{\binom{4}{3}\binom{6}{0}}{\binom{10}{3}}=\frac{1}{30}$$

The probability distribution of red marbles in a tabular form is:

X	0	1	2	3
$P(x)$	5/30	15/30	9/30	1/30

Example 8.4: Given the discrete probability mass function: -

$$P(x)=\binom{4}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{4-x}\qquad \text{for } x=0,1,2,3,4$$

Find probability distribution.

Solution:

$$P(x)\quad=\binom{4}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{4-x}\qquad \text{For } x=0,1,2,3,4$$

$$P(X=0)=\binom{4}{0}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{4}=\frac{1}{16},\quad P(X=1)=\binom{4}{1}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{3}=\frac{4}{16}$$

$$P(X=2)=\binom{4}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{2}=\frac{6}{16},\quad P(X=3)=\binom{4}{3}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)=\frac{4}{16}$$

$$P(X=4)=\binom{4}{4}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{0}=\frac{1}{16}$$

The probability distribution of X in tabular form is:

X	0	1	2	3	4
$P(x)$	1/16	4/16	6/16	4/16	1/16

Example 8.5: A random variable X has the following probability distribution:

X	–2	–1	0	1	2	3
$P(x)$	0.1	k	0.2	$2k$	0.3	$3k$

Find:

i) k ii) $P(X < 2)$ iii) $P(X = 2)$ iv) $P(-2<X<2)$ v) $P(X < 1)$.

Solution: Since $\sum_{x=-2}^{3} P(x) = 1$ gives

$$0.1 + k + 0.2 + 2k + 0.3 + 3k = 1 \text{ or } 0.6 + 6k = 1 \text{ or } 6k = 1 - 0.6 = 0.4$$

so that $k = \frac{0.4}{6} = \frac{4}{60} = \frac{1}{15}$

$P(x < 2) \quad = P(x= - 2\,) + P(x = -1) + P(x = 0) + P(x = 1)$

$$= 0.1 + k + 0.2 + 2k = 0.3 + 3k = 0.3 + 3\left(\frac{1}{15}\right) = 0.3 + 0.2 = 0.5$$

$P(x \geq 2) \quad = P(x = 2) + P(x = 3) = 0.3 + 3k = 0.3 + 3\left(\frac{1}{15}\right) = 0.3 + 0.2 = 0.5$

$P(-2 < x < 2) = P(x = -1) + P(x = 0) + P(x = 1) = k + 0.2 + 2k = 0.2 + 3k$

$$= 0.2 + 3\left(\frac{1}{15}\right) = 0.2 + 0.2 = 0.4$$

$P(x \leq 1) \quad = P(x = -2) + P(x = -1) + P(x = 0) + P(x = 1)$

$$= 0.1 + k + 0.2 + 2k = 0.3 + 3k = 0.3 + 3\left(\frac{1}{15}\right) = 0.8 + 0.2 = 0.5$$

Example 8.6: Check whether the function given by

$$f(y) = \frac{(y+1)}{14} \qquad \text{for } y = 1, 2, 3, 4$$

$$= 0 \quad \text{otherwise.}$$

is a probability function?

Solution: Here,

$$f(1) = 2/14,\ f(2) = 3/14,\ f(3) = 4/14 \text{ and } f(4) = 5/14$$

Since the values are all non negative and add up to 1 as

2/14+3/14+4/14/+5/14=1

So, both conditions are satisfied, concluding thereby that the function is a probability function.

Probability Density Function:

Let Y be a continuous random variable and $f(y)$ be its probability density function. Then $f(y)$ must satisfy the following conditions.

i) $f(y) \geq 0$ for all y ii) $P(-\infty < Y < \infty) = 1$

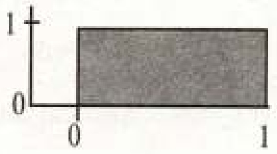
It means that the total area under the curve should be 1.

Of course, not every function defined for the values of a random variable can serve as a probability distribution unless it satisfies the above two conditions.

Example 8.7: Verify whether the function

$$f(y) = 1, \qquad 0 < y < 1$$

$$= 0 \text{ otherwise}$$



is a density function?

Solution: It is density function of a continuous random variable because y takes all values between 0 and 1. To calculate area we have width of rectangle as 1 and height of rectangle as 1, so

$$\text{Area} = (\text{width})\ (\text{height})$$

$$= (1)\ (1) = 1$$

The function is also positive, thus both conditions are satisfied and we conclude that the function is a density function.

8.4.2 Applications

Once the probability distribution for a random variable has been defined, very often, it becomes easier to calculate the probabilities. In case of discrete random variables, probabilities of the events are obtained by adding the corresponding probabilities but in case of continuous random variables the probability that a random variable falls in a certain interval is computed by calculating the area above that interval.

Example 8.8: The following table gives the probability distribution for the number of courses enrolled during spring semester 1995 by 50 M.Sc. Statistics students.

Y	1	2	3	4	5	6	7
$P(y)$	.02	.03	.16	.40	.25	.16	.05

i) Find the probability that

a) A student enrolled 3 courses?

b) A student enrolled less than 3 courses?

c) A student enrolled atleast 4 courses?

d) A student enrolled at the most 4 courses?

e) A student enrolled between 4 and 6 courses inclusive.

i.e., $P(4 \leq Y \leq 6)$

f) Student enrolled between 4 and 6 courses? (exclusive)

Solution:

a) The probability that a student enrolled three courses is
$P(Y=3) = 0.16$

b) The probability that a student enrolled less than 3 courses means by probability that he enrolled 1 course or 2 courses i.e.,

$P(Y = 1) + P(Y = 2) = 0.02 + 0.03$

$= 0.05$

c) The probability that a student enrolled atleast four courses means the probability of 4 or more courses i.e.,

$P(Y = 4) + P(Y = 5) + P(Y = 6) + P(Y = 7)$

$= 0.40 + 0.25 + 0.16 + 0.05$

$= 0.86$

d) $P(Y \leq 4) = P(Y = 1) + P(Y = 2) + P(Y = 3) + P(Y = 4)$

$=.02+0.3+0.16+.40=0.61$

e) $P(4\leq y\leq 6) = P(Y = 4) + P(Y = 5) + P(Y = 6)$

$= 0.40 + 0.25 + 0.16 = 0.81$

f) Between 4 and 6 there is only one number i.e., 5, so

$P(Y = 5) = 0.25$

Example 8.9: The amount of time (minutes) taken by a doctor to attend a patient is between 5 to 10 minutes. If we assume that the distribution followed is uniform, then calculate Probability that doctor

i) Takes between 6 and 8 minutes

ii) Less than 8 minutes.

Solution: Let Y denote the random variable. The amount of time taken by a doctor to attend a patient.

Here the value of $a = 10$, $b = 5$, so probability distribution is

$f(y) = 1/(b-a) \qquad a< y< b$

$= 1/5 = 0.2 \qquad 5<y<10$

i) The $P(6<y < 8)$ is the area between 6 and 8 and is shown in figure 8.2. Width is 8-6 = 2 and height is 0.2, so

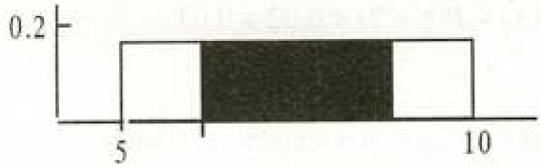


Fig. 8.2 The area between 6 and 8.

$P(6<Y<8)$ = (width of rectangle).(height)

$(2).(0.2) = 0.4$

ii) $P(Y<8) = P(5<Y<8)$

= (width of rectangle). (height)

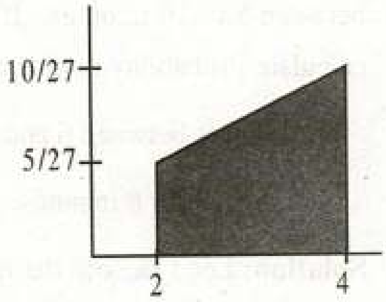
$= (3).(0.2) = 0.6$

Example 8.10: A continuous random variable X that can assume values between 2 and 5 has a density function given by: $f(x) = \frac{2(1+x)}{27}$

Find

i) $P(x<4)$ ii) $P(3 \leq x \leq 4)$.

Solution: We are given that:



i) $P(X<4)$?

$$f(x) = \frac{2(1+x)}{27}, \quad 2 \leq x \leq 5$$

$$f(2) = \frac{2\,(1+2)}{27} = \frac{6}{27}$$

$$f(4) = \frac{2(1+4)}{27} = \frac{10}{27}$$

Base $= 4-2 \quad = 2$

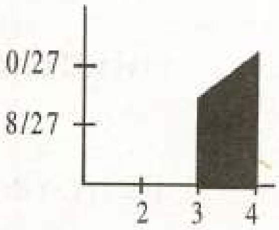
$$P(X<4) = \frac{\text{(Sum of parallel sides)}}{2} \times \text{Base}$$

$$= \frac{f(2) + f(4)}{2} \times \text{Base}$$

$$= \frac{\left(\frac{6}{27} + \frac{10}{27}\right)}{2} \times 2 = \frac{16}{27}$$

(ii) $P(3 \le X \le 4)$?

$$f(3) = \frac{2(1+3)}{27} = \frac{8}{27}$$



$$f(4) = \frac{2(1+4)}{27} = \frac{10}{27}$$

Base = 4 − 3 =.1

$$P(3 \le x \le 4) = \frac{f(3) + f(4)}{2} \times \text{Base}$$

$$= \frac{\left(\frac{8}{27} + \frac{10}{27}\right)}{2} \times 1 = \frac{18}{54} = \frac{1}{3}$$

Example 8.11:

i) A continuous random variable X has a density function $f(x) = 2x$ when $0 \le x \le 1$ and zero otherwise. Find

(a) $P\left(X < \frac{1}{2}\right)$, (ii) $P\left(\frac{1}{4} < X < \frac{1}{2}\right)$.

ii) If $f(x)$ has probability density kx^2, $0 < X < 1$, determine k and find the probability that $\frac{1}{3} < X < \frac{1}{2}$

Solution:

i) $f(x) = 2x, 0 \le X \le 1,$

$= 0,$ otherwise.

a) $$P\left(X<\frac{1}{2}\right)=\int_{0}^{1/2} f(x)\,dx=\int_{0}^{1/2} 2x\,dx=2\left[\frac{x^2}{2}\right]_0^{1/2}=\left[\left(\frac{1}{2}\right)^2-0\right]=\frac{1}{4}$$

b) $$P\left(\frac{1}{4}<x<\frac{1}{2}\right)=\int_{1/4}^{1/2} f(x)\,dx=\int_{1/4}^{1/2} 2x\,dx=2\left[\frac{x^2}{2}\right]_{1/4}^{1/2}$$

$$=\left[\left(\frac{1}{2}\right)^2-\left(\frac{1}{4}\right)^2\right]=\left[\frac{1}{4}-\frac{1}{16}\right]=\frac{3}{16}$$

ii) $f(x)$ will be a probability density function, if $\int_{1/4}^{1/2} f(x)\,dx=1$, i.e.,

$$1=\int_{0}^{1} f(x)\,dx=\int_{0}^{1} kX^2\,dx=k\left[\frac{x^3}{3}\right]_0^1=k\left[\frac{1}{3}-0\right]=\frac{k}{3}$$

So $1=k/3 \Rightarrow k=3$

Hence the probability density function $f(x)=3x^2,\ 0<X<1$

Now

$$P\left(\frac{1}{3}<x<\frac{1}{2}\right)=\int_{1/3}^{1/2} f(x)\,dx=\int_{1/3}^{1/2} 3x^2\,dx=3\left[\frac{x^3}{3}\right]_{1/3}^{1/2}$$

$$=\left[\left(\frac{1}{2}\right)^3-\left(\frac{1}{3}\right)^3\right]=\left[\frac{1}{8}-\frac{1}{27}\right]=\frac{19}{216}$$

Example 8.12: A continuous random variable x has probability density function. $f(x)=cx$ for $0<X<2$. Determine

i) c, ii) $P(1<X<1.5\ )$ iii) P (X < 1.5.)

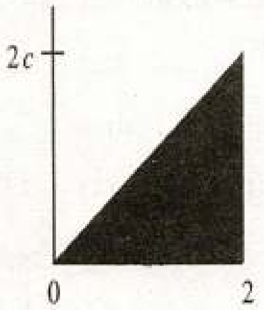
Solution: $f(x)=cx,\ 0<X<2$

i) We know that Area $=\dfrac{f(a)+f(b)}{2}\times$ Base $=1$

$f(0)=0, \quad f(2)=2c$

Base $=2-0=2$

$$\frac{\text{base}\times\text{height}}{2}$$



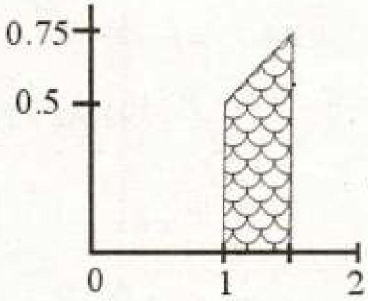
$$\frac{f(0)+f(2)}{2}\times\text{Base}=1$$

$$\Rightarrow\ \frac{0+2(c)}{2}\times 2=1$$

$$\Rightarrow\ 2c=1 \quad \text{or} \quad c=\frac{1}{2}$$

$$f(x)=\frac{1}{2}x,\ 0<X<2$$

ii) $P(1<X<1.5)=?$



$$f(1)=\frac{1}{2}=0.5,\quad f(1.5)=\frac{1.5}{2}=0.75$$

$$\text{Base}=1.5-1=0.5$$

$$P(1<X<1.5)=\frac{f(1)+f(1.5)}{2}\times\text{Base}$$

$$=\frac{0.5+0.75}{2}\times 0.5=0.3125$$

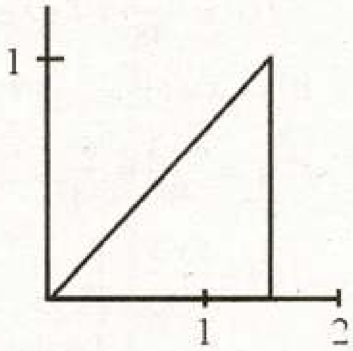
iii) $P(X<1.5)=\ ?$

$$f(0)=0,\ f(1.5)=\frac{1.5}{2}=0.75$$

$$\text{Base}=1.5-0=1.5$$

$$P(X<1.5)=\frac{f(0)+f(1.5)}{2}\times\text{Base}$$

$$=\frac{0+0.75}{2}\times 1.5=0.5625$$



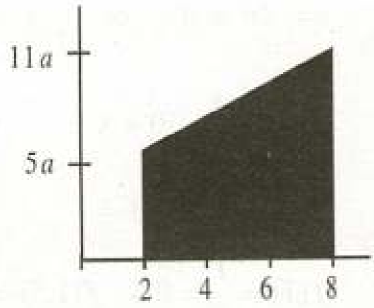
Example 8.13: A continuous random variable X, which can assume values between 2 and 8 inclusive has a density function given by $a\ (x + 3)$ where a is a constant then find:

i) a ii) $P\ (3 < X < 5)$

Solution: $f(x) = a\ (x + 3)$ $2 < X \leq 8$

i) We know that

$$\text{Area} = \frac{f(a)\ + f(b)}{2} \times \text{Base} = 1$$



$f(a) = f(2) = 5a$

$f(b) = f(8) = 11a$

$\text{Base} = 8 - 2 = 6$

$$\frac{f(2) + f(8)}{2} \times \text{Base} = 1$$

$$\frac{5a + 11a}{2} \times 6 = 1$$

or $48\ a = 1$

or $a = \dfrac{1}{48}$

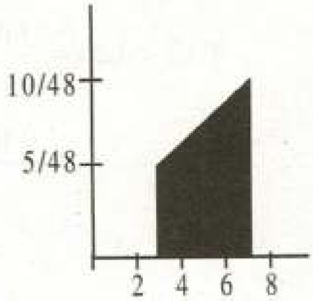
$$\therefore f(x) = \frac{x+3}{48},\ 2 < X \leq 8$$

(ii) $P\ (3 < X < 5) =$?

$$f(3) = \frac{3+3}{48} = \frac{6}{48}$$

$$f(5) = \frac{5+3}{48} = \frac{8}{48}$$

$$P(3 < X < 5) = \frac{[f(3) + f(5)]}{2} \times (5 - 3)$$



$$= \frac{\left(\frac{6}{48}+\frac{8}{48}\right)}{2}\times 2 = \frac{14}{48}$$

8.5 Drawing of Probability Mass Function and Probability Density Function

The Probability Function: The probability functions can be presented graphically in the following two ways:

i) Probability Histogram

To draw probability histogram, we take values of the random variable along x–axis and probabilities along y-axis. Adjacent rectangles are drawn against each value such that the height of each rectangle is equal to the probability at that point and the width of each rectangle is one taking 0.5 units to the left of value and 0.5 units to the right of value. Since, the width is one unit so the area of the rectagles is equals to their probabilities. The advantage of drawing probability histogram is that the discrete probability distribution can be approximated by a continuous curve.

Example 8.14: Consider the following probability distribution and draw a probability histogram.

Y	$P\,(Y=y)$
0	0.08
1	0.26
2	0.34
3	0.23
4	0.09

To construct a probability histogram, the first step is to mark values of the random variable i.e., 0,1,2,3 and 4 on x-axis as mid points. The second step is to draw adjacent rectangles of width 1 on each point going half way to each side from the mid points and with heights such that the heights represent the corresponding probabilities of 0.08, 0.26, 0.34, 0.23 and 0.09 respectively. It is shown in figure 8.3

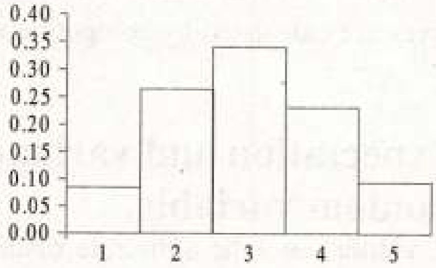


Figure 8.3: Probability Histogram

ii: Bar Chart

A bar chart is drawn with the values of the random variable along *x*-axis and probabilities along *y*-axis. The height of each bar equals the probability of the corresponding value.

Data of the above example is taken to explain the procedure. The values of the random variable are 0, 1, 2, 3 and 4. As a first step these are taken along *x*-axis. The second step is to draw bars against these points with a height equal to the corresponding probability. The probability corresponding to 0 is 0.08 so a bar is drawn on 0 with height 0.08: a bar of height 0.26 is drawn on point 1; a bar of height 0.34 is drawn on point 2; a bar of height 0.23 is drawn on point 3 and a bar of height 0.09 is drawn on point 4. It is shown in figure 8.4

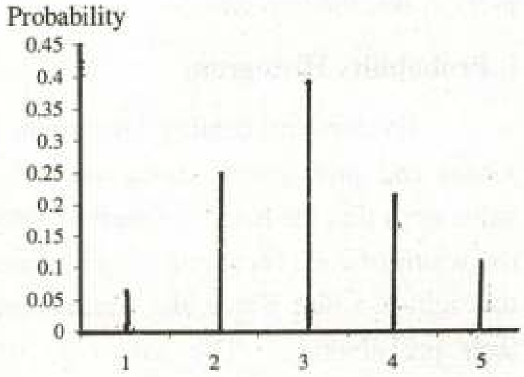


Fig 8.4: Bar Chart

Probability Density Function:

If values of the random variable are very close to each other then the probability histogram of the discrete random variable can be approximated by a smooth curve. So, the probability corresponding to each interval on *x*-axis will be the area under the curve. This is the situation in case of continuous random variable. So, the probability density functions are presented by smooth curves and the probabilities of the curves are calculated by computing area under the curves corresponding to the events.

8.6 Expectation and variance of the simple discrete random variable.

Expected value: Let *Y* be a discrete random variable with probability function *P*(*y*). The mathematical expectation or expectation of the discrete random variable *Y*. denoted by *E*(*Y*) is defined by:

$$E(Y) = \Sigma\, y\, P(y), \text{ sum is over all values of } Y \qquad (8.2)$$

This expectation of a random variable Y is the mean of its probability distribution i.e., $E(Y)$ is an alternative notation for the population mean μ. Similarly, $E(Y^2) = \Sigma\, y^2 P(y)$, sum over all y. (8.3)

Variance:

The variance of a random variable Y is the variance of the probability distribution and is

$$\sigma^2 = E(Y - \mu)^2 = \Sigma\,(Y - \mu)^2\, P(y) \text{ for all } Y \qquad (8.4)$$

Where σ^2 denotes variance.

The variance of a random variable Y is the variance of its probability distribution. It should be noted that the variance is expected value of $(Y - \mu)^2$.

We know that $E(y) = \Sigma Y\, P(Y) = \mu$ so

$$\begin{aligned}
\sigma^2 = E(Y-\mu)^2 &= \sum (Y-\mu)^2 P(y) \\
&= \Sigma (Y^2 + \mu^2 - 2\mu Y) P(y) \\
&= \Sigma Y^2 P(y) + \mu^2 \Sigma P(y) - 2\mu \Sigma Y P(y) \\
&= \Sigma Y^2 P(y) + \mu^2 - 2\mu^2 \text{ (as } \Sigma P(y) = 1 \text{ and } \Sigma YP(y) = \mu) \\
&= \Sigma (Y^2) P(y) - \mu^2 \\
&= \Sigma Y^2 P(y) - ((\Sigma YP(y))^2 \\
&= E(Y^2) - [E(Y)]^2
\end{aligned}$$

8.6.1 Properties of Expectation:

i) If c is a constant, then

$$E\,(c) = c$$

ii) If a and b are constants and Y is a random variable, then

$$E\,(b\,Y \pm a) = b\,E(Y) + a$$

If $a = -\mu$ and $b = 1$, then $E(Y - \mu) = 0$

iii) If X and Y are two random variables, then the expected value of their sum is the sum of their expected values i.e.,

$$E(X+Y) = E(X) + E(Y)$$

For the difference of two variables, the following result holds, true

$$E(X - Y) = E(X) - E(Y)$$

iv) If X and Y are two independent random variables, then the expected value of their product is the product of their expected values i.e.,

$$E(XY) = E(X)\,E(Y)$$

Example 8.15: The staff of the Department of Mathematics and Statistics at university of agriculture Faisalabad reckon that the number of microcomputers getting out of order in a year is well approximated by the following probability distribution.

No. out of order: y	0	1	2	3	4
Prob. P(y):	0.3	0.3	0.2	0.1	0.1

Find the average and variance of the number of computers getting out of order? Also find the $E(5 - 3Y)$, by using the properties of expectation.

Solution: Average = $E(Y) = \Sigma y P(y)$

Thus $\mu = \Sigma Y_i\, P(y_i) = 1.4$

Variance $= \sum_{i=0}^{4} [Y_i - E(Y_i)]^2\, P(y_i)$

Y_i	$P(y)$	$yP(y)$
0	0.3	0.0
1	0.3	0.3
2	0.2	0.4
3	0.1	0.3
4	0.1	0.4
	Total:	1.4

Calculations for the variance are:

y	$P(y)$	$y-E(y)$	$[y-E(y)]^2$	$[y-E(y)]^2\,P(y)$
0	0.3	-1.4	1.96	0.588
1	0.3	-0.4	0.16	0.048
2	0.2	0.6	0.36	0.072
3	0.1	1.6	2.56	0.256
4	0.1	2.6	6.76	0.676
-------	-------	-------	**Total:**	**1.640**

thus

Variance = 1.64

Using property ii) $E(5-3Y) = 5\text{-}3\,E(Y)$ we have, $E(Y) = 1.4$, so

$$E\,(5-3Y) = 5-3\,(1.4)$$

$$= 5-4.2$$

$$= 0.8$$

Example: 8.16 For the probability distribution of X given below find that;
(i) E (X), (ii) E(X^2)

x	**0**	**1**	**2**	**3**
$P\,(x)$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

Solution:

x	$P(x)$	$xP(x)$	$x^2P(x)$
0	3/10	0	0
1	4/10	4/10	4/10
2	2/10	4/10	8/10
3	1/10	3/10	9/10
Total	10/10 = 1	11/10	21/10

$$E(X) = \Sigma\, x\, P(x) = \frac{11}{10} = 1.1$$

$$E(X^2) = \Sigma x^2 P(x) = \frac{21}{10} = 2.1$$

Example 8.17: A random variable X has the probability distribution given below

X	0	1	2	3
$P(x)$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

Find i) $E(X)$ (ii) $E(3X + 5)$ iii) $E(X^2)$

iv) Show that $E(3X + 5) = 3E(X) + 5$

x	$P(x)$	$xP(x)$	$x^2P(x)$	$3x + 5$	$(3x+5)\,P(x)$
0	3/10	0	0	5	15/10
1	4/10	4/10	4/10	8	32/10
2	2/10	4/10	8/10	11	22/10
3	1/10	3/10	9/10	14	14/10
Total	10/10 = 1	**11/10**	**21/10**	-	**83/10**

i) $E(X) = \Sigma x\, P(x) = \frac{11}{10} = 1.1$

ii) $E(3X + 5) = \Sigma\,(3x + 5)\, P(x) = \frac{83}{10} = 8.3$

iii) $E(X^2) = \Sigma\, x^2\, P(x) = \frac{21}{10} = 2.1$

iv) $E(3X + 5) = 3E(X) + 5$

$8.3 = 3(1.1) + 5$

$8.3 = 3.3 + 5$

$8.3 = 8.3$

Example 8.18: *A, B* and *C* in the order cut a pack of cards, replacing them after each cut with condition that first who cuts a heart shall win a prize of Rs. 37. Find their respective expectation.

Solution: Let *P* be the probability of getting heart $= \frac{1}{4}$

And *q* be the probability of not getting heart $= \frac{3}{4}$

A can cut a heart on 1st, 4th, 7th... drawing with respective probabilities.

$p, q^3 p, q^6 p, \ldots..$

B can cut a spade on 2nd, 5th, 8th, ... drawing with respective probabilities.

$qp, q^4 p, q^7 p, \ldots.$

C can cut a spade on 3rd, 6th, 9th .... drawing with respective probabilities.

$q^2 p, q^5 p, q^8 p \ldots.$

Then the probability that *A* cuts the heart is

$$P(A) = P + q^3 p + q^6 p + \ldots..$$

$$P(A) = \frac{a}{1-r} = \frac{p}{1-q^3} = \frac{\frac{1}{4}}{1-\left(\frac{3}{4}\right)^3} = \frac{16}{37}$$. Expected amount for $A = 37 \times \frac{16}{37} = \text{Rs.}16$

$$P(B) = qp + q^4 p + q^7 p + \ldots.$$

$$P(B) = \frac{a}{1-r} = \frac{qp}{1-q^3} = \frac{\left(\frac{3}{4}\right)\left(\frac{1}{4}\right)}{1-\left(\frac{3}{4}\right)^3} = \frac{12}{37}$$, Expected amount for $B = 37 \times \frac{12}{37} = \text{Rs.}12$

$$P(C) = q^2 p + q^5 p + q^8 p + \ldots.$$

$$P(C) = \frac{a}{1-r} = \frac{q^2 p}{1-q^3} = \frac{\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)}{1-\left(\frac{3}{4}\right)^3} = \frac{9}{37}$$. Expected amount for $C = 37 \times \frac{9}{37} = \text{Rs.}9$

Example 8.19: A bag contains 6 Red and 4 white balls. A person draws 2 balls at random without replacement being promised 15 rupees for each red ball and 20 rupees for each white ball he draws. Find his expectation.

Solution: Red balls = 6, White balls = 4, Total = 10, Drawn balls = 2

Rupees for each red ball = 15

Rupees for each white ball = 20

The respective probabilities of the drawing are:

$$P\text{ (2 red balls)} = \frac{\binom{6}{2}\binom{4}{0}}{\binom{10}{2}} = \frac{15}{45}$$

$$P\text{ (one red and one white ball)} = \frac{\binom{6}{1}\binom{4}{1}}{\binom{10}{2}} = \frac{24}{10}$$

$$P\text{ (2 white balls)} = \frac{\binom{4}{2}\binom{6}{0}}{\binom{10}{2}} = \frac{6}{45}$$

Hence the required expectations is

$$E(X) = 30\left(\frac{15}{45}\right) + 35\left(\frac{24}{45}\right) + 40\left(\frac{6}{45}\right)$$

$$= 10 + 18.67 + 5.33$$

$$= 34$$

Example 8.20: If $f(x) = \dfrac{6 - |7 - x|}{36}$ for $X = 2, 3, 4, 5, \ldots, 12$ then find the mean and variance of the random variable x.

Solution:

x	$f(x)$	$xf(x)$	$x^2 f(x)$
2	1/36	2/36	4/36
3	2/36	6/36	18/36
4	3/36	12/36	48/36
5	4/36	20/36	100/36
6	5/36	30/36	190/36
7	6/36	42/36	294/36
8	5/36	40/36	320/36
9	4/36	36/36	324/36
10	3/36	30/36	300/35
11	2/36	22/36	242/36
12	1/36	12/36	144/36
Total	**36/36 = 1**	**252/36**	**1974/36**

$$\text{Mean} = E(X) = \Sigma x f(x) = \frac{252}{36} = 7$$

$$\text{Variance }(X) = \Sigma x^2 f(x) - [\Sigma x f(x)]^2$$

$$= \frac{1974}{36} - \left(\frac{252}{36}\right)^2 = 54.83 - 49 = 5.83$$

Example 8.21: Find the missing value such that the given distribution is a probability distribution of X.

X	2	3	4	5	6
$f(x)$	0.01	0.25	0.4	A	0.4

If $Y = 2X - 8$, then show that

i) $E(Y) = 2E(X) - 8$

ii) $\text{Var}(Y) = 4\,\text{Var}(X)$

Solution: We know that sum of probabilities is one

$$\Sigma f(X) = 1$$

$$0.01 + 0.25 + 0.4 + A + 0.04 = 1$$

$$A + 0.7 = 1$$

$$A = 1 - 0.7$$

$$A = 0.3$$

x	$f(x)$	$x\,f(x)$	$x^2 f(x)$	$y = 2x - 8$	$f(y)$	$y\,f(y)$	$y^2 f(y)$
2	0.01	0.02	0.04	–4	0.01	0.04	0.16
3	0.25	0.75	2.25	–2	0.25	–0.50	1.00
4	0.40	1.6	6.40	0	0.40	0	0
5	0.30	1.5	7.50	2	0.30	0.60	1.20
6	0.04	0.24	1.44	4	0.04	0.16	0.64
Total	**1.00**	**4.11**	**17.63**	**----**	**1.00**	**0.22**	**3**

$$E(X) = \Sigma x f(x) = 4.11$$

$$E(X^2) = \Sigma\, x^2 f(x) = 17.63$$

$$\text{Var}\,(X) = \Sigma\, x^2 f(x) - [\Sigma x\, f(x)]^2 = 17.63 - (4.11)^2 = 0.7379$$

$$E(Y) = \Sigma y\, f(y) = 0.22$$

$$E\,(Y^2) = \Sigma Y^2 f(y) = 3$$

$$\text{Var}\,(Y) = \Sigma y^2\, f(y) - (\Sigma y f(y))^2 = 3 - (0.22)^2 = 2.9516$$

i) $\therefore$ $E\,(Y) = 2E(X) - 8 = 2\,(4.11) - 8 = 8.22 - 8.00$

$0.22 = 0.22$

ii) $\therefore$ $\text{Var}(Y) = 4\ \text{Var}\,(X) = 4\,(0.7379)$

$2.9516 = 2.9516$

Example 8.22

i) Given a random variable X with $E(X) = 0.63$ and $\text{Var}\,(X) = 0.2331$.

Find $E(X^2)$.

ii) Given that $E(X^2) = 400$ and S.D. $(X) = 12$. Find $E(X)$

iii) Given the information that $E(X)=200$, C.V. $(X) = 7\%$. Find Var (X).

Solution:

a) We have, $\text{Var}(X) = E(X^2) - [E(X)]^2$

$$E(X^2) = \text{Var}(X) + [E(X)]^2 = 0.2331 + (0.63)^2 = 0.63$$

b) We have, $\text{S.D.}(X) = \sqrt{E(X^2) - [E(X)]^2}$

$$12 = \sqrt{400 - [E(X)]^2}$$

Squaring on both sides, we get

$$144 = 400 - [E(X)]^2$$

or $[E(X)]^2 = 400 - 144 = 256$

so that $E(X) = \sqrt{256} = 16$

c) We have, $\text{C.V.}(X) = \dfrac{\text{S.D.}(X)}{E(X)} \times 100$

$$7 = \frac{\text{S.D.}(X)}{200} \times 100 = \frac{\text{S.D}(X)}{2}$$

$\text{S.D.}(X) = 7(2) = 14$

so that $\text{Var}(X) = (14)^2 = 196$

8.9 Distribution Function

Very often we are interested in calculating the chances that the values of a random variable will remain at or below a certain fixed value. For example, what are chances that a student will get not more than 80% of marks? What are the chances that 5 tosses of coin will produce not more than three heads? In these situations, we are concerned with the probability that a given random variable Y will take on values that are less or equal to some fixed value of Y. Mathematically, it is written as $P(Y \leq y)$. This probability is called distribution function or cumulative distribution of the random variable Y.

The probability $P\ (Y \leq y)$ for the possible values of y is called distribution function (DF) or cumulative distribution function (cdf). It is usually denoted by $F(x)$, so we can write.

$F(Y) = P\ (Y \leq y)$ over possible values of Y (8.5)

In any density function the integral from - ∞ to y is called a distribution or cumulative distribution function is given as:

$$F(Y) = \int_{-\infty}^{Y} f(Y)\, dy$$

The function has the following properties:

i) $F(-\infty) = 0$

ii) $F(+\infty) = 1$

iii) $F(y_1) \leq F(y_2)$ if $y_1 \leq y_2$

(iv) F (y) is continuous atleast on the right of each y.

Example 8.23: A coin is tossed three times. Probability distribution for number of heads (Y) is given below:

Y	$P(y)$
0	1/8
1	3/8
2	3/8
3	1/8

Find distribution function $F(y)$?

Solution: From the definition of $F(y)$, we have

i) $F(Y) = P(Y \leq y)$ and Y takes values 0, 1, 2, 3. No value of Y is less than 0. So, $P(Y < 0) = 0$

ii) for $Y = 0$, $P(Y = 0)$ 1/8, there is no integral value between 0 and 1. So, this probability remains 1/8 till Y approaches 1.

iii) for $Y = 1$, $F(Y \leq 1) = P(Y = 0) + P(Y = 1) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8}$

iv) for $Y = 2$, $F(Y \leq 2) = P(Y = 0) + P(Y = 1) + P(Y = 2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}$

It remains 7/8 until Y reaches $Y = 3$

at $Y = 3$, $F(Y \leq 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3)$

$= 1/8 + 3/8 + 3/8 + 1/8 = 1$

So, this probability is 1 at $Y = 3$

As Y does not take any value beyond 3. So, There is no probability beyond 3. Thus $F(Y \geq 3)$ remains 1.

These results are summarized below and its graph is as in figure 8.5

$$\begin{aligned} F(y) &= 0 && \text{for } Y < 0 \\ &= 1/8, && \text{for } 0 \leq Y < 1 \\ &= 3/8, && \text{for } 1 \leq Y < 2 \\ &= 7/8, && \text{for } 2 \leq Y < 3 \\ &= 1, && \text{for } Y \geq 3 \end{aligned}$$

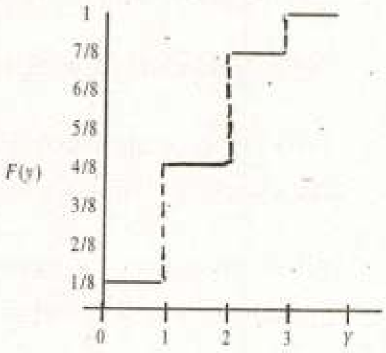


Figure 8.5: Distribution function

Exercise 8

Ans on Page 258

8.1 What are random numbers? How can they be generated? Explain the applications of random numbers.

8.2 Three balls are drawn from a bag containing 5 white and 3 black balls. If X denotes the number of white balls drawn from the bag, then find the probability distribution of X.

8.3 There are seven candidates for three positions of typist. Four of the candidates know Urdu typing while the other three do not know it. If the three candidates are selected at random, find the probability distribution of the number of persons knowing Urdu typing among those selected.

8.4 i) What is meant by probability distribution. Distinguish between discrete and continuous random variables by giving examples.

ii) A coin is tossed 4 times. If X denotes the number of tails, what is the probability distribution of X? Draw a probability histogram.

iii) A bag contains 4 red and 6 black balls. A sample of 4 balls is selected from the bag without replacement. Let X be the number of red balls, find the probability distribution of X.

8.5 i) Define probability function and give an example to explain it.

ii) Two fair dice are thrown and Y denote the product of the two scores. Obtain the probability distribution of Y

8.6 i) What properties a mathematical function should possess to be a probability function and probability density function?

ii) Check whether the following functions satisfy the conditions of a probability function?

a. $f(y) = 1/4$ for $y = 1,2,3,4,5$

b. $f(y) = 1/y$ for $y = 1,2,3,4$

c. $f(y) = y/15$ for $y = 0,1,2,3,4,5$

d. $f(y) = (5 - y^2)/6$ for $y = 0,1,2,3$

8.7 Determine the value of c so that the function can serve as a probability function of a random variable.

a) cy for $Y = 1,2,3,4,5$

b) $(1 - c)\, c^y$ for $Y = 0,1,2\ldots$

8.8 A random variable Y takes values 0, 1, 2, 3 with respective probabilities $1/4\,(1 + 3\theta)$, $1/4\,(1 - \theta)$, $1/4\,(1 + 2\theta)$, $1/4\,(1 - 4\theta)$. For what values of θ is this a valid probability function?

8.9 i) Given the following probability distribution:

x	0	1	2	3	4
$P(x)$	1/126	20/126	60/126	40/126	5/126

Verify that $E\,(2X + 3) = 2E\,(X) + 3$

ii) Let X be a random variable with probability distribution:

x	−1	0	1	2	3
$P(x)$	0.125	0.500	0.200	0.050	0.125

Find

a) $E(X)$ and $\text{Var}(X)$ b) The probability distribution of the random variable $Y = 2X + 1$.

Using the probability distribution of Y, determine $E(Y)$ and $\text{Var}\,(Y)$.

8.10 The following table gives the probability distribution of the random variable Y, the number of courses taught by a teacher during spring semester in the University of Agriculture Faisalabad.

y	1	2	3	4	5
$P(y)$	.1	.2	.3	.3	.1

Find the probability that:

i) Teacher taught 2 courses

ii) Teacher taught less than 4 courses.

iii) Teacher taught between 2 and 5 courses

iv) Teacher taught atleast 3 courses.

v) Teacher taught at the most 4 courses.

8.11 A box contains five slips of paper marked 1,2,3,4 and 5. Two slips are selected without replacement, list the possible values for each of the following random variables:

i) The sum of the two numbers on two slips.

ii) The difference between the first and second number.

8.12 Four randomly selected students from a class are asked their opinion about the teaching system as satisfactory (S) or not satisfactory (NS). Let Y denote the number of students saying satisfactory. Write down the possible outcomes and the possible Y values.

8.13 A point is randomly selected on the surface of a lake that has maximum depth of 30 feet. Let y denote the depth of the lake at the randomly selected point. What are possible values of Y ? Is Y a discrete variable or a continuous variable?

8.14 Calculate mean and variance of the following probability distribution:

y	0	1	2	3	4	5
$P(y)$	0.1	0.2	0.3	0.2	0.1	0.1

8.15 i) A continuous random variable X having values only between 0 and 4 has a density function given by: $f(x) = \dfrac{1}{2} - ax$, where a is a constant find a) a b) $P(1 < X < 2)$.

ii) A continuous random variable X has probability density function giving $f(x) = cx$ for $0 < X < 2$. Find

(a) c (b) Probability that $1 < X < 1.5$ (c) Probability that $X < 1.5$

iii) If $f(x)$ has probability density kx^2, $0 < X < 1$, determine its kind and find the probability that $\frac{1}{3} < X < \frac{1}{2}$

8.16 What is a random variable? Distinguish between discrete and continuous random variable, giving examples.

8.17 Find the probability distribution of the number of boys in families with three children, assuming equal probabilities for boys and girls.

8.18 From lot containing 12 items, 4 of which are defective, 5 are chosen at random. If X is the number of defective items found in the sample, write down

(i) The probability distribution of X (ii) $P(X \leq 1)$

iii) Verify $\sum_{x=0}^{4}[P(x)] = 1$

8.19 i) Define continuous random variable and its probability distribution.

ii) Find the constant k so that the function $f(x)$ defined as follows may be a density function.

$$f(x) = \frac{1}{k},\ a \leq X \leq b$$

$$= 0, \text{ elsewhere}$$

8.20 (i) What do you mean by expected value? What are the properties of expectation?

(ii) Given the following discrete probability distribution:

x	0	1	2	3	4	5
P(x)	6/36	10/36	8/36	6/36	4/36	2/36

Compute its mean, variance, standard deviation and coefficient of variation.

8.21 Let X be a random variable with probability distribution as follows:

x	1	2	3	4	5
$f(x)$	0.125	0.45	0.25	0.05	0.125

Find mean and variance.

8.22 i) A continuous random variable X has a density function

$$f(x) = \frac{x+1}{8} \text{ for } X = 2 \text{ to } X = 4. \text{ Find}$$

a) $P(X < 3.5)$ b) $P\,(2.4 < X < 3.5)$ c) $P(X = 1.5)$.

(ii) A continuous random variable X has a density function

$f(x) = 2x, \quad 0 \leq x \leq 1$. Find

a) $P(X = \frac{1}{2})$ b) $P(X > \frac{1}{4})$ c) $P(\frac{1}{4} < X < \frac{1}{2})$

8.23 A continuous random variable X which can assume values between $X = 2$ and $X = 8$ inclusive has a density function given by $a\,(x + 30)$, where a is a constant. Find i) a ii) $P\,(3 < X < 6)$ iii) $P\,(X \leq 6)$ iv) $P\,(X \geq 4)$.

8.24 In a summer season, a dealer of desert room coolers can earn Rs. 800 per day if the day is hot and can earn Rs. 200 per day if it is fair and loses Rs. 50 per day if it is cloudy. Find his mathematical expectation if the probability of the day being hot is 0.40 and for being cloudy it is 0.35

8.25 A committee of size 5 is to be selected from 3 female and 5 male members. Find the expected number of female members on the committee.

8.26 What are the properties of probability function and probability density function?

8.27 i) Explain the concept of distribution function.

ii) 10 vegetable cans, all of the same size, have lost their labels. It is known that 5 contain tomatoes and 5 contain corns. If 5 cans are selected at random, then find the probability distribution and the distribution function for the number of tomato cans in the sample.

8.28 i) Define the expected value for a random variable.

ii) The number of automobile accidents in a city are: 1, 2, 3, 4, 4 with corresponding probabilities 1/8, 2/8, 2/8 and 3/8. What is the expected number of daily accidents?

8.29 *A* and *B* throw a die for a prize of Rs. 11. Which is to be won by the player who first throws a 6. If *A* has the firs throw, what are their respective expectations?

8.30 A bag contains 2 white and 2 black balls. Three men *A, B* and *C* draw a ball and don't replace it. The person who draws the white ball first receives Rs. 12. What are their respective expectations?

8.31 Three balls are drawn from a bag containing 5 white and 3 black balls. If X denotes the number of white balls drawn from the bag, then find the probability distribution of X. Also find its mean and variance.

8.32 A coin is biased such that a head is thrice as likely to occur as a tail. Find the probability distribution of heads and also find the mean and variance of the distribution when it is tossed 4 times.

8.33 Approximately 10% of the glass bottles coming from a production line have serious defects. If two bottles are selected at random, find the expected number of bottles that having serious defects.

8.34 A random variable X takes the values –3, –2, 2, 3 and 4 with probabilities $P(X)$ equal to 1/5, 1/10, 1/10, 1/5 and 2/5 respectively. Compute $E(X)$ and show that $E\ (5X + 10) = 5E(X) + 10$. Also compute the variance (X) and variance $(5X + 10)$. Find the ratio of two variances.

8.35 If $f(x) = \dfrac{6-|7-x|}{36}$ for $X = 2,3,4,\ldots, 12$, then find the mean and variance of the random variable X.

8.36 For the following Probability distribution. Find

i) $E(X)$, ii) $E(X^2)$, iii) $E[X - E(X)]^2$.

x	-10	-20	30
$P(X)$	1/5	3/10	1/2

8.37 i) Define expectation of a random variable

ii) The probability distribution of a discrete random variable X is given by

$$f(x)=\binom{3}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{3-x}, x=0,1,2,3$$

Find $E(X)$ and $E(X^2)$.

8.38 Against each statement, write T for true and F for false statement.

i) A random variable is also named as a chance variable.

ii) The number of accidents occurring on G.T. road during one month is the example of continuous random variable.

iii) The Probability cannot exceed 1.

iv) The range of continuous Random variable is from '0' to 'n'.

v) The Distribution function is an increasing function.

vi) The expectation of a Random variable is also named as the Mean of a Random variable.

vii) The probability function can be negative.

viii) A discrete probability distribution is represented by area graph.

ix) If X and Y are independent random variables then $E(XY)=E(X)\,E(Y)$.

x) If X and Y are independent random variables, then

$$S.D(X-Y) = S.D(X) - S.D(Y).$$

9 Binomial and Hypergeometric Probability Distribution

9.1. Introduction

In an experiment of tossing a coin, drawing a card from a pack of playing cards repeatedly, etc., each drawing is called a **trial**. The results of each trial is classified as a success or failure. The probability of success is denoted by p and the probability of failure is denoted by q where, q=1-p or $q + p = 1$. Let an experiment be repeated n times, the number of successes obtained in a trial of the experiment is denoted by x and the number of failures by $n-x$.

A trial having two possible outcomes i.e., only success and failure is called **Bernoulli trials.** For each bernoulli trial, the probability of success remains the same and the successive trials are independent.

An experiment in which the outcomes can be classified as success or failure and in which the probability of success remains constant from trial to trial is called **Binomial experiment**. A binomial experiment possesses the following properties:

i) Each trial of the experiment results in an outcome which can be classified into two categories i.e., success and failure.

ii) The probability of success remains constant from one trial of the experiment to the next.

iii) The repeated trials are independent.

iv) The experiment is repeated a fixed number of times.

9.2 Binomial Probability Distribution

Suppose we have n independent trials for each of which the probability of success is p and the probability of failure is q and $q + p = 1$. The probability of exactly x success is given by

$$P(X=x)=\binom{n}{x}p^x q^{n-x} \qquad (9.1)$$

where $x = 0, 1, 2, 3, \ldots., n$. The random variable x is called the binomial variable and the distribution of x is called the binomial distribution, the quantities n and p are called the parameters of the binomial distribution.

The binomial probability distribution is generally denoted by $b(x, n, p)$. The probability $\binom{n}{x}q^{n-x}p^x$ are obtained by expanding the binomial expansion $(q + p)^n$

i.e.,

$$(q+p)^n=\binom{n}{0}q^{n-o}p^o+\binom{n}{1}q^{n-1}p^1+\binom{n}{2}q^{n-2}p^2+....+\binom{n}{n}q^{n-n}p^n.$$

where $\binom{n}{0}\binom{n}{1}, \binom{n}{2}, \ldots \binom{n}{n}$, are called the binomial co-efficients.

If $p=q=\frac{1}{2}$, the binomial distribution is a symmetrical distribution.

If $p\neq q$, the binomial distribution is a skewed distribution.

If $p>\frac{1}{2}$, the distribution is negatively skewed .

If $p<\frac{1}{2}$, the distribution is positively skewed .

Example 9.1: A fair coin is tossed 4 times. Find the probabilities of obtaining various number of heads.

Solution: $n = 4,\ p = \frac{1}{2},\ q = \frac{1}{2}$

$$\therefore \quad P(X=x)=\binom{n}{x}q^{n-x}\ p^x,\ x = 0, 1, 2, \ldots\ldots\ldots, n$$

where x denotes the number of heads.

$$\therefore P(X=0) = \binom{4}{0}\left(\frac{1}{2}\right)^{4-0}\left(\frac{1}{2}\right)^0=\frac{1}{16}$$

$$P(X=1) = \binom{4}{1}\left(\frac{1}{2}\right)^{4-1}\left(\frac{1}{2}\right)^{1} = \frac{4}{16}$$

$$P(X=2) = \binom{4}{1}\left(\frac{1}{2}\right)^{4-2}\left(\frac{1}{2}\right)^{2} = \frac{6}{16}$$

$$P(X=3) = \binom{4}{3}\left(\frac{1}{2}\right)^{4-3}\left(\frac{1}{2}\right)^{3} = \frac{4}{16}$$

$$P(X=4) = \binom{4}{4}\left(\frac{1}{2}\right)^{4-4}\left(\frac{1}{2}\right)^{4} = \frac{1}{16}$$

Example 9.2: A fair coin is tossed 5 times. What is the probability of getting

i) Exactly 3 heads ii) at least 3 heads iii) at the most two heads.

Solution: $n = 5,\ p = \frac{1}{2},\ q = \frac{1}{2},$

and $P(X = x) = \binom{n}{x} q^{n-x} p^{x}\quad x = 0, 1, 2, 3, 4, 5$

i) Exactly 3 heads

$$P(X=3) = \binom{5}{3}\left(\frac{1}{2}\right)^{5-3}\left(\frac{1}{2}\right)^{3} = \frac{10}{32}$$

ii) at least 3 heads

$$\begin{aligned} P(X \geq 3) &= P(X=3) + P(X=4) + P(X=5) \\ &= \binom{5}{3}\left(\frac{1}{2}\right)^{5-3}\left(\frac{1}{2}\right)^{3} + \binom{5}{4}\left(\frac{1}{2}\right)^{5-4}\left(\frac{1}{2}\right)^{4} + \binom{5}{5}\left(\frac{1}{2}\right)^{5-5}\left(\frac{1}{2}\right)^{5} \\ &= \frac{10}{32} + \frac{5}{32} + \frac{1}{32} \\ &= \frac{16}{32} = 0.5 \end{aligned}$$

iii) at the most two heads.

$$\begin{aligned} P(X \le 2) &= P(X=2)+P(X=1)+P(X=0) \\ &= \binom{5}{2}\left(\frac{1}{2}\right)^{5-2}\left(\frac{1}{2}\right)^{2}+\binom{5}{1}\left(\frac{1}{2}\right)^{5-1}\left(\frac{1}{2}\right)^{1}+\binom{5}{0}\left(\frac{1}{2}\right)^{5-0}\left(\frac{1}{2}\right)^{0} \\ &= \frac{10}{32}+\frac{5}{32}+\frac{1}{32} \\ &= \frac{16}{32}=0.5 \end{aligned}$$

Example 9.3: If you toss a fair die 6 times. What is the probability of getting no even number.

Solution: $n=6$, $p=\frac{3}{6}=\frac{1}{2}$, $q=\frac{1}{2}$

$$P(X=0) = \binom{6}{0}\left(\frac{1}{2}\right)^{6-0}\left(\frac{1}{2}\right)^{0}=\frac{1}{64}$$

9.2.1 Binomial frequency Distribution

If the binomial probability distribution is multiplied by the number of experiments N then the distribution is called the binomial frequency distribution. Expected frequency of x successes in N experiments of n trials is given by

$$f(X=x)\;\; Np(X=x)=N\binom{n}{x}q^{n-x}\,p^{x},$$

Example 9.4: Out of 800 families with 5 children each. How many would you expect to have

i) 4 boys ii) at least 3 boys iii) at the most one boy.

Solution: $n = 5, N=800, p=\frac{1}{2}$

$$q=1-p=\frac{1}{2}$$

i) With four boys

$$P(X=x) = \binom{n}{x} q^{n-x} \ p^{x}$$

$$P(X=4) = \binom{5}{4}\left(\frac{1}{2}\right)^{5-4}\left(\frac{1}{2}\right)^{4}$$

$$= \frac{5}{32}$$

Hence expected number of families with 4 boys = $800\times\frac{5}{32} = 125$

iii) with at least 3 boys

$$P(X \geq 3) = P(X=3)+P(X=4)+P(X=5)$$

$$= \binom{5}{3}\left(\frac{1}{2}\right)^{5-3}\left(\frac{1}{2}\right)^{3}+\binom{5}{4}\left(\frac{1}{2}\right)^{5-4}\left(\frac{1}{2}\right)^{4}+\binom{5}{5}\left(\frac{1}{2}\right)^{5-5}\left(\frac{1}{2}\right)^{5}$$

$$= \frac{10}{32}+\frac{5}{32}+\frac{1}{32}$$

$$= \frac{16}{32} = 0.5$$

Hence expected number of families with at least three boys

$$= 800\times\frac{16}{32} = 400$$

iii) with at the most one boy.

$$P(X \leq 1) = P(X=1)+ P(X=0)$$

$$= \binom{5}{1}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{5-1}+\binom{5}{0}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{5-0}$$

$$= \frac{5}{32}+\frac{1}{32}$$

$$= \frac{6}{32}$$

Hence expected number of families with at the most one boy will be $800\times\frac{6}{32} = 150$

Example 9.5: Five dice are tossed 96 times. Find the expected frequencies when throwing of a 4, 5 or 6 is regarded as a success.

Solution: Here,

$$n = 5,\ N = 96,\ p = \frac{3}{6} = \frac{1}{2},\ q = 1 - p = \frac{1}{2}$$

X	$P(x)=\binom{n}{x} q^{n-x}\, p^{x}$	Expected Frequencies $NP(x)=f_e$
0	$\binom{5}{0}\left(\frac{1}{2}\right)^{5-0}\left(\frac{1}{2}\right)^{0}=\frac{1}{32}$	$96\times\frac{1}{32}=3$
1	$\binom{5}{1}\left(\frac{1}{2}\right)^{5-1}\left(\frac{1}{2}\right)^{1}=\frac{5}{32}$	$96\times\frac{5}{32}=15$
2	$\binom{5}{2}\left(\frac{1}{2}\right)^{5-2}\left(\frac{1}{2}\right)^{2}=\frac{10}{32}$	$96\times\frac{10}{32}=30$
3	$\binom{5}{3}\left(\frac{1}{2}\right)^{5-3}\left(\frac{1}{2}\right)^{3}=\frac{10}{32}$	$96\times\frac{10}{32}=30$
4	$\binom{5}{4}\left(\frac{1}{2}\right)^{5-4}\left(\frac{1}{2}\right)^{4}=\frac{5}{32}$	$96\times\frac{5}{32}=15$
5	$\binom{5}{5}\left(\frac{1}{2}\right)^{5-5}\left(\frac{1}{2}\right)^{5}=\frac{1}{32}$	$96\times\frac{1}{32}=3$

9.2.2 Mean And Variance Of The Binomial Distribution

We find the mean and variance of the binomial distribution given by $(q+p)^n$. We know that

$$\text{Mean} = E(X) = \sum x P(x) \qquad (9.2)$$

$$\text{Variance} = E(X^2)-[E(X)]^2$$

$$= \sum x^2 P(x)-[\sum x\, P(x)]^2 \qquad (9.3)$$

The necessary calculations are shown in the table given below:

Number of Successes (x)	$P(x)=\binom{n}{x} q^{n-x} p^{x}$	$xP(x)$	$x^2P(x)$
0	$\binom{n}{0} q^{n-1} p^{0} = q^{n}$	0	0
1	$\binom{n}{1} p\, q^{n-1} = npq^{n-1}$	nqp^{n-1}	nqp^{n-1}
2	$\binom{n}{2} q^{n-2} p^{2} = \frac{n(n-1)}{2!} q^{n-2} p^{2}$	$n(n-1)p^{2}q^{n-2}$	$2n(n-1)p^{2}q^{n-2}$
3	$\binom{n}{3} p^{3} q^{n-3}$ $= \frac{n(n-1)(n-2)}{3!} q^{n-3} p^{3}$	$\frac{n(n-1)(n-2)}{2!} p^3q^{n-3}$	$\frac{3n(n-1)(n-2)}{2!} p^3q^{n-3}$
⋮	⋮	⋮	⋮
n	$\binom{n}{n} p^{n} q^{n-n} = p^{n}$	np^{n}	$n^{2}p^{n}$

$$E(X) = \sum x\, P(x)$$

$$= npq^{n-1} + n(n-1)p^{2}q^{n-2} + \frac{n(n-1)(n-2)}{2!} p^{3}q^{n-3} + \ldots + np^{n}$$

$$= np\left[q^{n-1} + (n-1)pq^{n-2} + \frac{(n-1)(n-2)}{2!} p^{2}q^{n-3} + \ldots + np^{n-1} \right]$$

$$= np\left[(q+p)^{n-1}\right]$$

$$= np\,(q+p)^{n-1} \qquad [\therefore\ p+q=1]$$

$$\therefore E(X) = np$$

$$E(X^{2}) = \sum x^{2}\ P(x)$$

$$= npq^{n-1} + 2n(n-1)p^{2}q^{n-2} + \frac{3n(n-1)(n-2)}{2!} p^{3}q^{n-3} + \ldots + n^{2}p^{n}$$

$$= np\left[q^{n-1} + 2(n-1)pq^{n-2} + \frac{3(n-1)(n-2)}{2!} p^2 q^{n-3} + \ldots. + p^{n-1} \right]$$

$$= np\left[\left\{ q^{n-1} + (n-1)pq^{n-2} + \frac{(n-1)(n-2)}{2!} p^2 q^{n-3} + \ldots. + p^{n-1} \right\}\right.$$

$$\left. + \left\{ (n-1)pq^{n-2} + \frac{2(n-1)(n-2)}{2!} p^2 q^{n-3} + \ldots. + (n-1)p^{n-1} \right\}\right]$$

$$E(X^2) = np\left[(q+p)^{n-1} + (n-1)p\{q^{n-2} + (n-2)pq^{n-3} + \ldots. + p^{n-2})\right]$$

$$E(X^2) = np\left[(q+p)^{n-1} + (n-1)\ p(q+p)^{n-2}\right]$$

$$= np\left[1 + (n-1)p\right]$$

$$E(X^2) = np + n^2p^2 - np^2$$

$$\text{Variance} = E(x^2) - [E(x)]^2$$

$$= np + n^2p^2 - np^2 - (np)^2$$

$$= np - np^2$$

$$= np(1-p) \quad \text{Variance} \quad = \quad npq \quad (\because q = 1-p)$$

And standard deviation $\sigma = \sqrt{npq}$

Example 9.6: In a binomial distribution $n = 20$ and $P = \frac{3}{5}$. Find the mean, variance and standard deviation of the binomial distribution.

Solution: Here $n = 20$, $p = \frac{3}{5}$, $q = 1 - p = \frac{2}{5}$

we know that

$$\text{Mean} = np$$

$$= 20\left(\frac{3}{5}\right)$$

$$= 12$$

$$\sigma^2 = \text{Variance} = npq$$

$$= 20\left(\frac{3}{5}\right)\left(\frac{2}{5}\right)$$

$$\sigma^2 = 4.8$$

$$\text{or } \sigma = S.D. = \sqrt{npq} = \sqrt{4.8}$$

Example 9.7: In a binomial distribution, mean and standard deviation were found to be 38 and 5.6 respectively find p and n.

Solution:

Mean $= np = 38$ (i)

Standard deviation $= \sqrt{npq} = 5.6$ (ii)

Squaring equation (ii)

$$npq = 31.36 \qquad \text{(iii)}$$

Dividing equation iii) by i), we get

$$\frac{npq}{np} = \frac{31.36}{38}$$

$$q = 0.83$$

$$\because p = 1-q$$

$$\therefore P = 1-0.83$$

$$P = 0.17$$

Putting the value of $p = 0.17$ in equation (i)

$$np = 38$$

$$n(0.17) = 38$$

$$n = 224$$

$$n = 224 \text{ and } p=0.17$$

Example 9.8: Is it possible to have a binomial distribution with mean = 5 and $S.D. = 4$.

Solution:

$$\text{mean} = np = 5 \qquad (i)$$

$$S.D. = \sqrt{npq} = 4 \qquad (ii)$$

$$\text{Variance} = npq = 16 \qquad (iii)$$

$$\frac{npq}{np} = \frac{16}{5}$$

$$q = 3.2$$

Since p or q cannot be greater than one. So it is not possible to have a binomial distribution with mean 5 and with standard deviation 4.

9.3 Hypergeometric Distribution and Hypergeometric Experiment

When the successive trials are without replacement then they are dependent and the probability of success changes from one trial of the experiment to the other. Such an experiment in which a random sample is chosen without replacement from a finite population is said to be hypergeometric experiment.

9.3.1 Properties

A hypergeometric experiment has the following properties:

i) The experiment is repeated a fixed number of times.

ii) The successive trials are dependent.

iii) The probability of success varies from trial to trial (it in not fixed).

iv) The outcome of an experiment can be classified as success or failures.

The random variable X representing the number of successes in a hypergeometric experiment is called a hypergeometric variable and probability distribution of the hypergeometric variable is called hypergeometric distribution.

9.3.2 Hypergeometric Probability Distribution:

Suppose that there are N total number of items out of which k are classified as successes and $(N\text{-}k)$ as failures. n items are to be selected at random without replacement $n \leq N$.

Let X denotes the number of successes and we can obtain exactly "x" successes and $(n\text{-}x)$ failures as follows:

$$P(X=x) \quad = \quad \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \qquad (9.4)$$

where $x = 0, 1, 2, \ldots\ldots, n$. The hypergeometric probability distribution has 3 parameters n, k and N.

$$\text{Mean of the hypergeometric distribution} = \frac{nk}{N} \qquad (9.5)$$

$$\text{variance of the hypergeometric distribution} = \frac{nk}{N}\left(1-\frac{k}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Example 9.9: Determine the probability distribution for the number of white beads. Among 5 beads drawn at random from a bowl containing 4 white and 7 black beads. Compute mean, variance and check the results by using the formula.

Solution:

$$N = 11, \quad n = 5$$

Taking white beads as success

$$k = 4$$

$$N-k = 11-4 = 7 \text{ (black beads)}$$

$$x = 0, 1, 2, 3, 4$$

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

$$P(X = 0) = \frac{\binom{4}{0}\binom{7}{5}}{\binom{11}{5}} = \frac{(1)(21)}{462} = \frac{21}{462}$$

$$P(X = 1) = \frac{\binom{4}{1}\binom{7}{4}}{\binom{11}{5}} = \frac{(4)(35)}{462} = \frac{140}{462}$$

$$P(X = 2) = \frac{\binom{4}{2}\binom{7}{3}}{\binom{11}{5}} = \frac{(6)(35)}{462} = \frac{210}{462}$$

$$P(X = 3) = \frac{\binom{4}{3}\binom{7}{2}}{\binom{11}{5}} = \frac{(4)(21)}{462} = \frac{84}{462}$$

$$P(X = 4) = \frac{\binom{4}{4}\binom{7}{1}}{\binom{11}{5}} = \frac{(1)(7)}{462} = \frac{7}{462}$$

Probability distribution is given below:

x	$f(x)$	$x\,f(x)$	$x^2\,f(x)$
0	21/462	0	0
1	140/462	140/462	140/462
2	210/462	420/462	840/462
3	84/462	252/462	756/462
4	7/462	28/462	112/462
Total	462/462=1	840/462	1848/462

$$\text{Mean} = E(X) = \sum x\,f(x)$$

$$= \frac{840}{462}$$

$$= 1.8182$$

$$\text{Variance} = E(X^2)-[E(X)]^2 = \sum x^2 f(x) - [\sum x\,f(x)]^2$$

$$= \frac{1848}{462} - \left(\frac{840}{462}\right)^2$$

$$= 0.6942$$

Checking The Results

$$\text{Mean} = \frac{nk}{N} = \frac{5\times 4}{11} = 1.8182$$

$$\text{Variance} = \left(\frac{nk}{N}\right)\left(\frac{N-k}{N}\right)\left(\frac{N-n}{N-1}\right)$$

$$= (1.8182)\left(\frac{7}{11}\right)\left(\frac{6}{10}\right)$$

$$\text{Variance} = 0.6942$$

Example 9.10 Ten vegetable cans, all of the same size, have lost their labels. It is known that 5 contain tomatoes and 5 contain corns. If 5 are selected at random, what is the probability that all contain tomatoes? What is the probability that 3 or more contain tomatoes?

Solution: $N = 10, \quad n = 5$

considering the tomatoes cans as success.

$$k = 7 \text{ (Tomatoes) and } N - k = 10 - 5 = 5$$

i) All contain tomatoes.

$$x = 5$$

$$n - x = 5 - 5 = 0$$

$$\because P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

$$\therefore P(X = 5) = \frac{\binom{5}{5}\binom{5}{0}}{\binom{10}{5}} = 0.005968$$

ii) 3 or more contain tomatoes

$$x = 3, 4, 5$$

$$n - x = 2, 1, 0$$

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)$$

$$= \frac{\binom{5}{3}\binom{5}{2}}{\binom{10}{5}} + \frac{\binom{5}{4}\binom{5}{1}}{\binom{10}{5}} + \frac{\binom{5}{5}\binom{5}{0}}{\binom{10}{5}}$$

$$= \frac{100}{252} + \frac{25}{252} + \frac{1}{252}$$

$$= \frac{126}{252}$$

$$\because P(X \geq 3) = 0.5$$

Exercise 9

Ans on Page 260

9.1 Define the Binomial Probability Distribution.

9.2 What is a Binomial experiment? Give its properties?

9.3 An event has the probability $p = \frac{3}{8}$. Find the complete binomial distribution for $n = 5$ trials.

9.4 Find the probability i.e., tossing a fair coin four times there will appear

i) 4 heads ii) 1 tail and 3 heads

iii) at least 2 heads iv) at the most 2 heads.

9.5 If 20% of the bolts produced by a machine are defective, determine the probability that out of 4 bolts chosen at random

i) zero ii) 2 bolts are defective.

9.6 Given that the probability of passing an examination is 0.75. What is the probability of

i) passing at least two examinations if you take six ?

ii) failing at least two examinations if you take four ?

9.7 The experience of a house agent indicates that he can provide suitable accommodation for 75% of the clients who come to him. If on a particular occasion 6 clients approach him independently, calculate the probability that

i) less than 4 clients will get satisfactory accommodation.

ii) at least 5 clients will get satisfactory accommodation.

9.8 The incidence of an occupational disease in an industry is such that the workers have 20% chance of suffering from it. What is the probability that out of 6 workmen:

i) not more than 2 will catch the disease ?

ii) 4 or more will catch the disease ?

9.9 If 8 coins are tossed what is the probability that there are

i) Exactly 5 heads ii) 1 to 7 heads.

9.10 If X = binomially distributed with $n = 10$, $P = 0.4$, then find Mean and variance of $Y = \frac{X-10}{6}$.

9.11 If X is the number of successes with probability of success as $\frac{1}{4}$ in each of 5 independent trials. Find

i) $P(X=0)$ ii) $P(X \leq 3)$.

9.12 If 5 true dice are thrown once, determine the probability of getting 0, 1, 2, 3, 4, 5 sixes. Find the mean and variance of the probability distribution so obtained.

9.13 Five dice are tossed 96 times. Find the expected frequencies when throwing of a 4, 5 or 6 is regarded as success.

9.14 Four dice are thrown and the number of sixes in each throw is recorded. This is repeated 180 times. Write down the theoretical frequencies 0,1,2,3 and 4 sixes. Calculate the mean number of sixes in a single throw.

9.15 The mean and variance of the binomial distribution are 6 and 2.4 respectively. Find p and n, the two parameters of the binomial distribution.

9.16 If the probability of a defective bolt is 0.1. Find the mean and standard deviation of the distribution of defective bolts in a total of 500.

9.17 If in a binomial distribution, the mean is 3 and the standard deviation is 1.5. Find its parameters.

9.18. Is it possible to have a binomial distribution with mean = 5 and S.D. = 3.7.

9.19 In binomial distribution with n=5, what is the value of other parameters of the binomial. If P(X = 0) = P(X = 1) find mean of the distribution.

9.20 Discuss the statement that in a binomial distribution $\mu = 6$ and $\sigma = 2.5$.

9.21 Find the binomial distribution whose mean is 12 and standard deviation is 3.

9.22 Find the mean and variance of the binomial $(q + p)^3$.

9.23 Find the mean and standard deviation of the Binomial distribution

$$(q + p)^n$$

9.24 Fill in the Blanks.

i) Mean of the binomial distribution is ______________ and its variance is __________________.

ii) Binomial distribution is symmetrical when _________.

iii) Binomial distribution is used when n is ___________.

iv) Binomial distribution has __________ parameters.

v) Binomial distribution is positively skewed when __________.

vi) Binomial variable is a _____________ variable which can assume any of the values $X = 0, 1, 2, 3, \ldots., n$.

vii) The shape of binomial distribution depends upon the values of ______.

viii) In a binomial distribution, the experiment consists of a ___________ number of _________ trials.

ix) Mean, median and mode for binomial distribution will be equal when __________________.

9.25 Five balls are drawn from a box containing 4 white and 7 black balls. If X denotes the number of black balls drawn, then obtain the probability distribution of X. Find the mean and variance of this distribution and verify the results using the formula.

9.26 Determine the probability distribution for the number of white beads among 5 beads drawn at random from a bowel containing 4 white and 7 black beads use this distribution to compute the mean and variance. And check the results by using the formula.

9.27 A committee of size 3 is selected from 4 men and 2 women. Find the probability distribution by hypergeometric experiment for the number of men on the committee.

9.28 A committee of size 5 is to be selected at random from 3 women and 5 men. Find the expected number of women on the committee.

9.29 Ten vegetable cans, all of the same size have lost their labels. It is known that 5 contain tomatoes and 5 contain corns. If 5 cans selected at random what is the probability that:

i) all contain tomatoes. **(ii)** 3 or more contain tomatoes.

9.30 Write T for true and F for f alse against each statement.

i) An experiment is called Bernolli experiment if it has two possible outcomes.

ii) The binomial distribution has three parameters *n, p* and *q*.

iii) A paper has 10 multiple-choice questions with 3 alternatives. Answering these questions by guesswork is a binomial experiment.

iv) If $p \neq q$, then the binomial Probability Distribution is skewed.

v) A binomial random variable is a continuous random variable.

vi) The binomial distribution is symmetrical distribution if $p = q = \frac{1}{2}$.

vii) The binomial distribution is negatively skewed distribution if $p > q$.

viii) The trials are independent in the hypergeometric distribution.

ix) Hypergeometric probability distribution has three parameters *n*, *N* and *k*.

x) The Binomial Distribution is used when '*n*' is large.

xi) The variance of the Binomial Distribution is '*npq*'

xii) The binomial random variable cannot assume the negative values

xiii) The binomial distribution is positively skewed distribution if $p < q$

Answers

Exercise 1

1.9 i) Population ii) less iii) parameter iv) statistic v) Constant

vi) zero vii) random viii) quantitative Ix) Inferential statistics x) primary data.

1.10 i) T ii) T iii) F iv) F v) F vi) T vii) F viii)F ix) T x) T

Exercise 2

2.8 ii. a) 3.45 – 3.95 class boundaries of 3.7

b) 3.455 – 3.945 class limits of 3.7

2.12 (i)

Class intervals	Frequency	Cumulative frequency
0.1 –0.8	5	5
0.9 –1.6	9	14
1.7 –2.4	15	29
2.5 –3.2	10	39
3.3 – 4.0	6	45
4.1 – 4.8	2	47
4.9 – 5.6	1	48
5.7 – 6.4	2	50

2.17 (i) process (ii) four (iii) lower, first, last

(iv) proportional (v) Cumulative frequency. (vi) Histogram

(vii) Vertical, no (viii) Proportional (ix) Frequency, Polygon

2.18 i) F ii) T iii) F iv) T v) F vi) F vii) T viii) T ix) F x) F

Exercise 3

3.2 Geometric Mean = 47.5675

3.3 Average: 14.2614

3.4 Mean = 1.5, Median =1, Mode = 1

3.5 Mean = 11.67, Median = 11.5, Mode= 12.8

3.6 (i) Mean = 4.405, (ii) Median = 4.533

3.9 (ii) Mode = 18

3.10 increase is 18.2%

3.11 Mean = 74.7024

3.12 Mean = 21.2, Geometric mean = 20.06

3.13 Mean = 42.12, Harmonic Mean = 35.56

3.14 Harmonic Mean = 50.51

3.15 Harmonic Mean = 37.5

3.16 Arithmetic Mean = 5; Numbers are $b = 2,8$ and $a = 2,8$

3.17 Three numbers are: $X_1 = 3,\ X_2 = 27,\ X_3 = 72$

3.18 (i) Harmonic Mean = 7.66

(ii) (Geometric Mean) increase is 15.98%

3.19 Mean = 11.09, Median = 11.07 Mode = 11.06

3.20 Mean = 0.7317, Median = 0.7318, Mode = 0.7319

$D_6 =$ 0.73192 $P_{74} =$ 0.7319

3.21 $Q_3 =$ 31.01 $D_5 =$ 23.783 $P_5 =$ 6.0125 Mode = 23.28

3.22 Mode = 0

3.23 Average = 24.003

3.24 Harmonic Mean = 134.607

3.25 Weighted Mean = 72.57

3.26 Weighted Mean = 203.4

3.27 Mean = 18 because $\sum(x-\bar{x}) = 0$ $\therefore$ $\sum(x - 18) = 0$

3.28 1st value is 30, 2nd value is 20, 3rd value is 10

3.29 i) T ii) T iii) F iv) T v) F vi) T vii) F viii) F ix) F
x) T xi) F xii) T xiii) T xiv) F xv) F xvi) T xvii) T

3.30 i) Arithmetic Mean ii) Harmonic Mean
iii) Arithmetic Mean iv) Median
v) Geometric Mean vi) Arithmetic Mean

3.31 (i) Measures of Central Tendency (ii) Extreme (iii) Zero (iv) two equal
(v) G.M (vi) Equal (vii) Mode (viii) Bimodal (ix) Identical (x) Median and Mode.

Exercise 4

4.5 ii. a) Q.D = 5.38 b) Coefficient of SK = 0.05

4.6 M.D = 5.73, Coefficient of M.D = 0.09, Variance = 47.855
Range = 30, Q.D = 5.3, Coefficient of Q.D = 0.08

4.7 Q.D = 33.25, Coefficient of Q.D = 0.302

4.8 S.D = 44.687 Variance = 1996.925 C. V = 9.914%

4.9 Combined SD=2.49

4.10 Combined SD=8.75, Combined mean = 57.06

4.11 $\overline{Y} = 2(\overline{X}) + 5$ and $S_Y = S_X$

4.12 $\overline{Y} = \overline{X} + 10$ $S_Y = S_X$

$\overline{Z} = \frac{110}{100}\overline{X}$ $S_Z = \frac{110}{100} S_X$

4.13 $\overline{X}_A = 44.47$ $C.V_A = 15.38$

$\overline{X}_B = 47.56$ $C.V_B = 13.32$

4.14 Median = 12.7, M.D=2.86

4.15 M.D $_{(\text{Med})}$ = 21.65

4.16 S.D = 7.99

4.17 (i) Tube B has greater absolute dispersion

(ii) Tube B has greater relative dispersion

4.19 C.V= 14.64%

4.20 C. V = 44.122%

4.21 $\mu_1 = 0$, $\mu_2 = 25.7$, $\mu_3 = 20.66$, $\mu_4 = 1189.7$

4.22 $\mu'_1 = 0.061$, $\mu'_2 = 2.64$, $\mu'_3 = 0.564$, $\mu'_4 = 28.38$

$\mu_1 = 0$, $\mu_2 = 2.637$, $\mu_3 = 0.0811$, $\mu_4 = 28.301$

4.23 C.V=34.64%, Distribution is Symmetric

4.24 Distribution is platy kurtic as $\beta_2 = 2.158$

4.25 (i) S.K=–1.118, – vely Skewed (ii) – vely Skewed

4.26 $\mu_1=0$, $\mu_2=7.139$, $\mu_3=-5.364$, $\mu_4=125.5705$

Distribution is +vely Skewed

4.27 -vely Skewed

4.28 Coefficient of SK=0.1218

4.29 Coefficient of SK=0.299

4.30 $\overline{X} = 3, C.V = 41\%$

4.31 $\mu_2=11$, $\mu_3=49$, $\mu_4=192$, $\beta_1=1.8$, $\beta_2=1.59$

4.33 LeptoKurtic

4.34 (i) –vely Skewed (ii) –vely Skewed (iii) +vely Skewed

4.35 (i) platy kurtic (ii) $S.D=3$

4.36 (i) $\mu_4=1875$

4.37 (i) +vely Skewed (ii) –vely Skewed (iii) +vely Skewed

4.38 (i) Symmetric (ii) + vely Skewed (iii) – vely Skewed

(iv) Symmetric (v) Lepto Kurtic

4.39

i)	F	ii)	F	iii)	T	iv)	F	v)	T
vi)	T	vii)	T	viii)	F	ix)	F	x)	F
xi)	T	xii)	F	xiii)	F	xiv)	T	xv)	T

4.40 (i) Called Scatter (ii) Relative (iii) Median

(iv) Change (v) Free measurement (vi) Consistent

(vii) Skewness (viii) Equidistant (ix) Zero

(x) Ratio

Exercise 5

5.7 48.39, 61.29, 67.74, 96.77, 119.35, 122.58, 129.03, 154.84

5.8 (i) 100, 149.37, 155.12, 106.3, 123.23, 172.05, 164.057, 155.91, 186.22, 193.37, 123.23, 127.17, 159.06, 198.82, 255.12

(ii) 63.31, 94.57, 98.20, 67.30, 78.08, 108.92, 104.19, 98.70, 117.9, 122.38, 78.02, 80.51, 100.70, 125.87, 161.52.

5.9 (i) 100, 111.11, 122.22, 133.33, 155.56, 144.44, 166.67, 211.11, 200

(ii) 66.94, 74.38, 81.82, 89.26, 104.14, 96.70, 111.57, 141.33, 133.89

5.10 (i) 100, 125, 150, 175, 200, 250, 225, 250, 275

(ii) 80, 100, 120, 140, 160, 200, 180, 200, 220.

(iii) 40, 50, 60, 70, 80, 100, 90, 100, 110.

5.11 100, 103.85, 100, 88.46, 94.23, 86.54, 80.77, 96.15, 78.84, 80.77

5.12 100, 104, 110, 114, 124, 144, 146, 150, 142, 140.

5.13 100, 100 , 101.07, 104.68, 108.90, 100.58.

5.14 100, 106.96, 103.175, 110.095

5.15 100, 109.0, 125.53, 145.94

5.16 100, 112.66, 132.49, 136.20, 146.74

5.17 (i) 100, 91.59, 82.08, 81.55, 115.57, 113.37

(ii) 100, 89.07, 68.48, 69.17, 114.32, 107.66

5.18 100, 114.3, 119.4, 128.7, 134.9

5.19 (i) 400 (ii) 400

5.20 101.2508

5.21 86.77

5.22 71.83

5.23 162.62

5.25 Laspeyre's=108.78 Paasche's=109.21

Fisher's=108.99

5.26 100, 99.1, 109.97, 104.1, 105.53, 126.66, 128.93, 122.67, 140.2

5.27 73.002

5.28 98.15

5.29 (i) Price-numbers

(ii) Volume index numbers

(iii) Aggregative index numbers

(iv) Forecasting seasonal, cycles

(v) General or special

(vi) Fixed base method, chain base method

(vii) A normal year

(viii) Chaining process, chain base method

(ix) Chain base method

(x) Simple aggregative, simple average of relatives

5.30 i) T ii) T iii) F iv) T v) T vi) T vii) F viii) F ix) F x) F

Exercise 6

6.1 (ii) (a) $A \cap B = \{\ \}$ (b) $\{1,2,3,4,5,7\}$ (c) $\{2,4,6,7,8\}$ (d) $\{2,4\}$

6.2 (i) 5040 (ii) 518918400 (iii) 367567200 (iv) 60480

(v) 3603600 (vi) 120 (vii) 10 (viii) 635013559600 (ix) 330

6.3 (i) {(1,2), (1,3), (4,2), (4,3)} (ii) {(2,1), (2,4), (3,1), (3,4)}

(iii){(1,4), (1,1), (4,1), (4,4)} (iv) {(2,2), (2,3), (3,2), (3,3)}

6.8 (i) $\frac{1}{2}$ (ii) $\frac{1}{6}$ (iii) $\frac{1}{2}$ **6.9** $\frac{1}{4}$

6.10 $P(\text{sum} < 7) = P(\text{sum} > 7) = \frac{15}{36}$ **6.11** $\frac{1}{22}$ **6.12** $\frac{1}{4}$

6.13 $\frac{4}{33}$

6.14 (i) $\frac{1}{22}$ (ii) $\frac{7}{22}$ (iii) $\frac{9}{44}$

6.15 (i) $\frac{5}{36}$ (ii) $\frac{8}{36}$ (iii) $\frac{9}{18}$

6.16 (i) $\frac{16}{1326}$ (ii) $\frac{650}{1326}$ (iii) $\frac{676}{1326}$

6.17 $\frac{792}{20349}$ **6.18** $\frac{13}{28}$ **6.19** $\frac{5}{8}$ **6.20** $\frac{16}{1326}$

6.21 (i) $\frac{33}{54145}$ (ii) $\frac{33}{54145}$

6.22 (i) $\frac{4}{364}$ (ii) $\frac{100}{364}$ **6.23** (i) $\frac{25}{75}$ (ii) $\frac{50}{75}$

(iii) $\frac{55}{75}$ (iv) $\frac{60}{75}$

6.26 0.92 **6.27** $\frac{2}{3}$ **6.28** $\frac{5}{18}$

6.29 $\frac{7}{429}$ **6.30** (i) 0.0355 (ii) 0.04395

6.31 (i) 0.25 (ii) 0.2083 **6.32** i) $\frac{207}{625}$

6.33 (i) $\frac{6}{15}$ (ii) $\frac{3}{15}$ (iii) $\frac{4}{15}$ (iv) $\frac{13}{15}$ (v) $\frac{2}{15}$ (vi) $\frac{9}{15}$

6.34 $\frac{64}{29100}$ **6.35** (i) $\frac{1}{169}$ (ii) $\frac{1}{221}$ **6.36** (i) $\frac{7}{16}$ (ii) $\frac{55}{784}$

6.38 i) $\frac{9}{14}$ ii) $\frac{2}{7}$ **6.37** i) 0.5 **6.39** 4 to 3

6.40 $\frac{53}{80}$ **6.41** i) $\frac{19}{42}$ ii) $\frac{23}{42}$ **6.42** $\frac{4}{9}$

6.43 0.23 **6.44** $\frac{7}{8}$ **6.45** (i) $\frac{1}{8}$ (ii) $\frac{5}{72}$

(iii) $\frac{5}{36}$ (iv) $\frac{19}{27}$ **6.46** 0.1

6.47 $\frac{1}{12}$ **6.48** $\frac{1}{17}$ **6.49** $\frac{36}{91}:\frac{30}{91}:\frac{25}{91}$

6.50 $\frac{38}{63}$ **6.51** $\frac{16}{36}$ **6.52** $\frac{13}{32}$

6.53 (i) $\frac{11}{25}$ ii) $\frac{60}{100}$ **6.54** i) well defined, distinct

ii) Singelton or unit iii) Power iv) Compound v) Mutually Exclusive

vi) Collectively exhaustive iiv) Exhaustive viii) Equally likely ix) 2^n

x) Permutation, ${}^{n}p_{r}$ xi) Combinatoin

6.55 i) T ii) F iii) F iv) F v) F vi) T vii) F viii) T ix) F x) T

Exercise 7

7.1 (iii)a) Discrete random variable

b) Discrete random variable

c) Continuous random variable

d) Continuous random variable

7.2 $m = 100$ $a = 21$ $b = 7$ $x = 10$

The numbers are: 17, 64, 51, 78 45 and 52

7.3 Let the one digit random numbers from table are:

8 3 2 6 9 8 2 8 1 2

so

H T H H T H H H T H

The frequency of 1 head = 7 and frequency of 0 head = 3

7.4 S={TT, HT, TH, HH}

7.6 S= {BB, BM, MB, MM} and

S={0, 1, 2}

7.7 S={TT, HT, TH, HH} and

S={0, 1, 2}

7.8

y	0	1	1	1	2	2	2	3
Outcomes	GGG	BGG	GBG	GGB	BBG	BGB	GBB	BBB

7.9

x	1	2	3	4
$f(x)$	5/70	30/70	30/70	5/70

7.11 (i) Generated, random (ii) Equal (iii) Chance or Stochastic

(iv) Discrete (v) Range, Continuous (vi) 1

(vii) A mathematical (viii) Mathematical expectation

(ix) Converges (x) Discrete, Continuous.

Exercise 8

8.2

x	0	1	2	3
$P(x)$	1/56	15/56	30/56	10/56

8.3

x	0	1	2	3
$P(x)$	1/35	12/35	18/35	4/35

8.4 (ii)

x	0	1	2	3	4
$P(x)$	1/16	4/16	6/16	4/16	1/16

(iii)

x	0	1	2	3	4
$P(x)$	15/210	80/210	90/210	24/210	1/210

8.6 (ii) a) No b) No c) Yes d) No

8.7 a) $C=\frac{1}{15}$ b) $0<C<1$

8.8 $-\frac{1}{3}<Q<\frac{1}{4}$

8.9 i) 7.44 ii. *a)* 0.55, 1.3475 b) 2.1, 5.39

8.10 i) 0.2 ii) 0.6 iii) 0.6 iv) 0.7 v) 0.9

8.11 i) {3, 4, 5, 6, 7, 8, 9} (ii) {1, 2, 3, 4}

8.13 y takes the value between 0 and 30, continuous

8.14 Mean = 2.3 Variance = 2.01

8.15 i. a) a=1/8 b) $\frac{5}{16}$

ii. a) c=1/2 b) 0.3125 c) 0.5625

iii. a) k=2, 13/216

8.17

x	0	1	2	3
$P(x)$	1/8	3/8	3/8	1/8

8.18 (i) Probability Distribution of x

x	0	1	2	3	4	Total
$P(x)$	56/792	280/792	336/792	112/792	8/792	792/792

(ii) 336/792

8.19 ii) $k=b-a$

8.20 ii) Mean=1.94, S.D (x) = 1.43, C.V. = 73.71%, 8.21 Mean= 2.6, Var= 1.34

8.22 i) a) 0.7031 b) 0.543125 c) 0

ii) a) $\frac{1}{4}$ b) $\frac{15}{16}$ c) $\frac{3}{16}$

8.23 i) a =1/210 ii) 0.4928 iii) 0.6476 iv) 0.6851

8.24 352.5

8.25 1.875

8.27 ii)

x	0	1	2	3	4	5
$P(x)$	1/252	26/252	126/252	226/252	251/252	252/252

8.28 ii) Mean= 1.875

8.29 A's expectation=Rs.6.0; B's expectation=5.0

8.30 A's expectation=Rs.6.0; B's expectation=4.0

C's expectation=Rs.2.0

8.31

x	0	1	2	3	Sum
$P(x)$	1/56	15/56	30/56	10/56	56/56

Mean = 105/56, Variance=0.5024

8.32 Probability Distribution

x	0	1	2	3	4
$p(x)$	1/256	12/256	54/256	108/256	81/256

Mean=3 Variance=0.75

8.33

x	0	1	2
$P(x)$	0.81	0.18	0.1

Mean = 0.2, Variance = 0.18

8.34 $E(X)$ =8/5 Variance (X)=8.24

$E(5X+10)$=18 Variance ($5X+10$)=206

1:25

8.35 Mean=7 Variance=5.83

8.36 i) $E(X)$=7 ii) $E(X^2)$=590 iii) Variance (X)=541

8.37 $E(X)$=0.75 $E(X^2)$=1.125

8.38 i) T ii) F iii) F iv) F v) F vi) T vii) F viii)T ix) F x) T

Exercise 9

9.3

x	0	1	2	3	4	5
$P(x)$	3125/32768	9375/32768	11250/32768	6750/32768	2025/32768	243/32768

9.4 (i) $\frac{1}{16}$ (ii) $\frac{1}{4}$ (iii) $\frac{11}{16}$ (iv) $\frac{11}{16}$

9.5 (i) 0.4096 (ii) 0.1536 **9.6** (i) 0.995 (ii) 0.2617

9.7 (i) 0.1694 (ii) 0.5340 **9.8** (i) 14080/15625 (ii) 265/15625

9.9 (i) 56/256 ~~(ii)~~ 254/256 **9.10** Mean = −1, Variance = 0.067

9.11 (i) 243/1024 (ii) 1008/1024

9.12

x	0	1	2	3	4	5
$P(x)$	3125/7776	3125/7776	1250/7776	250/7776	25/7776	1/7776

Mean = 0.833 Variance = 0.701

9.13 3, 15, 30, 30, 15, 3

9.14

x	0	1	2	3	4
f	87	69	21	3	0

Mean = 0.67

9.15 $P = 0.6$ $n = 10$ **9.16** Mean = 40 S.D = 6.

9.17 $P = 0.25$ $n = 12$ **9.18** No possible, q cannot be greater than 1

9.19 $P = \frac{1}{6}$ Mean = 5/6

9.20 Not possible, q cannot be greater than 1. **9.21** $p = 0.25$ $q = 0.75$ $n = 48$

$P(X = x) = {}^{48}C_x (0.25)^x (0.75)^{48-x}$

9.22 Mean = $3p$ Variance = $3pq$ **9.23** Mean = np S.D = $\sqrt{npq}$

9.24 (i) np , npq (ii) $p = q = \frac{1}{2}$ (iii) n is small (iv) Two (v) $p < q$

(vi) Discrete $r.v.$ (vii) $S.D.$ (viii) fixed, independent (ix) $p = q$

9.25

x	1	2	3	4	5
$P(x)$	1/66	12/66	30/66	20/66	3/66

Mean = 70/66 = $\frac{35}{11}$ Variance = 0.6942

Mean = $n.\frac{k}{N} = \frac{35}{11}$, Variance= $n.\frac{k}{N}(1-\frac{k}{N})\left(\frac{N-n}{N-1}\right)$ = 0.6942

9.26

x	0	1	2	3	4
$P(x)$	21/462	140/462	210/462	84/462	7/462

Mean = 1.8181 Variance = 0.6942

Mean = $n\frac{K}{N} = 1.8181$ Variance = 0.6942

9.27

x	1	2	3
$P(x)$	4/20	12/20	4/20

9.28 Expected number of women = 15/8 **9.29** (i) 1/252 (ii) 126/252

9.30 i) F ii) F iii) T iv) T v) F vi) T vii) F viii)T ix) T x) T

Look at the road! Not at your phone

Fasten your seat belt while driving to ensure your safety

Punjab Curriculum and Textbook Board provides standard textbooks at low price according to the approved curricula. Suggestions are requested for improvement of these books by pointing out any error in spellings, contents, etc.

Managing Director
Punjab Curriculum and Textbook Board
21-E-II, Gulberg-III, Lahore.

Fax No: 042-99230679
E-mail: chairman@ptb.gop.pk
Website: www.ptb.gop.pk

Designed By: ART CELL PCTB

Printed by Maryah Printers Islamabad.

امانت دیانت

پنجاب
کریکولم اینڈ
ٹیکسٹ بک بورڈ
لاہور

صداقت شرافت

PUNJAB CURRICULUM AND
TEXTBOOK BOARD, LAHORE