

INTRODUCTION TO STATISTICAL THEORY

Part 1

(A text book for Degree and Post-Graduate Students)

By

Prof. Sher Muhammad Qaudhry
*B.Sc. (Hons.), M.A. (Gold Medallist)
F.S.S. (Lahore)
Ex-Head, Department of Statistics
Government College (now GC University), Lahore*

Dr. Shahid Kamal
*M.Sc. (Gold Medallist), Ph.D. (U.K.)
Principal and Professor
College of Statistical & Actuarial Sciences
University of the Punjab, Lahore*

ILMI KITAB KHANA
Kabir Street, Urdu Bazar, Lahore 54000
(Pakistan)

© Copyright 1968, 1973, 1984, 1996, 2009 by Ilmi Kitab Khana, Lahore. All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means without the prior written permission of the author and the publisher.

Approved for College Libraries by the Government of the Punjab, Education Department vides letter No. S.O. (CD) Edu-2-23/72 dated 13.3.73.

First Edition	1968
Second Edition	1970
Third Edition	1971
Fourth Edition	1973
Fifth Edition	1984
Sixth Edition	1996
Seventh Edition	1999
Eighth Edition	2009
Reprint Edition	2013
Price	Rs. 400/- 350/-



BESTS/ANR/DR

Professor Sher Mohammad Chowdhry,
Head of Department of Statistics,
Government College,
Lahore,
WEST PAKISTAN.

THE UNIVERSITY OF ASTON IN BIRMINGHAM

Cotts Green, Birmingham B4 7ET/ Tel: 021 380 3813 Ext:

The Department of Mathematics
Head of Department: Professor N. M. Stephens
Professor C. F. Lawden

Date: February 1972

Dear Professor Sher Mohammad Chowdhry,

I am very thankful to you for a set of your books which I received a few days ago. I have found the books very well written and of great use to the student community and would like to congratulate you for producing such standard text books on "statistics".

With kind regards,

Yours sincerely,

A. H. Rajput

A. H. Rajput,
Senior Lecturer

	Page
Preface	
1. INTRODUCTION	
1.1 Meaning of Statistics	1
1.1.1 Use of Statistical Information	1
1.1.2 Characteristics of Statistics	2
1.1.3 Descriptive and Inferential Statistics	3
1.1.4 Populations and Samples	4
1.1.5 Importance of Statistics	4
1.2 Observations and Variables	5
1.2.1 Variables	5
1.2.2 Discrete and Continuous Variables	5
1.2.3 Measurement Scales	6
1.2.4 Errors of Measurement	7
1.2.5 Significant Digits	8
1.2.6 Rounding off a Number	8
1.3 Collection of Data	9
1.3.1 Collection of Primary Data	9
1.3.2 Collection of Secondary Data	10
1.3.3 Editing of Data	11
1.3.4 Uses and Misuses of Statistics	11
Exercises	11
2. PRESENTATION OF DATA	
2.1 Introduction	15
2.2 Classification	15
2.2.1 Aims of Classification	15
2.2.2 Basic Principles of Classification	15
2.3 Tabulation	15
2.3.1 Types of Tables	16
2.3.2 Main Parts of a Table and its Construction	16
2.4 Frequency Distribution	20
2.4.1 Class-limits	20
2.4.2 Class-boundaries	20
2.4.3 Class-Mark	20
2.4.4 Class Width or Interval	20
2.4.5 Constructing a Grouped Frequency Distribution	21
2.4.6 Cumulative Frequency Distribution	26
2.5 Stem-and-Leaf Display	27
2.6 Graphical Representation	28
2.7 Diagrams	29
2.7.1 Simple Bar Chart	30
2.7.2 Multiple Bar Chart	30
2.7.3 Component Bar Chart	31

2.7.4	Rectangles and Sub-divided Rectangles	32
2.7.5	Pictograms	33
2.7.6	Pie Diagrams	34
2.7.7	Profit and Loss Chart	35
2.8	Graphs	35
2.8.1	Graph of Time Series – Histogram	36
2.8.2	Histogram	37
2.8.3	Frequency Polygon	38
2.8.4	Frequency Curve	39
2.8.5	Cumulative Frequency Polygon or Ogive	40
2.8.6	Ogive for a Discrete Variable	41
2.8.7	Types of Frequency Curves	41
2.8.8	Ratio Charts or Semi-logarithmic Graphs	43
	Exercises	44
3.	MEASURES OF CENTRAL TENDENCY OR AVERAGES	
3.1	Introduction	51
3.2	Criteria of a Satisfactory Average	51
3.3	Types of Averages	53
3.4	The Arithmetic Mean	56
3.4.1	The Weighted Arithmetic Mean	57
3.4.2	Properties of the Arithmetic Mean	57
3.4.3	Mean from Grouped Data	59
3.4.4	Change of Origin and Scale	61
3.5	The Geometric Mean	62
3.6	The Harmonic Mean	64
3.7	The Median	67
3.7.1	Quantiles	68
3.8	The Mode	72
3.9	Empirical Relation between Mean, Median and Mode	74
3.10	The Box Plot	74
3.11	Relative Merits and Demerits of Various Averages	75
3.11.1	The Arithmetic Mean	75
3.11.2	The Geometric Mean	75
3.11.3	The Harmonic Mean	76
3.11.4	The Median	76
3.11.5	The Mode	76
	Exercises	77
4.	MEASURES OF DISPERSION, MOMENTS AND SKEWNESS	
4.1	Introduction	87
4.2	Range	87
4.3	The Semi-Interquartile Range or the Quartile Deviation	88
4.4	The Mean (or Average) Deviation	89

4.5	The Variance and Standard Deviation	91
4.5.1	Change of Origin and Scale	96
4.5.2	Interpretation of the Standard Deviation	97
4.5.3	Co-efficient of Variation	98
4.5.4	Properties of Variance and Standard Deviation	100
4.5.5	Standardised Variables	103
4.6	Trimmed and Winsorized Measures	104
4.7	Moments	105
4.7.1	Moments about the Mean in terms of Moments about an arbitrary origin, say α , and conversely	106
4.7.2	Sheppard's Corrections	108
4.7.3	Moment-Ratios	108
4.7.4	Change of Origin and Scale	110
4.7.5	Charlier Check	111
4.8	Skewness	114
4.9	Kurtosis	115
4.10	Describing a Frequency Distribution	116
	Exercises	116
5.	INDEX NUMBERS	
5.1	Introduction	131
5.1.1	Simple and Composite Index Numbers	131
5.1.2	Problems involved in Index Number Construction	131
5.2	Main Steps in the Construction of Index Numbers of Wholesale Prices	132
5.2.1	Selection of Commodities for Inclusion	132
5.2.2	Selection of the Base Period	132
5.2.3	Selection of Average	134
5.2.4	Selection of Appropriate Weights	135
5.3	Unweighted Index Numbers	135
5.3.1	Simple Aggregative Index	135
5.3.2	Simple Average of Relatives	136
5.4	Weighted Index Numbers	140
5.4.1	Weighted Aggregative Price Index Numbers	140
5.4.2	Weighted Average of Relatives Price Index Number	145
5.5	Quantity Index Numbers	147
5.6	Tests for Index Number Formulae	151
5.6.1	Time Reversal Test	151
5.6.2	Factor Reversal Test	153
5.6.3	Circular Test	155
5.7	Consumer Price Index Number	156
5.7.1	Measuring	156
5.7.2	Construction of Consumer Price Index Numbers	156

5.7.3	Shortcomings or Drawbacks of Consumer Price Index Numbers	159
5.8	Uses of Index Numbers	160
5.9	Limitations of Index Numbers	160
	Exercises	161
6.	PROBABILITY	
6.1	Introduction	173
6.2	An Aside - Sets	173
6.2.1	Subsets	174
6.2.2	Venn Diagram	175
6.2.3	Operations on Sets	175
6.2.4	The Algebra of Sets	177
6.2.5	Partition of Sets	177
6.2.6	Class of Sets	177
6.2.7	Cartesian Product Sets	177
6.2.8	Relation and Function	179
6.3	Random Experiment	179
6.3.1	Sample Space	179
6.3.2	Events	180
6.3.3	Events and Symbols: Representations	181
6.3.4	Counting Sample Points	181
6.4	Definitions of Probability	184
6.4.1	Subjective or Personalistic Probability	187
6.5	Laws of Probability	189
6.6	Conditional Probability	196
6.7	Independent and Dependent Events	203
	Exercises	214
7.	RANDOM VARIABLES	
7.1	Introduction	227
7.2	Distribution Function	227
7.3	Discrete Random Variables and its Probability Distribution	228
7.4	Continuous Random Variable and its Probability Density Function	231
7.5	Joint Distributions	237
7.5.1	Bivariate Distribution Function	237
7.5.2	Bivariate Probability Functions	238
7.5.3	Marginal Probability Functions	238
7.5.4	Conditional Probability Functions	239
7.5.5	Independence	239
7.5.6	Continuous Bivariate Distributions	243
7.6	Mathematical Expectation of a Random Variable	248
7.6.1	Expectation of a Function of a Random Variable	250

7.6.2	Properties of Expected Values	256
7.6.3	Covariance of Two Random Variables	263
7.6.4	Variance of the Sum or Difference of Two Random Variables	263
7.6.5	Correlation Co-efficient of Random Variables	263
7.7	Medians and Modes of Continuous Random Variables	266
7.8	Chebyshev's Inequality	268
7.9	Moment Generating Function	269
7.9.1	Cumulant Generating Function	271
7.9.2	Relation between Cumulants and Moments	272
7.9.3	Characteristic Function	273
	Exercises	273
8.	DISCRETE PROBABILITY DISTRIBUTIONS	
8.1	Introduction	285
8.2	Binomial Probability Distribution	285
8.2.1	Derivation of Binomial Probability Distribution	285
8.2.2	Binomial Frequency Distribution	291
8.2.3	Properties of the Binomial Probability Distribution	292
8.2.4	The Recurrence Formula for Binomial Distribution	297
8.2.5	Fitting a Binomial Distribution to Observed Data	298
8.2.6	Moment Generating and Cumulant Generating Functions of the Binomial Distribution	299
8.3	Hypergeometric Probability Distribution	302
8.3.1	Derivation of Hypergeometric Distribution	302
8.3.2	Properties of Hypergeometric Distribution	305
8.4	Poisson Distribution	308
8.4.1	Derivation of Poisson Approximation to the Binomial	309
8.4.2	Poisson Frequency Distribution	312
8.4.3	Properties of the Poisson Distribution	313
8.4.4	The Recurrence Formula for Poisson Distribution	316
8.4.5	Fitting a Poisson Distribution to Observed Data	317
8.4.6	Poisson Process	318
8.4.7	Moment Generating and Cumulant Generating Functions of the Poisson Distribution	320
8.5	Negative Binomial Distribution	321
8.5.1	Derivation of the Negative Binomial Distribution	322
8.5.2	Properties of the Negative Binomial Distribution	324
8.6	Geometric Distribution	325
8.6.1	Derivation of the Geometric Distribution	325
8.6.2	Properties of the Geometric Distribution	326
8.6.3	Moment Generating Functions of the Geometric Distribution	327
8.7	Multinomial Distribution	328
8.7.1	Derivation of the Multinomial Distribution	328

Exercises

9. CONTINUOUS PROBABILITY DISTRIBUTIONS

9.1	Introduction	38
9.2	Uniform Distribution	38
9.2.1	Properties of the uniform distribution	38
9.2.2	Moment-Generating Function of Uniform Distribution	38
9.3	Exponential Distribution	38
9.3.1	Properties	38
9.3.2	Moment-Generating Function of Exponential Distribution	38
9.4	Gamma and Beta Distributions	38
9.4.1	Gamma Function	38
9.4.2	Beta Function	38
9.4.3	Gamma Distribution	38
9.4.4	Properties of Gamma Distribution	38
9.4.5	Moment-Generating Function of Gamma Distribution	38
9.4.6	Beta Distribution of the First Kind	38
9.4.7	Properties of $\beta_1(m, n)$	38
9.4.8	Beta Distribution of Second Kind	38
9.4.9	Properties of $\beta_2(m, n)$	38
9.5	Normal Distribution	38
9.5.1	Standardized Normal Distribution	38
9.5.2	Properties of Normal Distribution	38
9.5.3	Moment-Generating and Cumulant Generating Functions of the Normal Distribution	38
9.5.4	Tabular Areas of Normal Distribution	38
9.5.5	Inverse Use of Table of Areas under Normal Curve	38
9.5.6	Normal Approximation to the Binomial Distribution	38
9.5.7	Normal approximation to the Poisson Distribution	38
9.5.8	Fitting a Normal Distribution	38
9.5.9	Ordinates of the Normal Distribution	38
	Exercises	38

10. SIMPLE REGRESSION AND CORRELATION

10.1	Introduction	395
10.2	Deterministic and Probabilistic Relations or Models	395
10.3	Scatter Diagram	396
10.4	Simple Linear Regression Model	396
10.4.1	An Aside – The Principle of Least Squares	396
10.4.2	Least-Squares Estimator in Simple Linear Regression	396
10.4.3	Properties of the Least-Squares Regression Line	400
10.4.4	Standard Deviation of Regression or Standard Error of Estimate	402
10.4.5	Coefficient of Determination	404

10.5	Correlation	406
10.5.1	Pearson Product Moment Correlation Co-efficient	406
10.5.2	Correlation and Causation	408
10.5.3	Properties of r	408
10.5.4	Correlation Co-efficient for Grouped Data	411
10.6	Rank Correlation	413
10.6.1	Derivation of Rank Correlation	414
10.6.2	Rank Correlation for Tied Ranks	416
10.6.3	Co-efficient of Concordance	417
	Exercises	419
11.	MULTIPLE REGRESSION AND CORRELATION	
11.1	Introduction	429
11.2	Multiple Linear Regression with two Regressors	429
11.2.1	Expression of Multiple Linear Regression in Deviation Form	421
11.2.2	Standard Error of Estimate	433
11.2.3	Co-efficient of Multiple Determination and Multiple Correlation	434
11.2.4	Subscript Notation	435
11.2.5	Properties of Residuals	436
11.2.6	Multiple Regression in Terms of Linear Correlation Coefficients	437
11.3	Multiple Correlation Co-efficient	438
11.4	Partial Correlation	441
11.4.1	Relationship between Multiple and Partial Correlation Co-efficients	446
11.5	Curvilinear Regression	447
	Exercises	447
12.	CURVE FITTING BY LEAST-SQUARES	
12.1	Introduction	455
12.2	Approximating Curves and the Principle of Least-Squares	455
12.2.1	Fitting a Straight Line	455
12.2.2	Fitting of a second degree Parabola	457
12.2.3	Fitting of Higher Degree Parabolic Curves	460
12.2.4	Change of Origin and Unit	461
12.3	Exponential Curves	462
12.4	Other Types of Curves	465
12.4.1	Modified Exponential Curve	465
12.4.2	The Gompertz Curve	467
12.4.3	The Logistic Curve	467
12.4.4	The Makeham Curve	467
12.5	Criteria for a Suitable Curve	468
12.6	Finding Plausible Values by LS Method	468

Exercises	431
13. TIME SERIES ANALYSIS	
13.1 Introduction	477
13.2 Components of a Time Series	479
13.2.1 Secular Trend	479
13.2.2 Seasonal Variations	479
13.2.3 Cyclical Fluctuations	479
13.2.4 Irregular or Random Variations	480
13.3 Time Series Decomposition	480
13.4 Analysing the Secular Trend	481
13.4.1 The Method of Freehand Curve	481
13.4.2 The Method of Semi-Averages	481
13.4.3 The Method of Moving Averages	481
13.4.4 The Method of Least-Squares	481
13.5 Detrending	491
13.6 Analysing the Seasonal Variations	492
13.6.1 The Percentage of Annual Average method	492
13.6.2 The Ratio-to-Moving-Average method	493
13.6.3 The Ratio-to-Trend Method	493
13.6.4 The Link-Relative Method	497
13.7 Desasonalization of Data	498
13.8 Analysing the Cyclical Variations	499
13.9 Analysing the Irregular Variations	500
13.10 Forecasting	501
13.10.1 Forecasting by Exponential Smoothing	501
13.11 Serial Correlation	502
Exercises	503
Answers to Exercises	515
Tables	539
Index	545

CHAPTER 1

INTRODUCTION

INTRODUCTION

1.1 MEANING OF STATISTICS

People view Statistics in many different ways. Generally it is considered to be a subject that deals with percentages, charts, graphs, averages and tables. Some people think that Statistics is a subject consisting of rules, methods and techniques of collecting and presenting large amount of numerical information, while other people think that it is a subject of making inferences about the population on the basis of sample information.

The word "Statistics" which comes from the Latin word *status*, meaning a political state, originally meant information useful to the state, for example, information about the sizes of populations and armed forces. But this word has now acquired different meanings.

In the first place, the word *statistics* refers to "numerical facts systematically arranged". In this sense, the word *statistics* is always used in the plural. We have, for instance, statistics of prices, statistics of road accidents, statistics of crimes, statistics of births, statistics of educational institutions, etc. In all these examples, the word *statistics* denotes a set of numerical data in the respective fields. This is the meaning the man in the street gives to the word *Statistic* and most people usually use the word *data* instead.

Example 1.1 In the following examples, the facts and figures usually called Statistics presented in the media almost every day are given:

- (i) Children who brush their teeth with brand XYZ toothpaste have 60% fewer cavities.
- (ii) The Bureau of census projects the population of Pakistan to be 170.1 million in the year 2010.
- (iii) Eight out of ten Pakistanis do not have mobiles.
- (iv) The prevalence of diabetes is nearly 3 times as high in over weight people as compared to normal weight people.
- (v) In 1980 it was estimated that 0.1% of people had tried any sort of drug; whereas in 2008 it was estimated that 10% had done so.

In the second place, the word *statistics* is defined as a discipline that includes procedures and techniques used to collect, process and analyse numerical data to make inferences and to reach decisions in the face of uncertainty. It should of course be borne in mind that uncertainty does not imply ignorance but it refers to the incompleteness and the instability of data available. In this sense, the word *statistics* is used in the singular. As it embodies more or less all stages of the general process of learning, sometimes called scientific method, statistics is characterized as a science. Thus the word *statistics* used in the plural refers to a set of numerical information and in the singular, denotes the science of basing decision on numerical data. It should be noted that statistics as a subject is mathematical in character.

Thirdly, the word *statistics* are numerical quantities calculated from sample observations. A single quantity that has been so calculated is called a statistic. The mean of a sample for instance is a statistic. The word *statistics* is plural when used in this sense.

1.1.1 Use of Statistical Information. The statistical information are and can be used for a variety of purposes. Some of them are:

- (i) to inform general public;
- (ii) to explain things that have happened;

- iii) to justify a claim;
- iv) to provide general comparisons;
- v) to predict the decision regarding future outcomes;
- vi) to estimate the unknown quantities;
- vii) to establish association / relationship between factors.

Hence Statistics is a subject which is much more than just numbers. It tells us what is done to numbers. The following three examples further explain how Statistics may be used:

Example 1.2: Suppose we want to determine the best teacher at Govt. College University, Lahore. How should we decide this? This could be done by asking Govt. College University students who the best teacher is. To do so, we collect the data, analyze the results and make the decision. Now various questions are:

- i) should we survey every student?
- ii) how will the survey be conducted?
- iii) how will the data be analyzed?
- iv) how will the best teacher be determined? etc.

In order to answer these and other questions, Statistical techniques are used.

Example 1.3: A TV station claims that an advertisement of a product on their channel attracts more customers compared to all other TV channels. Now, if this claim is based on data, then it can be used to market the TV channel. Suppose we have some doubts about the claim. In order to remove the doubts, we might gather relevant information, analyze the results using appropriate statistical technique and make decision regarding the claim.

Example 1.4: Suppose University of the Punjab is planning an expansion program of its physical facilities. To draw up an effective course of action, the University authorities decide that it needs to answer this question, how many college students will we need to accommodate over the next ten years. The question can be further broken down into many smaller questions. How many college students will then be in the Punjab? How many will want to attend the University of the Punjab? etc. Once again, Statistical methods can assist in evaluating and planning of expansion programs.

1.1.2 Characteristics of Statistics. The definition stated above indicates that statistics is a subject in its own right. It may therefore be desirable to know the characteristic features of statistics in order to appreciate and understand its general nature. Some of its important characteristics are given below:

- i) Statistics deals with the behaviour of aggregates or large groups of data. It has nothing to do with what is happening to a particular individual or object of the aggregate.
- ii) Statistics deals with aggregates of observations of the same kind rather than isolated figures.
- iii) Statistics deals with variability that obscure underlying patterns. No two objects in the universe are exactly alike. If they were, there would have been no statistical problem.

- (iv) Statistics deals with uncertainties as every process of getting observations whether controlled or uncontrolled, involves deficiencies or chance variation. That is why we have to talk in terms of probability.
- (v) Statistics deals with those characteristics or aspects of things which can be described numerically either by counts or by measurements.
- (vi) Statistics deals with those aggregates which are subject to a number of random causes, e.g. the heights of persons are subject to a number of causes such as race, ancestry, age, diet habits, climate and so forth.
- (vii) Statistical laws are valid *on the average* or in the long run. There is no guarantee that a certain law will hold in all cases. Statistical inference is therefore made in the face of uncertainty.
- (viii) Statistical results might be misleading and incorrect if sufficient care in collecting, processing and interpreting the data is not exercised or if the statistical data are handled by a person who is not well versed in the subject matter of statistics.

1.1.3 Descriptive and Inferential Statistics: Statistics as a subject, may be divided into *descriptive statistics* and *inferential statistics*.

Descriptive statistics is that branch of statistics which deals with concepts and methods concerned with summarization and description of the important aspects of data. This area of study consists of the *classification* of data, their graphical displays and the computation of a few numerical quantities that provide information about the centre of the data and indicate the spread of the observations.

Inferential statistics deal with procedures for making inferences about the characteristics that describe the large group of data as a whole, called the *population*, from the knowledge derived from only a part of the data, known as *sample*. This area includes the estimation of population parameters and testing of statistical hypotheses. This process of statistics is based on probability theory as the inferences which are made on the basis of sample evidence, cannot be absolutely certain.

Comparison

Descriptive Statistics

- i) A cricket player wants to find his score average for the last 20 games.
- ii) Aamir wants to describe the variation in his four test scores in Statistics.
- iii) Mrs. Rashid wants to determine the average weekly amount she spent on groceries in the past 6 months.

Inferential Statistics

- i) A cricket player wants to estimate his chance of scoring based on his current season average.
- ii) Based on the first four test scores, Aamir would like to predict the variation in his final Statistics test scores.
- iii) Based on last six months grocery bills, Mrs. Rashid would like to predict the average amount she will spend on groceries for the upcoming year.

1.1.4 Populations and Samples. A population or a statistical population is a collection or set of possible observations whether finite or infinite, relevant to some characteristic of interest. A statistical population may be real such as the heights of all college students or hypothetical such as all the possible outcomes from the toss of a coin. The number of observations in a finite population is called the size of the population and is denoted by the letter N . Numerical quantities describing a population are called parameters, customarily represented by Greek letters. It is important to note that in statistics the word population is a technical term not necessarily referring to all the people in a specified area, rather denoting the aggregate of measurements or counts of some characteristic for the entire group of objects or individuals.

A sample is a part or a subset of a population. Generally it consists of some of the observations in certain situations, it may include the whole of the population. The number of observations included in a sample is called the size of the sample and is denoted by the letter n . A numerical quantity computed from a sample, is called a statistic, which is usually represented by ordinary Latin letter. The information derived from sample data is used to draw conclusions about the population.

Example 1.5 State whether each of the following is a population or a sample.

- Total number of absentees by all students in a college during the last month.
- Number of colour TV sets owned by all families in Lahore.
- Monthly salaries of all employees of a company.
- What yield per acre for 5 pieces of land.
- Number of computers sold during the last month at all the computer stores in Lahore.

Solution

- Population
- Population
- Population
- Sample
- Population

1.1.5 Importance of Statistics. Statistics is perhaps a subject that is used by everybody. The following functions and uses of statistics in most diverse fields serve to indicate its importance.

- Statistics assists in summarizing the large sets of data in a form that is easily understandable.
- Statistics assists in the efficient design of laboratory and field experiments as well as surveys.
- Statistics assists in a sound and effective planning in any field of inquiry.
- Statistics assists in drawing general conclusions and in making predictions of how much of something will happen under given conditions.
- Statistical techniques being powerful tools for analysing numerical data, are used in almost every branch of learning. In the biological and physical sciences, Genetics, Agronomy, Anthropometry, Astronomy, Physics, Geology, etc. are the main areas where statistical techniques have been developed and are increasingly used.

- v) A businessman, an industrialist and a research worker all employ statistical methods in their work. Banks, Insurance companies and Governments all have their statistics departments.
- vi) A modern administrator whether in public or private sector, leans on statistical data to provide a factual basis for decision.
- vii) A politician uses statistics advantageously to lend support and credence to his arguments while elucidating the problems he handles.
- viii) A social scientist uses statistical methods in various areas of socio-economic life of a nation. It is sometimes said that "*a social scientist without an adequate understanding of statistics, is often like the blind man groping in a dark room for a black cat that is not there*".

1.2 OBSERVATIONS AND VARIABLES

In statistics, an **observation** often means any sort of numerically recording of information, whether it is a physical measurement such as height or weight; a classification such as heads or tails, or an answer to a question such as yes or no.

1.2.1 Variables. A characteristic that varies with an individual or object, is called a **variable**. For example, age is a variable as it varies from person to person. A variable can assume a number of values. The given set of all possible values from which the variable can take on a value, is called its **domain**. For a given problem, the domain of a variable contains only one value, then the variable is referred to as **constant**.

Variables may be classified into quantitative and qualitative according to the form of the characteristic of interest. A variable is called a **numerical variable** when a characteristic can be measured numerically such as age, weight, income or number of children. On the other hand, if the characteristic is non-numerical such as education, gender, eye-colour, quality, intelligence, poverty, pollution, etc. the variable is referred to as a **qualitative variable**. A qualitative characteristic is also called an **attribute**. An individual or object with such a characteristic can be counted or enumerated after having been assigned to one of the several mutually exclusive classes or categories.

1.2.2 Discrete and Continuous Variables. A quantitative variable may be classified as **discrete** or **continuous**. A **discrete variable** is one that can take only a discrete set of integers or whole numbers, that are values are taken by jumps or breaks. A discrete variable represents count data such as the number of children in a family, the number of rooms in a house, the number of deaths in an accident, the income of an individual, etc.

A variable is called a **continuous variable** if it can take on any value—fractional or integer—within a given interval, i.e. the domain is an interval with all possible values without gaps. A continuous variable represents measurement data such as the age of a person, the height of a plant, the weight of a toy, the temperature at a place, etc.

A variable whether countable or measurable, is generally denoted by some symbol such as X or Y , and X_i or Y_j represents the i^{th} or j^{th} value of the variable. The subscript i or j is replaced by a symbol such as $1, 2, 3, \dots$ when referred to a particular value.

Example 1.6 Identify each of the following as examples of (1) Attribute, (2) Discrete, or (3) Continuous variables.

- the hair colour of children;
- length of time required for a wound to heal;
- the number of telephone calls arriving at a switch board per 1 hour period;
- the breaking strength of a given type of a string;
- the number of questions answered correctly on a test;
- the number of stop signs in the city of Lahore;
- the colour of your eye;
- the number of children in your family.

Solution

- Attribute;
- Continuous;
- Discrete;
- Continuous;
- Discrete;
- Discrete;
- Attribute;
- Discrete

1.2.3 Measurement Scales. By measurement, we usually mean the assigning of numbers to observations or objects and scaling is a process of measuring. The four scales of measurements are briefly mentioned below:

Nominal Scale. The classification or grouping of the observations into mutually exclusive qualitative categories or classes is said to constitute a nominal scale. For example, students are classified as male and female. Numbers 1 and 2 may also be used to identify these two categories. Similarly, rainfall may be classified as heavy, moderate and light. We may use number 1, 2 and 3 to denote the three classes of rainfall. The numbers when they are used only to identify the categories of the given scale, carry no numerical significance and there is no particular order for the grouping.

Ordinal or Ranking Scale. It includes the characteristic of a nominal scale and in addition has the property of ordering or ranking of measurements. For example, the performance of students (or players) is rated as excellent, good, fair or poor, etc. Number 1, 2, 3, 4, etc. are also used to indicate ranks. The only relation that holds between any pair of categories is that of "greater than" (or more preferred).

Interval Scale. A measurement scale possessing a constant interval size (distance) but not a true zero point, is called an interval scale. Temperature measured on either the Celsius or the Fahrenheit scale is an outstanding example of interval scale because the same difference exists between 20°C (68°F) and 30°C (86°F) as between 5°C (41°F) and 15°C (59°F). It cannot be said that a temperature of 40 degrees is twice as hot as a temperature of 20 degree, i.e. the ratio 40/20 has no meaning. The arithmetic operations of addition, subtraction, etc. are meaningful.

Ratio Scale. It is a special kind of an interval scale where the scale of measurement has a true zero point at its origin. The ratio scale is used to measure weight, volume, length, distance, money, etc. The key to differentiating interval and ratio scale is that the zero point is meaningful for ratio scale.

Example of Measurement Scales

Nominal-level data	Ordinal-level data	Interval-level data	Ratio-level
Gender (Male, Female)	Grades (A, B, C, D, F)	Temperature	Age
Eye colour	Position (1 st , 2 nd , 3 rd etc.)	IQ score	Weight
Religion	Ranking of cricket player	SAT score	Height
Speculation	Rating (poor, good, excellent)		Time
Nationality	Socio-economic status (poor, middle class, rich)		Salary
			Distance

1.2.4 Errors of Measurement. Experience has shown that a continuous variable can never be measured with perfect fineness because of certain habits and practices, methods of measurements, instruments used, etc. The measurements are thus always recorded correct to the nearest units and hence are of limited accuracy. The actual or true values are, however, assumed to exist. For example, if a student's weight is recorded as 60 kg (correct to the nearest kg), his true weight in fact lies between 59.5 kg and 60.5 kg, whereas a weight recorded as 60.00 means the true weight is known to lie between 59.995 and 60.005 kg. Thus there is a difference, however small it may be, between the measured value and the true value. This sort of deviation from the true value is technically known as the **error of measurement**. In other words, if the observed value and the true value of a variable are denoted by x and $x + \varepsilon$ respectively, then the difference $(x + \varepsilon) - x$, i.e. ε is the error. This error involves the unit of measurement of x and is therefore called an **absolute error**. An absolute error divided by the true value is called the **relative error**. Then the relative error = $\frac{\varepsilon}{x + \varepsilon}$, which when multiplied by 100, is

percentage error. These errors are independent of the units of measurement of x . It ought to be noted that an error has both magnitude and direction and that the word error in statistics does not mean mistake which is a chance inaccuracy.

An error is said to be **biased** when the observed value is consistently and constantly higher or lower than the true value. Biased errors arise from the personal limitations of the observer, the imperfection in the instruments used or some other conditions which control the measurements. These errors are not avoided by repeating the measurements. They are cumulative in nature, that is, the greater the number of measurements, the greater would be the magnitude of error. They are thus more troublesome. These errors are also called **cumulative or systematic errors**.

An error, on the other hand, is said to be **unbiased** when the deviations, i.e. the excesses and deficits, from the true value tend to occur equally often. Unbiased errors are revealed when measurements are repeated and they tend to cancel out in the long run. These errors are therefore **compensating** and are also known as **random errors** or **accidental errors**.

A measurement free from all classes of errors is considered as an accurate measurement. This is why efforts are made to reduce the magnitude of errors to a minimum so that the level of accuracy at

which the measurements are recorded, is increased. To achieve this end, a clear understanding of the meaning of significant digits and the process of rounding off the numbers is very important in statistical computations.

1.2.5 Significant Digits. Accuracy in measurement is related to significant digits. The significant digits in a number, are those that represent accurate and meaningful information. For instance, the number 25 representing a continuous variable has two significant digits. In recorded measurements, all digits except zeros are always significant. For jets, we may state as:

- i) Zeros are significant if they follow a decimal point and conclude a number, e.g. the measurement 2.500 has four significant digits.
- ii) Zeros are non-significant when they follow a decimal point but commence a number, e.g. the measurements .04 and .000237 contain only 1 and 3 significant digits respectively.
- iii) Zeros may or may not be significant when they lie entirely to the left of the decimal point where they may not represent measurement but may be used to simply locate the decimal point. In such a case, a definite specification such as standard notation, becomes necessary. When any number is expressed as a product of a power of 10 and a number between 1 and 10, it is said to be written in standard notation. For example, the number 75400 can have 3 significant digits when written in standard notation as 7.54×10^4 . It can also have 5 significant digits if written as 7.5400×10^4 .
- iv) Zeros are always significant when occur within series of significant digits, e.g. the numbers 20.3, .1001, 4.00507, etc., have 3, 4 and 6 significant digits respectively.

It should be remembered that

- a) significant digits in a number are not disturbed by the location of the decimal point, e.g. the measurements recorded as 269, .269, 269 or .000269 have only 2 significant digits.
- b) in case of discrete data which are generated by the process of counting, the number of significant digits is considered indefinite because the level of accuracy cannot be improved, e.g. the number 15700 has indefinite significant digits.
- c) the rules regarding the determination of the number of significant digits, are applicable to continuous variables.
- d) in the operations of addition and subtraction, all digit positions which are not significant in any of the values being added or subtracted, are not significant in the total or difference.
- e) in the operations of multiplication and division, the number of significant digits in the result is determined by the value with the smallest number of significant digits that enters into the calculations.

1.2.6 Rounding off a Number. The process of rounding off is simply rounding a number down so that a certain number of digits counted from the left, are to be retained and the last few digits are to be (i) dropped in a decimal number or (ii) replaced with zeros in a whole number. The rules generally used for rounding decimal numbers are as follows:

- i) The last significant digit is increased by 1, if the first digit of the remainder to be dropped is more than 5 or 5 followed by digits not all of which are zero, e.g. the numbers 2.145001 and 5.3772 are rounded off to three significant digits as 2.15 and 5.38 respectively.

- ii) The last significant digit remains unaltered, if the first digit of the remainder to be dropped is 4 or less, e.g. the numbers 2.1548 and 7.3627 are rounded off to three significant digits as 2.15 and 7.36 respectively.
- iii) When the digit to be dropped is exactly 5, the accepted practice is to increase the last significant digit by 1, if it is odd and to leave unaltered if it is even e.g. the number 4.535 and 2.745 are rounded off to three significant digits as 4.54 and 2.74 respectively.

For rounding whole numbers, we can change the word "the first digit to be dropped" to "the first digit to be replaced by zero" in the rules stated above.

The point to be made here is the rules for identifying significant digits and the process of rounding the numbers should be applied to final calculations and not to the intermediate results.

1.3 COLLECTION OF DATA

The most important part of statistical work is perhaps the collection of data. Statistical data are collected either by a complete enumeration of the whole field, called *census*, which in many cases would be too costly and too time consuming as it requires large number of enumerators and supervisory staff, or by a *partial* enumeration associated with a sample which saves much time and money. The sampling methods explained at length in later chapters, are increasingly employed both in official and in private inquiries to collect data.

When data are classified according to source, it is customary to make the following distinction.

Data that have been originally collected (raw data) and have not undergone any sort of statistical treatment, are called *Primary data*, while data that have undergone any sort of treatment by statistical methods at least once, i.e. the data have been collected, classified, tabulated or presented in some form for a certain purpose, are called *Secondary data*.

The survey research includes the following important steps:

- a) to define the objectives of the survey.
- b) to define the variable(s) and the population of interest.
- c) to define the data collection and data measuring schemes.
- d) to determine the appropriate descriptive and inferential data analysis techniques.

A brief description of the methods generally adopted either on census basis or on sample basis for collecting data, is given below.

1.3.1 Collection of Primary Data. One or more of the following methods are employed to collect primary data:

Survey Research

- i) **Direct Personal Investigation.** In this method, an investigator collects the information personally from the individual concerned. Since he interviews the informant himself, the information collected is generally considered quite accurate and complete. This method may prove very costly and time consuming when the area to be covered is vast. However it is useful for laboratory experiments or localized inquiries. Errors are likely to enter the results due to personal bias of the investigator.

- iii) **Indirect Investigation or Personal Interviews.** Sometimes the direct sources do not exist or the informants hesitate to respond for some reasons or other. In such a case, third parties or witnesses having information are interviewed. As some of the informants are likely to deliberately give wrong information, so the reliance is not placed on the evidence of one witness only. Moreover, due allowance is to be made for the personal bias. This method is useful when the information desired is complex or there is reluctance or indifference on the part of the informants. It can be adopted for extensive inquiries.
- iv) **Collection through Questionnaires.** A questionnaire is an inquiry form comprising of a number of pertinent questions with space for entering information asked. The questionnaires are usually sent by mail and the informants are requested to return the questionnaires to the investigator, after doing the needful within a certain period. This method is cheap, fairly expeditious and good for extensive inquiries. But the difficulty is that the majority of respondents (persons who are required to answer the questions) does not care to fill the questionnaires in and to return them to the investigator. Sometimes, the questionnaires are returned incomplete and full of errors. In spite of these drawbacks, the method is considered as the standard method for routine business and administrative inquiries. The answers to the questionnaires are very often recorded by trained enumerators to overcome the difficulties these days. It is important to note that the questions should be few, brief, very simple, easy for all respondents to answer, clearly worded and not offensive to certain respondents.
- v) **Collection through Enumerators.** Under this method, the information is gathered by employing trained enumerators who assist the informants in making the entries in the schedules or questionnaires correctly. This method gives the most reliable information if the enumerator is well trained, experienced and tactful. It is considered the best method when a large scale governmental inquiry has to be conducted. This method cannot be adopted by private individual or institution as the cost would be prohibitive to them.
- vi) **Collection through Local Services.** In this method, there is no formal collection of data but the agents or local correspondents are directed to collect and to send the required information, using their own judgment as to the best way of obtaining it. This method is cheap and expeditious, but gives only the estimates.
- vii) **Computer Interviews.** Respondents enter data directly into a computer in response to questions presented on the monitor.

Experimental Research

- i) **Laboratory experiments.** Manipulation of the independent variable(s) in an artificial situation. Basic designs consider the impact of only one independent variable.
- ii) **Field experiments.** Manipulation of the independent variable(s) in a natural situation.

1.3.2 Collection of Secondary Data. The secondary data may be obtained from the following sources:

Secondary Research

Internal Secondary Data. Data generated within the organization itself, such as salesperson reports, sales invoices, accounting records.

External Secondary Data

- i) Official, e.g. the publications of the Statistical Division, Ministry of Finance, the Federal and Provincial Bureaus of Statistics, Ministries of Food, Agriculture, Industry, Labour, etc.
- ii) Semi-Official, e.g. State Bank of Pakistan, Railway Board, Central Cotton Committee, Boards of Economic Inquiry, District Councils, Municipalities, etc.
- iii) Publications of Trade Association, Chambers of Commerce, etc.
- iv) Technical and Trade journals and newspapers.
- v) Research organisation such as universities, and other institutions.

13.3 Editing of Data. The primary data should be intensively checked at an early stage in order to locate incomplete or inconsistent entries. If possible, the incomplete and defective questionnaires should be returned to the respondents for amendments. In order to accept the secondary data as authoritative, one should critically examine the reliability of the compiler and the suitability of the data. The scope and object of the inquiry, sources of information and the degree of accuracy should also be carefully scrutinized.

13.4 Uses and Misuses of Statistics. Statistics has numerous uses. It is difficult to find a field in which Statistics is not used. Statistics plays integral part in many disciplines, viz: Economics, Health, Planning, Astronomy, Management, Business, Psychology, Agriculture, Sociology, Education etc.

A few examples of how and where Statistics is used:

- i) In experimental science, the experiments generate data, it must be collected and analyzed.
- ii) In Government, many types of statistical data are collected all the time. This data can be used for various types of planning and also inform the general public.
- iii) In education, Statistics are used to describe the results and standards of education.

Statistical techniques are many times misused: to sell products that don't work, to prove something that is not really true, to get the attention of public by evoking fear and shock etc. There are two sayings about Statistics which explains the 'uses' of Statistics.

- a) Statistics can prove anything.
- b) "There are three types of lies—Lies, damned lies, and Statistics."

Statistics can also be misused in many ways such as using Non Representative Samples, Small Sample Size, Ambiguous Averages and dispersions, Detached facts, Implied Connections, Wrong and Misleading Graphs, Wrong use of Statistical techniques, Serious violation of assumptions behind the statistical techniques and Faculty Surveys etc.

EXERCISES

OBJECTIVE

Answer 'True' and 'False'. If the statement is not true then replace the underlined words with words that make the statement true:

- i) All numerical data are not Statistics.
- ii) A Statistic is a summary measure computed for a population.
- iii) A Parameter is a summary measure computed for a sample.
- iv) Descriptive Statistics are used to make projections or estimates about the population.
- v) Inferential Statistics is the study and description of data.
- vi) A sample is typically a very large collection of individuals or objects of our interest.
- vii) The thickness of the glass is an example of attribute data.
- viii) The number of students in a class is an example of continuous data.
- ix) The make of a car is an example of discrete data.
- x) The main objective of Statistics is to collect a sample, analyse it and make inferences about the unknown characteristics of the population from which the sample has been drawn.

SUBJECTIVE

- 1.1 Explain what is meant by Statistics. Give the important uses and limitations of statistics.
- 1.2 Define Statistics. Discuss, giving examples, the importance of the study of statistics and show how it can help the extension of scientific knowledge. (P.U., B.A. (Hons.), 1960)
- 1.3 a) What do you understand by the term Statistics? Give its chief characteristics.
b) Give a brief account of the importance of statistics in different fields. (P.U., B.A./B.Sc. 1971)
- 1.4 Comment upon the following statement of Sir Ronald Fisher: "Statistics may be regarded as (i) the study of populations, (ii) the study of variation and (iii) the study of the methods of reduction of data". (P.U., B.A. (Hons.), Part-I, 1964)
- 1.5 Comment on the statement given below:
"Statistics is concerned with understanding the real world through the information that we derive from classification and measurement. Its distinctive characteristic is that it deals with variability and uncertainty which is everywhere".
- 1.6 a) Define Statistics and explain its characteristics.
b) What are the uses of Statistics? (P.U., B.A. (Hons. in Econ.), 1960)
- 1.7 Explain the difference between the following?
 - a) Statistics and Statistic.
 - b) Population and Sample.
 - c) Descriptive and Inferential Statistics.
 - d) Quantitative and Qualitative Variables.

- e) Discrete and Continuous Variables.
- f) Biased and Unbiased errors.
- g) Primary data and Secondary data.
- h) Nominal and ordinal scale, Interval and ratio scale.

1.8 What is a statistical error? In what way does it differ from a mistake? Explain the difference between absolute and relative errors.

1.9 a) Define a Variable. Differentiate between a discrete and a continuous variable.

- b) Classify the following variables as discrete or continuous:

- i) The number of students attending a class.
- ii) The amount of milk produced by a cow.
- iii) The number of heads of the toss of 6 coins.
- iv) The yearly income of a College Professor.
- v) The age of a shopkeeper.
- vi) The weight of a college student.
- vii) The number of petals on a flower.
- viii) The life times of television tubes produced by a company.
- ix) Temperature recorded every half hour at a weather bureau.
- x) The number of shares sold on a day in the stock market.

1.10 Classify each variable as qualitative or quantitative

- i) Colour of eyes.
- ii) Number of computers sold in the last month.
- iii) Marital status of faculty members.
- iv) Student's weight.
- v) Lifetime of car batteries.
- vi) Number of burgers sold by a fast food shop.
- vii) Brand of cars.

1.11 Classify each as nominal-level, ordinal-level, interval-level or ratio-level measurement

- i) Weights of cars.
- ii) Rankings of squash players.
- iii) Temperature of the city.
- iv) Salaries of the top five executives in bank.
- v) Marital status.

- vi) Ages of students.
vii) Ratings of five players (Poor, Fair, Good, Excellent)
viii) IQ of student
ix) Rating of movies
x) Weights of suitcase on plane.
- 1.12 Round off the following continuous data to four significant digits each.
(i) 32.21705, (ii) 937.05002, (iii) 0.003599499,
(iv) 1.003599499, (v) 0.07000455, (vi) 22.2500001.
- 1.13 a) Distinguish between Primary and Secondary data, giving examples of each.
b) Describe the methods which can be used in the collection of statistical data, stating the advantages and disadvantages of each method.
c) Enumerate the main sources of errors in Statistics and give their effects.
(P.U., B.A./B.Sc., 1982-83)
- 1.14 What methods would you employ in the collection of statistical data when the field of inquiry is (i) small, (ii) fairly large and (iii) very large, if you are to pay due regard to accuracy, labour, cost and time?
- 1.15 What are the different methods employed in the collection of data for statistical enquiries? In what type of inquiry should each one of them be used?
(P.U., B.A./B.Sc., 1961)
- 1.16 Select a newspaper or a magazine article that involves a statistical study and answer the following questions.
a) Is this study descriptive or inferential? Explain your answer.
b) What variables are used in the study?
c) What level of measurements was used to obtain data?
d) Is population defined in the article? If not, how could it be defined?
e) How the data might have been collected?
f) Do you agree with the conclusions given in the article?

CHAPTER 2

PRESENTATION OF DATA

1.1 INTRODUCTION

The device of gathering data often results in a massive volume of statistical data, which are in the form of individual measurements or counts. It is difficult to learn anything by examining the unorganised mass which is more often confusing than clarifying. The mass of data is therefore to be organised and condensed into a form that can be more rapidly and easily understood and interpreted. For this purpose, techniques of classification, tabulation and graphic displays are presented in this chapter.

1.2 CLASSIFICATION

The term *classification* as defined as the process of dividing a set of observations or objects into classes or groups in such a way that (i) observations or objects in the same class or group are similar, (ii) observations or objects in each class or group are dissimilar to observations or objects in other class or groups.

Classification is thus the sorting of data into homogeneous classes or groups according to their alike or not. When the data are sorted according to one criterion only, it is called a *simple classification* or a *one-way classification*. Classification is called a *two-way classification* when the data are sorted according to two criteria. A *multiple classification* or *cross-classification* is made according to three criteria.

Data may also be classified according to qualitative, temporal and geographical characteristic element of data according to the values of a variable characteristic, called a *distribution*. When the variable is expressed in terms of location, we get a spatial or geographical distribution. Such arrangement of values is referred to as a *map series*.

1.2.1 Aims of Classification. The main aims of classification are

- to reduce the large sets of data to an easily understand summary;
- to display the points of similarity and dissimilarity;
- to save mental strain by eliminating unnecessary details;
- to reflect the important aspects of the data; and
- to prepare the ground for comparison and inference.

1.2.2 Basic Principles of Classification. While classifying large sets of data, the following points be taken into consideration.

The classes or categories into which the data are to be divided, should be mutually exclusive and no overlap should exist between successive classes. In other words, classes should be arranged so that each observation or object can be placed in one and only one class.

The classes or categories should be all inclusive. All inclusive classes are classes that include all the data.

As far as possible, the conventional classification procedure should be adopted.

The classification procedure should not be so elaborate as to lead to trivial classes nor it should be so crude as to concentrate all the data in one or two classes.

1.3 TABULATION

By tabulation, we mean a systematic presentation of data classified under suitable heads and sub-heads, and placed in columns and rows. This sort of logical arrangement makes the data easy to

understand, facilitates comparisons and provides an effective way to convey information to a reader. British statistician, Professor Bowley (1869-1957), refers to tabulation as "the intermediate stage between the accumulation of data, in whatever form they are obtained, and the final reasoned account of the results shown by the statistics."

2.3.1 Types of Tables. Statistical tables classified according to purpose, are of two types: General purpose (primary) tables and Specific purpose (derived or test) tables. The general purpose tables are large in size, are extensive with vast coverage and are constructed for reference purposes. The specific purpose tables are simpler in structure and deal with one or two criteria of classification only. Such tables are used to analyse or to assist in analysing data.

When the classification corresponds to one, two or many criteria or characteristics, the tabulation is called a single, double or manifold tabulation respectively. Tabulation of a dependent variable (e.g. number of students) against the independent variable (say, weight) provides an example of a single tabulation. Tables with two criteria of classification, e.g. gender and marital status or height and weight etc., are examples of double tabulation. An example of manifold tabulation is the presentation of a population of a country by age, by gender, by residence, by literacy, by livelihood classes, etc.

The main parts of a statistical table are the title, the boxhead, the stub, the body, one or more prefatory notes, footnotes and a source, etc. They are described in the next section.

2.3.2 Main Parts of a Table and its Construction. The main parts of a table and the general rules to be observed in constructing any table are described below:

a) **Title.** A table must have a self-explanatory title which should usually tell us the "what, who, how classified and when" of the data, in that order. Some other important points are stated below:

- i) Titles should be brief in the form of clauses. Complete sentences are unnecessary.
- ii) Abbreviations should not be used.
- iii) Main titles should be in caps throughout. Sub-titles, if any, should be in lower case letters with major words capitalized and should indicate clearly what the table describes.
- iv) The different parts of a title should be separated by commas but no full-stop at the end.
- v) Words in titles should not be hyphenated except when really necessary.
- vi) If a title needs to use the use of two or more lines, an inverted pyramid arrangement of lines should be used.

b) **Column Captions and Boxhead.** The heading of each column is called a Column Caption while the section of a table that contains the column captions, is referred to as Boxhead. Points to note here are given below:

- i) The heading should be clear but concise.
- ii) They should be arranged in such a way that the most important characteristic is placed in the first column. The column of totals is usually placed at extreme right, but some people prefer the totals on the left.
- iii) Only the first word in each column caption should be capitalized. No full-stop should be put at the end.
- iv) Abbreviations, when clear, may be used.
- v) Main caption should be centred over the column it is to span.
- vi) Extra lines should be used to avoid crowding in caption box.

- (i) Whenever possible, caption width should be made roughly proportional to the size of numbers to be inserted.
- (ii) **Row Captions and Stub.** The heading or title for a row, it is called the *Row Caption* and the containing the row captions is known as *Stub*. The necessary points in this respect are given below:
 - (i) The principles for column captions apply to row captions in stub.
 - (ii) If the stub is long and has several levels of classification, the major classification should be capitalized to separate the table into parts.
 - (iii) Whenever the figures have more than four or five significant digits, the digits should be grouped in threes or four. For example, one should write 23.178.327, not 23178327.
 - (iv) In long tables, some space should be left after every five or ten rows.
 - (v) Totals should usually be placed at the bottom, but some prefer to place them at the top.
 - (vi) Items in the stub should be arranged so as to facilitate easy reading.
 - (vii) Every stub should have an appropriate heading describing its contents. This heading should be centred in the upper left box of the table.

(d) **Prefatory Notes and Footnotes.** Explanatory notes incorporated in the table beneath the title and below the body, are called prefatory notes and footnotes respectively.

Prefatory notes give additional specifications of the data, i.e., nature of items included or excluded from the table, statements of the box, etc. They are placed between the title and the head-line. The heading should be in lower case alphabet. Footnotes are used to clarify anything in the table by giving a brief description, by drawing attention to incompleteness or by stating any special circumstances among the data. The footnotes should be specific in nature. They are placed immediately below the head-line of the table, above the source. Footnote symbols should be placed as follows:

- (i) If they refer to an entire column or set of columns, place them at the end of the appropriate caption.
- (ii) If they refer to an entire row or a set of rows, place them at the end of the appropriate stub title.
- (iii) If they refer to a single cell in the table, place beside the cell entry in the body of the table.

The footnotes should be indicated either by lower case alphabet enclosed in parentheses or by symbols *, †, ‡, etc.; never by a number.

(e) **Source Notes.** Every table should have a source note, unless the table is an original tabulation and its source is clear from the context. It is placed immediately below the table and below the footnotes. The source notes must include the compiling agency, publication, date of publication and page as they are used as a means of verification and reference.

(f) **Body and Arrangement of Data.** The body of a table is the most important part, which contains the entire data arranged in columns and rows. A rough-sketch enables us to have an idea about the number of columns and rows required.

Arrangement of the data is made by taking into consideration the *basis of classification* and the *nature of the table*. Thus the data may be arranged either (i) according to the alphabetical order or (ii) according to the time of occurrence or (iii) according to location or (iv) according to magnitude or

importance, or (iv) by a customary classification, e.g. classifying as men, women and children, etc. Whatever arrangements are used, the table should be neat, simple and attractive to the eye.

g) **Spacing and Rulings.** A proper and judicious use of spacing and ruling enhances the effectiveness of a table and helps in separating or emphasizing certain items in it. Thick or double lines (rulings) are used for emphasis and for separating the title, the boxhead, the stub, etc., while parts under captions and related columns are separated by thin or single lines.

h) **General.** There are some other considerations too, that are enumerated below:

- A table should be simple. A complex table if possible, may be broken into relatively simple tables.
- Units of measurements and nature of the data should be specified in title, captions, etc. in parentheses.
- Percentages should be clearly indicated as 'per cent of total' etc. and their total should be shown as 100.0.
- If the figures entered in the table are rounded off, this should be indicated in the preface note or in the stub or caption.
- Zeros need not be entered.
- Minus signs are a part of the table and precede the numbers.
- The relationship of the parts to the whole should be shown by thin or heavy rulings.
- The item or items to be emphasized should be placed in the most prominent position of the table.

The general sketch of a table is given below:

TITLE		COLUMN CAPTIONS				
		Prefatory notes				
Boxhead		COLUMNS				
	→		Units	B	O	D
↑				Y		
STUB	↓					

Footnotes.....

Source notes.....

Example 2.1 A district is divided into two areas, viz. Urban area and Rural area. Total population of the district is 271,076 out of which only 46,740 live in the urban area. Total male population

district is 139,699 and that of urban area is 23,083. Total unmarried population of the district is 112,352 out of which 36,864 are rural females. In the urban area, unmarried people number 21,072 out of which 12,149 are males.

Prepare a table showing the population of the district by marital status, by residence and by gender.

A rough table which will probably need amending later, might look as follows:

AREA	BOTH GENDERS		MALE		FEMALE	
	Married	Unmarried	Married	Unmarried	Married	Unmarried
Urban						
Rural						

We first compute the relevant figures as below:

$$\text{Rural population} = \text{Total population} - \text{Urban population}$$

$$= 271,876 - 46,740 = 224,336$$

$$\text{Female population} = \text{Total population} - \text{Male population}$$

$$= 271,876 - 139,699 = 131,377$$

$$\text{Rural male population} = \text{District male population} - \text{Urban males}$$

$$= 139,699 - 23,083 = 116,616$$

$$\text{Similarly, Urban females} = 46,740 - 23,083 = 23,657$$

$$\text{Rural females} = 224,336 - 116,616 = 107,720$$

$$\text{Married population of District} = 271,876 - 112,352 = 158,724$$

$$\text{Rural unmarried population} = 116,616 - 21,072 = 91,280$$

$$\text{Rural unmarried males} = 116,616 - 36,864 = 79,752$$

$$\text{Urban unmarried females} = 21,072 - 12,149 = 8,923 \text{ etc.}$$

Having computed all these figures, they are presented in the final table that appears below:

Table

POPULATION OF DISTRICT "A" BY GENDER, MARITAL STATUS AND RESIDENCE

Secthead Category	Area	Both Genders			Male			Female		
		Total	Married	Un- marred	Total	Married	Un- marred	Total	Married	Un- marred
District	District	271,876	158,724	112,352	139,699	73,134	66,565	131,377	83,290	48,787
Sub-	Urban	46,740	25,668	21,072	23,083	10,934	12,149	23,657	14,734	8,923
	Rural	224,336	133,056	91,280	116,616	54,200	54,416	107,720	70,856	36,864

Source

2.4 FREQUENCY DISTRIBUTION

The organization of a set of data in a table showing the distribution of the data into classes or groups together with the number of observations in each class or group is called a *Frequency Distribution*. The number of observations falling in a particular class is referred to as the *class frequency* or simply *frequency* and is denoted by f . Data presented in the form of a frequency distribution are also called *grouped data* while the data in the original form are referred to as *ungrouped data*. The data are said to be arranged in an *array* when arranged in ascending or descending order of magnitude. The purpose of a frequency distribution is to produce a meaningful pattern for the overall distribution of the data from which conclusions can be drawn. A fairly common frequency pattern is the rising to a peak and then declining. In terms of its construction, each class or group has lower and upper limits, lower and upper boundaries, an interval and a midfix value.

2.4.1 Class-limits. The class-limits are defined as the numbers or the values of the variables which describe the classes; the smaller number is the lower class-limit and the larger number is the upper class-limit. Class-limits should be well defined and there should be no overlapping. In other words, the limits should be exclusive, i.e. the values corresponding exactly to the lower limit or the upper limit be included in that class. The class-limits are therefore selected in such a way that they have the same number of significant places as the recorded values. Suppose the data are recorded to the nearest integers. Then an appropriate method for defining the class limits without overlaps, for example, may be 10 – 14, 15 – 19, 20 – 24, etc. The class limits may be defined as 10.0 – 14.0, 15.0 – 19.9, 20.0 – 24.9, etc., when the data are recorded to nearest tenth of an integer. Sometimes, a class has either no lower class limit or no upper class-limit. Such a class is called an *open-end class*. The open-end classes, if possible, should be avoided as they are a hindrance in performing certain calculations. A class indicated as 10 – 15 will include 10 but not 15, i.e. $10 \leq X < 15$.

2.4.2 Class-boundaries: The class-boundaries are the precise numbers which separate one class from another. The selection of these numbers solves the difficulty, if any, in knowing the class in which a particular value should be assigned. A class boundary is located midway between the upper limit of a class and the lower limit of the next higher class, e.g. 9.5 – 14.5, 14.5 – 19.5, 19.5 – 24.5, or 9.95 – 14.95, 14.95 – 19.95, etc. The class-boundaries are thus always defined more precisely than the level of measurements being used so that the possibility of any observation falling exactly on the boundary is avoided. That is why the class-boundaries carry one more decimal place than the class limits or the observed values. The upper class-boundary of a class coincides with the lower boundary of the next class.

2.4.3 Class Mark: A class mark, also called class midpoint, is that number which divides each class into two parts. In practice, it is obtained by dividing either the sum of the lower and upper limits of a class, or the sum of the lower and upper boundaries of the class by 2 but in a few cases, it does not hold, particularly in modern practice of age grouping. For purposes of calculations, the frequency in a particular class is assumed to have the same value as the class-mark or midpoint. This assumption may introduce an error, called the grouping error, but statistical experience has shown that such errors usually tend to counterbalance over the entire distribution. The grouping errors may also be minimized by selecting a class (group) in such a way that its midpoint corresponds to the mean of the observed values falling in that class.

2.4.4 Class Width or Interval: The class-width or *interval* of a class is equal to the difference between the class boundaries. It may also be obtained by finding the difference either between two successive lower class-limits, or between two successive class marks. The lower limit of a class should not be subtracted from its upper limit to get the class interval. An equal-class interval, usually denoted by h or c , facilitates the calculations of statistical constants such as the mean, the standard deviation,

ments, etc. That is why, in practice, it is desirable to have equal class-intervals. But in some types of economic and medical data, it is wise to use unequal class-intervals on account of greater concentration of measurements in certain classes. Such class intervals usually become uniform when logarithms of class marks are taken. It should be noted that some people use the terms "class" and "class-interval" interchangeably and the width of the class is referred to as the size or length of the class-interval.

2.4.5 Constructing a Grouped Frequency Distribution. The following are some basic rules that should be kept in mind when constructing a grouped frequency distribution:

- i) **Decide on the number of classes into which the data are to be grouped.** There are no hard and fast rules for deciding on the number of classes which actually depends on the size of data. Statistical experience tells us that no less than 5 and no more than 20 classes are generally used. Use of too many classes will defeat the purpose of condensation and too few will result in too much loss of information. H.A. Sturges has proposed an empirical rule for determining the number of classes into which a set of observations should be grouped. The rule is

$$k = 1 + 3.3 \log N$$

where k denotes the number of classes and N is the total number of observations. For example, if there are 100 observations, then by applying Sturges' rule, we should have:

$$k = 1 + 3.3 (2.0000) = 7.6, i.e., 8 \text{ classes}$$

Thus eight classes are required but this rule is rarely followed in practice.

- ii) **Determine the range of variation in the data.** This is the difference between the largest and the smallest values in the data.
- iii) **Divide the range of variation by the number of classes to determine the approximate width or size of the equal class-interval.** In case of fractional results, the next higher whole number is usually taken as the size or width of class-interval. If equal class-intervals are inconvenient or may be undesirable, then classes of unequal size are used. But in practice, intervals that are multiples of 5 or 10, are commonly used as people can understand them more readily.
- iv) **Decide where to locate the class-limit of the lowest class and then the lower class boundary.** The lowest class usually starts with the smallest data value or a number less than it. It is better if it is a multiple of class-interval. Find the upper class boundary by adding the width of the class-interval to the lower class-boundary and write down the upper class limits too. The open-end classes, i.e., classes with the lowermost or uppermost class boundary unknown, should be avoided if possible.
- v) **Determine the remaining class-limits and class boundaries by adding the class-interval repeatedly.** The lowest class should be placed at the top and the rest should follow according to size. In some cases, the highest class is placed at the top.
- vi) **Distribute the data into the appropriate classes.** This is best done by using a "Tally Column" where values are tabulated against appropriate classes by merely making short bars or tally marks to represent them. It is customary for convenience in counting to place the first four bars vertically and the fifth one diagonally and to leave a space. The number of tallys is then written in the frequency column. The tally column is usually omitted in the final presentation of the frequency distribution. But in case of small number of values, the actual values should be shown against each class to mitigate chances of error.

- vii) Finally, total the frequency column to see that all the data have been accounted for.

These rules are applied to group raw data which are assumed to be continuous. In case of discrete data which carry only integral values, the concept of a class boundary is unrealistic as there can be no points where the adjoining classes meet. In spite of this logical difficulty, when the discrete data are sufficiently large, they are treated for convenience of calculations as continuous and hence are grouped in the same way as the continuous data.

Example 2.2 Make a grouped frequency distribution from the following data, relating to the weight recorded to the nearest grams of 60 apples picked out at random from a consignment:

106	107	76	82	109	107	115	93	187	95	123	125
111	92	86	70	126	68	130	129	139	119	115	128
100	186	84	99	113	204	111	141	136	123	90	115
98	110	78	183	162	171	140	152	173	146	138	194
148	96	107	181	131	75	184	104	110	80	118	82

By scanning the data, we find that the largest weight is 204 grams and the smallest weight is 68 grams so that the range is $204 - 68 = 136$ grams.

Suppose we decide to take 7 classes of equal size. The size or width of the equal class intervals would be $\frac{136}{7} = 19.47$. But we take $k = 20$, the next integral value higher than 19.47 to facilitate the numerical work.

Let us decide to locate the lower limit of the lowest class at 65. With this choice, the class limits will be 65 – 84, 85 – 104, 105 – 124, ..., the class boundaries become 64.5 – 84.5, 84.5 – 104.5, 104.5 – 124.5, ..., and the class-marks are 74.5, 94.5, 114.5, ... The grouped frequency distribution is then constructed as follows:

- i) By listing the actual values

FREQUENCY DISTRIBUTION OF WEIGHTS OF 60 APPLES

Weight	Entries	Frequency
65 – 84	76, 82, 70, 68, 84, 78, 75, 80, 82	9
85 – 104	93, 95, 92, 86, 100, 99, 90, 98, 90, 104	10
105 – 124	106, 107, 109, 107, 115, 123, 111, 119, 115, 113, 111, 122, 115, 110, 107, 110, 118	17
125 – 144	125, 126, 130, 129, 139, 128, 141, 126, 140, 131	10
145 – 164	162, 152, 146, 158, 148	5
165 – 184	178, 173, 181, 184	4
185 – 204	187, 186, 204, 185, 194	5
Total		60

This table is sometimes known as an *array*. The values against each class may be arranged in an array.

(i) By using a Tally-Column:

FREQUENCY DISTRIBUTION OF WEIGHTS OF 60 APPLES

Classes (weight)	Class- boundaries	Mid-points or Class- Marks	Tally	Frequency
65 - 84	64.5 - 84.5	74.5		9
85 - 104	84.5 - 104.5	94.5		10
105 - 124	104.5 - 124.5	114.5		17
125 - 144	124.5 - 144.5	134.5		10
145 - 164	144.5 - 164.5	154.5		5
164 - 184	164.5 - 184.5	174.5		4
185 - 204	184.5 - 204.5	194.5		1
Total	-----	-----	-----	60

Example 2.3: Given below are the mean annual death rates per 1,000 at ages 20 - 65 in each of 88 occupational groups. Construct a grouped frequency distribution.

7.5	8.2	6.2	8.9	7.8	5.4	9.4	9.9	10.9	10.8	7.4	L ₁
9.7	11.6	12.6	10.2	9.2	12.0	9.9	7.3	7.3	8.4	1.2	
10.3	10.1	10.0	7.1	6.5	12.5	7.8	6.5	8.7	9.3	12.4	1.9
10.4	9.1	9.3	6.2	10.3	6.6	7.4	8.6	7.7	9.4	2.3	
7.7	12.8	8.7	5.5	8.6	9.6	11.9	10.4	7.8	7.6	12.1	1.8
4.6	14.0	8.1	11.4	10.6	11.6	10.4	8.1	4.6	6.6	12.8	1.2
6.8	7.1	6.6	8.8	8.8	10.7	10.8	6.0	7.9	7.3	9.3	
9.3	8.9	10.1	3.9	8.0	6.9	9.0	8.8	9.4	11.4	10.9	1

(B.I.S.E., Lahore, 1971)

A scan of the data shows that the largest value is 14.0 and the smallest value is 3.9 so that the range = 14.0 - 3.9 = 10.1.

As the data are recorded to one decimal place, we may therefore locate the lower limit of the first group at 3.5. Let us choose a class interval of 1.0. Then the class limits are specified as 3.5 - 4.4

4.5 - 5.4, 5.5 - 6.4, ... With this choice, the class-boundaries are 3.45 - 4.45, 4.45 - 5.45, 5.45 - 6.45, ..., which do not coincide with the given values.

The following table shows the required frequency distribution:

FREQUENCY DISTRIBUTION OF MEAN DEATH RATES

Death Rates	Class-boundaries	Tally	Frequency
3.5 - 4.4	3.45 - 4.45		1
4.5 - 5.4	4.45 - 5.45		4
5.5 - 6.4	5.45 - 6.45		5
6.5 - 7.4	6.45 - 7.45		13
7.5 - 8.4	7.45 - 8.45		12
8.5 - 9.4	8.45 - 9.45		19
9.5 - 10.4	9.45 - 10.45		13
10.5 - 11.4	10.45 - 11.45		10
11.5 - 12.4	11.45 - 12.45		6
12.5 - 13.4	12.45 - 13.45		4
13.5 - 14.4	13.45 - 14.45		1
Total			88

Example 2.4. Construct a frequency distribution for the data below. Indicate the class boundaries and class limits clearly.

41.78	39.32	31.47	35.35	32.82	39.42
61.30	28.31	44.63	22.78	44.44	48.12
81.71	33.47	50.35	29.19	51.26	50.32
26.84	18.95	48.19	43.72	43.89	47.15
60.20	44.43	41.17	37.50	22.35	29.17

By scanning the data, we find that the largest value is 81.71 and the smallest value is 18.95 so that the range is $81.71 - 18.95 = 62.76$.

Suppose we decide to take 5 classes of equal size. Then size or width of the equal class intervals would be $\frac{62.76}{5} = 12.55$. But we take $h = 13.00$, the next integral value higher than 12.55 to take less numerical work.

As the data are recorded to two decimal places, we may locate the lower limit of the first group at 18.00. With this choice, the class limits will be 18.00 – 30.99, 31.00 – 43.99, ..., the class boundaries become 17.995 – 30.995, 30.995 – 43.995. The grouped frequency distribution is then constructed as follows:

Classes	Class boundaries	Tally	f
18.00 – 30.99	17.995 – 30.995		8
31.00 – 43.99	30.995 – 43.995		10
44.00 – 56.99	43.995 – 56.995		9
57.00 – 69.99	56.995 – 69.995		2
70.00 – 82.99	69.995 – 82.995		1
Total	30

Example 2.5 A survey of 50 retail establishments had assistants, excluding proprietors, as follows:

2, 3, 9, 0, 4, 4, 1, 4, 5, 4, 3, 5, 3, 6, 6, 0, 2,
2, 7, 6, 4, 8, 4, 3, 3, 1, 0, 8, 7, 5, 1, 3, 4, 2, 4, 7,
3, 2, 6, 3, 1, 7, 5, 4, 6, 4, 2, 5, 3.

Arrange the values as a frequency distribution.

By scanning the data, we find that the number of assistants is a *discrete variable* and the range is small, so the data can be conveniently sorted by taking the classes as 0, 1, 2, etc. The frequency distribution is then constructed as shown below:

FREQUENCY DISTRIBUTION OF ASSISTANTS IN 50 RETAIL ESTABLISHMENTS

Number of Assistants (x)	Tally	Number of Establishments (f)
0		3
1		4
2		6
3		7
4		10
5		6
6		4
7		2
8		1
9		1
Total	50

Such a frequency distribution in which each class consists of a single value is sometimes called a *discrete* or *ungrouped* frequency distribution.

2.4.6. Cumulative Frequency Distribution. The total frequency of a variable from its one end to a certain value (usually upper class boundary in grouped data), called the base, is known as the *cumulative frequency*, *less than or more than* the base of the variable. A table that shows the cumulative frequencies is called a *cumulative frequency distribution*. The cumulative frequency of the last class is the sum of all frequencies in the distribution. If the cumulation process is from the lowest value to the highest, it is referred to as "less than" type cumulative frequency distribution. For example, let us consider a frequency distribution having k classes, each of width h . Let us denote the midpoint of the i th class by x_i

with frequency f_i such that $\sum_{i=1}^k f_i = n$. Now the lower class-boundary of the first group is $x_1 - h/2$ and the upper class boundaries are $x_i + h/2$, ($i=1, 2, \dots, k$). The cumulative frequency distribution is then obtained by adding each successive frequency to the cumulative total of frequencies for the preceding classes as shown below:

Class-boundary	Cumulative Frequency
less than $x_1 - h/2$	0
less than $x_1 + h/2$	f_1
less than $x_2 + h/2$	$f_1 + f_2$
less than $x_3 + h/2$	$f_1 + f_2 + f_3$
...	...
less than $x_k + h/2$	$\sum_{i=1}^k f_i = n$

It should be noted that a *less than* type cumulative frequency distribution starts with the lower class boundary of the first group indicating that there is no frequency below $x_1 - h/2$.

When the frequencies are cumulated from the highest value to the lowest value, it is called a "*more than*" type cumulative frequency.

If the class frequencies of all various classes are divided by the total frequency, we get the *relative frequencies* which always add to one. The class frequencies may also be expressed as percentages, the total of which would be 100. A percentage cumulative distribution is useful to read off the percentage of values falling between certain specified values.

Example 2.6. Construct (i) a "less than" type cumulative distribution, and (ii) a "more than" type cumulative distribution from the frequency distribution of weights of 60 apples of Example 2.2.

(i) A "less than" type cumulative frequency distribution is shown below

Weight (grams)	Cumulative Frequency (F)
Less than 64.5	0
Less than 84.5	9
Less than 104.5	19
Less than 124.5	36
Less than 144.5	46
Less than 164.5	51
Less than 184.5	55
Less than 204.5	60

- i) A "more than" type cumulative frequency distribution is given below:

Weight (grams)	Cumulative Frequency (F)
More than 64.5	60
More than 84.5	51
More than 104.5	41
More than 124.5	24
More than 144.5	14
More than 164.5	9
More than 184.5	5
More than 204.5	0

2.5 STEM-AND-LEAF DISPLAY

A clear disadvantage of using a frequency table is that the identity of individual observations is lost in grouping process. To overcome this drawback, John Tukey (1977) introduced a technique known as the *stem-and-leaf Display*. This technique offers a quick and novel way of simultaneously sorting and displaying data sets where each number in the data set is divided into two parts, a *stem* and a *leaf*. A *stem* is the leading digit(s) of each number and is used in sorting, while a *leaf* is the rest of the number or the trailing digit(s) and shown in display. A vertical line separates the leaf (or leaves) from the stem. For example, the number 243 could be split two ways:

leading digit	trailing digits	OR	leading digit	trailing digits
2	43		24	3
stem	leaf		stem	leaf

All possible stems are arranged in order from the smallest to the largest and placed on the left hand side of the line.

The *stem-and-leaf display* is a useful step for listing the data in an array, leaves are associated with the stem to know the numbers. *stem-and-leaf* table provides a useful description of the data set and can easily be converted to a frequency table. It is a common practice to arrange the trailing digits in each stem from smallest to highest.

Example 2.7 The ages of 30 patients admitted to a certain hospital during a particular week were as follows:

48, 31, 54, 37, 18, 64, 61, 43, 40, 71, 51, 12, 52, 65, 53

42, 39, 62, 74, 48, 29, 67, 30, 49, 68, 35, 57, 36, 27, 58

Construct a stem-and-leaf display from the data and list the data in an array.

A scan of the data indicates that the observations range (in age) from 12 to 74. We use the first (or leading) digit as the *stem* and the second (or trailing) digit as the *leaf*. The first observation is 48, which has a stem of 4 and a leaf of 8, the second a stem of 3 and a leaf of 1, etc. Placing the leaves in the order in which they appear in the data, we get the *stem-and-leaf* display as shown on next page:

Stem (leading digit)	Leaf (trailing digit)
1	8 2
2	9 6 7
3	1 2 9 0 5
4	8 3 0 1 8 9
5	4 1 2 3 7 8
6	4 3 5 2 7 8
7	1 4

To get the array, we associate the leaves in order of size with the stems as shown below:

12, 18, 26, 27, 29, 30, 31, 35, 37, 39, 40, 42, 43, 48, 48, 49,

51, 52, 53, 54, 57, 58, 61, 62, 64, 65, 67, 68, 71, 74

Example 2.8 Construct a stem-and-leaf display for the data of annual death rates given in Example 2.5.

Using the decimal part in each number as the leaf and the rest of the digits as the stem, we get the following stem-and-leaf display (leaves are ordered):

Stem	Leaf
3	9
4	6 6
5	0
6	2 2 5 5 6 6 6 8 9
7	1 3 3 3 4 4 5 6 7 7 8 8 8 9
8	1 1 2 4 6 6 7 7 8 8 8 9 9
9	0 1 2 3 3 3 4 4 4 6 7 7 9 9
10	0 1 1 2 3 3 4 4 6 6 7 8 8 9 9
11	1 4 4 6 6 9
12	0 1 4 5 6 8 8
13	0

2.6 GRAPHICAL REPRESENTATION

Tabulation, we know, is a good method of condensing and representing statistical data in a more understandable form, but many people have no taste for figures. They would prefer a visual representation where figures could be avoided. This purpose is achieved by the presentation of statistical data in a visual form. The visual display of statistical data in the form of points, lines, areas and geometrical forms and symbols, is in the most general terms known as *Graphical Representation*. Statistical data can be studied with this method without going through figures, presented in the tables.

Such visual representation can be divided into two main groups, graphs and diagrams to be described in the sections that follow. The basic difference between a graph and a diagram is that a graph is a representation of data by a continuous curve, usually shown on a graph paper while a diagram is any other one, two or three-dimensional form of visual representation.

2.7 DIAGRAMS

Diagrammatic representation is best suited to spatial series and data split into different categories. Whenever a comparison of the same type of data at different places is to be made, diagrams will be the best way to do that. Diagrammatic representation has several advantages over tabular representation of figures. Beautifully and neatly constructed diagrams are more attractive than simple figures. Diagrams, being a visual display, leave more effective and long lasting impression on the mind of a reader. They make unwieldy data intelligible at a glance. Comparison is made easier with diagrams. Diagrams have some disadvantages too. Diagrams are less accurate than tables; cost money and time and the amount of information conveyed is limited. However, this method of representation is excessively used in business and administration.

Different types of diagrams or charts commonly used for displaying statistical data are described below:

- i) **Linear or One-Dimensional Diagrams.** They consist of Simple Bars, Multiple Bars and Component Bar charts. Here the values are represented only by one dimension, generally the length of the bar.
- ii) **Areal or Two-Dimensional Diagrams.** They consist of Rectangles, Sub-divided Rectangles and Squares, the areas of which are proportional to the values of the given quantities. This device is used to represent data having nodal or large variations.
- iii) **Cubic or Three-Dimensional Diagrams.** They are in the form of Cubes and cylinders, whose volumes are proportional to the values they represent. These diagrams are used when the variation among the values of the data to be portrayed is so large that even the square roots of the values concerned fail to reduce the variation appreciably.
- iv) **Pie-Diagrams.** They are in the form of Circles and Sectors. Here the areas of circles or sectors are in proportion to the values they represent or compare.
- v) **Pictograms.** They consist of pictures or small symbolic figures representing the statistical data. A pictogram is an effective way of visual communication. For example, we can compare the armed strength of various countries by drawing pictures of the number of soldiers, where each pictorial soldier may denote, say, 1,000 soldiers. In a similar way, the production of wheat can be compared by means of the pictures of wheat bags of a specified size. It is essential to repeat the pictures a number of times to represent the differences in magnitudes.

In drawing diagrams, the following points should be kept in mind:

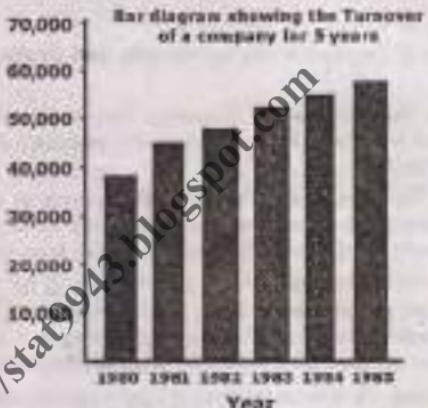
- An appropriate scale consistent with the size of paper available and the size of the data to be represented, should be chosen and indicated either at the side or at the bottom of the diagram. This scale must start at zero.
- A diagram like a table, must have a title, which should be brief and self-explanatory. A key, footnote or source will also be necessary.
- A diagram should be shaded, coloured or cross-hatched to show the different parts, if any. Lettering should be shown horizontally.

2.7.1 Simple Bar Chart. A simple bar chart consists of horizontal or vertical bars of equal width and lengths proportional to the values they represent. As the basis of comparison is linear or one-dimensional, the widths of these bars have no significance but are taken to make the chart look attractive. The space separating the bars should not exceed the width of the bar and should not be less than half its width. The bars should neither be exceedingly long and narrow nor short and broad. The vertical bar chart is an effective way for presenting a time series and qualitatively classified data whereas horizontal bars are useful for geographical or spatial distributions. The data when do not relate to time, should be arranged in ascending or descending order before charting.

Example 2.9 Draw a simple bar diagram to represent the turnover of a company for 6 years.

Year:	1980	1981	1982	1983	1984	1985
Turnover (Rupees):	38,000	45,000	48,000	52,500	55,000	58,000

The bar chart is drawn below:



2.7.2 Multiple Bar Chart. A multiple bar chart shows two or more characteristics corresponding to the values of a common variable in the form of grouped bars, whose lengths are proportional to the values of the characteristics, and each of which is shaded or coloured differently to aid identification. This is a good device for the comparison of two or three kinds of information. For example, imports, exports and productions of a country can be compared from year to year by grouping the three together.

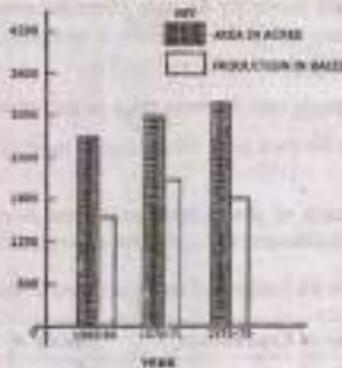
Example 2.10 Draw multiple bar charts to show the area and production of cotton in the Punjab from the following data:

Year	Area (000 acres)	Production (000 bales)
1965 - 66	2866	1588
1970 - 71	3233	2229
1975 - 76	3420	1937

(Source: Statistical Wing, Agriculture Deptt. Lahore)

The multiple bar charts are drawn below:

AREA AND PRODUCTION OF COTTON IN THE PUNJAB



2.7.3 Component Bar Chart. A *component bar chart* is an effective technique in which each bar is divided into two or more sections, proportional in size to the components parts of a total being displayed in each bar. The various component parts shown as sections of the bar are shaded or coloured differently to increase the overall effectiveness of the diagram. Component bar charts are used to represent the composition of the various components of data and the percentage. They are also known as sub-divided charts.

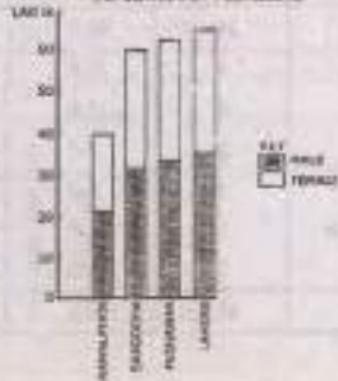
Example 2.11 Draw a component bar chart for the following data.

(Population in lakhs)

Division	Bal. Leetders	Male	Female
Peshawar	54	33	31
Rawalpindi	40	21	19
Sargodha	60	32	28
Lahore	65	35	30

The appropriate component bar chart after arranging the population figures in ascending order is shown below:

COMPONENT BAR CHART SHOWING POPULATION OF 4 DIVISIONS



2.7.4 Rectangles and Sub-divided Rectangles. The area of a rectangle is equal to the product of its length and breadth. To represent a quantity by a rectangle, both length and breadth of the rectangle are used. Sub-divided rectangles are drawn for the data where the quantities along with their components are to be compared. These diagrams are generally drawn to compare the budgets of various families. In construction of sub-divided rectangles, we are required to

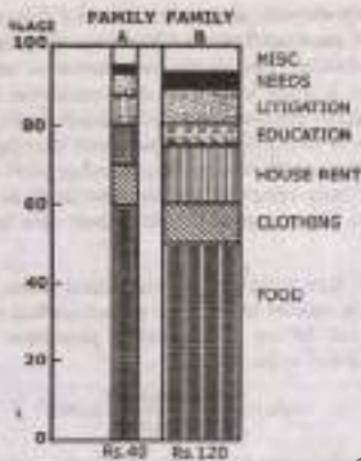
- change each component into the percentage of the corresponding total,
- draw one rectangle for each total, taking equal lengths (100 units) and breadths proportional to the totals,
- divide every rectangle so drawn into parts equal in number to the number of components. Each part shaded or coloured will represent percentage size of one component.

Example 2.12 Compare the budgets of families A and B with a suitable diagram.

<i>Items of Expenditure</i>	<i>Family A</i>	<i>Family B</i>
Food	24	60
Clothing	4	14
House Rent	4	16
Education		6
Litigation	2	10
Conventional Needs	1	6
Miscellaneous	2	8
Total	40	120

The necessary computations required for the drawing of sub-divided rectangles are given below and the diagram is shown on page 33.

<i>Items of Expenditure</i>	<i>Family A</i>		<i>Family B</i>	
	<i>Actual Expenses</i>	<i>Percentage Expenses</i>	<i>Actual Expenses</i>	<i>Percentage Expenses</i>
Food	24	60.0	60	50.0
Clothing	4	10.0	14	11.7
House Rent	4	10.0	16	13.3
Education	3	7.5	6	5.0
Litigation	2	5.0	10	8.3
Conventional Needs	1	2.5	6	5.0
Miscellaneous	2	5.0	8	6.7
Total	40	100.0	120	100.0

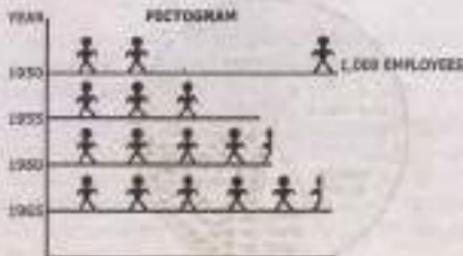


2.7.5 Pictograms. A pictogram is a popular device for portraying the statistical data by means of pictures or small symbols. It is said that a picture is worth ten thousand words. It is customary to represent a unit value of the data by a standard symbol or a picture and the whole quantity by an appropriate number of repetitions of symbol concerned. This means the larger quantities should be represented by a larger number of symbols and not by large symbols. A quantity smaller than the unit is represented by a part of the picture or symbol used. The symbols or pictures to be used, must be simple and clear. A pictogram is virtually a bar chart constituted in pictorial way as the number of symbols or pictures corresponds to the length of a bar.

Example 2.13: The following table shows the number of employees in a certain Textile Mills. Represent the data by means of a pictogram.

Year	No. of Employees
1950	2,004
1955	2,990
1960	4,240
1965	5,380

Representing 1,000 employees by one picture, the pictogram is drawn below:



2.7.6 Pie Diagrams. A pie-diagram, also known as sector diagram, is a graphic device consisting of a circle divided into sectors or pie-shaped pieces whose areas are proportional to the various parts in which the whole quantity is divided. The sectors are shaded or coloured differently to show the relationship of parts to the whole. If space permits, the descriptive titles of the constituent parts should be placed horizontally on each sector, otherwise a key becomes necessary. It is a convenient way of displaying the component parts in proportion to the total and therefore is used as an alternative to component bar chart. It is an effective way of showing percentage parts when the whole quantity is taken as 100. It is also used when the basic categories are not quantifiable such as expenditure, classified as food, clothing, fuel and light, etc. The arrangement of the sectors must be made uniform in comparing pie charts.

To construct a pie chart, draw a circle of any convenient radius. As a circle consists of 360° , the whole quantity to be displayed is equated to 360 . The proportion that each component part or category bears to the whole quantity will be the corresponding proportion of 360° . These corresponding proportions, i.e., angles, are calculated by the formula

$$\text{Angle} = \frac{\text{component part}}{\text{whole quantity}} \times 360^\circ$$

Then divide the circle into different sectors by constructing angles at the centre by means of a protractor and draw the corresponding radii.

Example 2.1.4 Represent the total expenditure and expenditures on various items of a family by a pie diagram.

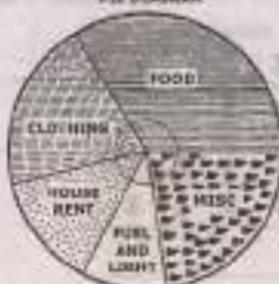
Items	Food	Clothing	House Rent	Fuel and Light	Misc.
Expenditure (in Rs.)	50	30	20	15	35

The corresponding angles needed to draw a chart are computed below.

Items	Expenditure (in Rs.)	Angles of the Sectors (in Degrees)
Food		
Clothing	30	
House Rent	20	
Fuel and Light	15	120
Miscellaneous	35	72
Total	150	360

"The pie diagram consisting of a circle divided into five sectors defined by angles 120° , 72° , 48° and 84° , is drawn below."

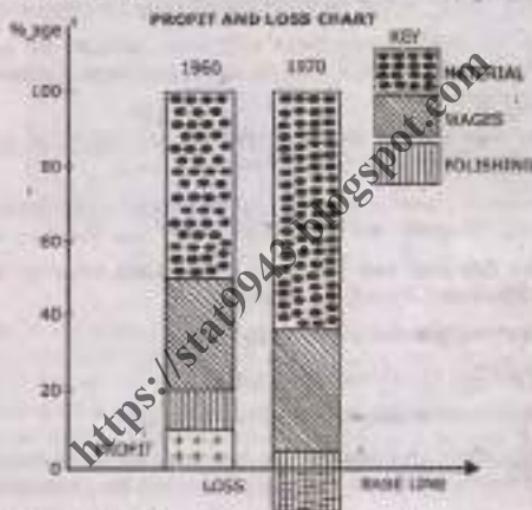
PIE DIAGRAM



2.7.7 Profit and Loss Chart. This is virtually a percentage component bar chart in which profits can be shown above the normal base line and losses below the base line. Since the bars are to be extended from the zero line to show losses, we start from the top. For an illustration, the following data are represented:

COST, PROCEEDS, PROFIT OR LOSS PER CHAIR

Particulars	1968	1970
i) Materials	Rs. 10	Rs. 16
ii) Wages	6	8
iii) Polishing, etc.	2	4
Total cost	18	28
Proceeds	20	25
Profit (+) or loss (-)	+2	-3



A pie chart may also be used for this purpose.

GRAPHS

As already stated, diagrams are useful for representing spatial series. Diagrams fail when we want to represent a statistical series spread over a period of time, or a frequency distribution or two-related series in visual form. For such representations, graphs are employed.

Graphs present the data in a simple, clear and effective manner, facilitate comparison between two or more than two statistical series, and help us in appreciating their significance readily. Another advantage of graphs is that they provide an overall picture of a statistical series. Graphs are also used to make predictions and forecasts. Certain particular values can also be located easily. But graphs are less accurate as they do not give minute details. Moreover, they cost considerable expenditure and time.

Construction of Graphs. In the construction of a graph, the first step is to take a starting point, known as the origin, in the left-hand bottom corner of the graph paper. Two straight lines perpendicular to each other are drawn through the origin. The horizontal line is called the *X-axis* or abscissa and the vertical line is labeled as *Y-axis* or ordinate. The two lines together are known as co-ordinate axes. Some suitable scales are selected along *X-axis* and *Y-axis*. Independent variable is taken along *X-axis* and dependent variable along *Y-axis*. Points are plotted and joined to get the required graph. While constructing a graph, the following points should be kept in mind:

- A scale and the form of representation is to be selected in such a way that the true impression of the data to be represented is given by the graph.
- Every graph must have a clear and comprehensive title at top. Where necessary, sub-title should be added.
- The source of the data must be given. A key and footnotes should be provided when necessary.
- The independent variable should always be placed on the horizontal axis.
- The vertical scale should always begin with zero, otherwise the graph will give a false impression. If, however, the first item of the data is quite large, a scale-break should be shown between zero and next member.
- The horizontal axis does not have to begin with zero unless of course, the independent variable or the lower limit of the first class interval is zero.
- The axes of the graph should be properly labelled. Labels should clearly state both the variable and the units, e.g. "Distance" and "Kilometre", "Sales" and "Rupees", etc.
- Curves if more than one, must be clearly distinguished either by different colours or by differentiated lines (solid, dashed and dotted).
- The graph should not be loaded with too many curves.

Graphs can be divided into two main categories, namely:

- Graphs of Time-Series or Graphs of Historical Data, and
- Graphs of Frequency Distributions. The important graphs of frequency distributions are Histogram, Frequency Polygon, Frequency Curve and the Cumulative Frequency Curve or Ogive.

2.8.1 Graph of Time Series-Histogram. A curve showing changes in the value of one or more items from one period of time to the next is known as the graph of a time series. This curve is also called a Histogram. Thus a histogram displays the variations in time series dealing with prices, production, imports, population, etc. To construct a histogram, time is taken along *X-axis* and the values of the variables along *Y-axis*. Points are plotted and are then connected by means of straight line segments to get the "Histogram".

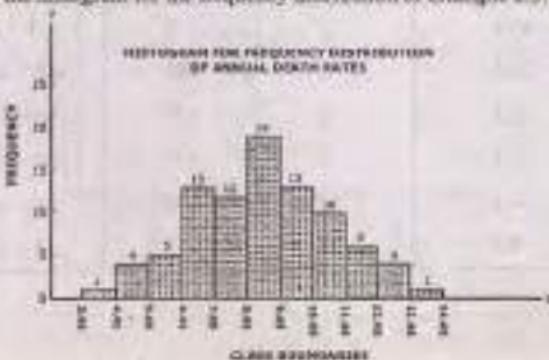
Example 2.15 The following table gives the number of cars produced in Germany during the years 1929-1936. Draw a suitable graph, i.e. Histogram of the series

Years:	1929	1930	1931	1932	1933	1934	1935	1936
No. of Cars:	98	74	68	50	99	172	245	302

The histogram is drawn for the data by taking years on horizontal axis and the number of cars on vertical axis as below:



2.8.3 Histogram. A histogram consists of a set of adjacent rectangles whose bases are marked off on boundaries (not class limits) on the x -axis and whose heights are proportional to the frequencies and with respective classes. The area of each rectangle represents the respective class frequencies. This is one of the most important graphical representation of a frequency distribution. When the class widths are equal, the rectangles have the same width and their heights directly represent the class frequencies, that is they are numerically proportional to the frequencies in the respective classes. The figure shows the histogram for the frequency distribution of Example 2.3.



If the class-intervals are not all equal, the height of the rectangle over an unequal class-interval, is to be adjusted because it is area and not height that measures frequency. This means that the height of the rectangle must be proportionally decreased if the length of the corresponding class-interval increases. For example, if the length of a class-interval becomes double, then the height of the rectangle is to be halved so that the area, being the fundamental property of the rectangle of a histogram, remains unchanged. This sort of rescaling is necessary so that the correct pattern of the distribution is to be conveyed.

When the frequencies in a frequency distribution are given against the class-marks x_i of equal class-intervals of width h , a histogram is constructed by drawing vertical lines (dotted) whose heights correspond to the respective class-frequencies at the class-marks marked off on the axis of X and erecting a series of adjacent rectangles with widths equal to $x_i \pm h/2$ (i.e. half of the width is taken on either side of x_i).

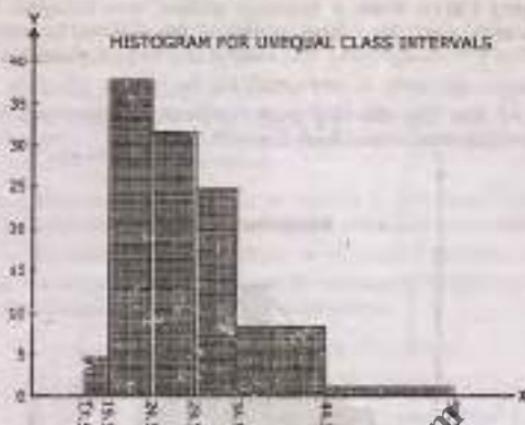
It is important to note that in the construction of a histogram, we assume that within any one class the values of the variable are evenly spread out between the class-boundaries. A histogram which may not be confused with the *autovogram* (graph of a time series) is useful in forming a rough idea of the overall pattern and shape of the frequency distribution.

Example 2.16 Construct a Histogram for the following frequency distribution relating to the age (to nearest birthday) of telephone operators.

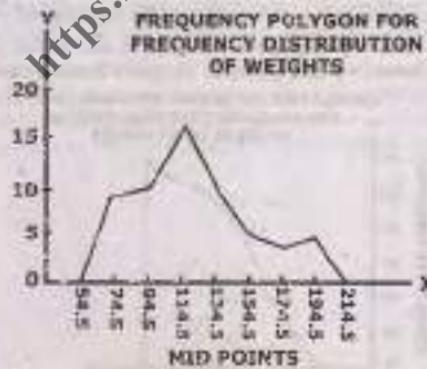
Age (Years)	18-19	20-21	25-29	30-34	35-44	45-59
No. of Operators	9	188	160	123	84	15

As the class-intervals are unequal, the height of each rectangle cannot be made equal to the frequency. The height of a rectangle is therefore calculated by dividing the frequency (the area) by the corresponding class interval (the width). The necessary calculations and the histogram follow:

Class boundaries	Class-Interval (h)	Frequency	Proportional Heights
17.5 - 19.5	2	9	$9 \div 2 = 4.5$
19.5 - 24.5	5	188	$188 \div 5 = 37.6$
24.5 - 29.5	5	160	$160 \div 5 = 32.0$
29.5 - 34.5	5	123	$123 \div 5 = 24.6$
34.5 - 44.5	10	84	$84 \div 10 = 8.4$
44.5 - 59.5	15	15	$15 \div 15 = 1.0$

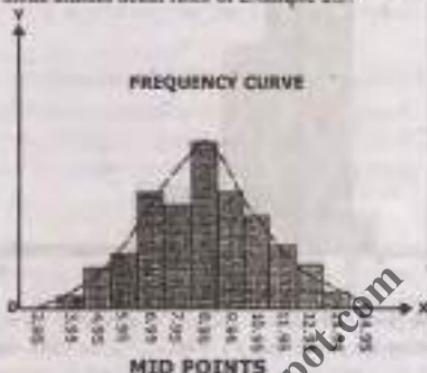


2.8.3 Frequency Polygon. A frequency polygon is a graphic form of a frequency distribution, which is constructed by plotting the points (x_i, f_i) where x_i is the class-mark of the i th class and f_i is the corresponding frequency, and then connecting them by straight line segments provided the class-intervals are equal. In case of unequal class-intervals, heights of unequal classes are adjusted by using the same technique that was used for histogram. It can also be obtained by joining the tops of the successive rectangles in the histogram by means of straight line segments. The graph drawn in this way does not touch the horizontal axis. But a polygon, as we know, is a closed figure having many sides. It is therefore necessary to add "extra" class-marks at both ends of the distribution with zero class-frequencies so that the polygon does form a closed figure with the horizontal axis. This should be done even if the curve ends in the middle part of the graph. The frequency polygon for the frequency distribution of weights in Example 2.2 is given below.

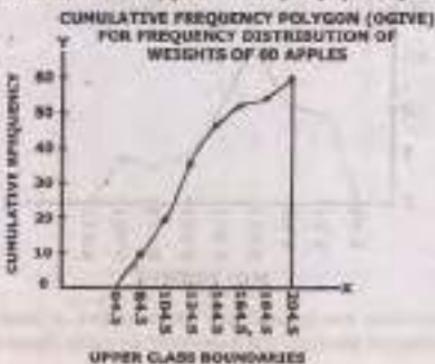


A frequency polygon which can be used for comparing two or more data sets, gives roughly the idea of mode, some idea of skewness and kurtosis of the curve (these terms are defined later).

2.3.4 Frequency Curve. When a frequency polygon or a histogram constructed over the intervals made sufficiently small for a large number of observations, is smoothed, it approaches a continuous curve, called a *frequency curve*. The concept of a frequency curve is of great importance in statistics. Mathematically, the curve is represented by the relation $y = f(x)$ and has an important property concerning its area. The following graph represents histogram and frequency curve for the frequency distribution of the mean annual death rates of Example 2.3.



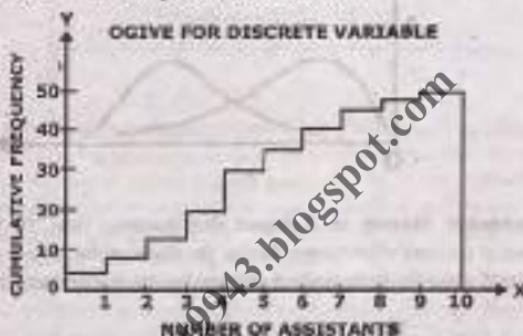
2.4.5 Cumulative Frequency Polygon or Ogive. A cumulative frequency polygon, popularly known as *Ogive* (rhymes with "alive" and pronounced 'o'jev') is a graph obtained by plotting the cumulated frequencies of a distribution against the upper or lower class boundaries depending upon whether the cumulation is of the "less than" or "more than" type, and the points are joined by straight line segments. Because of its likeness to an irregular moulding called an *ogee*, a cumulative frequency polygon is called an *Ogive*. An *Ogive*, whose cumulation is of less-than type, is constructed by plotting the points $(x_i + h/2, F_i)$ where $x_i + h/2$ is the upper class-boundary of the i th class and F_i is the cumulative frequency for the i th class, and connecting the successive points by straight line segments. The polygon should start from zero at the lower boundary of the first interval, i.e. the point $(x_1 - h/2, 0)$ is plotted and joined, and to close a polygon, the last point is also joined with the last upper class-boundary. In case of unequal classes, we merely join the unequally spaced points.



If relative frequencies are used, the cumulative frequency polygon rises from the value 0 at the left to the value 1 at the right. A smoothed Ogive is called an *Ogive curve*, which is often used to locate the partition values such as the median, quartiles, percentiles, etc. of a frequency distribution.

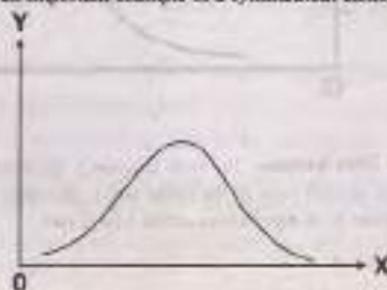
A percentage cumulative frequency polygon or *curve* may also be drawn by expressing the cumulative frequencies as percentages of the total frequency and then connecting the plotted percentages against upper class boundaries. This graphic device is useful for comparing two or more frequency distributions as they are adjusted to a uniform standard.

2.8.6 Ogive for a Discrete Variable. When a variable X is discrete, its cumulative frequency polygon consists of horizontal line segments between any two successive values and has a jump of height f_i at each value of x_i . In other words, the cumulative distribution increases only in jumps and is constant between jumps. For the purpose of illustration, the cumulative frequency polygon drawn for the frequency distribution of assistants in Example 2.5, is shown below:

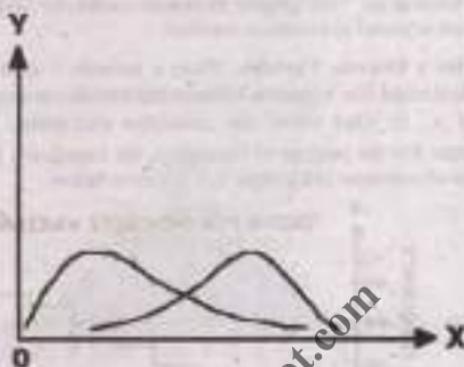


2.8.7 Types of Frequency Curves. The frequency distribution occurring in practice, usually belongs to one of the following four types:

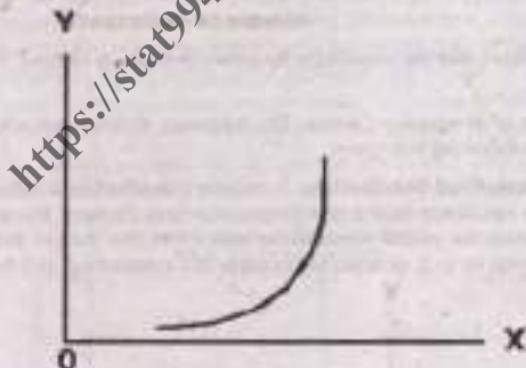
- i) **The Symmetrical Distributions.** A frequency distribution or curve is said to be symmetrical if values equidistant from a central maximum have the same frequencies, i.e. the curve can be folded along the central maximum in such a way that the two halves of the curve coincide. The *Normal* curve is an important example of a symmetrical distribution.



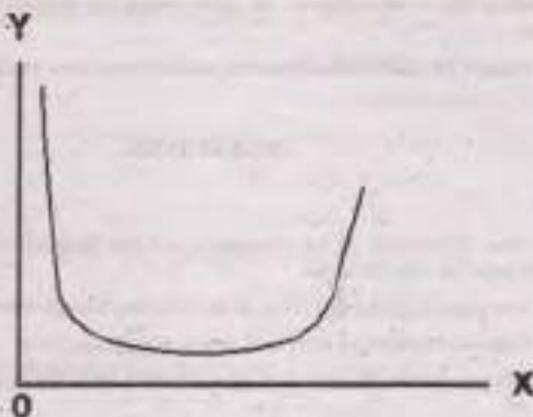
- a) **The Moderately Skewed or Asymmetrical Distributions.** A frequency distribution or curve is said to be *skewed* when it departs from symmetry. Here the frequencies tend to pile up at one end or the other end of the distribution or curve. This is the most common pattern encountered in practice.



- iii) **The Extremely Skewed or J-shaped distributions.** Here the frequencies run up to a maximum at one end of the range, having the shape of the letter J or its reverse. Most of the distributions in economic or medical statistics belong to extremely skewed distributions.



- iv) **The U-shaped Distributions.** In such frequency distributions or curves, the maximum frequencies occur at both ends of the range and a minimum towards the centre, shaped more or less like the letter U. A distribution of this type is rare.



2.8.8 Ratio Charts or Semi-logarithmic Graphs. In the ordinary types of graph, the scales used are called the *natural scales* or the *arithmetic scales*. These graphs can only be used to compare the absolute changes in values because the ordinary graph paper, also known as arithmetic paper, is so ruled that equal intervals anywhere on the paper represent equal differences or amounts. More often we are interested in studying the relative changes or ratios. The relative changes or ratios can be displayed and compared by the slope of straight line when the logarithms of the values are plotted on an arithmetic paper. In practice, the difficulty of looking up logarithms can be dispensed with by using another type of graph paper, called *Semi-logarithmic paper* or *ratio paper*. A semi-logarithmic paper or ratio paper is so constructed that equal intervals on the vertical axis indicate equal ratios or rates of change, while equal intervals on the horizontal axis represent real differences or amounts of change. Thus the essential feature of a *Semi-logarithmic chart* is that one axis has a logarithmic scale and the other has arithmetic scale.

Graphs obtained by plotting the values on a semi-logarithmic paper or ratio paper and joining the successive points by means of straight line segments are called *Semi-logarithmic graphs* or *Ratio charts*. They are generally used when

- (i) the relative rates of change are to be compared;
 - (ii) visual comparisons are to be made between two or more series which differ widely in magnitude; and
 - (iii) the data are to be examined to see whether they are characterized by a constant rate of change.
- A ratio chart possesses the following characteristics:

- (i) There is no zero line on the logarithmic scale as the logarithm of zero is minus infinity.
- (ii) A geometric progression when plotted on semi-logarithmic paper, forms a straight line, as the logarithms of a geometric progression form an arithmetic progression.
- (iii) The slope of the logarithmic scale variable indicates the rate at which the variable is changing (i.e. increasing or decreasing).

- v) In case of two or more curves, the curve having the steepest slope, has the largest rate of change.
- v) Equal slopes (in case of parallel curves) indicate equal rates of change.

EXERCISES

OBJECTIVE

- a) Answer 'True' and 'False'. If the statement is not true then replace the underlined words with words that make the statement true:
 - i) The term cross-sectional data refers to data that may change overtime.
 - ii) The frequency distribution represents data in a condensed form.
 - iii) The data presented in an array does not allow us to locate the largest and smallest values in the data set.
 - iv) The classes in any frequency distribution are generally not mutually exclusive.
 - v) For nominally or ordinaly scaled data the frequency distribution cannot be constructed.
 - vi) Frequency distribution of continuous data may be represented diagrammatically.
 - vii) Frequency distribution can be presented graphically by using both histogram and histogram.
 - viii) Time series data can be tabulated using frequency distributions.
 - ix) Simple bar diagram is used for two-dimensional comparisons.
 - x) The width of a bar in histogram represents the frequency rather than the value of a variable.
 - xi) Class marks are the lower limit of each class.
 - xii) The lower class limit is the middle possible data value for a class.
 - xiii) The sum of relative frequencies in a relative frequency distribution should always equal 100.
 - xiv) A pie chart can be used to display quantititative data.
 - xv) A shape of the frequency distribution and the relative frequency distribution always will be different.
- b) MULTIPLE CHOICE QUESTIONS:
 - i) Which of following is not an example of condensed data?
 - a) frequency distribution
 - b) data array
 - c) histogram
 - d) polygon
 - ii) In the construction of a frequency distribution the steps are to:
 - a) decide the number of classes
 - b) arranging the data in ascending / descending order
 - c) locate the smallest and largest values in a data set
 - d) all of above

- i) The number of classes in a frequency distribution generally should be:
- less than five
 - more than twenty
 - between five and twenty
 - between ten and twenty
- ii) As the number of observations and classes increase, the shape of the frequency polygon:
- remains same
 - tends to smooth
 - become more erratic
 - none of them
- iii) A cumulative frequency distribution is graphically represented by:
- frequency curve
 - frequency polygon
 - pie chart
 - ogive
- iv) A relative frequency distribution presents frequencies in terms of:
- whole numbers
 - percentages
 - fractions
 - all of above
- v) A diagram that presents properties that look like slices of a pizza is known as:
- a bar diagram
 - a component bar diagram
 - a histogram
 - a pie diagram
- vi) Observed data organized into tabular form is called:
- a bar chart
 - a pie chart
 - a frequency polygon
 - a frequency distribution
- vii) The number of occurrences of a particular value is called:
- the frequency
 - the cumulative frequency
 - the relative frequency
 - all of above
- viii) Is the following a stem-and-leaf diagram:

Stem	Leaf
3	2 3
4	1 2 2 2 3
5	1 1 3 5 5 5 5 6
6	4 5 6 7
7	2 8
8	6

The number that occurred the most is:

- 2
- 55
- 42
- 5

EXERCISE

Explain what is meant by classification. What are its basic principles?

- 2.2 Define the terms "Classification" and "Tabulation". Outline the main steps in tabulation. What do you mean by *caption, stubs, title and prefatory notes?* (P.U., B.A/B.Sc., 1983)
- 2.3 What is a statistical table? What are different types of tables? Explain the different parts of a table and the main points to be kept in mind in their construction. (P.U., B.A/B.Sc., 1979)
- 2.4 Represent the data given in the following paragraph in the form of a table, so as to bring out clearly all the facts, indicating the source and bearing suitable title:

"According to the census of Manufactures Report 1945, the John Smith Manufacturing Company employed 400 non-union and 1,250 union employees in 1941. Of these 220 were females of which 140 were non-union. In 1942, the number of union employees increased to 1,475 of which 1,300 were males. Of the 250 non-union employees 200 were males. In 1943, 1,700 employees were union members and 50 were non-union. Of all the employees in 1944, 250 were females of which 240 were union members. In 1944, the total number of employees was 2,000 of which one percent were non-union. Of all the employees in 1944, 300 were females of which only 5 were non-union."

- 2.5 a) Write short notes on:
 Class-frequency, Class-Interval, Class limits, Class Marks, Size of Class-Interval and Sturges' rule.
- b) Determine class boundaries, class limits and class marks for the first and last classes in respect of the following:
- Weights of 300 entering freshmen ranged from 98 to 226 pounds, correct to the nearest pound.
 - The thickness of 460 meters ranged from 0.421 to 0.563 inches.
 - A sample consists of 100 observations, each recorded as correct to the nearest integer, ranging in value from 201 to 537. If it is decided to use seven classes of width 20 units each to begin the first class at 199.5, find the class boundaries, limits and marks of the new classes.
- (I.U. M.A Econ., 1983)
- 2.6 a) What is meant by a frequency distribution? Describe briefly the main steps in the preparation of a frequency table from raw data.
 (P.U., B.A/B.Sc., 1983)
- ~ b) Prepare a frequency table for the price data given below, taking 5 units as the width of class-interval.

100	96	92	88	86	84	82	80	78	91
87	83	79	77	75	73	71	69	58	56
73	50	57	55	53	51	48	46	63	59
55	31	49	47	45	43	41	38	34	30
56	44	42	40	38	36	46	53	50	43

- 27 a) Why are frequency distributions constructed? What are the rules to be observed in making a frequency distribution from ungrouped data?
- b) A record was made of the number of absences per day from a factory over 35 days with the following results:

Absentees (x)	0	1	2	3	4	5	6	7
No. of days (f)	3	7	9	6	4	2	1	1

- i) On how many days were there fewer than 4 people absent? $\frac{27}{35}$
- ii) On how many days were there at least 4 people absent? $\frac{10}{35}$
- iii) What is the total number of absences over the whole 35 days? $\frac{28}{35}$

(M.A. Econ. II Semester, 1980)

- 28 a) Describe the steps you would take to construct a frequency distribution.

- b) Tabulate the following marks in a grouped frequency distribution.

74	49	103	95	90	118	52	88	101	96	72	56	64	110	97
59	62	96	82	65	85	105	116	91	83	99	53	76	84	89
77	104	96	84	62	58	66	100	80	54	75	55	99	104	78
86	96	83	57	60	51	114	120	121	92	64	64	63	95	78

- 29 The following data give the index numbers of 100 commodities in a certain year. Form a grouped frequency distribution, taking 5 as class interval.

91	124	189	129	141	143	86	76	118	111
99	99	114	100	88	108	87	101	101	71
63	121	122	111	77	127	61	133	68	7
77	177	110	91	96	96	86	106	119	79
81	127	86	103	79	129	151	39	143	147
90	142	118	94	125	96	99	138	145	113
129	87	165	110	144	91	106	104	97	115
100	117	73	134	106	102	123	106	119	104
101	120	112	138	140	103	96	136	78	83
75	100	113	114	109	116	109	116	104	128

- 30 Arrange the data given below in an array and construct a frequency distribution, using a class interval of 5.00. Indicate the class boundaries and class limits clearly.

79.4	71.6	95.5	73.0	74.2	81.8	90.6	55.9
75.2	81.9	68.9	74.2	80.7	65.7	67.6	82.9
88.1	77.8	69.4	83.2	82.7	73.8	64.2	63.9
68.3	48.6	83.5	70.8	72.1	71.6	59.4	77.6

(B.I.S.E. Lahore, 1972)

- 2.11 The following figures give the number of children born to 30 women:

2	0	1	5	4	3	3	8	3	1
4	3	3	0	5	2	1	4	2	3
5	1	1	6	3	3	3	2	7	3
1	4	2	4	4	4	6	5	10	7
7	3	6	3	3	2	1	9	2	2

Construct an ungrouped frequency distribution of these data.

- 2.12 Count the number of letters in each word of the following passage, hyphenated words also being treated as single words and make a frequency distribution of word length.

"To forgive an injury is often considered to be a sign of weakness; it is really a sign of strength. It is easy to allow oneself to be carried away by resentment and hate into an urge of vengeance, but it takes a strong character to restrain those natural passions. The man who forgives an injury proves himself to be the superior of the man who wronged him, and *your wrongs never turn you into a saint*. Forgiveness may even turn a foe into a friend. So mercy is the noblest form of revenge."

- 2.13 The weights of 50 football players are listed below:

193	240	217	283	268	212	251	263	275	208
230	288	259	225	257	218	243	247	280	234
250	236	277	218	245	218	231	269	224	259
258	231	255	228	242	246	271	249	255	265
269	223	243	219	255	242	238	257	254	284

Make a stem-and-leaf display of the data and convert it to a frequency table with 10 class beginning with 100.

- 2.14 Make a stem-and-leaf plot for the following data. Using 8.8 as the lower limit of the first class and with a width of 1 unit, convert it to a frequency distribution.

9.9	10.2	11.3	12.1	10.7	13.8	10.8
13.6	16.4	11.0	15.3	9.2	12.7	
12.7	11.0	8.0	12.0	11.3	9.7	11.6
10.3	14.1	10.0	9.9	13.4	15.7	11.3
12.1	9.8	15.0	9.1	8.1	12.9	14.0
10.8	13.2	10.5	10.6	12.5	15.1	12.8
10.4	11.2	9.3	11.7	12.7	13.9	16.9
13.4	11.8	10.8	14.2	13.8	9.6	11.9
8.7	14.7	10.9	17.9	11.5	14.7	15.9
11.8	10.6	12.6	12.6	15.7	14.9	9.9

- 2.15 Describe the advantages and disadvantages of diagrammatic representation. Describe the important types of diagrams.

- 2.16 Describe each of the diagrams listed below and give an illustration in each case.
 Bar diagrams, Multiple bar diagrams,
 Pie-diagrams, Pictograms, and Profit and Loss charts.

2.17 Give a description of various graphic and pictorial aids for representing data. Mention particular uses of some methods. (P.U., B.A./B.Sc., 1961)

2.18 Describe briefly the different types of diagrams generally used for presenting statistical data. State advantages and disadvantages of any two of them giving illustrations where possible.

2.19 Represent the following yield per acre data by a bar diagram.

Years	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
Yield Per Acre	5	7	9	6	10	12	8	11	12	10

2.20 Following table gives the birth rates and death rates per thousand of a few countries. Represent them by multiple bar charts.

Country	Birth Rate	Death Rate
India	33	24
Japan	32	19
Germany	16	10
Egypt	44	24
Australia	10	9
New Zealand	18	1
France	21	1
Russia	38	16

Represent the following data by rectangular diagrams showing percentage of income spent by two families on different items of expenditure.

Family-budget of two families

Items of Expenditure	Family A Through Rs. 80	Family B Through Rs. 40
	Actual Expenses	Actual Expenses
Food	Rs. 32	Rs. 20
Clothing	Rs. 20	Rs. 8
Shelter	Rs. 8	Rs. 4
Fuel and Light	Rs. 4	Rs. 2
Miscellaneous	Rs. 16	Rs. 6
Total	Rs. 80	Rs. 40

The following table gives the details of monthly expenditure of three families. Represent the data by a suitable diagram in percentage basis.

Items of Expenditure	Family A (Rs.)	Family B (Rs.)	Family C (Rs.)
Food Articles	43	87	126
Clothing	18	17	25
Recreation	3	10	12
Education	5	9	15
Rent	10	21	17
Miscellaneous	6	19	11

- 2.23 Represent the following data by means of a pictogram:

(a)	Industry	No. of Employees (000)
	Murine	96
	Forest	182
	Mineral	290
	Farm	635

(b)	Year	Production of Van
	1982	2040
	1983	2996
	1984	4319
	1985	6324

2.24

- a) Draw a Pie-diagram and also a Component Bar-diagram for the following data:

Item	Expenditure in Rs.
Food	190
Clothing	64
Rent	100
Medical care	46
Other items	88

- b) Graph the following data showing the areas in millions of square miles of the oceans of the world, using (i) a bar chart, (ii) a pie chart.

Ocean	Pacific	Atlantic	Indian	Antarctic	Arctic
Area	70.8	41.2	28.5	7.6	4.8

- 2.25 a) The area sown in Rabi Crops is as follows. Prepare a Pie-chart.

Wheat	105	lakh acres
Gram	30	lakh acres
Barley	15	lakh acres
Pulses	10	lakh acres
Fodder	25	lakh acres
Other crops	14	lakh acres

- b) Calculate the per cent contribution of each crop to the total Rabi crops.

(P.U., B.A.B.Sc., 1988)

- 3.26 Represent the following data by sub-divided bars drawn on a percentage basis or by a Pie-diagram.

Cost per ton disposed commercially

Particulars	1924	1928
Wages	12.74	7.95
Other costs	5.46	4.51
Royalties	0.56	0.50
Total	18.76	12.96
Sale proceeds per ton	19.91	12.16
Profit (+) or loss (-) per ton	+1.15	0.80

- 3.27 a) Describe the Graphical Methods used in Statistics, giving their respective advantages and disadvantages.
 (P.U., M.A. Econ. 1981)
- b) Draw up a list of rules for the construction of graphs.
 (P.U., B.A. (Optional), 1969)
- 3.28 State the general rules which should be borne in mind in the construction of graphs. Draw a suitable graph of the following time series.

Year	Gross Profit (Rs.)	Expenses (Rs.)
1931	7,900	2,700
1932	5,550	1,700
1933	4,800	1,500
1934	4,200	1,150
1935	6,550	4,300
1936	9,100	5,800
1937	8,500	3,700
1938	7,300	3,900
1939	6,500	4,800
1940	6,200	3,650

- 1.29 Show graphically the following monthly imports and exports of a particular commodity during the year 1960-61. Also show graphically the balance of trade. Imports and exports are given in crores of rupees.

Month	Imports	Exports
April	14	12
May	13	19
June	10	9
July	11	12
August	12	10
September	13	9
October	10	13
November	9	12
December	10	
January	11	12
February	12	10
March	11	13

- 1.30 a) Explain with the help of diagram, the difference between a frequency polygon, histogram and an ogive. (P.U., B.A. (Part-I), 1968)
- b) Construct (i) a Histogram, (ii) a Relative Frequency polygon and (iii) an Ogive for the following frequency distribution of the heights of 100 male students at Jhilmil University, Bahawalpur.

Height (inches)	60-62	63-65	66-68	69-71	72-74
No. of Students	5	18	42	27	8

(U.T. M.A., Econ., 1969)

- 2.31 What is meant by a Histogram? Draw a histogram for the distribution of earnings (Rs.) given below:

Earnings	180-184	185-189	190-194	195-199	200-204	205-209	210-214	215-219
Workers	10	14	30	36	40	26	23	8

Show how you would construct the histogram if the class-intervals were unequal in size. (Engg. University, B.Sc. Final, 1977)

- 2.32 a) Define the statistical term Histogram.
 b) Explain the method of constructing histograms when the class intervals are unequal.

- c) In a savings group, there are 400 members and the number of savings certificates held by them are shown in the following table.

No. of certificate held	No. of members
1 - 50	10
51 - 100	15
101 - 150	30
151 - 200	40
201 - 250	120
251 - 300	100
301 - 350	85

Construct a histogram of the distribution of savings certificates.

(P.U., B.Com., 1961)

- 2.33 Draw a Histogram and a Frequency Polygon for the following distribution.

Degree of Cloudiness	10	9	8	7	6	5	4	3	2	1	0
Frequency	180	150	190	75	55	40	45	65	75	130	220

(P.U., B.A. (Part-I), 1962)

- 2.34 a) Draw a histogram illustrating the following data:

Age (nearest birth day)	20-24	25-29	30-39	40-44	45-49	50-54	55-64
Number of men	1	2	16	22	20	15	14

(P.U., B.A. (Part-I), 1973, 75, 78; M.A. Econ., 1980; P.C.S., 1971)

- b) Given the following frequency distribution of the heights in centimeters of 1,000 students, draw a histogram showing the class intervals.

155-157, 158-160, 161-163, 164-166, 167-169, 170-172, 173-175, 176-178, 179-181, 182-184, 185-187, 188-190, 191-193, 194-196, 197-199, 200-202, 203-205, 206-208, 209-211, 212-214, 215-217, 218-220, 221-223, 224-226, 227-229, 230-232, 233-235, 236-238, 239-241, 242-244, 245-247, 248-250, 251-253, 254-256, 257-259, 260-262, 263-265, 266-268, 269-271, 272-274, 275-277, 278-280, 281-283, 284-286, 287-289, 290-292, 293-295, 296-298, 299-301, 302-304, 305-307, 308-309, 310-311, 312-313, 314-315, 316-317, 318-319, 320-321, 322-323, 324-325, 326-327, 328-329, 330-331, 332-333, 334-335, 336-337, 338-339, 340-341, 342-343, 344-345, 346-347, 348-349, 350-351, 352-353, 354-355, 356-357, 358-359, 360-361, 362-363, 364-365, 366-367, 368-369, 370-371, 372-373, 374-375, 376-377, 378-379, 380-381, 382-383, 384-385, 386-387, 388-389, 390-391, 392-393, 394-395, 396-397, 398-399, 399-400.

(P.U., B.A./B.Sc., 1971)

- c) What is a cumulative frequency curve? How does it differ from the ordinary curve?

(P.U., M.A. Econ., 1985)

- d) Describe the common types of frequency curves. Indicate their shapes.

Construct an Ogive from the following table.

Weights	Frequency
118-126	3
127-135	5
136-144	9
145-153	12
154-162	5
163-171	4
172-180	2

(P.U., B.A./B.Sc., 1967)

- 1.37 Pupils were asked how long it took them to walk to school on a particular morning. A cumulative frequency distribution was formed.

Time taken (minutes)	< 5,	< 10,	< 15,	< 20,	< 25,	< 30,	< 35,	< 40,	< 45
Cum. Frequency (f^C)	28	45	81	143	289	349	374	395	430

- a) Draw a cumulative frequency curve and estimate how many pupils took less than 35 minutes.
 - b) 6% of the pupils took x minutes or longer. Find x .
 - c) Take equal class-intervals of 0-, 5-, 10-, etc., construct a frequency distribution and draw a histogram.
- 2.38 Describe in detail the method of drawing Ratio Charts and explain their uses in economic statistics. (P.U., M.A. Econ., 2007)
- 2.39 Explain what is meant by a ratio chart and discuss its advantages over the natural log diagram. Describe and illustrate two practical applications of a ratio chart.
- 2.40 Toss five coins together and note the number of heads. Do this 64 times and count the number of times that $x = 0, 1, 2, 3, 4, 5$. Construct a frequency polygon and an Ogive to represent these results.
- 2.41 a) Distinguish between primary and secondary data. Describe briefly the methods of collecting primary data.
 b) Find the missing entries in the following frequency distribution table:

Class Limits	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Percentage
8 to -	-	-	-	25
- to -	-	0.05	-	-
- to -	-	-	9	-
- to -	-	0.30	15	-
- to 32	-	-	-	-

(P.U., B.A./B.Sc., 2007)

- 2.42 The following data give the annual earnings (rounded to thousands of dollars) of 20 households.

30	5	29	7	30	31	28	10
17	24	25	7	22	15	15	22
78	79	99	30	80	80	75	63
24	55	25	90	35	33	70	63
35	17	9					

- i) Prepare a stem-and-leaf display for these data.
- ii) Condense the stem-and-leaf display by grouping the stems as 0-2, 3-5 and 6-9.

(P.U., B.A./B.Sc., 2007)



CHAPTER 3

MEASURES OF CENTRAL TENDENCY OR AVERAGES

MEASURES OF CENTRAL TENDENCY OR AVERAGES

3.1 INTRODUCTION

For practical purposes, the condensation of data set into a frequency distribution and the visual presentation are not enough, particularly, when two or more different data sets are to be compared. A data set can be summarized in a single value. Such a value, usually somewhere in the centre and representing the entire data set, is a value at which the data have a tendency to concentrate. The tendency of the observations to cluster in the central part of the data set is called *Central Tendency* and the summary value is a *measure of central tendency*. Since a measure of central tendency indicates the location or general position of the distribution of the data set in the range of observations, it is also known as a *measure of location or position*. The measures of central tendency or location are generally known as *Averages*. But in everyday language, 'the average' is often understood to refer to the arithmetic mean (a form of average to be discussed in section 3.4); it is for this reason that when anyone speaks of 'the average' (without qualification) of a set of observations, it may, as a rule, be assumed that the arithmetic mean is meant. The use of the term *average* has been traced to the time of Pythagoras (570–500 B.C.). Two points should be noted. First, a measure of central tendency should be somewhere within the range of the data, and secondly, it should remain unchanged by a rearrangement of the observations in a different order.

Since the late nineteenth century, the practice has been to make a distinction between a sample and a population from which the sample is drawn, by using Latin letters for numerical quantities describing the sample and Greek letters for corresponding quantities characterizing the population. It should be noted that population parameters are rarely calculated directly as all observations from the population are not usually available. The measures corresponding to population parameters are generally calculated from sample data and are regarded as the *estimates* of population parameters.

3.2 CRITERIA OF A SATISFACTORIAL AVERAGE

Several types of averages are defined to measure the representative or "typical" value of a set of data or distribution. It is therefore desired that an average should be

- (i) rigorously defined,
- (ii) based on all the observations made,
- (iii) simple to understand and easy to interpret,
- (iv) quickly and easily calculated,
- (v) amenable to mathematical treatment,
- (vi) relatively stable in repeated sampling experiments, and
- (vii) not unduly influenced by abnormally large or small observations.

An average that possesses all or most of the conditions stated above, is considered a *satisfactory average*.

3.3 TYPES OF AVERAGES

The most common types of averages are (i) the arithmetic mean or simply the mean, (ii) the geometric mean, (iii) the harmonic mean, (iv) the median and (v) the mode. The first three types are numerical in character and give an indication of the magnitude of the observed values. The fourth type indicates the middle position while the last provides information about the most frequent value in the distribution or the data set.

3.4 THE ARITHMETIC MEAN

The arithmetic mean or simply the mean is the most familiar average. It is defined as a value obtained by dividing the sum of all the observations by their number, that is

$$\text{Mean} = \frac{\text{Sum of all the observations}}{\text{Number of the observations}}$$

The mean may correspond to either a population or a sample from the population. If the given set of observations represents a population, the mean is called the population mean which is traditionally denoted by μ (the Greek letter mu). Thus the population mean of a set of N observations x_1, x_2, \dots, x_N is given as

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum x_i}{N}, \quad (i = 1, 2, \dots, N)$$

where Σ , the Greek capital sigma, is a convenient symbol for summation.

If, instead, the given set of observations represents a sample, the mean is called the sample mean usually denoted by placing a bar over the symbol used to represent the observations or the variable. Thus the mean of a set of n observations x_1, x_2, \dots, x_n is defined as

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x_i}{n}, \quad (i = 1, 2, \dots, n)$$

where \bar{x} is the mean of a sample of size n .

It is worthwhile to note that the population mean is a fixed quantity, whereas \bar{x} , the sample mean is a variable because different samples from the same population tend to have different means.

In order to interpret the meaning of the arithmetic mean, let x_i denote the marks obtained by the i th student in a class. Then $\sum x_i$ stands for the total marks obtained by all students and \bar{x} , the mean, represents the number of marks the child *could have obtained by each student if everyone in the class had obtained the same number of marks*. Geometrically the mean represents a point at which the distribution of the set of observations would balance.

Example 3.1 The marks obtained by 9 students are given below:

$$45, 32, 37, 46, 39, 36, 41, 48, 36$$

Calculate the arithmetic mean.

The mean is given by

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} \\ &= \frac{45 + 32 + 37 + 46 + 39 + 36 + 41 + 48 + 36}{9} \\ &= \frac{360}{9} = 40 \text{ marks}\end{aligned}$$

It is relevant to note that, if these marks represent the entire set of observations for the population, the above calculation gives the population mean, i.e. μ , would equal to 40 marks.

3.4.1 The Weighted Arithmetic Mean. The multipliers or a set of numbers which express more or less adequately the relative importance of various observations in a set of data are technically called the weights. We assign weights w_1, w_2, \dots, w_n to the observations in a set of data according to their relative importance, when the observations are not of equal importance. The weighted mean, denoted by \bar{x}_w , of a set of n values x_1, x_2, \dots, x_n with corresponding weights w_1, w_2, \dots, w_n is then defined as

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} \quad (i=1, 2, \dots, n)$$

A weighted average is generally employed in the calculation of index numbers, birth and death rates, etc.

Example 3.2 Calculate the weighted mean from the following data:

Item	Expenditure (Rs.)	Weights
Food	290	7.5
Rent	54	2.0
Clothing	98	1.5
Fuel and Light	75	1.0
Other items	75	0.5

(P.U., B.A. (Optional), 1969, 94)

We calculate the weighted mean as below:

Item	Expenditure (x_i)	Weights (w_i)	$x_i w_i$
Food	290	7.5	2175.0
Rent	54	2.0	108.0
Clothing	98	1.5	147.0
Fuel and Light	75	1.0	75.0
Other items	75	0.5	37.5
Total	—	12.5	2542.5

$$\text{Hence } \bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} = \frac{2542.5}{12.5} = \text{Rs. } 203.4$$

3.4.2 Properties of the Arithmetic Mean. The arithmetic mean has the following four properties:

- For a set of data, the sum of the deviations of the observations x_i 's from their mean, \bar{x} , taken with their proper signs, is equal to zero.

$$\begin{aligned} \text{The sum of the deviations} &= \sum (x_i - \bar{x}) \\ &= \sum x_i - n\bar{x} \quad (\because \bar{x} \text{ is constant}) \\ &= \sum x_i - \sum \bar{x}_i = 0 \quad (\because \bar{x} = \sum x/n) \end{aligned}$$

- ii) The sum of squared deviations of the x_i 's from the mean, \bar{x} , is minimum. In other words, $\sum (x_i - \bar{x})^2 \leq \sum (x_i - a)^2$, where a is an arbitrary value other than the mean.

$$\begin{aligned} \text{Now } \sum (x_i - a)^2 &= \sum (x_i - \bar{x} + \bar{x} - a)^2 \\ &= \sum [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - a) + (\bar{x} - a)^2] \\ &= \sum (x_i - \bar{x})^2 + 2(\bar{x} - a) \sum (x_i - \bar{x}) + n(\bar{x} - a)^2 \\ &= \sum (x_i - \bar{x})^2 + n(\bar{x} - a)^2 \quad [\because \sum (x_i - \bar{x}) = 0] \end{aligned}$$

It is obvious that $\sum (x_i - a)^2 \geq \sum (x_i - \bar{x})^2$ by $n(\bar{x} - a)^2$. The equality sign holds only when $\bar{x} = a$.

Hence $\sum (x_i - \bar{x})^2$ is always less than $\sum (x_i - a)^2$ if $a \neq \bar{x}$.

This property is usually called the *minimum property* of the mean.

- iii) If k subgroups of data consisting of n_1, n_2, \dots, n_k , ($\sum n_i = n$) observations have respective means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$, then \bar{x} , the mean of all the data, is given by

$$\begin{aligned} \bar{x} &= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n_1 + n_2 + \dots + n_k} \\ &= \frac{\sum n_i \bar{x}_i}{n} \quad (i = 1, 2, \dots, k) \end{aligned}$$

i.e., a weighted mean of all the subgroup means.

- iv) If $y_i = ax_i + b$ ($i = 1, 2, \dots, n$), where a and b are any two numbers and $a \neq 0$, then $\bar{y} = a\bar{x} + b$.

Now summing over all values of i , we obtain

$$\sum y_i = a \sum x_i + nb$$

Dividing both sides by n , we get

$$\bar{y} = a\bar{x} + b$$

As the equation $y = ax + b$ represents a linear transformation from x to y , this property is usually called the *invariance of the mean under a linear transformation* and it provides the basis for so-called codis-

which refers to the operation of subtracting (or adding) a constant from each observation and then dividing (or multiplying) by another constant for computational convenience.

Example 3.3 The mean heights and the number of students in three sections of a statistics class are given below:

Section	Number of boys	Mean height
A	46	62"
B	37	58"
C	43	61"

Find the overall mean height of 120 boys.

$$n_1 = 46, n_2 = 37, n_3 = 43, \text{ and}$$

$$\bar{x}_1 = 62", \bar{x}_2 = 58", \bar{x}_3 = 61"$$

The mean height of the combined class is given as

$$\begin{aligned}\bar{x} &= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3}{n_1 + n_2 + n_3} \\ &= \frac{(46 \times 62) + (37 \times 58) + (43 \times 61)}{46 + 37 + 43} = \frac{7249}{120} = 60\end{aligned}$$

3.4.3 Mean From Grouped Data: When the number of observations is very large, the data are arranged into a frequency distribution, which is used to calculate the approximate values of descriptive measures. As the identity of the observations is lost, to calculate the approximate value of the mean, the observations in each class are assumed to be identical with the class midpoint so that the product of the midpoint by the number of observations in a frequency, would be approximately equal to the sum of the products for each class. Thus, if a frequency distribution has k classes with midpoints x_1, x_2, \dots, x_k and corresponding frequencies f_1, f_2, \dots, f_k ($\sum f_i = n$), the mean is then given by the formula

$$\begin{aligned}\bar{x} &= \frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{f_1 + f_2 + \dots + f_k} \\ &= \frac{\sum f_i x_i}{n}, \quad (i = 1, 2, \dots, k)\end{aligned}$$

where f_i indicates the number of times an observation is to be counted, the mean calculated from a frequency distribution may also be regarded as the weighted mean where each class midpoint x_i , takes the average value of the observations in that class, it is weighted by the respective frequency f_i and the sum of the weighted products is divided by the sum of frequencies, i.e. weights.

Sometimes, there may be a slight difference in the values of \bar{x} on account of errors caused by the assumption that all observations in any class may be treated as approximately the midpoint of that class, which once tells us that this error is usually small and never serious. The following example illustrates this.

Example 3.4 Calculate the mean weight of apples from the data given in Example 2.2 (a) from the observed values and from the data grouped into a frequency distribution.

The calculations are outlined below:

Weight (gram)	Sum of actual observations	f_i	Mean Weight of each class (\bar{x}_i)	$f_i \bar{x}_i$	Midpoints (x_i)	$f_i x_i$
65 - 84	895	9	77.2	694.8	74.5	670.5
85 - 104	947	10	94.7	947.0	94.5	945.0
105 - 124	1919	17	112.9	1919.3	114.5	1946.5
125 - 144	1325	10	132.5	1325.0	134.5	1345.0
145 - 164	766	5	153.2	766.0	154.5	722.5
165 - 184	716	4	179.0	716.0	174.5	698.0
185 - 204	956	5	191.2	956.0	194.5	972.5
Total	7324	60	-	7324.1	-	7350.0

Calculation based on Ungrouped data

We calculate the mean weight, \bar{x} , directly from all the observed values, which add to 7324. The second column consists of subtotal of actual observations in any class).

$$\therefore \bar{x} = \frac{\sum x_i}{n}, \quad (i = 1, 2, \dots, 7)$$

$$= \frac{7324}{60} = 122.07 \text{ gram}$$

This is the exact mean of the given data.

Next, we find the mean weight, \bar{x}_i , by multiplying the actual mean of the observations in any class by the corresponding frequency, adding the products and then dividing by n (column 5).

$$\text{Thus } \bar{x} = \frac{\sum f_i \bar{x}_i}{n}, \quad (i = 1, 2, \dots, 7)$$

$$= \frac{7324.1}{60} = 122.07 \text{ gram}$$

Calculation based on Grouped data

Here we calculate the mean weight from grouped data, assuming that all observations in any class are identical with the midpoint of that class. The sixth column consists of class midpoints, x_i , and the products are given in column 7.

$$\text{Then } \bar{x} = \frac{\sum f_i x_i}{n}, \quad (i = 1, 2, \dots, 7)$$

$$= \frac{7350.0}{60} = 122.5 \text{ grams}$$

It should be noted that the numerical value of \bar{x} , calculated from the frequency distribution is slightly different from the value obtained directly from the ungrouped data.

3.4.4 Change of Origin and Scale. To reduce the computational labour and to save time, a change of the origin and scale can be made. If x_i denotes an observed value, a and b are two constants with $b \neq 0$, then the operations $x_i + a$, bx_i , and $bx_i + a$ are known respectively as the *change of origin*, the *change of scale* and both *change of origin and scale*.

Let a be an arbitrary origin, sometimes called *assumed mean*, and let $x_i = a + bu_i$, where b denotes the unit of measurement. Then its corresponding coded value is $u_i = \frac{x_i - a}{b}$.

$$\text{Now } \bar{x} = \frac{1}{n} \sum x_i = \frac{1}{n} \sum (a + bu_i) \\ = \frac{na}{n} + \frac{b \sum u_i}{n} = a + bu$$

Thus the arithmetic mean can be calculated from any origin we may choose and using any scale we want. This transformation is particularly useful for calculations based on grouped data, where b is the width of class interval and a is usually chosen the class midpoint lying in the region of the higher frequencies so that the larger frequencies may be multiplied by smaller values of u . This procedure gives a short method for hand calculations.

Example 3.5 Given the following frequency distribution of weights, calculate the mean weight by short method.

Weight (grams)	65–84	85–104	105–124	125–144	145–164	165–184	185–204
f_i	9	10		10	5	4	5

To calculate the mean weight, let us take $x_i = \frac{x_i - 114.5}{20}$, where $x = 114.5$ is the mid-point belonging to the largest frequency and $b = 20$ is the width of the uniform class interval. The calculations are shown in the table below:

Weight (grams)	Midpoints (x_i)	f_i	u_i	$f_i u_i$
65–84	74.5	9	-2	-18
85–104	94.5	10	-1	-10
105–124	114.5	17	0	-28
125–144	134.5	10	1	10
145–164	154.5	5	2	10
165–184	174.5	4	3	12
185–204	194.5	5	4	20
Total	—	60	—	452 24

Thus $\bar{x} = a + bu$, where $u = \frac{\sum f_i u_i}{n}$

$$= 114.5 + \frac{(24)(20)}{60} = 114.5 + 8.0 = 122.5 \text{ grams.}$$

Example 3.6 Compute the mean for the following frequency distribution of annual death rates.

Death Rate	Frequency
3.5 - 4.4	1
4.5 - 5.4	4
5.5 - 6.4	3
6.5 - 7.4	13
7.5 - 8.4	12
8.5 - 9.4	19
9.5 - 10.4	13
10.5 - 11.4	10
11.5 - 12.4	6
12.5 - 13.4	4
13.5 - 14.4	1
Total	88

The necessary calculations are given below:

Death Rate	Midpoints (x_i)	f_i	$u_i (= x_i - 8.95)$	$f_i u_i$
3.5 - 4.4	3.95	1	-5	-5
4.5 - 5.4	4.95	4	-4	-16
5.5 - 6.4	5.95	3	-3	-9
6.5 - 7.4	6.95	13	-2	-26
7.5 - 8.4	7.95	12	-1	-12
8.5 - 9.4	8.95	19	0	0
9.5 - 10.4	9.95	13	1	13
10.5 - 11.4	10.95	10	2	20
11.5 - 12.4	11.95	6	3	18
12.5 - 13.4	12.95	4	4	16
13.5 - 14.4	13.95	1	5	5
Total	-	88	-	+72
				-2

Hence

$$\bar{x} = a + \frac{\sum f_i u_i}{n}, \text{ where } a \text{ is assumed mean and } h = 1.$$

$$= 8.95 + \frac{(-2)}{88} = 8.95 - 0.02 = 8.93$$

3.5 THE GEOMETRIC MEAN

The geometric mean, G , of a set of n positive values x_1, x_2, \dots, x_n is defined as the positive n th root of their product, i.e.

$$G = \sqrt[n]{x_1 x_2 \dots x_n}, \quad \text{where } x > 0$$

When n is large, the computation of the geometric mean becomes laborious, as we have to multiply all the values and then extract the n th root. The arithmetic is simplified by using logarithms to the base 10. Thus, taking logarithms, we get

$$\log G = \frac{1}{n} [\log x_1 + \log x_2 + \dots + \log x_n]$$

$$= \frac{1}{n} \sum \log x_i$$

Hence $G = \text{antilog} \left[\frac{1}{n} \sum \log x_i \right]$

It means that geometric mean is the anti-logarithm of the arithmetic mean of the logarithms of the values themselves.

For data organized into a grouped frequency distribution, having k classes with class-marks x_1, x_2, \dots, x_k and the corresponding frequencies f_1, f_2, \dots, f_k ($\sum f_j = n$), the formula for the geometric mean is given by

$$G = [x_1^{f_1} \cdot x_2^{f_2} \cdots x_k^{f_k}]^{1/n}$$

In terms of logarithms, the formula becomes

$$\log G = \frac{1}{n} [f_1 \log x_1 + f_2 \log x_2 + \dots + f_k \log x_k]$$

$$= \frac{1}{n} \sum f_i \log x_i$$

or $G = \text{antilog} \left[\frac{1}{n} \sum f_i \log x_i \right]$

The weighted geometric mean of the values x_1, x_2, \dots, x_k with corresponding weights w_1, w_2, \dots, w_k is given by

$$\log G_w = \frac{1}{\sum w_i} [\sum w_i \log x_i]$$

The geometric mean is appropriate to average ratios and rates of change.

Example 3.7 Find the geometric mean of 45, 32, 37, 46, 39, 36, 41, 48 and 36.

The geometric mean, G , is calculated as

$$G = \sqrt[9]{(45 \times 32 \times 37 \times 46 \times 39 \times 36 \times 41 \times 48 \times 36)}$$

Taking log, we have

$$\log G = \frac{1}{9} [\log 45 + \log 32 + \log 37 + \log 46 + \log 39 + \log 36 + \log 41 + \log 48 + \log 36]$$

$$= \frac{1}{9} [1.65321 + 1.5015 + 1.56630 + 1.66276 + 1.49106 + 1.55630 + 1.61278 + 1.68124 + 1.55630]$$

$$\log G = \frac{1}{9} [14.38700] = 1.59856$$

Hence $G = \text{anti-log}(1.59856) = 19.68$

Example 3.8 Given the following frequency distribution of weights, calculate the geometric mean.

Weight (grams)	65-84	85-104	105-124	125-144	145-164	165-184	185-204
f	9	10	17	10	5	4	5

The computation of the geometric mean is shown below:

Weight (grams)	x_i	f_i	$\log x_i$	$f_i \log x_i$
65 - 84	74.5	9	1.8722	16.8498
85 - 104	94.5	10	1.9754	19.7540
105 - 124	114.5	17	2.0589	35.0013
125 - 144	134.5	10	2.1177	21.2870
145 - 164	154.5	5	2.189	10.9445
165 - 184	174.5	4	2.2418	8.9672
185 - 204	194.5	5	2.2889	11.4445
Σ	-	60	-	124.2483

$$\log G = \frac{1}{n} \sum f_i \log x_i = \frac{124.2483}{60} = 2.0708$$

Hence $G = \text{Anti-log}(2.0708) = 117.7 \text{ grams}$

3.6 THE HARMONIC MEAN

The harmonic mean, H , of a set of n values x_1, x_2, \dots, x_n is defined as the reciprocal of the arithmetic mean of the reciprocals of the values. In symbols,

$$H = \text{Reciprocal of } \frac{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}{n}$$

$$= \frac{n}{\sum \frac{1}{x_i}}, \quad \text{where } x \neq 0$$

The harmonic mean is an appropriate type to be used in averaging certain kinds of ratios or rates of change. To illustrate this formula, let us take an example. Suppose a car is running at the rate of 15 km during the first 30 km; at 20 km/hr during the second 30 km; and at 25 km/hr during the third 30 km. Then

distance is constant but the times are variable. Therefore, the harmonic mean is the correct average. In this case, the harmonic mean is:

$$\begin{aligned}
 H &= \text{Reciprocal of } \frac{\frac{1}{15} + \frac{1}{20} + \frac{1}{25}}{3} \\
 &= \frac{3}{\frac{1}{0.06667} + \frac{1}{0.5000} + \frac{1}{0.4000}} \\
 &= \frac{3}{0.15667} = 19.15 \text{ km/hr approximately.}
 \end{aligned}$$

Care should be exercised to apply the harmonic mean. The following rule will help determine the nature of the harmonic mean.

"When rates are expressed as x per y , and x is constant, the harmonic mean is required; but if y is constant, the arithmetic mean is required."

For data organised into a frequency distribution having k classes with classmarks x_1, x_2, \dots, x_k and corresponding frequencies f_1, f_2, \dots, f_k ($\sum f_i = n$) the harmonic mean of the distribution is given by

$$\begin{aligned}
 H &= \text{Reciprocal of } \frac{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \dots + \frac{f_k}{x_k}}{f_1 + f_2 + \dots + f_k} \\
 &= \frac{n}{\sum f_i \left(\frac{1}{x_i} \right)}
 \end{aligned}$$

Similarly, the weighted harmonic mean is defined as

$$\begin{aligned}
 H_w &= \frac{w_1 + w_2 + \dots + w_k}{w_1 \left(\frac{1}{x_1} \right) + w_2 \left(\frac{1}{x_2} \right) + \dots + w_k \left(\frac{1}{x_k} \right)} \\
 &= \frac{\sum w_i}{\sum w_i \left(\frac{1}{x_i} \right)}
 \end{aligned}$$

Example 3.9 Find the harmonic mean from the following frequency distribution of weights:

Weight (grams)	65-84	85-104	105-124	125-144	145-164	165-184	185-204
f	9	10	17	10	5	4	5

We calculate the harmonic mean as below:

Weight (grams)	x_i	f_i	$f_i \left(\frac{1}{x_i} \right)$
65 - 84	74.5	9	0.12081
85 - 104	94.5	10	0.10582
105 - 124	114.5	17	0.14847
125 - 144	134.5	10	0.07435
145 - 164	154.5	5	0.03236
165 - 184	174.5	4	0.02292
185 - 204	194.5	5	0.02571
Σ	-	60	0.53044

Hence $H = \frac{n}{\sum f_i \left(\frac{1}{x_i} \right)} = \frac{60}{0.53044} = 113.11$ grams

Example 3.10 Compute the Geometric and Harmonic means for the following distribution of annual death rates:

x_i	3.95	4.95	5.95	6.95	7.95	8.95	9.95	10.95	11.95	12.95	13.95
f_i	1	4	1	13	12	19	13	10	6	4	1

(B.I.S.E. Lahore 1997)

We can construct the following table to compute the geometric and harmonic means:

x_i	f_i	$\log x_i$	$f_i \log x_i$	$\frac{1}{x_i}$	$f_i \left(\frac{1}{x_i} \right)$
3.95		0.59660	0.59660	0.25316	0.25316
4.95	4	0.69461	2.77844	0.20202	0.20202
5.95	1	0.77452	0.77452	0.16807	0.16807
6.95	13	0.84198	10.94574	0.14388	1.87044
7.95	12	0.90037	10.80444	0.12579	1.50948
8.95	19	0.95182	18.08458	0.11173	2.12287
9.95	13	0.99782	12.97166	0.10050	1.30650
10.95	10	1.03945	10.39450	0.09132	0.91220
11.95	6	1.07740	6.46440	0.08348	0.50208
12.95	4	1.11229	4.44916	0.07722	0.38888
13.95	1	1.14459	1.14459	0.07168	0.07168
Total	88	-	82.50671	-	10.50671

Now $\log G = \frac{1}{n} \sum f_i \log x_i = \frac{82.50671}{88} = 0.93758$

Hence $G = \text{anti-log} (0.9375) = 8.66$, and

$$\text{Harmonic mean} = \frac{n}{\sum f_i \left(\frac{1}{x_i} \right)} = \frac{88}{10.50672} = 8.38$$

3.7 THE MEDIAN

The median is defined as a value which divides a data set that have been ordered, into two equal parts, one part comprising of observations greater than and the other part smaller than it. Or more precisely, the median is a value at or below which 50% of the ordered data lie.

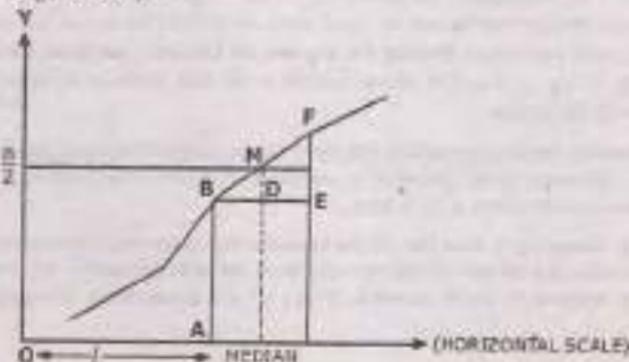
Thus the sample median of the n observations x_1, x_2, \dots, x_n , when arranged in order from smallest to largest, is the middle value if n is odd, and the average of two middle values if n is even. Stated differently, when $\frac{n}{2}$ is not an integer, the median is $\left(\frac{n+1}{2}\right)$ th observation, and when $\frac{n}{2}$ is an integer, the median is the average of $\frac{n}{2}$ th and $\left(\frac{n}{2} + 1\right)$ th observations.

The median in case of a discrete or ungrouped frequency distribution can be found as above by forming a cumulative frequency distribution.

For data grouped into a frequency distribution, the median is a value or a point on the horizontal axis through which a vertical line divides the histogram of the distribution into two parts of equal area. In other words, the median is that value on the horizontal scale which corresponds to a cumulative frequency $\frac{n}{2}$. This value would lie in a certain group, called the median group, but a single value for the median is often desirable. To obtain this single value, we assume that the observations are evenly distributed within the median group. This may be obtained as follows:

Let us consider a relevant portion of the cumulative frequency polygon as drawn below. Then the median is the abscissa of the point. That is

$$\text{Median} = OA = BD \text{ (see figure below).}$$



Since $\triangle DM$ and $\triangle BEF$ are similar triangles; therefore $\frac{BD}{BE} = \frac{DM}{EF}$

$$\text{or } BD = BE \cdot \frac{DN}{EF}$$

Now evidently BE is the width of the class-interval containing median and hence is equal to $\frac{n}{2}$.

$DM = \frac{n}{2} - AB$, where AB represents the cumulative frequency corresponding to the group preceding the median group. Let it be equal to C . Then $DM = \frac{n}{2} - C$.

EF is the difference between two cumulative frequencies, which is clearly the frequency corresponding to the median group and is denoted by f . $OA = l$ (which is the lower boundary of median group), then substituting these values, we get the following formula

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

This process of determining the median is called *arithmetic interpolation* and it does not require uniform class interval. If this arithmetical process is not used, but the value of median (on the X -axis) corresponding to a cumulative frequency $\frac{n}{2}$ is read directly from the graph of Ogive curve, the process is called the *location of the median graphically*. The median is an average of position. It is also known as partition value. The population median may be read in the same way from all the observations of a population.

3.7.3 Quantiles. When the number of observations is quite large, the principle according to which a distribution or an ordered data set is divided into two equal parts, may be extended to any number of divisions. The three values which divide the distribution into four equal parts, are called the *Quartiles*. These values are denoted by Q_1 , Q_2 and Q_3 , respectively. Q_1 is called the *first or lower quartile* and Q_3 is known as the *third or upper quartile*. In other words, the quartiles Q_1 , Q_2 and Q_3 are the values of X below which lie respectively, the lowest 25, 50 and 75 per cent of the data. Similarly, the nine values which divide the distribution into ten equal parts, are called *Deciles* and are denoted by D_1 , D_2 , ..., while the ninety nine values dividing the data into one hundred equal parts, are called *Percentiles* and are denoted by P_1 , P_2 , ..., P_{99} . The second quartile or the fifth decile or the fiftieth percentile is obviously identical with the median.

Quantiles, deciles, percentiles and other values obtained by equal subdivision of the given set of data, are collectively called *Quantiles* or sometimes *Fractiles*. The Quantiles should be calculated when the number of observations is quite large.

It is interesting to note that all the Quantiles are percentiles. For example, the 3^{rd} quartile is the 75^{th} percentile and the 6^{th} decile corresponds to the 60^{th} percentile. We therefore use the following formula to compute P_j , the j^{th} percentile from a set of n observations, arranged in order from smallest to largest.

i) When $\frac{J\pi}{100}$ is not an integer, the j th percentile is given as

P_j = Observation with ordinal number $\left(\frac{J\pi}{100}\right) + 1$; and

ii) When $\frac{J\pi}{100}$ is an integer, the j th percentile is

P_j = Average of two observations with ordinal numbers

$$\left(\frac{J\pi}{100}\right) \text{ and } \left(\frac{J\pi}{100} + 1\right)$$

where $[x]$ stands for the largest integer in x .

In case of grouped data, Quantiles are calculated in the same way as the median.

Sir Francis Galton (1822–1911) is considered the originator of the terms *median* and *percentiles*, although it was Gauss (1777–1855) who actually suggested *median*.

Example 3.11 Given below are the marks obtained by 9 students:

45, 32, 37, 46, 39, 36, 41 and 36.

Find the median and the quartiles.

To find the median and the quartiles, we first arrange the marks in order from lowest to highest. The ordered marks are;

32, 36, 36, 37, 39, 41, 45, 46, 48.

Hence $n = 9$ and $\frac{n}{2}$, i.e., $\frac{9}{2}$ is not an integer; therefore

Median = Marks obtained by $\left[\frac{n}{2}\right] + 1$ th student

= Marks obtained by $(4 + 1)$, i.e. 5th student in ordered data,

= 39 marks

Q_1 = Marks obtained by $\left(\frac{n}{4}\right)$ th student as $\frac{n}{4}$ is not an integer.

= Marks obtained by $\left(\frac{9}{4}\right)$ th, i.e. 3rd student.

= 36 marks.

Similarly,

Q_1 = Marks obtained by $\left(\frac{3n}{4} + 1\right)$ th student

= Marks obtained by $\left(\frac{27}{4} + 1\right)$ th, i.e. 7th student

= 45 marks.

Example 3.12 The following distribution relates to the number of assistants in 50 establishments.

No. of assistants	0	1	2	3	4	5	6	7	8	9
f	3	4	6	7	10	6	5	3	3	1

Find the median number of assistants. Also calculate the quartiles and the 75th decile.

This is an example of ungrouped frequency distribution with unit class interval. To locate the median, the quartiles and the 75th decile for such a distribution, we cumulate the frequencies as shown in the table.

No. of assistants (x)	0	1	2	3	4	5	6	7	8	9
Frequency	3	4	6	7	10	6	5	3	3	1
Cumulative Frequency	3	7	13	20	30	36	41	46	49	50

Since $\frac{n}{2}$, i.e. $\frac{50}{2}$ is an integer, therefore the median is the average number of assistants

$\left(\frac{n}{2}\right)$ th and $\left(\frac{n}{2} + 1\right)$ th, i.e., 25th and 26th establishment. Looking at the cumulative frequency we find that these two values correspond to the same value of x, i.e. 4.

Hence median number of assistants = 4.

For Q_3 , we see that $\frac{3n}{4}$, i.e. $\frac{30}{4}$ is not an integer, therefore

Q_3 = No. of assistants in $\left(\frac{3n}{4} + 1\right)$ th establishment.

= No. of assistants in $(12 + 1)$, i.e. 13th establishment

= 2 assistants.

Similarly,

Q_1 = No. of assistants in $\left(\frac{3n+30}{4} + 1\right)$ th establishment as $\frac{30}{4}$ is also not an integer.

= No. of assistants in 38th establishment.

= 6 assistants.

again $D_2 = \text{Average number of assistants} = \frac{7 \times 50}{10} = 35$ and $\left(\frac{7 \times 50}{10} + 1\right) = 36$ establishment in $\frac{70}{10} = 7$ is an integer.

= Average number of assistants in 35th and 36th establishments

= 5 assistants (since both values correspond to 5)

Example 3.13 Find the median, the quartiles and the 8th decile for the distribution of examination marks given below:

Marks	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Number of students	3	87	190	304	211	85	20

(P.U., B.A.B.Sc., 1970)

To find the median marks, quartiles, etc. by the process of linear interpolation, the data are to be continuous. Thus we need the class boundaries as the cumulative frequencies correspond to upper class boundaries, i.e. to 39.5, 49.5, etc., and not to 39, 49, etc., the upper class limit. Hence the boundaries are shown in the first column and the last column of the table below consists of frequencies.

Class-boundaries (marks)	Midpoints (x_i)	Frequencies (f_i)	Cumulative frequency (F)
29.5 - 39.5	34.5		8
39.5 - 49.5	44.5	87	95
49.5 - 59.5	54.5	190	285
59.5 - 69.5	64.5	304	589
69.5 - 79.5	74.5	211	800
79.5 - 89.5	84.5	85	885
89.5 - 99.5	94.5	20	905

Median = Marks obtained by $\left(\frac{n}{2}\right)^{\text{th}}$ student

= Marks obtained by $\frac{505}{2}$, i.e. 452.5th student which corresponds to marks in the class 59.5 - 69.5. Therefore

$$= l + \frac{f}{f} \left(\frac{n}{2} - C \right) \quad \text{where the letters have their usual significance.}$$

$$= 59.5 + \frac{10}{304} (452.5 - 285)$$

$$= 59.5 + \frac{1675}{304} = 59.5 + 5.5 = 65 \text{ marks.}$$

And $Q_3 = \text{Marks of } \left(\frac{3n}{4} \right) \text{th student}$

= Marks of $\frac{905}{4}$, i.e. 226.25th student which corresponds to a value in the class 49.5 - 59.5. Therefore:

$$Q_3 = l + \frac{h}{f} \left(\frac{3n}{4} - C \right) = 49.5 + \frac{10}{190} (226.25 - 95)$$

$$= 49.5 + 6.4 = 56.4 = 56 \text{ marks}$$

Again $Q_3 = \text{Marks of } \left(\frac{3n}{4} \right) \text{th student}$

= Marks of $\frac{3 \times 905}{4}$, i.e. 678.75th student which lies in the class 69.5 - 79.5

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

$$= 69.5 + \frac{10}{211} (678.75 - 589) = 69.5 + 4.2 = 74 \text{ marks}$$

And $D_8 = \text{Marks of } \left(\frac{8n}{10} \right) \text{th student}$

= Marks of $\frac{8 \times 905}{10}$, i.e. 724th student which also lies in the class 69.5 - 79.5

Hence $D_8 = l + \frac{h}{f} \left(\frac{8n}{10} - C \right)$

$$= 69.5 + \frac{10}{211} (724 - 589) = 69.5 + 6.4 = 76 \text{ marks}$$

3.8 THE MODE

The French word mode meaning fashion, has been adopted to convey the idea of "most frequent". The mode is defined as a value which occurs most frequently in a set of data, that is it indicates the most common result. A set of data may have more than one mode or no mode at all when each observation occurs the same number of times.

In an ungrouped frequency distribution with classes consisting of single values, the mode can be immediately located by examining the distribution. For example, in the distribution relating to the number of assistants in 50 retail establishments (Example 3.12) the mode is 4, as the frequency for $x = 4$ is greater than for any other value of x .

When the data are organised into a grouped frequency distribution, the mode would lie in the class that carries the highest frequency. This class is called the *modal class*. For most practical purposes, it is sufficient to take the midpoint of the modal class as the mode but generally it is a poor approximation. It therefore becomes desirable to decide at what point of the modal class, the mode should be located. To meet this requirement a method based on three adjacent rectangles of the histogram, with the tallest in the middle, has been developed. The method is:

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

where l = lower class boundary of the modal class,

f_m = frequency of the modal class,

f_1 = frequency associated with the class preceding the modal class,

f_2 = frequency associated with the class following the modal class, and

h = width of class interval.

The mode can also be calculated by the following formula:

$$\text{Mode} = l + \frac{f_2}{f_1 + f_2} \times h,$$

where the letters have their usual meaning. It should be noted that the first formula is more accurate and should be generally used in calculating the mode.

When a frequency distribution is displayed as a smooth curve, the mode is the abscissa of the highest ordinate. A distribution having a single mode, is called a *unimodal* distribution, while a distribution with two or more modes, is called a *bimodal* or *multimodal* distribution. It has no meaning for skewed distributions. It should be remembered that, when a frequency distribution has classes of unequal widths, the modal class is the class with maximum frequency per unit. The mode should be calculated if the frequency distribution has a class that carries more frequencies than the others and this should not be at the extremity of the distribution.

Example 3.14 Calculate the mode for the distribution of examination marks given in Example

The class that carries the highest frequency is 59.5 - 69.5, which is thus the modal class.

Also $l = 59.5$, $f_1 = 190$, $f_2 = 211$, $f_m = 304$ and $h = 10$.

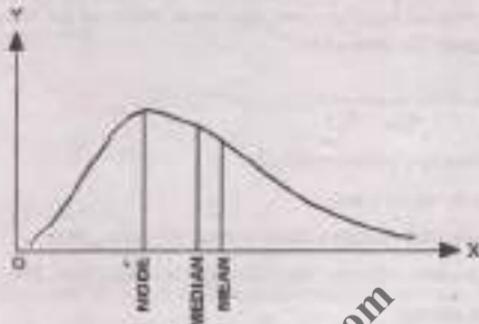
$$\text{Mode} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h,$$

$$= 59.5 + \frac{304 - 190}{(304 - 190) + (304 - 211)} \times 10$$

$$\Rightarrow 59.5 + 5.8 = 65.3 = 65 \text{ marks.}$$

3.9 EMPIRICAL RELATION BETWEEN MEAN, MEDIAN AND MODE

In a single-peaked frequency distribution, the values of the mean, median and mode coincide if the frequency distribution is absolutely symmetrical. But if these values differ, the frequency distribution is said to be skewed or asymmetrical.



Experience tells us that in a unimodal curve of moderate skewness, the median is usually sandwiched between the mean and the mode and between them, the following approximate relation holds good.

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

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

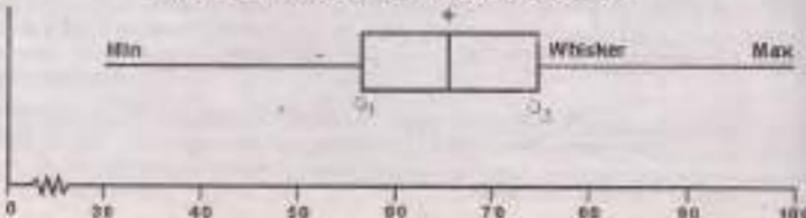
This empirical relation does not hold in case of a J-shaped or an extremely skewed distribution.

3.10 THE BOX PLOTS

The box plots, which are graphically very simple, are based on the Median, a measure of location and the Interquartile Range (IQR), a measure of data's variability. They are informative and effective in comparing two or more data sets distributions.

A box plot is constructed by drawing a rectangle (the box) with the ends (called the hinges) drawn at the lower and upper quartiles (Q_1 and Q_3). The median of the data is shown in the box usually by a '+' sign. The straight lines (called the whiskers) are drawn from each hinge to the most extreme observations. The entire graph is called a Box and Whiskers plot. If one whisker is longer, the distribution is skewed in the direction of the longer whisker. The box plot given below represents the distribution of examination marks given in Example 3.13.

BOX PLOT FOR DATA IN EXAMPLE 3.13



When two or more distributions are to be compared by drawing box plots, the scale of measurement is usually plotted vertically. Sometimes, two sets of limits, called *inner fences* and *outer fences*, are also used.

RELATIVE MERITS AND DEMERITS OF VARIOUS AVERAGES

It is necessary to understand the merits and demerits of each one of the averages in order that it may be appropriately employed.

3.1.1.1 The Arithmetic Mean. The advantages of the mean are:

- It is rigorously defined by a mathematical formula.
- It is based on all the observations in the data.
- It is easy to calculate and simple to comprehend. *understandable*
- It is determined for almost every kind of data.
- It is a relatively stable statistic with the fluctuations of sampling. That is why it is universally used.
- It is amenable to mathematical treatment.

The disadvantages of the mean are:

- It is greatly affected by extreme values in the data.
- It gives sometimes fallacious conclusions.
- In a highly skewed distribution, the mean is not an appropriate measure of average.
- If the grouped data have "open-end" classes, mean cannot be calculated without assuming the limits.

3.1.1.2 The Geometric Mean. The advantages of the geometric mean are:

- It is rigorously defined by a mathematical formula.
- It is based on all observed values.
- It is amenable to mathematical treatment in certain cases.
- It gives equal weightage to all the observations.
- It is not much affected by sampling variability.
- It is an appropriate type of average to be used in case rates of change or ratios are to be averaged.

The disadvantages are:

- It is neither easy to calculate nor to understand.
- It vanishes if any observation is zero.
- In case of negative values, it cannot be computed at all.

3.11.3 The Harmonic Mean. The advantages of the harmonic mean are:

- i) It is rigorously defined by a mathematical formula.
- ii) It is based on all the observations in the data.
- iii) It is amenable to mathematical treatment.
- iv) It is not much affected by sampling variability.
- v) It is an appropriate type for averaging rates and ratios.

The disadvantages of the harmonic mean are:

- i) It is not readily understood.
- ii) It cannot be calculated, if any one of the observations is zero.
- iii) It gives too much weightage to the smaller observations.

3.11.4 The Median. The advantages of the median are:

- i) It is easily calculated and understood.
- ii) It is located even when the values are not capable of quantitative measurement.
- iii) It is not affected by extreme values. It can be computed even when a frequency distribution involves "open-end" classes like those of income and prices.
- iv) In a highly skewed distribution, median is an appropriate average to use.

The median has the following disadvantages:

- i) It is not rigorously defined.
- ii) It is not capable of leading itself to further statistical treatment.
- iii) It necessitates the arrangement of data into an array which can be tedious and time consuming for a large body of data.

3.11.5 The Mode. The advantages of the mode are:

- i) It is simply defined and easily calculated. In many cases, it is extremely easy to locate the mode.
- ii) It is not affected by abnormally large or small observations.
- iii) It can be determined for both the quantitative and the qualitative data.

The disadvantages of the mode are:

- i) It is not rigorously defined.
- ii) It is often indeterminate and indefinite.
- iii) It is not based on all the observations made.
- iv) It is not capable of leading itself to further statistical treatment.
- v) When the distribution consists of a small number of values, the mode may not exist.

EXERCISES

OBJECTIVE

- Answer 'True' and 'False'. If the statement is not true then replace the underlined words with words that make the statement true.
 - i) A measure of central tendency is a quantitative value that tends to locate in some sense the middle of a set of data.
 - ii) The mean of a sample always divides the data into two equal halves – half larger and half smaller.
 - iii) For any distribution, the sum of the deviations from the mean equals zero.
 - iv) The arithmetic mean is not affected by the extreme values.
 - v) A distribution always has exactly one median score, but it can have more than one mode score.
 - vi) The value that occurs most frequently is known as median.
 - vii) The mean may not exist for some sets of data.
 - viii) Another name for the median is the score at the third quartile.
 - ix) The Harmonic mean, the geometric mean and ~~the~~ arithmetic mean are equal only if all the numbers x_1, x_2, \dots, x_n are identical.
 - x) The Harmonic mean is the N^{th} root of the product of the numbers.
 - xi) The first quartile is also referred to as the 25th percentile.
 - xii) If a distribution of scores is symmetric, then the median and the mode will be the same.
 - xiii) If a distribution is positively skewed, then the mean is larger than the median.
 - xiv) If a distribution is skewed to left, generally mean > median > mode.
 - xv) If the mean, median, and mode of a distribution are 5, 6, and 8, respectively, the distribution is positively skewed.

MULTIPLE CHOICE QUESTIONS.

- Half the observations are always larger than the
 - a) Mean
 - b) Total
 - c) Median
 - d) Mode
- The value that occurs most often in a set of data is called the
 - a) Mean
 - b) Mode
 - c) Geometric mean
 - d) Harmonic mean
- In case of an open-end class,
 - a) A median cannot be computed.
 - b) The arithmetic mean and the median will always be exactly equal.
 - c) A mean cannot be computed.
 - d) The distribution is always positively skewed.

- iv) Which of the following is a true statement about the median?
- It is always one of the data values.
 - It is influenced by extreme values.
 - Fifty percent of the observations are larger than the median.
 - It is the middle value of the data values.
- v) Which of the following is not a characteristic of the arithmetic mean?
- It is influenced by extreme values.
 - The sum of the observations from the mean is zero.
 - Fifty percent of the observations are larger than the mean.
 - The sum of the squared deviations from mean is always minimum.
- vi) Find the mean of the following sample of distances of stars from the earth: 18.2, 56.9, 24.6, 13.5
- $\bar{x} = 28.30$
 - $\bar{x} = 43.40$
 - $\mu = 28.30$
 - $\mu = 43.40$
- vii) In a positively skewed distribution, the mean is always
- Smaller than the median
 - Equal to the median
 - Larger than the median
 - Equal to the mode
- viii) The median is larger than the arithmetic mean when
- The distribution is positively skewed.
 - The distribution is negatively skewed.
 - The data is organized into a frequency distribution.
 - The distribution is symmetrical.
- ix) The geometric mean of the numbers 2, 4 and 8 is
- 3.67
 - 4
 - 3.43
 - 5
- x) Which of the following statement is not true for Harmonic Mean?
- Harmonic mean is smaller than the mean.
 - It is based on all the values.
 - It is an appropriate type for averaging rates and ratios.
 - It gives equal weightage to all the values.

SUBJECTIVE

- 3.1 What is a statistical average? Name the important types of averages. Discuss the advantages and disadvantages of each average. (P.U., B.A. (Hons.) 1988)

- 3.2 What is a measure of "central tendency"? What is the purpose served by it? What are its desirable qualities?
- 3.3 What are the principal criteria for a satisfactory average? State giving reasons the circumstances in which it would be preferable to use (i) the mean, (ii) the median (iii) the mode, (iv) geometric mean and (v) harmonic mean.
- 3.4 What criteria do you apply to judge the merits of an average? Discuss the merits and demerits of the different averages in common use with special reference to these criteria.
- 3.5 In what circumstances would you consider the Arithmetic mean, the Geometric mean and the Harmonic mean respectively, the most suitable statistic to describe the central tendency of distributions? (P.U., B.A./B.Sc. 1989)
- 3.6 What are the different measures of central tendency? Describe the manner of computation of any three of them with suitable illustrations. (P.U., M.A. Econ. 1967)
- 3.7 Define weighted average and explain how it differs from simple mean. Give the method of its computation and discuss the use of weighted mean in Statistics. (P.U., M.A. Econ. 1974)
- 3.8 What is the median? What are its advantages and disadvantages? Give reasons why the statistician usually prefers the arithmetic mean to the median. (P.U., M.A. Econ. 1981)
- 3.9 Define, and explain how to compute, the following quantities for a grouped distribution: \bar{x} , Q_1 , Q_3 , D_4 and Mode.
- 3.10 Define the arithmetic mean, the mode and the median. Discuss the relationship of these three measures of location in a skewed distribution. State the chief advantages of the arithmetic mean as a form of average.
- 3.11 Define Mean, Median and Mode. What are their advantages and limitations in the analysis of data? Give various methods of calculating Arithmetic mean, with illustrations. (P.U., B.A./B.Sc. 1958, 1960)
- 3.12 a) Define Mean, Median and Mode. Give an empirical relation between them. Does this relation give correct value for the mode?
b) Criticise the following statements:
i) An average does not reveal all the information about the data.
ii) The median is described as the value of the average rather than the average value.
- 3.13 Comment on the following statements:
i) The depth of a river at four different points is 2, 7, 5, 6 feet respectively. The average depth is 5 feet. Therefore all the people with heights above 5 feet can cross it.
ii) The average marks of one class of students are 30. Therefore every student is hopeless.
iii) The average income of a king and his household servants is £20,000 per month, therefore all the household servants must be fabulously poor.
iv) On an average, the number of accidents occurring in the middle of the road are 5 per thousand. The number of deaths at other places is 30 per thousand. Therefore, it is safer to walk in the middle of the road.
v) In a country, 2,000 vaccinated persons died. Therefore vaccination is useless.

- 3.14 Define the mean, the median and the mode of a frequency distribution. It is commonly true that the median lies between the other two measures and is approximately twice as far from the mode as from the mean. State with reasons, whether you expect this relationship to hold, and which of the three statistics is likely to be the most useful single statistic for each of the following distributions:
- The annual earnings of employed males in Pakistan.
 - The percentage of sky, to the nearest 10 per cent, covered by cloud at Karachi at mid-day.
 - The exact length of rods cut to a standard size by machine.
- (P.C.S., 1971)
- 3.15 a) Define arithmetic mean and describe its properties.
 b) If the arithmetic mean of n numbers x_1, x_2, \dots, x_n is M and A is any arbitrary number, then show that

$$\sum(x_i - A)^2 = \sum(x_i - M)^2 + n(M - A)^2$$
- (P.U., B.A./B.Sc., 1977, 82)
- c) A distribution consists of three components with frequencies 3, 4 and 5, and having means 2, 5.5 and 10. Find the mean of the combined distribution. (P.U., B.A./B.Sc., 1977)
- 3.16 a) State the properties of the arithmetic mean.
 b) Show that $\sum(x_i - \bar{x})^2 = \sum(x_i - \bar{x})^2 + n(\bar{x} - a)^2$. In other words, show that the sum of squares about $a - \bar{x}$ is smallest.
 c) A distribution consists of 3 components with frequencies 45, 40 and 65, having their means 2, 2.5 and 2 respectively. Show that the mean of the combined distribution is 2.1 approximately.
- 3.17 a) The number of cars crossing a certain bridge in a big city in 10 intervals of five minutes each was recorded as follows:
 25, 15, 18, 30, 10, 12, 9, 16, 13
 Calculate (i) the arithmetic mean, (ii) the median and (iii) the geometric mean.
 b) Explain why the mean calculated for a set of ungrouped data might differ from the mean if the same data were grouped into a frequency distribution.
- 3.18 a) Define Arithmetic mean, Geometric mean and Harmonic mean; and prove that for any two positive numbers a and b ,

$$A.M. \geq G.M. \geq H.M.$$
- (P.U., B.A./B.Sc., 1991)
- b) The mainline incomes of ten families in rupees in a certain locality are given below:
 Family: A B C D E F G H I J
 Income (Rs.): 85, 70, 10, 75, 500, 8, 42, 250, 40, 36
 Calculate the arithmetic mean, the geometric mean and the harmonic mean of the above incomes. Which one of the above three averages represents the above figures best?
- 3.19 Calculate the arithmetic mean, the geometric mean and the harmonic mean of the annual incomes of fifteen families as given below:
 Rs. 60, 80, 90, 96, 120, 150, 200, 360, 480, 520, 1060, 1200, 1450, 2500, 7200.

- Ques. 6) In a company having 30 employees, 60 earn Rs.3.00 per hour and 20 earn Rs.2.00 per hour. (i) Determine the mean earnings per hour. (ii) Do you consider this mean hourly wage to be typical? (P.U., B.A./B.Sc. 1980-S)
- b) An examination candidate's percentages are: English, 73; French, 82; Mathematics, 57; Science, 62; History, 60. Find the candidate's weighted mean if weights of 4, 3, 3, 1, 1 respectively are allotted to the subjects.
- Find (i) the simple average of prices in column 2 and (ii) the weighted average, using the quantities in column 3 as weights, and explain the difference between the two results.

(1) Piece goods	(2) Price per metre (Rs.)	(3) Quantity (millions metres)
Unbleached	8.37	286
Bleached	9.50	355
Printed flags	9.16	64
Other sorts	9.84	172
Dyed in piece	13.65	165
Of dyed yarn	11.95	80

The following are the monthly salaries in rupees of 30 employees of a firm.

139	126	134	100	88	62	77	103	108
144	129	148	63	69	148	132	142	116
123	104	95	80	85	106	123	133	140

The firm gave bonuses of Rs. 10, 15, 20, 25, 30 and 35 for individuals in the respective salary groups exceeding 60 but not exceeding 75, exceeding 75 but not exceeding 90 and so on upto exceeding 135 but not exceeding 150. Find the average bonus paid per employee.

(U., M.A., Econ. 1974; B.Z.U. M.A., Econ. 1991)

The following table shows the age distribution of 1,143 horses.

Age (year)	Number of horses (f_i)	Average age (\bar{x}_i)
1 - 4	12	2.7
5 - 9	223	7.6
10 - 14	435	12.0
15 - 19	272	16.3
20 - 24	161	20.8
25 - 29	34	25.8
30 and over	6	31.8

Compute the average age of these horses (a) from the first two columns of the table by the usual short method, (b) from the last two columns by weighting the group averages by the number of horses in the groups. Compare the two results. Which one is more nearly the real average age?

Find the arithmetic and geometric means of the series 1, 2, 4, 8, 16, ..., 2^n . Find also the harmonic mean. (P.U. D. St. 1960)

Find (i) arithmetic mean, (ii) geometric mean, and (iii) harmonic mean of the series 1, 3, 9, 27, 81, ..., 3^n . (P.U., B.A./B.Sc. 1973, 82)

- 3.26 a) Define Geometric mean and describe its advantages and disadvantages.
 b) Given two sets, each of n positive values, $x_{11}, x_{12}, \dots, x_{1n}; x_{21}, x_{22}, \dots, x_{2n}$, prove that the geometric mean of the ratios of corresponding values in the two sets is equal to the ratio of the geometric means of the two sets. (P.U., B.A./B.Sc.)

Hint: Let a ratio be defined as $X = \frac{x_1}{x_2}$.

Then $\log X = \log x_1 - \log x_2$.

Sum for all pairs of x_1 's and x_2 's.

$$\text{Hence } G = \frac{\sum \log x_i}{n}$$

- 3.27 A man gets a rise of 10% in salary at the end of his first year of service, and further rises 20% and 25% at the end of the second and third years respectively. The rise in each case is calculated on his salary at the beginning of the year. To what annual percentage increase is this equivalent?

- 3.28 a) Define Harmonic mean. How does it differ from arithmetic mean? What are its advantages and disadvantages?
 b) A man travels from A to B at average speed of 10 miles per hour and returns from B to A along the same route at an average speed of 15 miles per hour. Find the average speed for the entire journey. (P.U., B.A./B.Sc.)
 c) Find out the average speed of person who rides the first mile at the rate of 8 miles per hour, the next mile at the rate of 10 miles an hour and the third mile at the rate of 12 miles an hour. (P.U., B.A./B.Sc.)

- 3.29 a) A bus traveling 200 kilometers has 10 stages at equal intervals. The speed of the various stages was observed to be 10, 15, 20, 25, 20, 30, 40, 50, 30, 40 kilometers per hour. Find the average speed at which the bus travels.
 b) Find out the average speed of (i) motion in the case of a person who rides the first mile at the rate of 10 miles an hour, the next mile at the rate of 8 miles per hour, and the third mile at the rate of 6 miles per hour; (ii) increase in population, which in the first year has increased by 20%, in the next 25% and in the third 44%. (P.U., B.A. (Part-I))

- 3.30 Find the geometric mean and the harmonic mean of the following frequency distribution:
- | Weekly Income (Rs.) | 35-39 | 40-44 | 45-49 | 50-54 | 55-59 | 60-64 | 65-69 |
|---------------------|-------|-------|-------|-------|-------|-------|-------|
| No. of workers | 15 | 13 | 17 | 29 | 11 | 10 | 5 |
- (P.U., B.A. (Home in Econ.))

- 3.31 Calculate the geometric and the harmonic means for the distribution given below:

Variable	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	2	5	7	13	21	16	8	3

(P.U., M.A. Econ.)

- 3.32 Find the mean or median, whichever you think more suitable, in each of the following:
 i) Salaries of 5 men in an industrial concern: Rs. 950, Rs. 2100, Rs. 1500, Rs. 1000, Rs. 10,000.
 ii) Heights of 6 boys: 64", 65", 65", 66", 66", 67".
 iii) Handicaps of four golfers: 4, 18, 15, 20.

- 3.33 The following data relate to sizes of shoes sold at a store during a given week. Find the median of the shoes. Also calculate the quartiles, the 7th decile and the 64th percentile.

Size of Shoes	5	5½	6	6½	7	7½	8	8½	9	9½
No. of Pairs	2	5	15	30	60	40	23	11	4	1

(P.U., B.A. (Hons.) 1962)

- 3.34 Calculate the Mean, Median and Modal numbers of persons per house from the data:

No. of persons per house	1	2	3	4	5	6	7	8	9	10
No. of houses	26	113	120	95	60	42	21	14	5	4

(P.U., B.A. (Hons.) 1969)

- 3.35 Draw (i) a Histogram and (ii) an Ogive from the following data:

Daily wages (Rs.)	4-6	6-8	8-10	10-12	12-14	14-16
No. of employees	13	111	182	105	19	7

Find approximate value of the median from the Ogive and check your answer by calculation.

(B.I.S.E. Sargodha, 1969-S)

- 3.36 Estimate graphically and by formula the median and quartile ages of head of household from the following distribution:

Age of head (yrs)	Number of households
under 25	44
25 and under 30	
30 and under 40	2
40 and under 50	122
50 and under 60	141
60 and under 65	100
65 and under 70	58
70 and under 75	32
75 and over 85	28

- 3.37 Compute the median and quartiles of the following distribution of heights and check the results on a graph.

Heights (inches)	57.5-	60.0-	62.5-	65.0-	67.5-	70.0-	72.5-
Number	5	26	190	281	412	127	38

(P.U., M.A. Econ. 1969)

- 3.38 Explain when median is more representative than mean. Calculate the median of the following distribution.

Class	Number	Class	Number	Class	Number
100-104	4	124-128	298	150-154	260
105-109	14	130-134	380	155-159	128
110-114	60	135-139	450	160-164	66
115-119	138	140-144	500	165-169	28
120-124	236	145-149	400	170-174	12

(P.U., B.A./B.Sc. 1960)

- 3.39 The frequency distribution of a group of persons according to age is given below:

Age in years	< 1	1-4	5-9	10-19	20-29	30-39	40-59	60-79
No. of persons	5	10	11	12	22	18	8	7

Calculate the Mean and the Median ages of the distribution.

- 3.40 a) Describe the merits and demerits of mean and median.
 b) Calculate the median, the upper and lower quartiles from the following data. Also draw a box plot.

Class-Interval	Frequency
under 25	222
25 - 29	405
30 - 34	508
35 - 39	520
40 - 44	525
45 - 49	490
50 - 54	457
55 - 59	416
60 and over	166

(P.U., B.A./B.Sc., 1946)

- 3.41 The following distribution shows Kilowatt-Hours of Electricity used in one month by 70 residential consumers in a certain locality of Lahore.

Consumption in Kilowatt hours	0-34	35-48	49-62	63-76	77-108	109-128	129-144	145-160
No. of consumers	4	11	18	22	14	5	7	3

Estimate the mean, the median and the two quartiles.

- 3.42 The yields of grain (x lb) from 500 small plots are grouped in classes with a common class interval (0.2 lb) as in the table below, the values of x given being the midvalues of the classes. Show that the mean of the distribution is 3.95 lb; the median is 3.95 lb.; and quartiles are 3.6 lb. and 4.23 lb.

	f		f
3.8	4	4.2	69
3.9	15	4.4	59
3.2	20	4.6	35
3.4	47	4.8	10
3.6	63	5.0	8
3.8	78	5.2	4
4.0	88	Total	500

- 3.43 The weights in milligrams of 2538 seeds of the long leaf pine were as follows:

Weight (milligrams)	Number of Seeds	Weight (milligrams)	Number of Seeds
10 - 24.9	16	85 - 99.9	655
25 - 39.9	68	100 - 114.9	803
40 - 54.9	204	115 - 129.9	294
55 - 69.9	233	130 - 144.9	21
70 - 84.9	280	145 - 159.9	4

MEASURES OF CENTRAL TENDENCY OR AVERAGES

- a) Find the average weight, the median weight and the most common weight (mode) of the seeds.
- b) Find the first and third quartiles. Find the third decile and the 45th percentile.
- c) Explain your answers as you would to a person who had never studied statistics.
- 3.48 In a group of 500 wage-earners, the weekly wages of 4% were under Rs. 60 and those of 15% were under Rs. 62.50. 15% of the workers earned Rs. 95 and over, and 5% of them got Rs. 100 and over.

The median and quartile wages were Rs. 82.25, Rs. 72.75 and Rs. 50.50; the fourth and sixth decile wages were Rs. 78.75 and Rs. 65.25 respectively.

Put the above information in the form of a frequency distribution and estimate the mean wage of the 500 wage-earners therefrom.

How. First put the information in the form of a cumulative frequency table.

- 3.49 a) Describe the advantages and disadvantages of the mean, the median and the mode. Explain the empirical relation between them.
- b) The weight of the 40 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	172-180
Frequency	3	5	9	6	5	4	2

Calculate the mean, median and the mode.

(P.U., B.A./B.Sc. 1969)

The following table shows the distribution of the maximum loads in short tons supported by certain cables produced by a company.

Max. loads (Short ton)	9.8-10.2	10.3-10.7	10.8-11.2	11.3-11.7	11.8-12.2	12.3-12.7
No. of cables	7		17	14	6	4

Determine the mean, the median and the mode.

The following is the distribution of wages per thousand employees in a certain factory.

Wages (Rs.)	22	24	26	28	30	32	34	36	38	40	42	44
No. of employees	2	13	43	102	175	220	204	139	69	23	6	1

Calculate the Modal and Median wages and explain why there is a difference between the two.

- a) Define the mode of a frequency distribution. How does it compare with other types of averages?
- b) Write down the empirical relation between mean, median, and mode for unimodal distributions 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. (P.U., B.A./B.Sc. 1972, 80-5)
- c) For a certain frequency distribution, the mean was 40.5 and median 36. Find the mode approximately using the formula connecting the three.

(P.U., B.A./B.Sc. Optional, 1971-S)

- 3.49 a) What types of averages would be suitable for the following cases? Give reasons.
- i) Size of agricultural holdings.
 - ii) Heights of students.
 - iii) Marks obtained in any examination.
 - iv) Income of workers in a factory.
 - v) Per capita income in Pakistan.
 - vi) Comparison of intelligence.
 - vii) Volumes of sales of ready-made shirts, shoes and collars.
 - viii) Number of petals of flowers.
- b) What measures of central tendency would you recommend for the following cases? Give reasons in support of your answer.
- i) Symmetrical Distribution.
 - ii) A J-shaped Distribution.
 - iii) Distribution having "open-end" classes at the end of the classes.
 - iv) Frequency distribution of a quantitative variable

(P.U., M.A. Econ., 1995)

- 3.50 a) A distribution $x_1, x_2, \dots, x_r, \dots$, with frequencies $f_1, f_2, \dots, f_r, \dots, f_k$ is transformed into the distribution $X_1, X_2, \dots, X_r, \dots, X_k$, with the same corresponding frequencies by the relation $X_r = ax_r + b$, where a and b are constant. Show that the mean, mode and medians of the new distribution are given in terms of those of the first distribution by the same transformation.
- b) A distribution of values of the variable x_1, x_2, \dots, x_k with corresponding frequencies f_1, f_2, \dots, f_k , a new distribution with the same frequencies is formed by the relation $X_r = 2x_r^r - 3$ for values of r ($r = 1, 2, \dots, k$). If the values of the mean, median and mode of the original distribution are \bar{x} , \tilde{x} and c respectively, what are these values of the new distribution?

(P.U., B.A., B.Sc., 1995)



CHAPTER 4

**MEASURES OF
DISPERSION,
MOMENTS AND
SKEWNESS**

MEASURES OF DISPERSION, MOMENTS AND SKEWNESS

4.1 INTRODUCTION

It is quite possible that two or more sets of data may have the same average (mean, median or mode) but their individual observations may differ considerably from the average. Thus a value of central tendency does not adequately describe the data. We therefore need some additional information concerning with how the data are dispersed about the average. This is done by measuring the dispersion by which we mean the extent to which the observations in a sample or in a population vary about their mean. A quantity that measures this characteristic, is called a measure of dispersion, scatter or variability. It is desirable to have the measure of dispersion (i) in the same units as the observations, (ii) zero when all the observations are equal, (iii) independent of origin, (iv) multiplied or divided by a constant. It is also desirable that it should satisfy the conditions similar to those laid down for an average in previous chapter (see section 3.2).

There are two types of measures of dispersion: absolute and relative. An absolute measure of dispersion is one that measures the dispersion in terms of the same units as in the square of units, as the units of the data. For example, if the units of the data are rupees, metres, kilograms, etc., the units of the measures of dispersion will also be rupees, metres, kilograms, etc. A relative measure of dispersion is one that is expressed in the form of a ratio, co-efficient or percentage and is independent of the units of measurement. It is useful for comparison of data of different nature. A measure of central tendency together with a measure of dispersion gives an adequate description of data.

The main measures of dispersion are the following:

- i) The Range.
- ii) The Semi-Interquartile Range or the Quartile Deviation.
- iii) The Mean Deviation or the Average Deviation.
- iv) The Variance and the Standard Deviation.

4.2 THE RANGE

The range R , is defined as the difference between the largest and the smallest observations in a set of data. Symbolically, the range is given by the relation:

$$R = x_m - x_0,$$

where x_m stands for the largest observation and x_0 denotes the smallest one. When the data are grouped in a frequency distribution, the range is estimated by finding the difference between the upper boundary of the highest class and the lower boundary of the lowest class. The range cannot be computed if there are any open-end classes in the distribution.

The range is a simple concept and is easy to compute. It has, however, two serious disadvantages. First, it ignores all the information available from the intermediate observations; and second, as its value is based only on the two extreme (unusually large or small) observations, it might give a misleading picture of the spread in the data. It is therefore an unsatisfactory measure of dispersion. However, it is appropriately used in statistical quality control charts of manufactured products, daily temperatures, stock prices, etc. This is an absolute measure of dispersion. Its relative measure known as the co-efficient of dispersion, is defined by the following relation:

$$\text{Co-efficient of Dispersion} = \frac{x_m - x_0}{x_m + x_0}.$$

This is a pure (i.e. dimensionless) number and is used for the purposes of comparison.

Example 4.1 The marks obtained by 9 students are given below:

$$45, 32, 37, 46, 39, 36, 41, 48, 36.$$

Find the range and the co-efficient of dispersion.

Here the highest marks, i.e. $x_m = 48$,

and the lowest marks, i.e., $x_0 = 32$.

$$R = x_m - x_0 = 48 - 32 = 16 \text{ marks, and}$$

$$\begin{aligned}\text{Co-efficient of Dispersion} &= \frac{x_m - x_0}{x_m + x_0} \\ &= \frac{48 - 32}{48 + 32} = \frac{16}{80} = 0.2\end{aligned}$$

4.3 THE SEMI-INTERQUARTILE RANGE OR THE QUARTILE DEVIATION

The interquartile range is a measure of dispersion, defined by the difference between the third and first quartiles; and half of this range is called the semi-interquartile range (S.I.Q.R.) or the quartile deviation (Q.D.). Symbolically, we have

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

where Q_1 and Q_3 are the first and the third quartiles of the data. The quartile deviation has an attractive feature that the range "Median \pm Q.D." covers approximately 50% of the data. The quartile deviation is superior to range as it is not affected by extremely large or small observations. It is simple to understand and easy to calculate. It has certain advantages. It gives no information about the position of observations lying outside the two quartiles; is not amenable to mathematical treatment and is greatly affected by sampling variability. The quartile deviation is not as widely used as other measures of dispersion. It is, however, used in situations where extreme observations are thought to be unrepresentative.

The quartile deviation is also an absolute measure of dispersion. Its relative measure called the Co-efficient of Quartile Deviation or of Semi-Interquartile Range, is defined by the relation

$$\text{Co-efficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1},$$

which is a pure number and is used for comparing the variation in two or more sets of data.

Example 4.2 Find the quartile deviation and the co-efficient of quartile deviation for (i) the data in Example 3.11 and (ii) the frequency distribution in Example 3.13.

- i) Using the data of Example 3.11, we find that

$$Q_1 = 36 \text{ marks, } Q_3 = 45 \text{ marks, and therefore}$$

$$Q.D. = \frac{45 - 36}{2} = 4.5 \text{ marks}$$

$$\text{Co-efficient of Q.D.} = \frac{45 - 36}{45 + 16} = \frac{9}{81} = 0.11$$

Values of Q_1 and Q_3 calculated in Example 3.13 are respectively 56 and 74 marks. Therefore

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{74 - 56}{2} = 9 \text{ marks, and}$$

$$\text{Co-efficient of Q.D.} = \frac{74 - 56}{74 + 56} = \frac{18}{130} = 0.14$$

— THE MEAN (OR AVERAGE) DEVIATION

The mean deviation ($M.D.$) of a set of data is defined as the arithmetic mean of the deviations either from the mean or from the median, all deviations being counted as positive. The reason for taking the deviations as positive, i.e. to disregard the algebraic signs (+ and -) is to avoid the difficulty from the property that the sum of deviations of the observations from their mean is zero. The definition of the mean deviation from the mean is

$$M.D. = \frac{\sum |x_i - \bar{x}|}{n}, \quad \text{for sample data,}$$

$$M.D. = \frac{\sum |x_i - \mu|}{N}, \quad \text{for population data.}$$

$|\bar{x}|$ and $|\mu - \bar{x}|$ (pronounced "med. deviations") indicate the absolute deviations of the data from the mean of a sample and population respectively. It is more appropriate to call it the *mean deviation (M.A.D.)*.

If the data organised into a grouped frequency distribution having k classes with midpoints x_i and the corresponding frequencies f_1, f_2, \dots, f_k ($\sum f_i = n$), the mean deviation of the data is given by

$$M.D. = \frac{\sum f_i |x_i - \bar{x}|}{n}$$

Mean deviation is also defined in terms of absolute deviations from the median in a similar way, so that the mean deviation is least when the deviations are measured from the median. But it is generally calculated from the arithmetic mean. The mean deviation gives more information than the range or the quartile deviation as it is based on all the observed values. It is easily calculated and understood. As it is not amenable to mathematical treatment, its usefulness is limited. We artificially in its calculation by ignoring the algebraic signs of the deviations and this step is statistically defensible. As the mean deviation does not give undue weight to occasional large deviations, it is used in situations where such deviations are likely to occur. It is unsatisfactory for inference.

Mean deviation is an absolute measure of dispersion. Its relative measure, known as the co-efficient of variation, is obtained by dividing the mean deviation by the average used in the calculation of M.D. Thus

$$\text{Co-efficient of M.D.} = \frac{M.D.}{\text{Mean}} \text{ or } \frac{M.D.}{\text{Median}}$$

Example 4.3 Calculate the mean deviation from (i) the mean, (ii) the median, of the following examination marks:

45, 32, 37, 46, 39, 36, 41, 48 and 36.

Also calculate the co-efficient of mean deviation.

We first arrange the given marks in an increasing sequence to find the median. The ordered set is

32, 36, 36, 37, 39, 41, 45, 46, 48.

- ∴ Median = Marks obtained by $\left[\left[\frac{n}{2} \right] + 1 \right]$ th student in ordered data as $\frac{n}{2}$ is not an integer.
 - Marks obtained by $\left[\left[\frac{n}{2} \right] + 1 \right]$ th, i.e. 5th student
 - 39 marks

and $\bar{x} = \frac{\sum x_i}{n} = \frac{360}{9} = 40$ marks

The necessary calculations are given below:

x_i	\bar{x}	$ x_i - \bar{x} $	$ x_i - \text{median} $
32	40	8	7
36	40	4	3
36	40	4	3
37	40	3	2
39	40	1	0
41	40	1	2
45	40	5	6
46	40	6	7
48	40	8	9
Σ	360	0	40
			39

$$\text{M.D. (from mean)} = \frac{\sum |x_i - \bar{x}|}{n} = \frac{40}{9} = 4.4 \text{ marks}$$

$$\text{and M.D. (from median)} = \frac{\sum |x_i - \text{median}|}{n} = \frac{39}{9} = 4.3 \text{ marks}$$

$$\text{Co-efficient of M.D.} = \frac{\text{M.D.}}{\bar{x}} \text{ or } \frac{\text{M.D.}}{\text{median}}$$

$$= \frac{4.4}{40} \text{ or } \frac{4.3}{39} = 0.11 \text{ or } 0.11$$

Example 4.4 Calculate the mean deviation of the following frequency distribution showing the weight of apples.

Weight (grams)	65-84	85-104	105-124	125-144	145-164	165-184	185-204
f	9	10	17	10	5	4	5

The calculation of the mean deviation (*M.D.*) from the mean is illustrated below:

Weight	x_i	f_i	$f_i x_i$	$x_i - \bar{x}$	$f_i x_i - \bar{x} $
65 - 84	74.5	9	670.5	-48.0	432.0
85 - 104	94.5	10	945.0	-28.0	280.0
105 - 124	114.5	17	1946.5	-8.0	136.0
125 - 144	134.5	10	1345.0	+12.0	120.0
145 - 164	154.5	5	772.5	32.0	160.0
165 - 184	174.5	4	698.0	52.0	208.0
185 - 204	194.5	5	972.5	72.0	360.0
Total	...	60	7350.0	-	1696.0

$$\bar{x} = \frac{\sum f_i x_i}{N} = \frac{7350.0}{60} = 122.5 \text{ grams}$$

$$M.D. = \frac{\sum f_i |x_i - \bar{x}|}{N} = \frac{1696.0}{60} = 28.27 \text{ grams.}$$

THE VARIANCE AND STANDARD DEVIATION

The variance of a set of observations is defined as the mean of the squares of deviations of all the observations from their mean. When it is calculated from the entire population, the variance is called the population variance, traditionally denoted by σ^2 (σ is the Greek lowercase "sigma"). If, instead, the observations in a sample are used to calculate the variance, it is referred to as the sample variance and is denoted by s^2 in order to distinguish between the two. The symbolic definition for variance is

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}, \quad \text{for population data,}$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n}, \quad \text{for sample data.}$$

Variance is also denoted by $\text{Var}(X)$. The term variance was introduced in 1918 by R.A. Fisher.

It should be noted that the variance is in square of units in which the observations are expressed because it is a large number compared to observations themselves. The variance because of its mathematical properties, assumes an extremely important role in statistical theory.

Standard Deviation. The positive square root of the variance is called standard deviation. Symbolically,

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}, \quad \text{for population data,}$$

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}, \text{ for sample data.}$$

The standard deviation is expressed in the same units as the observations themselves and is a measure of the average spread around the mean. Karl Pearson (1857–1936), "founder of the science of Statistics", credited with the name standard deviation, the most useful measure of dispersion. The sample variance in some texts is defined as

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

where n is replaced by $n-1$ on the basis of the argument that knowledge of any $n-1$ observations automatically determines the remaining deviation as the sum of n deviations must be zero. This is an unbiased estimator of the population variance σ^2 , the explanation for which is deferred to chapter 10. In estimating where we shall learn that sample variance $S^2 = \frac{\sum(x_i - \bar{x})^2}{n}$, for small samples, underestimates the population variance σ^2 .

When the data are grouped into a frequency distribution having k classes with midpoints x_1, x_2, \dots, x_k and the corresponding frequencies f_1, f_2, \dots, f_k ($\sum f_i = n$), the sample variance and standard deviation are given by

$$s^2 = \frac{\sum f_i(x_i - \bar{x})^2}{n}, \text{ and}$$

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

It should be noted that for a frequency distribution, as the number of observations or the frequency n is usually large, dividing the sum of squared deviations by $n-1$ is practically equivalent to dividing it by n .

The standard deviation has a definite mathematical meaning, utilizes all the observed values amenable to mathematical treatment but is affected by extreme values. The standard deviation is an absolute measure of dispersion. Its relative measure called coefficient of standard deviation, is defined as

$$\text{Coefficient of S.D.} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

The quantity $\sqrt{\frac{\sum(x_i - \sigma)^2}{n}}$, where σ is some arbitrary origin, is called the root-mean-square deviation which becomes the standard deviation when this arbitrary origin coincides with the mean.

To calculate the variance and standard deviation on an electronic calculator, the above formulas for use are obtained by showing that $\sum(x_i - \mu)^2 = \sum x_i^2 - (\sum x_i)^2 / n$.

$$\begin{aligned}
 \text{Now } \sum(x_i - \mu)^2 &= \sum(x_i^2 - 2x_i\mu + \mu^2) \\
 &= \sum x_i^2 - 2\mu \sum x_i + N\mu^2 \\
 &= \sum x_i^2 - 2N\mu^2 + N\mu^2 \quad (\because \mu = \frac{\sum x_i}{N}) \\
 &= \sum x_i^2 - N\mu^2 = \sum x_i^2 - \left(\frac{\sum x_i}{N}\right)^2
 \end{aligned}$$

Thus the sum of squares of the deviations from the mean is equal to the sum of the squares of all x_i 's minus a correction factor which is the $(1/N)$ th of the square of the sum of all x_i 's.

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

The variance is the mean of the squares minus the square of the mean. The corresponding formula for sample variance is

$$S^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2$$

The alternative formulas for standard deviations are

$$\sigma = \sqrt{\frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N}\right)^2}, \text{ and}$$

$$S = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$$

The following alternative formulas for the sample variance and standard deviation of a frequency distribution are obtained in a similar way.

$$s^2 = \frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2, \text{ and}$$

$$s = \sqrt{\frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2}$$

Example 4.5. A population of $N = 10$ has the observations 7, 8, 10, 13, 14, 15, 20, 25, 26 and 28. Find the variance and standard deviation.

Calculations appear in the following table:

	x_i	$x_i - \mu$	$(x_i - \mu)^2$	x_i^2
	7	-10	100	49
	8	-9	81	64
	10	-7	49	100
	13	-4	16	169
	14	-3	9	196
	19	+2	4	361
	20	3	9	400
	25	8	64	625
	26	9	81	676
	28	11	121	784
Σ	170	0		3424

Now $\mu = \frac{\sum x_i}{N} = \frac{170}{10} = 17$

Therefore $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} = \frac{324}{10} = 32.4$

and $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} = \sqrt{32.4} = 7.31$

Using the alternative method

$$\begin{aligned}\sigma^2 &= \frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N} \right)^2 \\ &= \frac{3424}{10} - \left(\frac{170}{10} \right)^2 = 342.4 - 289 = 53.4\end{aligned}$$

and $\sigma = \sqrt{\frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N} \right)^2} = \sqrt{53.4} = 7.31$

Example 4.6: Calculate the variance and standard deviation from the following marks obtained by 9 students.

45, 32, 37, 46, 39, 36, 41, 48, 36

The variance S^2 and the standard deviation S for the sample are calculated as below:

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
	45	5	25	2025
	32	-8	64	1024
	37	-3	9	1369
	46	6	36	2116
	39	-1	1	1521
	36	-4	16	1296
	41	1	1	1681
	48	8	64	2304
	36	-4	16	1296
Σ	360	0	232	14632

$$\bar{x} = \frac{\sum x_i}{n} = \frac{360}{9} = 40 \text{ marks}$$

$$\text{Therefore } S^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{232}{9} = 25.78 \text{ (marks)}^2$$

$$\text{and } S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{25.78} = 5.08 \text{ marks}$$

Using the alternative method:

$$S^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n} \right)^2$$

$$= \frac{14632}{9} - \left(\frac{360}{9} \right)^2 = 1625.78 - 1600 = 25.78 \text{ (marks)}^2$$

$$\text{and } S = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n} \right)^2} = \sqrt{25.78} = 5.08 \text{ marks}$$

Example 4.7 Calculate the variance and standard deviation from the data of Example 4.4.

The necessary calculations may be carried out on an electronic calculator as below:

x_i	f_i	$f_i x_i$	$f_i x_i^2$
74.5	9	670.5	49 952.25
94.5	10	945.0	89 302.50
114.5	17	1946.5	222 874.25
134.5	10	1345.0	180 902.50
154.5	5	772.5	119 351.25
174.5	4	698.0	121 801.00
194.5	5	972.5	189 151.25
Σ	60	7350.0	973 335.00

Thus we find $s^2 = \frac{\sum f x_i^2}{n} - \left(\frac{\sum f x_i}{n} \right)^2$

$$= \frac{973335}{60} - \left(\frac{7350}{60} \right)^2 = 16222.25 - 15006.25 \\ = 1216 \text{ (grams)}^2$$

and $s = \sqrt{\frac{\sum f x_i^2}{n} - \left(\frac{\sum f x_i}{n} \right)^2} = \sqrt{1216} = 34.87 \text{ grams}$

4.5.1 Change of Origin and Scale. The computational labour can be reduced by using the transformation as was used for computing the arithmetic mean.

Let $u_i = \frac{x_i - a}{h}$. Then $x_i = a + hu_i$ and $\bar{x} = a + h\bar{u}$.

Therefore $\sum (x_i - \bar{x})^2 = \sum [a + hu_i - (a + h\bar{u})]^2 \\ = h^2 \sum (u_i - \bar{u})^2$

Further $\sum (u_i - \bar{u})^2 = \sum u_i^2 - \frac{(\sum u_i)^2}{n}$

Hence $s^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{h^2}{n} \left[\sum u_i^2 - \frac{(\sum u_i)^2}{n} \right] \\ = h^2 \left[\frac{\sum u_i^2}{n} - \left(\frac{\sum u_i}{n} \right)^2 \right]$

and $s = h \sqrt{\frac{\sum u_i^2}{n} - \left(\frac{\sum u_i}{n} \right)^2}$

This gives us a *short method* for hand calculations.

When the data are grouped into a frequency distribution, the corresponding *short method* for calculations, is

$$s = h \sqrt{\frac{\sum f_i x_i^2}{n} - \left(\frac{\sum f_i x_i}{n} \right)^2}$$

where h is the width of the class-interval, f_i is the frequency of the i th class and x_i is the deviation from an assumed mean in terms of class intervals. This method is also known as the *step-deviation method* or *coding method*.

Example 4.8 Find the standard deviation by the short method from the data of Example 4.4.

Let $u_i = \frac{x_i - 114.5}{20}$, where $x_i = 114.5$, value corresponding to the highest frequency, and $h = 20$,

in class-interval. Then $u_i = -2, -1.0, 0, 1, 2, 3, 4$. Other calculations appear below:

x_i	f_i	u_i	$f_i u_i$	$f_i u_i^2$
74.5	0	-2	-18	36
94.5	10	-1	-10	10
114.5	17	0	-28	0
134.5	10	1	10	10
154.5	5	2	10	20
174.5	4	3	12	36
194.5	5	4	20	80
Σ	60		-52	192
			+24	

$$\text{Hence } s = h \times \sqrt{\frac{\sum f u^2}{n} - \left(\frac{\sum f u}{n} \right)^2}$$

$$= 20 \times \sqrt{\frac{192}{60} - \left(\frac{-52}{60} \right)^2} = 20 \times \sqrt{20 - 0.16}$$

$$= 20 \times \sqrt{19.84} = 20 \times 4.456 = 34.87 \text{ grams.}$$

4.5.2 Interpretation of the Standard Deviation. The standard deviation (σ or s) has not a simple interpretation like the arithmetic mean (μ or \bar{x}) that is interpreted as the balancing point for the distribution. The standard deviation is a very important concept that serves as a basic measure of variability. A smaller value of the standard deviation indicates that most of the observations in a data set are close to the mean while a large value implies that the observations are scattered widely about the mean. However, a connection between the standard deviation and fraction of data included in intervals centered around the mean, was discovered by the Russian mathematician P.L. Chebyshev (1894). This result, generally known as Chebyshev's rule, is stated below:

"For any set of data, the interval $\bar{x} - k\sigma$ to $\bar{x} + k\sigma$, where k is any number greater than 1, contains at least the fraction $\left(1 - \frac{1}{k^2}\right)$ of the data." For example, the intervals $\bar{x} \pm 2\sigma$ and $\bar{x} \pm 3\sigma$ will contain at least the fractions $\left(1 - \frac{1}{2^2}\right)$, i.e., $\frac{3}{4}$ and $\left(1 - \frac{1}{3^2}\right)$, i.e., $\frac{8}{9}$ of the data.

This rule is applied to any distribution (Population or Sample) and guarantees the inclusion of a minimum fraction of the data in the constructed interval whereas the actual fraction of the inclusion (especially in bell-shaped distributions) will exceed $\left(1 - \frac{1}{k^2}\right)$.

4.5.3 Co-efficient of Variation. The variability of two or more than two sets of data cannot be compared unless we have a relative measure of dispersion. For this purpose, Karl Pearson (1857-1930) introduced a relative measure of variation, known as the *co-efficient of variation*, abbreviated C.V. which expresses the standard deviation as a percentage of the arithmetic mean of a data set. Symbolically, it is defined as

$$C.V. = \frac{\sigma}{\bar{x}} \times 100, \text{ for sample data,}$$

$$= \frac{\sigma}{\mu} \times 100, \text{ for population data.}$$

As the coefficient of variation is a pure number without units, it is therefore used to compare the variations in two or more data sets or distributions that are measured in different units, e.g. one may be measured in hours and the other in kilograms or rupees. A large value of C.V. indicates that the variability is great and a small value of C.V. indicates less variability.

The coefficient of variation is also used to compare the performance of two candidates or two players given their scores in various papers or games. The smaller the coefficient of variation the more consistent is the performance of the candidates or players. Thus it is used as a criterion for the consistent performance of the candidates or the players. It should be noted that this co-efficient is unreliable when the arithmetic mean is very small.

Example 4.9 Using the co-efficient of variation, determine whether or not there is greater variation among the prices of certain similar commodities given, than among the life in hours under test.

Price in Rupees: 8, 13, 18, 23, 30

Life in hours: 130, 150, 180, 250, 345

We have to compute the mean and the standard deviation for each set so that the corresponding coefficient of variation can be obtained. The necessary arithmetic is shown below:

Price in Rupees (X)		Life in hours (Y)	
X	X^2	Y	Y^2
8	64	130	16900
13	169	150	22500
18	324	180	32400
23	529	250	62500
30	900	345	119025
92	1984	1055	233325

Price of Commodities

$$\bar{x} = \text{Rs. } \frac{92}{5} = \text{Rs. } 18.4$$

$$S_x = \sqrt{\frac{1985}{5} - \left(\frac{92}{5}\right)^2}$$

$$= \sqrt{397.2 - 338.56}$$

$$= \sqrt{58.64} = \text{Rs. } 7.64$$

$$C.V. = \frac{7.64}{18.4} \times 100 = 41.63\%$$

Life in Hours

$$\bar{Y} = \frac{1055}{5} = 211 \text{ hours}$$

$$S_Y = \sqrt{\frac{253325}{5} - \left(\frac{1055}{5}\right)^2}$$

$$= \sqrt{50665 - 44521}$$

$$= \sqrt{6144} = 78.28 \text{ hours}$$

$$C.V. = \frac{78.28}{211} \times 100 = 37.15\%$$

We see that the co-efficient of variation for the prices of commodities (X) is larger than that for the items in hours (Y). Hence the prices of certain similar commodities are showing greater variation than that among the life in hours under test.

Example 4.18: Goals scored by two teams A and B in a football season were as follows:

No. of goals scored in a match (x_i)	Number of matches	
	A	B
0	27	17
1	9	9
2	8	6
3	5	5
4	4	3

By calculating the co-efficient of variation in each case, find which team may be considered more consistent.

The necessary arithmetic is shown below:

No. of goals (x_i)	Team A			Team B		
	f_i	$f_i x_i$	$f_i x_i^2$	f_j	$f_j x_i$	$f_j x_i^2$
0	27	0	0	17	0	0
1	9	9	9	9	9	9
2	8	16	32	6	12	24
3	5	15	45	5	15	45
4	4	16	64	3	12	48
Total	53	56	150	40	68	126

Team A:

$$\text{Mean} = \frac{\sum f_i x_i}{n} = \frac{56}{53} = 1.06, \text{ and}$$

$$\begin{aligned}s &= \sqrt{\frac{\sum f_i x_i^2}{n} - \left(\frac{\sum f_i x_i}{n}\right)^2} \\&= \sqrt{\frac{150}{53} - \left(\frac{56}{53}\right)^2} = \sqrt{1.7138} = 1.308\end{aligned}$$

$$\text{C.V.} = \frac{s}{\bar{x}} \times 100 = \frac{1.308}{1.06} \times 100 = 123.4\%$$

Team B:

$$\text{Mean} = \frac{\sum f_i x_i}{n} = \frac{48}{40} = 1.20, \text{ and}$$

$$\begin{aligned}s &= \sqrt{\frac{\sum f_i x_i^2}{n} - \left(\frac{\sum f_i x_i}{n}\right)^2} \\&= \sqrt{\frac{126}{40} - \left(\frac{48}{40}\right)^2} = 1.71 = 1.308\end{aligned}$$

$$\text{Thus C.V.} = \frac{s}{\bar{x}} \times 100 = \frac{1.308}{1.20} \times 100 = 109.0\%$$

We see that the co-efficient of variation for the team B is smaller than that for the team A. Hence team B is more consistent than team A.

4.5.4 Properties of Variance and Standard Deviation. The variance and standard deviation have the following useful and interesting properties:

- i) The variance of a constant is equal to zero. If a is any constant, then

$$\begin{aligned}\text{Var}(a) &= \frac{1}{N} \sum (a - a)^2 \quad (\because \text{mean of a constant is constant itself}) \\&= 0\end{aligned}$$

- ii) The variance is independent of the origin, i.e. it remains unchanged when a constant is added or subtracted from each observation of the variable X . Symbolically,

$$\text{Var}(X + a) = \text{Var}(X)$$

$$\text{Now } \text{Var}(X + a) = \frac{1}{N} \sum [(x_j + a) - (\mu + a)]^2 \quad (\because \frac{\sum (x_j + a)}{N} = \mu + a)$$

$$= \frac{1}{N} \sum (x_i - \mu)^2 = \text{Var}(X)$$

Hence $\text{Var}(X)$ is invariant to change of the origin.

- (ii) The variance is multiplied or divided by the square of the constant, when each observation of the variable X is either multiplied or divided by a constant.

$$\text{Var}(aX) = \frac{1}{N} \sum (ax_i - a\mu)^2$$

$$= a^2 \frac{\sum (x_i - \mu)^2}{N} = a^2 \text{Var}(X)$$

This may also be interpreted as that the variance increases by a^2 when the scale of X is changed by a .

- (iii) The variance of the sum or difference of two independent variables is equal to the sum of their respective variances.

If X and Y are two independent variables, then

$$\begin{aligned}\text{Var}(X \pm Y) &= \frac{1}{N} \sum [(x_i \pm y_i) - (\mu_x \pm \mu_y)]^2 \\ &= \frac{1}{N} \sum [(x_i - \mu_x) \pm (y_i - \mu_y)]^2 \\ &= \frac{1}{N} \sum (x_i - \mu_x)^2 + \frac{1}{N} (y_i - \mu_y)^2 \pm \frac{2}{N} \sum (x_i - \mu_x)(y_i - \mu_y)\end{aligned}$$

The quantity $\frac{1}{N} \sum (x_i - \mu_x)(y_i - \mu_y)$ is called the covariance and is denoted by $\text{Cov}(X, Y)$. We shall show at some later stage that the covariance of two independent variables is equal to zero. Thus we are left with

$$\begin{aligned}\text{Var}(X \pm Y) &= \frac{1}{N} \sum (x_i - \mu_x)^2 + \frac{1}{N} \sum (y_i - \mu_y)^2 \\ &= \text{Var}(X) + \text{Var}(Y).\end{aligned}$$

- (iv) If k subgroups of data consisting of N_1, N_2, \dots, N_k ($\sum N_i = N$) observations have respective means $\mu_1, \mu_2, \dots, \mu_k$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$, then the variance σ^2 of the combined observations is given by

$$\sigma^2 = \frac{1}{N} \sum N_i (\sigma_i^2 + D_i^2), \quad i = 1, 2, \dots, k$$

where $D_i = \mu_i - \mu$ and μ is the mean for all the data.

Let for the i th subgroup with mean μ_i, μ , the general mean, be considered as an arbitrary origin. Then the sum of squares of deviations of the observations in the i th subgroup from μ is given by

$$\begin{aligned} \sum_{j=1}^{N_i} (x_j - \mu)^2 &= \sum_{j=1}^{N_i} (x_j - \mu_i + \mu_i - \mu)^2 \\ &= \sum_{j=1}^{N_i} (x_j - \mu_i)^2 + N_i(\mu_i - \mu)^2. \quad (\text{product term vanishes}) \\ &= N_i \sigma_i^2 + N_i D_i^2 \\ &= N_i(\sigma_i^2 + D_i^2) \end{aligned}$$

But the variance σ^2 of the combined observations is the mean of the sum of the deviations of all observations in k subgroups from the general mean μ . Hence summing over k -subgroups, we get

$$N\sigma^2 = \sum N_i(\sigma_i^2 + D_i^2)$$

It is relevant to note that all these properties are valid for standard deviation (S.D.), which is the positive square root of variance. In other words,

- i) $S.D. (a) = 0$
- ii) $S.D. (X + a) = S.D. (X)$
- iii) $S.D. (aX) = |a| S.D. (X)$, as S.D. cannot be negative
- iv) $S.D. (X \pm Y) = \sqrt{\text{Var}(X) + \text{Var}(Y)}$
- v) $\sigma = \sqrt{\frac{1}{N} \sum N_i(\sigma_i^2 + D_i^2)}$

For sample data, the corresponding results may be obtained in the same way.

Example 4.11 Let \bar{x}_1 and s_1^2 be the mean and variance respectively of n_1 observations, \bar{x}_2 and s_2^2 be the mean and variance respectively of n_2 observations. Then, if the variance of all $n_1 + n_2$ observations is S^2 , prove that

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^2 \quad (\text{P.U., B.A.B.Sc. 1986})$$

Let \bar{X} denote the general mean and be regarded as an arbitrary origin for the set of n_1 observations and set of n_2 observations. Then the variance of all $n_1 + n_2$ observations, by definition, is given by

$$S^2 = \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} (x_i - \bar{X})^2$$

$$\begin{aligned}
 &= \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x})^2 \right] \\
 &= \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} (x_i - \bar{x}_1 + \bar{x}_1 - \bar{x})^2 + \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2 + \bar{x}_2 - \bar{x})^2 \right] \\
 &= \frac{1}{n_1 + n_2} [n_1(S_1^2 + (\bar{x}_1 - \bar{x})^2) + n_2(S_2^2 + (\bar{x}_2 - \bar{x})^2)] \\
 &= \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2} + \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2}
 \end{aligned}$$

Since \bar{x} is the mean of all $n_1 + n_2$ observations, i.e. $\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$,

$$\begin{aligned}
 \bar{x}_1 - \bar{x} &= \bar{x}_1 - \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{n_2(\bar{x}_1 - \bar{x}_2)}{n_1 + n_2}, \text{ and} \\
 \bar{x}_2 - \bar{x} &= \bar{x}_2 - \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = -n_1(\bar{x}_1 - \bar{x}_2)
 \end{aligned}$$

After substituting and simplifying, we get

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^2$$

4.5.5 Standardized Variables. A variable is defined to be **standardized** or in standard units if it is expressed in terms of deviations from its mean and divided by its standard deviation. It is denoted by Z . In symbols, this means that

$$Z_i = \frac{x_i - \mu}{\sigma}, \text{ for population data.}$$

$$z_i = \frac{x_i - \bar{x}}{S}, \text{ for sample data.}$$

This is a very important concept in advanced statistics as the mean of a standardized variable is equal to zero and its variance is equal to one. Thus

$$\bar{Z} = \frac{1}{N} \sum \left(\frac{x_i - \mu}{\sigma} \right) = \frac{1}{\sigma} \frac{\sum (x_i - \mu)}{N} = 0,$$

$$\text{and } \text{Var}(Z) = \frac{1}{N} \sum \left[\frac{(x_i - \mu)^2}{\sigma^2} - 0 \right]^2$$

$$= \frac{1}{\sigma^2} \frac{\sum (x_i - \mu)^2}{N} = \frac{1}{\sigma^2} \cdot \sigma^2 = 1$$

The *Z*-values, being independent of the units of measurement, provide a basis for comparison between individual values, even though they belong to different distributions. That is why they are often used in psychological and education testing, where they are known as *standard scores*. The negative numbers are avoided by multiplying the *Z* values by 10, an arbitrary *S.D.*, and adding 50, an arbitrary mean, to them. The values so obtained are called the *standard Z-scores*. Thus a standard *Z* score is given by the relation

$$Z = 50 + 10 \left(\frac{x - \bar{x}}{S} \right)$$

4.6 TRIMMED AND WINSORIZED MEASURES

Data sets often contain *extreme* (unusually large or small) observations which may be very different from the main body of the data set and may seem to be incorrect. Such extreme observations are generally called *wild observations* or *Outliers*. These outliers can cause problems. In the presence of outliers, the mean and the standard deviation, being affected by the extreme observations, are therefore misleading measures of central tendency and variability. The appropriate measures then may be the Median and the Interquartile Range, which are much less sensitive to extremes or wild observations. However, it is important to examine a data set for outliers and if present should be excluded.

For this purpose, we either remove a certain percentage of the smallest and largest observations to get the so-called *Trimmed data set* or replace trimmed values by those next in magnitude to obtain what is known as *Winsorized data set* (proposed by C.P. Winsor). The mean and the standard deviation of such data sets are known as the *Trimmed Mean* and the *Trimmed Standard Deviation*, and the *Winsorized Mean* and the *Winsorized Standard Deviation* respectively.

Generally, the *Trimmed mean* is obtained from the data set after having removed all observations below the first quartile and all observations above the third quartile. The *Winsorized mean* is calculated from the modified data set obtained by replacing each observation below the first quartile with the value of the first quartile and each observation above the third quartile with the value of the third quartile. The Trimmed standard deviation and the Winsorized standard deviation are computed from the *trimmed data set* and *Winsorized data set* respectively. The trimmed and Winsorized measures have gained importance in recent years as they are not disturbed by the presence of a few wild observations and have been found almost as good as the corresponding measures in symmetric distributions with no unusual observations.

Example 4.12: Calculate the trimmed and Winsorized means and standard deviations for the data set given in Example 3.11.

The data ordered from smallest to largest and the two quartiles were found to be

32, 36, 36, 37, 39, 41, 45, 46, 48



To find the trimmed mean and the trimmed standard deviation, we remove the two observations 32 and 36 below the first quartile and the two observations 46 and 48 above the third quartile. Thus we have five observations 36, 37, 39, 41, 45 as trimmed data set.

$$\text{Trimmed mean} = \frac{36 + 37 + 39 + 41 + 45}{5} = \frac{198}{5} = 39.6 \text{ and}$$

$$\text{Trimmed S.D.} = \sqrt{\frac{(36)^2 + \dots + (45)^2}{5} - \left(\frac{138}{5}\right)^2}$$

$$= \sqrt{1398.4 - 1368.16} = \sqrt{30.24} = 5.5$$

To find the Winsorized mean and standard deviation, we replace the two values 32, 36 below the first quartile with 36, and the two values 46, 48 above Q_3 with 45 to get the Winsorized data set as 36, 36, 36, 39, 41, 45, 45, and 45. Thus

$$\text{the Winsorized mean} = \frac{\sum X_i}{n} = \frac{360}{9} = 40, \text{ and}$$

$$\begin{aligned}\text{the Winsorized S.D.} &= \sqrt{\frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2} = \sqrt{\frac{14534}{9} - \left(\frac{360}{9}\right)^2} \\ &= \sqrt{1614.89 - 1600} = \sqrt{14.89} = 3.86\end{aligned}$$

MOMENTS

A moment designates the power to which deviations are raised before averaging them, e.g. the $\frac{1}{N} \sum (x_i - \mu)$ is called the first population moment and is denoted by μ_1 . Similarly, the quantity $\sum (x_i - \mu)^2$ is called the second population moment and is denoted by μ_2 . The corresponding sample moments are denoted by m_1 and m_2 . In general, a r th moment about the mean is the arithmetic mean power of the deviations of the observations from the mean. In symbols, this means that

$$\mu_r = \frac{1}{N} \sum (x_i - \mu)^r, \text{ for population data.}$$

$$m_r = \frac{1}{n} \sum (x_i - \bar{x})^r, \text{ for sample data.}$$

These moments are also called the central moments or the mean moments and are used to describe the data.

In a similar way, moments about an arbitrary origin, say a , are defined by the relation

$$\mu'_r = \frac{1}{N} \sum (x_i - a)^r, \text{ for population data.}$$

$$m'_r = \frac{1}{n} \sum (x_i - a)^r, \text{ for sample data.}$$

If we put $r = 0$, we see that

$$\mu_0 = \mu'_0 = 1 \quad \text{and} \quad m_0 = m'_0 = 1$$

For $r = 1$, we have

$$\mu_1 = \frac{1}{N} \sum (x_i - \mu) = \frac{\sum x_i}{N} - \mu = \bar{x} - \mu = 0, \text{ and}$$

$$\mu'_1 = \frac{1}{N} \sum (x_i - a) = \frac{\sum x_i}{N} - a = \bar{x} - a.$$

The corresponding sample results are $m_1 = 0$ and $m'_1 = \bar{x} - a$.

Putting $r = 2$ in the relation for mean moments, we see that

$$\mu_2 = \frac{1}{N} \sum (x_i - \mu)^2 = \sigma^2, \text{ which is the population variance,}$$

$$\text{and } m_2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = S^2, \text{ which is the sample variance,}$$

When $a = 0$, the moment $m_r' = \frac{1}{n} \sum x_i^r$ is called the r th moment about zero.

The moments about the mean or about the arbitrary origin are also called the power moments.

When the sample data are grouped into a frequency distribution having k classes with midpoints x_1, x_2, \dots, x_k and the corresponding frequencies f_1, f_2, \dots, f_k ($\sum f_i = n$), the r th sample moment is given by

$$m_r = \frac{1}{n} \sum f_i (x_i - \bar{x})^r, \text{ and}$$

$$m'_r = \frac{1}{n} \sum f_i (x_i - a)^r$$

4.7.1 Moments about the Mean in terms of Moments about the arbitrary origin, say a , conversely. It is easier to calculate the moments in the first instance, about an arbitrary origin. They are then transformed to the new moments. This is done by using the relationships obtained as follows:

By definition, the r th sample moment about the mean is given by

$$m_r = \frac{1}{n} \sum f_i (x_i - \bar{x})^r$$

The quantity within brackets may be written as

$$(x_i - \bar{x}) = (x_i - a + a - \bar{x}) = (x_i - a) + (\bar{x} - a)$$

$$= D_i - m'_1 \text{ where } D_i = (x_i - a) \text{ and } m'_1 = (\bar{x} - a)$$

$$\text{Thus, we have } m_r = \frac{1}{n} \sum f_i (D_i - m'_1)^r$$

By means of Binomial expansion, we have

$$m_r = \frac{1}{n} \sum f_i D_i^r - \left[\binom{r}{1} D_1^{r-1} m'_1 + \binom{r}{2} D_1^{r-2} (m'_1)^2 + \dots + (-1)^r (m'_1)^r \right]$$

$$\binom{r}{j} = \frac{r!}{j!(r-j)!} \text{ and } r! = r(r-1)(r-2)\dots 3 \times 2 \times 1.$$

$$m_r = \frac{1}{n} \sum f_i D_i^r - \left(\binom{r}{1} \frac{1}{n} \sum f_i D_i^{r-1} m'_1 - \binom{r}{2} \frac{1}{n} \sum f_i D_i^{r-2} (m'_1)^2 - \dots + (-1)^r (m'_1)^r \frac{1}{n} \sum f_i \right) \\ = m'_r - \left(\binom{r}{1} m'_{r-1} m'_1 + \binom{r}{2} m'_{r-2} (m'_1)^2 + \dots + (-1)^r (m'_1)^r \right)$$

For $r = 1, 2, 3$ and 4 , we get

$$m_1 = m'_1 - m_1' = 0,$$

$$m_2 = m'_2 - \left(\binom{2}{1} m'_1 \cdot m'_1 + \binom{2}{2} (m'_1)^2 \right) m'_0 \\ = m'_2 - (m'_1)^2,$$

$$m_3 = m'_3 - 3m'_2 m'_1 + 2(m'_1)^2, \text{ and}$$

$$m_4 = m'_4 - 4m'_3 m'_1 + 6m'_2 (m'_1)^2 - 3(m'_1)^4$$

The corresponding results for population data are:

$$\mu_1 = 0;$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2;$$

$$\mu_3 = \mu'_3 - 3\mu'_2 \mu'_1 + 2(\mu'_1)^3;$$

$$\mu_4 = \mu'_4 - 4\mu'_3 \mu'_1 + 6\mu'_2 (\mu'_1)^2 - 3(\mu'_1)^4$$

It should be noted that in each of these relations, the sum of the coefficients of various terms on both sides equals zero. Each term on the right is of the same dimension as the term on the left.

Conversely, the r th sample moment about an arbitrary origin, a , is given by

$$m'_r = \frac{1}{n} \sum f_i (x_i - a)^r \\ = \frac{1}{n} \sum f_i (x_i - \bar{x} + \bar{x} - a)^r \\ = \frac{1}{n} \sum f_i (d_i + m'_1)^r, \text{ where } d_i = x_i - \bar{x} \text{ and } m'_1 = \bar{x} - a \\ = \frac{1}{n} \sum f_i d_i^r + \binom{r}{1} m'_1 \frac{1}{n} \sum f_i d_i^{r-1} + \binom{r}{2} (m'_1)^2 \frac{1}{n} \sum f_i d_i^{r-2} + \dots + (m'_1)^r \frac{1}{n} \sum f_i \\ = m_r + \binom{r}{1} m'_{r-1} (m'_1) + \binom{r}{2} m'_{r-2} (m'_1)^2 + \dots + (m'_1)^r$$

$$m_1 = m'_1 - m_1' = 0$$

$$m_2 = m'_2 - (m'_1)^2$$

$$m_3 = m'_3 - 3m'_2 m'_1$$

$$m_4 = m'_4 - 6m'_3 m'_1$$

$$6m'_1 (m'_1)^2 = 36(-1)^2$$

Putting $r = 2, 3$ and 4 , we get

$$m_2' = m_2 + (m_1')^2$$

$$m_3' = m_3 + 3m_2(m_1') + (m_1')^3$$

$$m_4' = m_4 + 4m_3(m_1') + 8m_2(m_1')^2 + (m_1')^4.$$

For a population of size N , the corresponding relations are

$$\mu_2' = \mu_2 + \mu_1'^2,$$

$$\mu_3' = \mu_3 + 3\mu_1'\mu_2 + \mu_1'^3,$$

$$\mu_4' = \mu_4 + 8\mu_1'\mu_3 + 6\mu_1'^2\mu_2 + \mu_1'^4.$$

4.7.2 Sheppard's Corrections. In the calculation of moments from a grouped frequency distribution, certain errors are introduced by the assumption that the frequencies associated with a class are located at the mid-point of the class interval. These errors therefore need corrections. It has been shown by W.F. Sheppard that, if the frequency distribution (i) is continuous and (ii) tails off to zero at each end, the corrected moments are as given below:

$$m_2 \text{ (corrected)} = m_2 \text{ (uncorrected)} - \frac{h^2}{12};$$

$$m_3 \text{ (corrected)} = m_3 \text{ (uncorrected)},$$

$$m_4 \text{ (corrected)} = m_4 \text{ (uncorrected)} - \frac{h^2}{12}m_2 \text{ (uncorrected)} + \frac{7}{240}h^4;$$

where h denotes the uniform class-interval and m 's are the moments about the mean, of the grouped frequency distribution. The important point to note here is that these corrections are not applicable to highly skewed distributions and to distributions having unequal class-intervals.

4.7.3 Moment-Ratios. There are certain ratios in which both the numerators and the denominators are moments. The most common of these moment-ratios are β_1 and β_2 defined by the ratios

$$\beta_1 = \frac{\mu_3^2}{\mu_2^2}, \quad \beta_2 = \frac{\mu_4}{\mu_2^2}. \quad \text{They are independent of origin and units of measurement, i.e., they are dimensionless.}$$

Actually, β_1 is the square of the third population moment expressed in standard units and β_2 is the fourth standardized moment for a population, where a standardized variable has been defined as

$$Z = (x - \mu)/\sigma.$$

For symmetrical distributions, β_1 is equal to zero. It is, therefore, used as a measure of skewness and is used to explain the shape of the curve and is a measure of peakedness. For the normal distribution discussed later, $\beta_2 = 3$.

The moment-ratios (or the standardized moments) for sample data are similarly defined as

$$b_1 = \frac{(m_3)^2}{(m_2)^3} \quad \text{and} \quad b_2 = \frac{m_4}{(m_2)^2}.$$

Example 4.13. Calculate the first four moments about the mean for the following set of values marks: 45, 32, 37, 46, 39, 36, 41, 43 & 36.

For convenience, the observed values are written in an increasing sequence. The necessary notes appear in the table below:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
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	100

Now $\bar{x} = \frac{\sum x_i}{n} = \frac{360}{9} = 40$ marks

$$m_1 = \frac{\sum (x_i - \bar{x})}{n} = 0$$

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{232}{9} = 25.78 \text{ (marks)}^2$$

$$m_3 = \frac{\sum (x_i - \bar{x})^3}{n} = \frac{186}{9} = 20.67 \text{ (marks)}^3$$

$$m_4 = \frac{\sum (x_i - \bar{x})^4}{n} = \frac{10708}{9} = 1189.78 \text{ (marks)}^4$$

$\beta_1, \beta_2 = ?$

Example 4.14. Compute the first four moments for the following distribution of wages after Sheppard's corrections.

Weekly Earnings (Rupees)	5	6	7	8	9	10	11	12	13	14	15
No. of men	1	2	3	10	20	51	22	11	5	3	1

(P.U., B.A./B.Sc. (Part-I), 1962)

We first calculate moments about an arbitrary origin. The necessary calculations are shown below. The moments about $x = 10$ are obtained by dividing the column sums by n .

Earnings in Rs. (x_i)	Men f_i	D_i ($x_i - 10$)	$f_i D_i$	$f_i D_i^2$	$f_i D_i^3$	$f_i D_i^4$
5	1	-5	-5	25	-125	625
6	2	-4	-8	32	-128	512
7	5	-3	-15	45	-135	405
8	10	-2	-20	40	-80	160
9	20	-1	-20	20	-20	20
10	51	0	-68	0	-488	0
11	22	1	22	22	22	22
12	11	2	22	44	88	176
13	5	3	15	45	135	405
14	3	4	12	48	192	768
15	1	5	5	25	125	625
Sums	131	-	-76	346	562	3718
Sum $+n$	1	-	0.06	1.6	0.56	28.38
			= m_1'	= m_2'	= m_3'	= m_4'

Moments about the mean are:

$$m_1 = 0$$

$$m_2 = m_2' - (m_1')^2 = 2.64 - (0.06)^2 = 2.56$$

$$\begin{aligned} m_3 &= m_3' - 3m_2'm_1' + 2(m_1')^3 \\ &= 0.56 - 3(2.64)(0.06) + 2(0.06)^3 = 0.08. \end{aligned}$$

$$\begin{aligned} m_4 &= m_4' - 4m_3'm_1' - 6m_2'(m_1')^2 + 3(m_1')^4 \\ &= 28.38 - 4(0.56)(0.06) + 6(2.64)(0.06)^2 - 3(0.06)^4 = 28.30 \end{aligned}$$

Applying Sheppard's corrections, we have

$$m_2 (\text{corrected}) = m_2 (\text{uncorrected}) - \frac{h^2}{12} = 2.64 - 0.08 = 2.56,$$

$$m_3 (\text{corrected}) = m_3 (\text{uncorrected}) = 0.08,$$

$$\begin{aligned} m_4 (\text{corrected}) &= m_4 (\text{uncorrected}) - \frac{h^2}{3} \cdot m_2 (\text{uncorrected}) + \frac{7h^4}{240} \\ &= 28.30 - 1.32 + 0.03 = 27.01 \end{aligned}$$

4.7.4 Change of Origin and Scale. Let α and h denote the arbitrary origin and the class-interval width. Then we define a new variable u as

$$u_j = \frac{x_j - \alpha}{h}$$

so that $x_j - \alpha = hu_j$, $\bar{x} - \alpha = h\bar{u}$, and hence $x_j - \bar{x} = h(u_j - \bar{u})$.

Substituting these values in the n th sample moments, we get

$$m'_r = \frac{1}{n} \sum f_i (x_i - u)^r = N' \cdot \frac{1}{n} \sum f_i u_i^r;$$

$$\text{and } m'_r = \frac{1}{n} \sum f_i (x_i - \bar{x})^r = N' \cdot \frac{1}{n} \sum f_i (u_i - \bar{u})^r.$$

This shows that the r th moments of the variable X are N' times the corresponding moments of the variable u , and are independent of the origin 'u'. In other words, the moments are not affected by a change of origin but are affected by a change of scale.

4.7.5 Charlter Check. We have seen that the computation of the moments depends upon the sum of the products of the frequencies by the corresponding values of the variable. It is, therefore, desirable to make these computations so that arithmetic mistakes, if any, are avoided. For this purpose, L.V. Charlter, a Norwegian statistician, introduced a check known as *Charlter check*. This check actually consists in writing the assumed origin in the coded form by one interval. The relations used for this purpose are given below:

$$\sum f(n+1) = \sum f_i + n$$

$$\sum f(n+1)^2 = \sum f u^2 + 2 \sum f u + n$$

$$\sum f(n+1)^3 = \sum f u^3 + 3 \sum f u^2 + 3 \sum f u + n$$

$$\sum f(n+1)^4 = \sum f u^4 + 4 \sum f u^3 + 6 \sum f u^2 + 4 \sum f u + n$$

Example 4.15 Calculate the first four moments about the mean from the data of Example 4.4.

The necessary calculations by taking $u_i = \frac{x_i - 125}{20}$, are set out in the following table. The last column is used for *Charlter's* check and the column sums are divided by n to get m'_r .

Data		Computations					
x_i	f_i	u_i	$f u^2$	$f u^3$	$f u^4$	$f(u+1)^4$	
74.5	9	-18	36	-72	144	9	
94.5	10	-10	10	-10	10	0	
114.5	17	0	0	0	0	0	
134.5	10	1	10	10	10	10	160
154.5	5	2	10	20	40	80	405
174.5	4	3	12	36	108	324	1024
194.5	2	4	20	80	320	1280	3125
Sum	60	—	192	396	1848	4740	
Sums+n	1	—	0.4	3.2	6.6	30.8	Fo: Charlter's check
			= m'_1	= m'_2	= m'_3	= m'_4	

For check,

$$\sum f(u+1)^4 = \sum f u^4 + 4 \sum f u^3 + 6 \sum f u^2 + 4 \sum f u + n$$

$$= 1848 + 4(396) + 6(192) + 4(24) + 60$$

$$= 1848 + 1584 + 1152 + 96 + 60 = 4740, \text{ which is the sum in the last column.}$$

Hence the moments about the mean and in place-interval units are obtained as below:

$$m_1 = 0$$

$$m_2 = m'_2 - (m'_1)^2$$

$$= 3.2 - (0.4)^2 = 3.04$$

$$m_3 = m'_3 - 3m'_2m'_1 + 2(m'_1)^3$$

$$= 6.6 - 3(3.2)(0.4) + 2(0.4)^3 = 2.89$$

$$m_4 = m'_4 - 4m'_2m'_2 + 6m'_1(m'_1)^2 - 3(m'_1)^4$$

$$= 10.8 - 4(3.2)(0.4) + 6(3.2)(0.4)^2 - 3(0.4)^4 = 23.24$$

To get the moments about the mean in ordinary units, we multiply m_2 by b^2 , i.e. 400 , m_3 by $(20)^2$ and m_4 by $(20)^4$. Thus $m_2 = 1216$, $m_3 = 23120$ and $m_4 = 3718400$.

Example 4.16 The first three moments of a distribution about the value 2 of the variable are 1216 and -40 . Show that the mean is 3, the variance 15 and $m_3 = -80$. Also show that the first three moments about $x = 0$ are 3, 24 and 76.

$$\text{Here we are given } m'_1 = \frac{1}{n} \sum f_i(x - 2) = 1 \quad \dots \quad (1)$$

$$m'_2 = \frac{1}{n} \sum f_i(x - 2)^2 = 15 \quad \dots \quad (2)$$

$$m'_3 = \frac{1}{n} \sum f_i(x - 2)^3 = -40 \quad \dots \quad (3)$$

We also know that $\bar{x} = \bar{z} - \alpha$, so that

$$\bar{x} = m'_1 + \alpha = 1 + 2 = 3 \quad (\because \alpha = 2)$$

And variance, $S^2 = m'_2 - (m'_1)^2 = 16 - 1 = 15$, and

$$\begin{aligned} m_3 &= m'_3 - 3m'_2m'_1 + 2(m'_1)^3 \\ &= -40 - 3(16)(1) + 2(1)^3 = -80 \end{aligned}$$

To find the moment about $x = 0$, we need the values of $\frac{1}{n} \sum f_i$, $\frac{1}{n} \sum f_i x^2$ and $\frac{1}{n} \sum f_i x^3$, which are obtained from relations (1), (2) and (3).

$$\text{From (1), } \frac{\sum f_i}{n} = 3$$

$$\text{From (2), } \frac{1}{n} = \sum f(x-2)^2 = 16$$

$$\text{or } \frac{1}{n} \sum f(x^2 - 4x + 4) = 16$$

$$\text{or } \frac{1}{n} \sum fx^2 - 4 \frac{\sum fx}{n} + 4 = 16$$

$$\text{or } \frac{1}{n} \sum fx^2 = 16 - 4 + 4(3) = 24$$

(3) on expansion can be written as

$$\frac{1}{n} \sum fx^2 = 6 \frac{1}{n} \sum fx^2 + 12 \frac{\sum fx}{n} - 8 = -40$$

$$\text{or } \frac{1}{n} \sum fx^2 = -40 + 8 - 12(3) + 6(24) = 76$$

Hence the moments about $x = 0$ are

$$m_1 = \frac{\sum fx}{n} = 3$$

$$m_2 = \frac{\sum fx^2}{n} = 24, \text{ and}$$

$$m_3 = \frac{\sum fx^3}{n} = 76$$

Example 4.17 Show that for discrete distributions, $\beta_3 > 1$.

$$\text{Distribution, } \beta_3 = \frac{\mu_3}{\mu_2^2}$$

(P.U., D.Sc. 1962)

Now β_3 will be greater than one if the numerator is greater than the denominator, i.e. if $\mu_3 > \mu_2^2$, or if $\mu_3 - \mu_2^2 > 0$

$$\text{Now } \mu_3 - \mu_2^2 = \frac{1}{N} \sum f(x-\mu)^3 - \sigma^2 \quad (\because \mu_2 = \sigma^2)$$

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

$$= \frac{1}{N} \sum f(x-\mu)^3 + \frac{\sigma^4 \sum f}{N} - 2\sigma^2 \cdot \frac{\sum f(x-\mu)^2}{N}, \left[\sigma^2 = \frac{\sum f(x-\mu)^2}{N} \right]$$

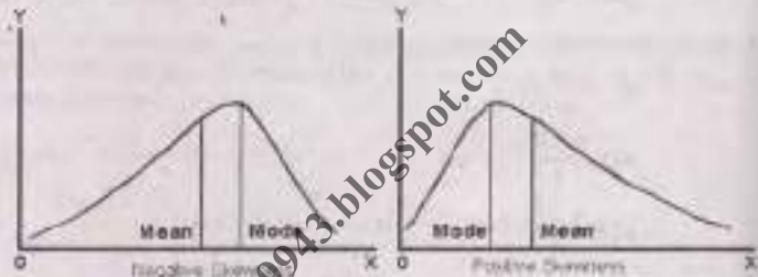
$$= \frac{1}{N} \sum f[(x-\mu)^3 + \sigma^4 - 2\sigma^2(x-\mu)^2] \quad (N\sigma^4 = \sigma^4 \sum f)$$

$$= \frac{1}{N} \sum f[(x - \mu)^2 - \sigma^2]^2 \text{ which is essentially positive.}$$

Hence $\beta_2 \geq 1$ because $\mu_4 = \mu_2^2$ is always positive.

4.8 SKEWNESS

A distribution in which the values equidistant from the mean have equal frequencies is defined to be *symmetrical* and any departure from symmetry is called *skewness*. It is important to note that in a perfectly symmetrical distribution, the mean, median and mode coincide and that the two tails of the distribution are equal in length from the mean. These values are pulled apart when the distribution departs from symmetry and consequently one tail becomes longer than the other. If the right tail is longer than the left tail, the distribution is said to have *positive skewness*. If the left tail is longer than the right tail, it is said to be *negatively skewed* or to have *negative skewness*. In a positively skewed distribution, the mean is greater than the median and the median is greater than the mode, i.e. $\text{mean} > \text{median} > \text{mode}$ and in a negatively skewed distribution, $\text{mode} > \text{median} > \text{mean}$.



The difference between the measures of location, being an indication of the amount of skewness or asymmetry, is used as a measure of skewness. A measure of skewness is defined in such a way that (i) the measure should be zero when the distribution is symmetric and (ii) the measure should be a pure number, i.e. independent of origin and units of measurements.

According to me, the degree of skewness of a distribution or curve, Karl Pearson (1857-1947) introduced a coefficient of skewness denoted by Sk and defined by

$$Sk = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

We know that mode is sometimes ill-defined and is difficult to locate by simple methods so therefore, replaced by its equivalent from empirical relation holding good in moderately skewed distributions. The Pearsonian co-efficient of skewness then becomes

$$Sk = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

The coefficient usually varies between -3 (negative skewness) and +3 (positive skewness) and sign indicates the direction of skewness. The formula satisfies both the requirements considered above for a measure of skewness.

Arthur Lyon Bowley (1869–1957), a British statistician, has also proposed a measure of skewness that is based on the median and the two quartiles. In a symmetrical distribution, the two quartiles are equidistant from the median but in an asymmetrical distribution, this will not be the case. The Bowley's coefficient of skewness is

$$Sk = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Its values lies between 0 and ± 1 .

Another measure of skewness that is often used, is the third moment express in standard units (or moment ratio) and thus is given by

$$Sk = \frac{\mu_3}{\sigma^3}, \text{ for population data,}$$

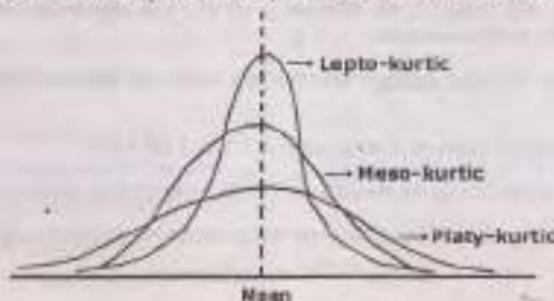
$$= \frac{s_3}{s^3}, \text{ for sample data,}$$

The coefficient for most distributions, will be between -2 and +2. Some statisticians denote x by α_3 or β_3 . If the coefficient is greater than zero, the distribution or curve is positively skewed. If $Sk < 0$, there is negative skewness. For symmetrical distributions or curves, the coefficient is zero.

KURTOSIS

Karl Pearson (1857–1936) introduce the term Kurtosis (literally the amount of hump) for the degree of peakedness or flatness of a unimodal frequency curve. When the values of a variable are closely clustered round the mode in such a way that the curve is leptokurtic. If, on the other hand, the curve is less peaked, we say that the curve is platykurtic. Since the normal curve (to be described later) is neither very peaked nor very flat-topped, it makes as basis for comparison. The normal curve itself is called mesokurtic.

Kurtosis is usually measured by the fourth standardized moment or the moment-ratio $\beta_2 = (\mu_4 / \mu_2^2)$ whose value for a normal distribution is equal to 3. When β_2 is greater than 3, the curve is more sharply peaked and has wider tails than the normal curve and is said to be leptokurtic. When it is less than 3, the curve has a flatter top and relatively narrower tails than the normal curve and is said to be platykurtic.



The corresponding measure of kurtosis for the sample data is $D_3 = \frac{n_4}{n_2^2}$. It should be noted that the value of D_3 for a large sample from the normal population is very nearly 3.

Another measure of kurtosis not widely used, is given by

$$K = \frac{Q.D.}{P_{90} - P_{10}},$$

where Q.D. is the semi-interquartile range and P 's are the percentiles. This is known as the Percentile co-efficient of kurtosis. It has been shown that K for a normal distribution is 0.263 and that it lies between 0 and 0.50.

4.10 DESCRIBING A FREQUENCY DISTRIBUTION

To describe the major characteristics of a frequency distribution, we need the calculations of the following five quantities:

- i) The number of observations that describes the size of the data.
- ii) A measure of central tendency such as the mean or median that provides information about the centre or average value.
- iii) A measure of dispersion such as standard deviation that indicates the variability of the data.
- iv) A measure of skewness that shows the lack of symmetry in the frequency distribution.
- v) A measure of kurtosis that gives information about its peakedness.

It is interesting to note that all these measures can be derived from the first four moments. For example, the first moment about $x = 0$ is the arithmetic mean, the second moment about the mean is the variance and its positive square root is the standard deviation. The third mean moment is a measure of skewness while the fourth central moment is used to measure kurtosis. Thus the first four moments play a key role in describing frequency distributions.

EXERCISES

OBJECTIVE

- i) Answer 'True' and 'False'. If the statement is not true then replace the underlined words with words that make the statement true.
 - i) A measure of central tendency describes how widely the data are dispersed about a central value.
 - ii) The standard deviation for a set of values 5, 5, 5, 5 and 5 is 5.
 - iii) The unit of measure for the standard score is always in standard deviation.
 - iv) For a bell shaped distribution, the range will be approximately equal to three standard deviations.

- v) The difference between the largest and the smallest observations is called the Quartile Deviation.
- w) The square of variance gives the standard deviation.
- x) The coefficient of variation is an absolute measure of dispersion.
- y) The coefficient of variation is measured in different units as the data.
- z) The Quartile Deviation is based on only two values in the series.
- aa) The square root of the variance of a distribution is the absolute deviation.

MULTIPLE CHOICE QUESTIONS

- i) The main disadvantage of the range is that:
 - a) It does not use all the observations in its calculation.
 - b) It can be influenced by an extreme value.
 - c) Both a and b are correct.
 - d) None of the above.
- ii) Which one of the following is not a measure of dispersion?
 - a) Range.
 - b) Standard deviation.
 - c) Second quartile.
 - d) Coefficient of variation.
- iii) Which of the following is not a measure of dispersion?
 - a) Interquartile range.
 - b) Difference between the values of the largest and smallest items.
 - c) Mean of the values of the largest and smallest items.
 - d) Standard deviation.
- iv) The standard deviation is:
 - a) The square of the variance.
 - b) Two times the standard deviation.
 - c) Half the variance.
 - d) The square root of the variance.

- v) The coefficient of variation is measured in
- The same units as the mean and the standard deviation.
 - Percent.
 - Squared units.
 - None of the above.
- vi) If the original units are measured in pounds, the variance is
- Also measured in pounds.
 - Measured in pounds squared.
 - Measured in half pounds.
 - None of the above.
- vii) If the tail of a frequency distribution is in positive direction (to the right), the coefficient of skewness is
- Zero.
 - Positive.
 - Negative.
 - None of the above.
- viii) The standard deviation of a frequency distribution is 10, the mean is 250, the median is 250 and the mode is also 250. The coefficient of skewness is
- Zero
 - Positive.
 - Negative.
 - None of the above.
- ix) Which of the following statement is true?
- The standard deviation is less than the range.
 - The range is less than the interquartile range.
 - The arithmetic mean always exceeds the median.
 - The arithmetic mean always exceeds the mode.
- x) Which of the following is not a property of the standard deviation?
- It is always negative number.
 - It is affected by extreme values in a data set.
 - It is based on all the values in the data set.
 - It is the most widely used measure of dispersion.

- (i) If a distribution has zero standard deviation, then which of the following is true?
- All observations are negative.
 - All observations are positive.
 - All observations are equal.
 - Number of positive values and negative values are equal.
- (ii) The empirical rule generally can be applied to
- Bell shaped distribution.
 - Any distribution.
 - Only continuous distribution.
 - Any skewed distribution.
- (iii) Symmetrical distribution will always have skewness equal to
- Negative.
 - Positive.
 - Zero.
 - Approximately zero.
- (iv) For a normal distribution the measure of kurtosis equals to
- Zero.
 - 3.
 - Positive number.
 - Negative number.
- (v) For the given sample set 2, 8, 10, 15, 20, 9, 18, 0, 7, 10, which is the value of coefficient of variation?
- 70.00 percent.
 - 15.50 percent.
 - 145.00 percent.
 - 61.21 percent.

OBJECTIVE

- (i) Explain clearly the meaning of the term Dispersion. What are the most usual methods of measuring dispersion? Indicate the advantages and disadvantages of these methods.
(P.U., B.Com. 1960; B.A. (Hons.), 1960; B.A. (Part-I), 1961)
- (ii) Discuss the different measures of dispersion. Describe the method of computation of any two of them with suitable examples.
(P.U., M.A., Econ. 1960)

- 4.3 Describe carefully how Mean Deviation, Standard Deviation and Quartile Deviation of a given distribution are obtained. In what problems, should each be used?

(P.U., B.A. (Part-I), 1962-63)

- 4.4 a) What is Range and how is it calculated? What are its uses?

- b) Define Quartile Deviation. Find the quartile deviation from the following data
(i) graphically; (ii) using an appropriate formula.

Income per week (Rs.)	41-50	51-60	61-70	71-80	81-90	91-100	Total
No. of Earners	30	36	43	304	73	14	300

(P.U., B.A./B.Sc. 1962-63)

- 4.5 The members of a sports club, 50 male adults, had their weights recorded, in pounds. The weights are given below:

171	160	144	132	154	160	160	158	148	160	131	153
131	165	139	163	149	149	140	149	150	161	136	144
165	174	153	149	157	169	147	156	151	171	149	154
153	149	147	154	145	158	160	152	156	138	167	142
165	155	140	155	158	147	149	148	174	150	144	

Construct a cumulative frequency table for the above weights, using classes of width 5 lb, starting at 129.5 lb. Hence draw a cumulative frequency graph, and use this to find the median and semi-interquartile range.

Use the grouped frequency table to calculate the mean and standard deviation, and compare them with the values obtained using the original ungrouped data.

(M.A. Econ. I.U. 1990; B.Z.U. 1990)

- 4.6 a) Define Mean Deviation and its co-efficient. Discuss its advantages and uses.

- b) Estimate the mean deviation from the arithmetic mean of the following set of examination marks.

Marks	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79
No. of students	2	3	8	24	27	40	11	5

- 4.7 When originally collected, the data in the following distribution had been in 5-year age groups throughout the whole range 20-64. Using central values of these age-groups, the average age had been calculated to be 44.5 years and the mean deviation without collecting the data was found to be 7.15 years. Reconstruct the original table with 5-yearly age groups from the information given.

Age nearest birthday	20-24	25-29	30-39	40-44	45-49	50-54	55-59
Number of men	1	2	20	22	20	15	14

(P.C.S., 1971, 93; P.U., B.A./B.Sc. 1962-63)

Solution: Let f_1 and f_2 be the frequencies corresponding to the groups 30-34 and 55-59 in 5-yearly age groups. Then the calculations will proceed as below:

Age-group	C.V. (x_i)	f	f_i	$ x - \bar{x} $	$f x - \bar{x} $
20-24	22	1	22	22.5	22.5
25-29	27	2	54	17.5	35.0
30-34	32	f_1	$32f_1$	12.5	$12.5f_1$
35-39	37	$(26-f_1)$	$962-37f_1$	7.5	$195.0-7.5f_1$
40-44	42	22	924	2.5	55.0
45-49	47	20	940	2.5	50.0
50-54	52	15	780	7.5	112.5
55-59	57	f_2	$57f_2$	12.5	$12.5f_2$
60-64	62	$(14-f_2)$	$868-62f_2$	17.5	$245.0-17.5f_2$
Total	—	100	$4550-5(f_1+f_2)$	—	$715+5(f_1-f_2)$

$$\text{Now, Mean} = \frac{1}{n} \sum f x_i$$

$$\text{i.e., } 44.5 = \frac{1}{100} [4550 - 5(f_1 + f_2)]$$

$$\text{or, } f_1 + f_2 = 20 \quad \dots (A)$$

$$\text{and Mean Deviation} = \frac{1}{n} \sum f |x - \bar{x}|$$

$$\text{i.e., } 7.15 = \frac{1}{100} [715 + 5(f_1 - f_2)]$$

$$\text{or, } f_1 - f_2 = 0$$

From (A) and (B) we find, $f_1 = 10$ and $f_2 = 10$.

Find the quartile deviation and mean deviation as well as their co-efficients from the following data and comment.

Height (inches)		58	59	60	61	62	63	64	65
No. of Persons	Group A	10	18	30	42	35	28	18	8
	Group B	15	20	32	35	33	22	20	10

(i) Define Variance and Standard Deviation. Describe their properties.

(ii) For a population of numbers 10, 8, 7, 9, 5, 12, 3, 6, 8, 2, calculate σ^2 and σ .
(P.U., M.A. Econ. 1972)

(iii) Prove that the variance remains unchanged when a constant is added to or subtracted from every value of the variable.

(iv) The scores obtained by five students on a set of examination papers are 70, 50, 60, 73, 50. These scores are changed by (i) adding 10 points to all scores, (ii) increasing all scores by 10%. What effect will these changes have on the mean and on the standard deviation?
(P.U., B.A./B.Sc., 1971)

- 4.11 a) Define the mean and standard deviation of a distribution. \bar{x} is the mean and S , the standard deviation. When a provisional mean "a" is chosen, the corresponding provisional standard deviation is found to be S_1 . Prove that $S_1^2 = S^2 + (\bar{x} - a)^2$. Explain briefly the advantages of this procedure in numerical work.
 (P.U., B.A./B.Sc. 1978)
- b) For a set of ungrouped values, the following sums are found: $n = 20$, $\sum x = 480$, $\sum x^2 = 15735$. Find the mean and the standard deviation.
 (B.Z.U., B.A./B.Sc. 1978)
- c) The mean and standard deviation of a sample of 20 observations were found to be 75 and 2.5 respectively. On checking the original figures, it was discovered that one observation which was actually 68, was copied down as 86. Find the correct mean and standard deviation.
- 4.12 a) Describe the properties of the standard deviation.
- b) By multiplying each of the numbers 3, 6, 2, 1, 7, 5 by 3 and then adding 5, we obtain 17, 9, 7, 19, 15. What is the relation between the standard deviations and the means of the two sets?
 (P.U., B.A./B.Sc. 1972, 78)
- c) A child is born to Mrs. X every year for consecutive years. Compute the standard deviation of the children's age when (i) the youngest is 1 year old, and (ii) the youngest is 8 years old. Why do the answers (i) and (ii) coincide?
 (P.U., B.A./B.Sc. 1978)
- 4.13 Show how to compute the Range and the Standard deviation of a sample of n observations and explain briefly the meaning of these statistics, giving examples of situations in which they would be used.
 Show that when $n = 2$, the two estimates are simply related.
- *4.14 It is often stated that in frequency distributions there exists the approximate relation $\frac{\text{Mean Deviation}}{\text{Standard Deviation}} = 0.8$. Test this statement in the following distribution:
- | Weight (grams) | 65-84 | 85-104 | 105-124 | 125-144 | 145-164 | 165-184 | 185-204 |
|----------------|-------|--------|---------|---------|---------|---------|---------|
| f | 9 | 10 | 17 | 10 | 5 | 4 | 5 |
- *4.15 Calculate the mean and standard deviation for the following distribution of lengths of metal bars.
- | x | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 |
|-----|----|----|----|----|----|----|----|----|----|----|
| f | 4 | 8 | 23 | 35 | 62 | 44 | 18 | 4 | 1 | 1 |
- (P.I.U., D.St. 1963; W.P.C.S. —)
- 4.16 Explain how Chebyshev's rule can be used to answer questions about a data set? Using the rule, find and interpret the interval $\bar{x} \pm 2\sigma$ for question 4.15.

- The following table gives the frequency distribution of expenditures on food per family per month among working class families in two localities. Find the arithmetic mean and the standard deviation of the expenditure at both places.

Range of expenditure in rupees per month	No. of Families	
	Place A	Place B
30-60	28	39
60-90	292	284
90-120	380	401
120-150	212	202
150-180	59	48
180-210	18	31
210-240	2	5

(P.U., B.A., (Pan-I), 1961; M.A. Econ. 1970)

What do you understand by Variance? The wages of 1,000 employees range from Rs.4.50 to Rs.19.50. They are grouped in 15 classes with a common class interval of Rs.1, and the class frequencies from the lowest class to the highest class are: 6, 7, 15, 48, 65, 90, 131, 173, 155, 117, 75, 52, 21, 9 and 6. Find the mean wage and its standard deviation.

(P.U., D.Sc., 1965, P.C.S., 1986)

The breaking strength of 20 test pieces of a certain alloy is given as under:

95	103	97	130	96	102	78	95	89	68
82	79	89	67	83	108	94	87	93	117

Calculate the average breaking strength of the alloy and the standard deviation. Calculate the percentage of observations lying within the limits: mean ± 5 , mean $\pm 2S$, mean $\pm 3S$, where S stands for the standard deviation.

(P.U., B.Com., 1961, I.I.U., M.A. Econ. 1989)

A collar manufacturer is considering the production of a new style of collar to attract young men. The following statistics of neck circumferences are available based upon measurements of a typical group of college students:

size-inches	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0	16.5
no. of students	4	19	30	63	66	29	18	1	1

Compute the standard deviation and use the criterion $Z = 3$ (standard deviation) to determine the largest and smallest size of collars he should make in order to meet the needs of practically all his customers, bearing in mind that collars are worn, on average, $\frac{1}{4}$ inches larger than neck sizes.

(P.U., B.A./B.Sc. 1960)

The values of the arithmetic mean and standard deviation of the following frequency distribution of a continuous variable derived from the use of working origin and scale are 11 and 9.6 respectively. Determine the actual class-intervals.

Y	-4	-3	-2	-1	0	1	2	3	Total
f	2	5	8	18	22	13	8	6	89

(P.U., D.Sc., 1964; B.A./B.Sc. 1973)

- * 4.22 In the manufacture of a certain scientific instrument great importance is attached to the life of a particular critical component. This component is obtained in bulk from two sources, A and B, and in the course of inspection, the lives of 1,000 of the components from each source were determined. The following frequency tables are obtained:

Age Group	Source A		Source B	
	Life (hours)	No. of components	Life (hours)	No. of components
10-12	1,000 - 1,020	40	1,030 - 1,040	105.5
12-14	1,020 - 1,040	96	1,040 - 1,050	104.5
14-16	1,040 - 1,060	364	1,050 - 1,060	105.5
16-18	1,060 - 1,080	372	1,060 - 1,070	106.5
18-20	1,080 - 1,100	85	1,070 - 1,080	107.5
20-22	1,100 - 1,120	43	1,080 - 1,090	108.5

Examine the effectiveness of the measures of dispersion with which you are familiar in comparing the dispersions of the two distributions.

- 4.23 Show that, for any discrete distribution, the mean deviation about the mean is not greater than the standard deviation. (P.U., B.A./B.Sc.)

Solution: By definition,

$$\text{Mean Deviation} = \frac{1}{n} \sum f_i |x_i - \bar{x}|, \text{ and}$$

$$s = \sqrt{\frac{1}{n} \sum f_i (x_i - \bar{x})^2}, \text{ where } \sum f_i = n$$

We are required to show that

$$\frac{1}{n} \sum f_i d_i^2 \geq \left(\frac{1}{n} \sum f_i d_i \right)^2 \quad \text{here } d_i = |x_i - \bar{x}|$$

$$\text{i.e. } s[\sum f_i d_i^2] \geq (\sum f_i d_i)^2$$

$$\text{i.e. } (f_1 + f_2 + \dots + f_n)(d_1^2 + d_2^2 + \dots + d_n^2) \geq (f_1 d_1 + f_2 d_2 + \dots + f_n d_n)^2$$

$$\text{i.e. } \sum f_i^2 d_i^2 + \sum_{i \neq j} f_i f_j (d_i^2 + d_j^2) \geq \sum f_i^2 d_i^2 + 2 \sum_{i \neq j} f_i f_j d_i d_j$$

$$\text{i.e. } \sum_{i \neq j} f_i f_j (d_i - d_j)^2 \geq 0$$

which is true because $(d_i - d_j)^2$ is always positive.

- 4.24 a) What is the co-efficient of variation? What purpose does it serve?
 b) The following data have been obtained from a frequency distribution of a variable making the substitution $N = 62 + 5n$; $\sum f = 120$, $\sum f u^2 = 398$, calculate co-efficient of variation, using corrected standard deviation. (P.U., B.A./B.Sc.)
- 4.25 a) What do you mean by absolute and relative measures of dispersion? State the uses of co-efficient of variation in statistical analysis. (P.U., B.A./B.Sc.)

- (b) Given below is the distribution of weekly income (in the nearest rupee) of 100 households in a locality A. Calculate the standard deviation.

Income	35-39	40-44	45-49	50-54	55-59	60-64	65-69
f	13	15	17	28	12	10	5

If 123 households in a different locality B had a mean weekly income of Rs.52.28 and a standard deviation of Rs.4.96, then compare the variability of the weekly income of two localities. (B.Z.U., M.A. Econ. 1986; P.U., B.A./B.Sc. 1974)

- a) Define Range, Mean Deviation and Standard Deviation, and compare their merits as descriptive measures of dispersion.
 b) Two candidates X and Y at the B.A. (Hons.) Examination obtained the following marks in ten papers. Which of the candidate showed a more consistent performance?

Paper	I	II	III	IV	V	VI	VII	VIII	IX	X
X	58	49	76	80	47	72	61	59	77	48
Y	39	38	86	72	75	69	57	83	56	66

(B.Z.U., M.A. Econ. 1986; P.U., B.A./B.Sc. 1974)

- a) Calculate the corrected co-efficient of variation when mean = 67.45, variance (uncorrected) = 8.5275 and the class-interval = 10. (P.U., B.A./B.Sc. 1958)
 b) The following are the scores made by two batsmen A and B in a series of innings:

A	12	15	6	73	7	19	199	36	34	29
B	47	12	76	48	7	51	37	48	13	0

Who is better as a run getter? Who is the more consistent player?

(P.U., B.A./B.Sc. 1970; B.Z.U., M.A. Econ. 1984)

- a) Explain the difference between absolute dispersion and relative dispersion. Describe the properties of the standard deviation.
 b) Find the co-efficient of variation from the following data using both uncorrected and corrected standard deviations.

Weight (kg)	118-125	127-135	136-144	145-153	154-162	163-171	172-180
f	3	5	- 9	12	5	4	2

- A manufacturer of television tubes has two types of tubes A and B. The tubes have respective mean life-times $T_A = 1495$ hours and $T_B = 1895$ hours and standard deviations $S_A = 280$ hours and $S_B = 310$ hours. Which tube has the greater (i) absolute dispersion, (ii) relative dispersion? (P.U., B.A./B.Sc. 1969, 71)

- 4.29 Compare the variability of expenditure of families in two towns as given below:

Expenditure (Rupees)	No. of Families	
	Town A	Town B
21 - 30	3	2
31 - 40	61	14
41 - 50	132	20
51 - 60	153	27
61 - 70	140	28
71 - 80	51	7
81 - 90	2	2

(B.Z.U., M.A. Econ.)

- 4.30 Compare the variations in the following frequency distributions of weights of boys computing the co-efficient of variation in each case. Also draw a box plot.

Weight (kilograms)	Classes		
	A	B	C
20 \leq x $<$ 30	7	5	6
30 \leq x $<$ 40	10	9	25
40 \leq x $<$ 50	20	21	24
50 \leq x $<$ 60	18		4
60 \leq x $<$ 70	7	6	3

- 4.31 If in a series of measurements, we obtain n_1 values of magnitude x_1, n_2 of magnitude x_2 and so on, and if \bar{X} is the mean value of all the measurements, prove that the s.d. is

$$\sqrt{\frac{\sum n_i (x_i - \bar{X})^2}{\sum n_i}} = \delta^2, \text{ where } \bar{X} = k + \delta.$$

(B.Z.U., B.A./B.Sc.)

- 4.32 a) For a group of 50 boys, the mean score and standard deviation of scores on a test are 54.5 and 8.5. For a group of 40 girls, the mean and standard deviation are 54.0 and 8.2 for the same test. Find the mean and standard deviation for the combined group of children.
 b) A distribution consists of three components with frequencies 200, 250 and 350 with means of 25, 10 and 15, and standard deviations of 3, 4 and 5 respectively. Show that the mean of the combined distribution is 16 and its s.d. is 7.2 approximately. Find the coefficient of variation.
 c) What is meant by a standardized variable?
 d) Show that for any distribution expressed in standard measures, the mean is zero and standard deviation is one.

(P.U., B.A./B.Sc.)

- c) The mean of scores for a group of students on a certain test was 63.7 with a standard deviation of 12.3. Find the Z score for the top student, with a score of 98 and the bottom student, with a score of 21.
- 4.24 Three tests had the values in the following table for mean and standard deviations.

Test	\bar{x}	s
1	70	5
2	75	8
3	60	12

$$\hat{X} = \bar{X} + \frac{s}{\sqrt{n}}$$

Student A received grades of 70, 90, 70 on the three tests, while student B received grades of 90, 70, 70. Assuming that all three tests should carry equal weight, change the grades to z-scores and average each student's scores.

(P.U., B.Sc. (Hons.) Part-I, 1971)

- 4.25 a) Explain what is meant by the trimmed and Winsorized measures of central tendency and dispersion.
 b) Calculate the trimmed and the Winsorized measure and standard deviations for the following set of 15 scores:
 80, 75, 42, 63, 65, 43, 78, 96, 82, 58, 79, 72, 67, 55.
- 4.26 a) Define Moments about an arbitrary origin and about the mean. Express the moments about the mean in terms of the moments about any point and conversely.
 b) Give Sheppard's corrections to moments and explain where they are used.
- (P.U., B.A. (Part-I), 1961, 1982-S)
- 4.27 a) Prove that the second moment about an arbitrary origin equals the second moment about the mean increased by the square of the distance between the arbitrary origin and the mean.
 (B.Z.U., B.A./B.Sc. 1976)
 b) Derive the relations which give the third and fourth moments about the mean in terms of moments about the origin.
- 4.28 a) What aspects or characteristics of a frequency distribution are measured by the moments?
 (P.U., B.A./B.Sc. 2006)
 b) Prove the general formula connecting the moments about the mean with the moments about the origin.

$$\mu_r = \mu'_r - r\mu'_r \mu'_{r-1} + \frac{r(r-1)}{2!} \mu'_1^2 \mu'_{r-2} - \dots$$

(P.U., B.A./B.Sc. 1985, 89, 06)

- 4.29 a) Explain how changes of origin and of scale affect the following:
 Mean, standard deviation, moments and β_1 .
 (P.U., B.A./B.Sc. 1985, 89)
 b) The first four moments of a distribution about $x = 2$ are 1, 2.5, 5.5 and 16. Calculate the four moments about the mean and about zero.
 (P.U., B.A./B.Sc. 1973, 76, 09)

- b) The following data have been obtained from a frequency distribution of a variable X after making the substitution $x = 10 + 5u$.

$$\sum f = 125, \sum fu = -46, \sum fu^2 = 806, \sum fu^3 = -242 \text{ and } \sum fu^4 = 1962.$$

Calculate the Mean, Variance, b_1 and b_2 (moment-ratios). Would you consider the distribution normal? (P.U., B.A./B.Sc. 1973, 88)

- 4.40 Calculate the first four moments about the mean for the following data.

x	1	2	3	4	5	6	7	8	9
f	1	6	13	23	30	22	9	5	2

- 4.41 Calculate the first four moments and apply the test of normality to the following data which pertain to weekly earnings in rupees of 200 labourers.

Weekly wages	15	16	17	18	19	20	21	22	23	24
No. of Labourers	6	19	13	18	20	25	28	34	22	15

- 4.42 Compute the first four moments about the mean from the following frequency distribution. Also calculate b_1 and b_2 .

Groups	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20
Frequency	18	24	47	80	102	66	40	21	15

(P.U., D.S.L. 1967)

- 4.43 a) Define the moment-ratios β_1 and β_2 . What information do they give?
 b) Eight coins were tossed together 256 times and the frequency of the occurrence of 0, 1, 2, ..., 8 heads in a throw was:

Number of heads	0	1	2	3	4	5	6	7	8
Frequency	1	26	54	74	52	32	9	1	

Calculate the values of moment-ratios b_1 and b_2 .

- 4.44 Calculate the values of b_1 and b_2 from the following data. Use Charlier check and Shepard's correction.

x	10-12	12-14	14-16	16-18	18-20	20-22	22-24	24-26	26-28
f	3	0	160	218	275	223	108	32	2

- 4.45 a) What is meant by skewness? How would you find it in a non-symmetrical distribution? Distinguish between positive and negative skewness.

- b) Define the moment-ratios β_1 and β_2 , and state the purpose for which they are used. (P.U., B.A. (Part-I), 1967)

- 4.46 a) State the measures commonly employed to define Skewness and Kurtosis. What aspects of the frequency curve are measured by them?

- b) What can you say of the skewness in each of the following cases?

i) The median is 49.21 while the two quartiles are 37.15 and 61.27.

ii) Mean = 1403 and Mode = 1487.

iii) The first three moments about 16 are respectively -0.35, 2.09 and -1.93.

(P.U., B.A./B.Sc. 1974, 78; M.A. Econ. 1977)

- Q.47 Calculate skewness by (i) the Pearsonian method, and (ii) by the Bowley's formula from the following frequency distribution and interpret the result.

Age (years)	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54
No. of Men	29	176	208	173	82	40	15	3

(P.U., M.A. Econ. 1986)

- Q.48 Calculate the first four moments about the mean and provide one estimate each of skewness and kurtosis of the following distribution.

Age:	25	30	35	40	45	50	55	60
Frequency:	2	8	18	27	25	16	7	2

(P.U., M.A. Econ. 1967, I.U. 1992)

- Q.49 The following values have been obtained from two different frequency distributions of weights (lb) having 125 and 200 observations respectively after making the substitution:

$$X = 16 + 5w, Y = 20 + 2v,$$

$$\text{a)} \sum f_w = -46, \sum f_w^2 = 306, \sum f_w^3 = -341, \sum f_w^4 = 1962$$

$$\text{b)} \sum f_v = 21, \sum f_v^2 = 1265, \sum f_v^3 = -627, \sum f_v^4 = 14169$$

Find i) which of the distributions is more consistent.

ii) which of the distributions is negatively skewed.

iii) which of the distributions is Mesokurtic.

(P.U., B.A./B.Sc. 1993)

- Q.50 a) The fourth mean moment of a symmetric distribution is 243. What would be the value of the standard deviation in order that the distribution may be mesokurtic?

(P.U., B.A./B.Sc. 1981)

- b) In a certain distribution, the first four moments about the point 4 are -1.5, 17, -30 and 100. Calculate k_1 and k_2 , and state whether the distribution is leptokurtic or platykurtic.

(P.U., D.Sc. 1962)

- c) The second moments about the mean of two distributions are 9 and 16, while the fourth moments about the mean are 230 and 780 respectively. Which of the distribution is (i) leptokurtic, (ii) mesokurtic, (iii) platykurtic?

- d) The second moment about the mean of a symmetrical distribution is 25. What must be the value of the fourth moment about the mean in order that the distribution be (a) leptokurtic, (ii) mesokurtic, (iii) platykurtic?

(P.U., B.A./B.Sc. 1971)

- Q.51 Discuss the various measures or quantities by which the characteristics of frequency distributions are measured and compared.

(P.C.S. 1993)

- Q.52 Explain briefly how averages, measures of dispersion, skewness and kurtosis are complementary to one another in understanding a frequency distributions.

- Q.53 The mean and standard deviation of a variable X are 60 and 8.944 respectively. Find the mean and standard deviation of a new variable if

- i) All the values of X are increased by 20 points.

- ii) All the values of X are increased by 25%.

(P.U., B.A./B.Sc. 2008-S)

- 4.55 The mean, mode and standard deviation of the weekly earnings of a random sample of women workers from a locality are 3133.33, 2804.35 and 796.70 respectively.
- Calculate Skewness of the distribution and interpret the result. Also find coefficient of variation.
 - What will happen to the values of the mean and standard deviation if every woman has an increase of Rs.500 per week?
 - What will the mean and standard deviation be if every woman has an increase of 10% in previous earnings?

(P.U., B.A./B.Sc. 2007)

CHAPTER 5

INDEX NUMBERS

INDEX NUMBERS

5.1 INTRODUCTION

An **index number** is a statistical measure of average change in a variable or a group of variables with respect to time or space. The variable may be the enrolment of students in an institution, the cost of education for college students, prices of a particular commodity or a group of commodities, wages of workers, volume of trade, sales, exports and imports, production, unemployment, gross health, Government securities, etc. Index numbers are obtained by expressing the data for various periods or places as percentage of some specific period or place selected for the purposes of comparison and technically called the **base**. Index numbers may be computed on weekly or monthly basis but generally they are computed on annual basis.

The classical definition of an index number is provided by the English economist F.Y. Edgeworth (1843–1926) who states that "**an index number is a quantity which shows by its variations, the changes over time or space of a magnitude which is not susceptible either of accurate measurement in itself or direct valuation in practice.**"

In short, an index number is a device that measures the changes occurring in data from time to time or from place to place.

5.1.1 Simple and Composite Index Numbers. Index numbers are generally classified into Simple indices and Composite Indices. An index number is called a simple index when it is computed for a single variable. Index numbers of enrolment in colleges, index numbers of gold prices, etc. are examples of simple indices. A simple index can be very easily computed. The value of the variable for each period is divided by the value in the base period and the result is multiplied by 100. For example, the wages paid to the workers in a certain institution in 1980 and 1983 were Rs. 9,650 and Rs. 11,380 respectively. Now taking 1980 as the base year and 1983 as the given year, we have

$$\text{Wage Index for 1983} = \frac{\text{wages paid in 1983}}{\text{wages paid in 1980}} \times 100$$
$$\frac{\text{Rs. } 11,380}{\text{Rs. } 9,650} \times 100 = 120$$

This result indicates that if the wage level in 1980 were denoted by 100, it is 120 in 1983. In other words, wages have increased by 20% in 1983 by comparison with 1980.

An index that is computed from two or more variables is referred to as a composite index. Examples of composite indices are the wholesale price index numbers, consumer price index numbers, etc. Composite indices are constructed in many of the index numbers in commerce because they are composite in nature.

Composite indices may further be classified into *Unweighted* and *Weighted* index numbers. Before we discuss these index numbers and the methods of their computation, let us first consider the problems, we have to face, in the compilation of these index numbers.

5.1.2 Problems Involved in Index Number Construction. In the compilation of index numbers, many problems are involved. The first problem is to understand the purpose which an index is to serve. The purpose of the index may be to compare the scores of two students, or to measure the changes in the general price level or to measure the changes in the production of scooters or to compute the changes in wages of factory workers over different places, etc. The next problem is to decide what data should be included. The data to be included should relate to purpose for which the index is to be used. This step also

involves the collection of data on items, production process, wages or whatever is being compared. Another problem is to decide what period should be chosen as the base period, i.e. the period with which the other periods are to be compared. In case of composite index numbers, another problem is to decide what method of averaging should be used to arrive at a single index for each period. The method of averaging usually includes the system of weighting but sometimes one faces the problem of assigning some explicit weights to the various items of the data so that their relative importance is taken into account.

These problems are discussed in somewhat detail in section 5.2 with reference to the construction of price index numbers as the indices used to measure the change in price level are more important.

5.2 MAIN STEPS IN THE CONSTRUCTION OF INDEX NUMBERS OF WHOLESALE PRICES

The usual method of compilation of an index number of wholesale prices involves the following steps:

- i) Selection of commodities to be included, their number and price quotations.
- ii) Selection of the base period and calculation of price relatives.
- iii) Selection of average to be used.
- iv) Selection of appropriate weights.

5.2.1 Selection of Commodities for Inclusion. The first step is to decide on the number of commodities to be included. There is no hard and fast rule for this purpose. Though in statistical theory there is a well recognised principle that *the larger the number of items included, the greater would be the accuracy*, but we know that a very large number of commodities would involve complications, expense and delay in the construction. Hence a reasonable number of commodities on the basis of their evaluated importance should be used.

As pointed out by Dr. Irving Fisher (1867-1947), *index numbers of prices are seldom of much value unless they consist of more than 30 commodities and 50 is a much better number*. However, it is important to bear in mind that while deciding on their number, the commodities to be selected must be (i) representative of the tastes, habits and requirements of the people concerned, (ii) unlikely to vary in quality or grade and (iii) comparable.

Having decided on the number of commodities, arrangements are made to collect the wholesale prices of the commodities chosen. The prices should be obtained from the various sources, e.g. from exchanges and big markets where they are quoted; from price bulletins, trade journals and newspapers; and from leading firms. The price quotations obtained should be representative, reliable and comparable as regards the quality of the commodities and the units in which the commodities are expressed.

5.2.2 Selection of the Base Period. The next step is the selection of a base period-a period from which the changes are measured. The prices of all other periods are then expressed as percentages of the base period prices. Two methods of selecting the base period are available. They are the Fixed base method and the Chain base method.

Fixed Base Method. A fixed base method is one in which a particular year is generally chosen as the base period that remains unchanged during the life term of the index. It is relevant to note that the base year should not be too far distant in the past and should be a "normal" year. By a "normal" year, we generally mean a year of economic stability and free from any major financial crisis caused by inflation.

inflation, wars, labour unrest, lock-outs, famines, etc. In other words, it is a year during which prices have remained more or less stable. If a single year of normal conditions is not available, an average of prices of several years is used as the base period price. This average minimizes the influence of unusual and disrupted economic conditions.

It is customary to denote the base period by the subscript 0, e.g. p_0 (or q_0) will denote the price (or quantity) of a commodity in the base period, while the subscripts, 1, 2, ..., n denote the other time periods in chronological order. The average price of the base period chosen is then set equal to 100. Index numbers (or price relatives) for other periods denoted by $P_{01}, P_{02}, \dots, P_{0n}$ are then computed as relative to the base period. Thus the price relative for the given year n , will be

$$\text{Price relative} = \frac{\text{Price of a commodity in the given year}}{\text{Price of the commodity in the base year}} \times 100$$

$$= \frac{P_n}{P_0} \times 100$$

This relative expresses the price of a commodity in a given year as a fraction of the price in the base year. It has multiplied by 100 to make it a percentage but is usually expressed without the percent symbol. This relative is independent of any unit of measurement.

Chain Base Method. A chain base method is one in which the base period is not fixed but moves with the given year. That is, the relatives are computed with the immediately preceding year as the base period. Such relatives are called link relatives. Thus a

$$\text{Link relative} = \frac{\text{Price of a commodity in the given year}}{\text{Price of the commodity in the preceding year}} \times 100$$

$$= \frac{P_{n-1,n}}{P_{n-1}} \times 100$$

The link relatives are converted back to a fixed base by multiplying together all the link relatives with the factor 100 involved between the two years. This process of conversion is called the *chainning* and the indices thus determined are the *chain indices*. In other words, we say that the link relatives are "chained" back to a fixed base period by a process of successive multiplication and hence, the name "chain index". For example, if $P_{01}, P_{02}, P_{03}, \dots, P_{0n}$ denote the link relatives (or average of relatives) without the factor 100, then the indices on fixed base are obtained as below:

$$P_{01} = P_{01}$$

$$P_{02} = P_{01} \times P_{02}$$

$$P_{03} = P_{01} \times P_{02} \times P_{03}$$

$$P_{0n} = P_{01} \times P_{02} \times P_{03} \times \dots \times P_{0n-1,n}$$

If link relatives are in percentages, the product considered pair-wise, is divided by 100. It is interesting to note that this chain index was originally suggested in 1887 by Alfred Marshall (1842-1924).

The chain index method has several advantages. They are:

- i) The chain method provides a direct comparison between each year and the preceding year as it is in such terms that a business man often thinks.
- ii) The chain base method allows the addition of new commodities, removal of old commodities or the substitution of one commodity for another.
- iii) It is possible to change the geographical coverage or the weight of a commodity in the changing conditions.
- iv) It satisfies the so-called circular test (to be explained later).
- v) An index with a fixed period can be computed by the product of link relatives.

It suffers, however, from the following disadvantages:

- i) The computational procedure of chain indices is relatively cumbersome.
- ii) If an error is committed during the changing process, it will be carried through the entire series.
- iii) The changes in two years separated by a long interval cannot be compared.

Since the merits of the chain index outweigh its demerits, it is therefore considered a more reliable index.

5.2.3 Selection of Average. The next step involves the choice of an appropriate average to get a single index number for each year. For this purpose, any of the following averages may be used:

- (a) The arithmetic mean, (b) the median and (c) the geometric mean.

The advantages and disadvantages of these averages in the construction of index numbers are given below:

The Arithmetic Mean. It is easily understood and is easier to compute. It is amenable to mathematical treatment, i.e. the means of subgroups can be averaged to find the mean of all the data. Disadvantages of the mean are that it is greatly affected by extreme values; it gives too much weight to increasing prices and too little to decreasing ones, i.e. it is biased upward. Moreover, the mean of relatives is not reversible, i.e. changing the base cannot be made without affecting the proportionate change in the index number.

The Median. It is easy to understand as well as to compute. It is less affected by extreme values than the mean and does not overemphasize increases. The defects of the median are that it is not amenable to algebraic treatment, i.e. the medians of sub-groups cannot be averaged to obtain the mean of all the data, and the median of relatives is not reversible.

The Geometric Mean. The geometric mean is a suitable average for ratios as it gives equal importance to equal ratios of change. It makes possible to replace the commodities which have ceased to be representative by those which have become representative. The geometric mean of relatives is reversible and hence we can change a series of index numbers with any year as base to any other year if the series is base by dividing each index number of the series by the index number of the new time period selected for base. But it has two disadvantages. It is an unfamiliar type of average and it involves considerable computational labour.

Theoretically, the geometric mean is the most suitable average but, in practice, the arithmetic mean is generally employed.

5.2.4 Selection of Appropriate Weights. The last important step in the construction of wholesale price numbers is to decide how to select weights which would indicate the relative importance of various commodities in the group. It is evident that all the commodities selected are not equally important. For example, eggs and coffee cannot be given the same importance as wheat and rice. Wheat is much more important than coffee; it is therefore desirable that wheat must be given more importance. To meet this requirement of taking into account the relative importance of each commodity, a sample survey should be conducted. On the basis of this survey, each commodity should be assigned a multiplier which expresses more or less adequately its relative importance in the group. Such a multiplier is usually called a weight. The weights could be either a set of suitable numbers adding to 100, or the values of the various commodities actually consumed, produced or sold, or their money values. It is necessary to apply quantity weights when dealing with prices themselves and value weights when using price relatives. The quantity or value weights may relate to the base period or to the given period. Then the base period quantity (q_0) is used as weight, the price index number, P_{0n} for the year n is obtained by the formula:

$$P_{0n} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

This is called the base-year weighting. With the given period quantity (q_n) as weight, the corresponding formula becomes:

$$P_{0n} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$$

This is also known as the given or current year weighting. It is relevant to note that different systems of weighting would not generally lead to identical results.

5.3 UNWEIGHTED INDEX NUMBER

Unweighted indices are generally classified into Simple Aggregative indices and Simple Average of Price relatives.

5.3.1 Simple Aggregative Index is one that indicates the percentage change in the aggregate price of a number of commodities, (say k) at different periods. It is obtained by dividing the sum of the current year prices of all commodities by the sum of the base year prices of the same commodities and expressing the result as a percentage. Symbolically, we have

$$P_{0n} = \frac{\sum p_n}{\sum p_0} \times 100,$$

P_{0n} denotes the price index for the given year n relative to the base year 0.

$\sum p_n$ denotes the sum of prices for the given year; and

$\sum p_0$ denotes the sum of prices for the base year.

A simple aggregative index is easy to understand as well as to apply but it has a disadvantage of not taking into account the relative importance of the various commodities. Moreover, the units of prices of different commodities being different, influence the price index, which then becomes an inappropriate measure. That is why it is seldom used in practice.

5.3.2 Simple Average of Relatives. A simple average of price relatives is an index obtained taking the average of the price relatives of the given commodities for each year and expressing these as a percentage. If we take the arithmetic mean, then we have

$$P_{\text{Avg}} = \frac{1}{k} \sum \left(\frac{P_n}{P_0} \right) \times 100,$$

where k denotes the number of commodities whose price relatives are thus combined.

The simple average of relatives index is superior to simple aggregative index. It suffers from disadvantage that each price relative carries equal influence and gives no consideration to the relative importance of each commodity. Moreover, the use of arithmetic mean, which is not an appropriate average to use with ratios, results in an upward bias. This sort of drawback may be got rid of by using geometric average, in which case the formula becomes

$$P_{\text{Avg}} = \left[\prod \left(\frac{P_n}{P_0} \right) \right]^{1/k} \times 100,$$

where k denotes the number of commodities whose price relatives are being averaged.

The median may also be used for averaging the price relatives.

Example 5.1 Following are the prices of a commodity for the ten years ending with 1957. Calculate the index numbers with (i) 1948 as a base; and (ii) average of first five years as a base.

Year	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957
Price in Ru.	5.25	5.87	6.12	5.50	6.25	6.62	6.75	7.12	6.50	7.50

(P.U., B.A. (Hons. in Econ.) 1996)

Let P_0 denote base year's price and P_n the given year's price. Then the prices are converted into price relatives or price indices by the formula

$$\text{Price relative for a given year} = \frac{P_n}{P_0} \times 100,$$

Year	Prices (rupees)	(i) Index No. $\left(= \frac{P_n}{P_0} \times 100 \right)$	(ii) Index No. $\left(= \frac{P_n}{P_0} \times 100 \right)$
		(1948 as base)	(average as base)
1948	5.25	100	91
1949	5.87	112	101
1950	6.12	117	106
1951	5.50	105	95
1952	6.25	119	108
1953	6.62	126	114
1954	6.75	129	116
1955	7.12	136	123
1956	6.50	124	112
1957	7.50	143	129

Example 5.2 The average prices of certain commodities are given below:

Commodity	Unit	Price (Rs.) in	
		1957	1958
Wheat	Trn	351.00	325.00
Rice	40 kg	35.00	32.00
Salt	40 kg	10.00	11.00
Sugar	kg	1.25	1.40
Cloth	Meter	2.25	2.60
Milk	Litre	0.75	0.85
Oil	Gallons	1.25	1.35

Compute a simple aggregative price index number for the year 1958 based on 1957 prices.

(P.U., M.A. Econ. 1969-S)

The simple aggregative price index number is given by

$$\begin{aligned} P_{0w} &= \frac{\sum p_w}{\sum p_0} \times 100 \\ &= \frac{325.00 + 32.00 + \dots + 1.35}{351.00 + 35.00 + \dots + 1.25} \times 100 = \frac{384.20}{401.50} \times 100 \end{aligned}$$

The simple aggregative price index for 1958 on the base of 1957 is 95.7. This means that the prices have decreased by 4.3%.

Example 5.3 Compute price index for the year 1958 based on 1957 prices, using the simple relatives method for the data in Example 5.2.

The computation of the simple average price relatives index involves the calculation of price relatives for 1958 by dividing the prices in 1958 by the corresponding prices in 1957. The price relatives are shown below:

Commodity	Price relatives	
	1957	1958
Wheat	100	95.44
Rice	100	91.43
Salt	100	110.00
Sugar	100	112.00
Cloth	100	115.56
Milk	100	113.33
Oil	100	108.00
Σ	-	745.76

Then the simple average of price relatives for 1958 is

$$\begin{aligned} P_{0w} &= \frac{1}{K} \sum \left(\frac{p_w}{p_0} \right) \times 100 \\ &= \frac{745.76}{7} = 106.5 \end{aligned}$$

This indicates an increase in prices from 1957 to 1958 by 6.5%.

It should be noted that the index numbers obtained by the two methods show considerable difference because the simple aggregative index has been affected by the units of prices of all commodities, and although the influence due to different units has been eliminated by the simple average of relatives, commodities such as salt, sugar, cloth, etc. have exercised an influence out of all proportion of their economic importance.

Example 5.4 From the data given below, compute the index numbers of prices, taking 1980 as base. Use (i) simple average of price relatives and (ii) the median of price relatives.

Year	Commodity (Prices in Rs.)			
	A	B	C	D
1980	16.25	20.00	2.40	10.30
1981	17.22	22.40	2.64	12.50
1982	19.55	16.00	3.00	12.60
1983	18.70	20.00	3.80	14.65

We calculate the price index numbers as in the following table:

Year	Price Relatives				Total	Index Numbers by	
	A	B	C	D		(i) Mean	(ii) Median
1980	100	100	100	100	400	100	100
1981	106	112	110	119	447	112	111
1982	120	80	125	116	441	111	120
1983	115	100	158	140	513	128	128

The first entry in row 1981 is $\frac{P_1}{P_0} = \frac{\text{Rs } 17.22}{\text{Rs } 16.25} = 106$, and the price index for 1981 is 112.

$$\frac{1}{4} \times \left(\frac{P_1}{P_0} \right) \times 100 = \frac{447}{4} = 111.75$$

The other entries have been computed in a similar way.

The price index numbers appearing in the last column of the table, have been obtained by finding the medians of the price relatives.

Example 5.5. Construct chain indices for the following years, taking 1940 as the base and using the simple average of relatives.

Year	Price in Rs. per maund (1 maund = 37.2 kg)		
	Wheat	Rice	Maize
1940	2.80	10.50	2.70
1941	3.40	10.80	3.20
1942	3.60	10.60	3.50
1943	4.00	11.00	3.80
1944	4.20	11.50	4.00

Now we calculate the link relatives by the formula

$$P_{n-1, n} = \frac{\text{Price of a commodity in the given year}}{\text{Price of the commodity in the preceding year}} \times 100$$

$$= \frac{P_n}{P_{n-1}} \times 100$$

Next, we multiply the averages of link relatives successively (taking two at a time) and divide the product by 100 to chain the relatives back to the base 1940. The calculations appear in the table below:

Year	Link Relatives			Sum of relatives	Simple average of relatives	Chain Indices
	Wheat	Rice	Maize			
1940	100	100	100	300	100	100
1941	$\frac{3.40}{2.80} \times 100 = 121$	103	119	343	114	$\frac{100 \times 114}{100} = 114$
1942	$\frac{3.60}{3.40} \times 100 = 106$	98	109	313	104	$\frac{114 \times 104}{100} = 118.6$
1943	$\frac{4.00}{3.60} \times 100 = 111$	104	109	324	106	$\frac{118.6 \times 106}{100} = 126.1$
1944	$\frac{4.20}{4.00} \times 100 = 105$	105	105	315	105	$\frac{126.1 \times 105}{100} = 134.5$

The Chain Index Numbers are 100, 114, 118.6, 126.1, 134.5.

Example 5.6 Find the chain indices from the following price relatives of four commodities, using the geometric mean of the relatives for each year.

Year	Commodity		
	B	C	D
1991	77	119	55
1992	62	54	128
1993	104	87	111
1994	93	75	154
1995	66	43	165

First we calculate the link relatives. The link relatives for commodity A (for instance) are given below:

$$\text{Link relative for 1992} = \frac{62}{81} \times 100 = 76$$

$$\text{Link relative for 1993} = \frac{104}{62} \times 100 = 168$$

$$\text{Link relative for 1994} = \frac{93}{104} \times 100 = 89$$

$$\text{Link relative for 1995} = \frac{66}{93} \times 100 = 65$$

The link relatives for other commodities are obtained in a similar way. The chain indices, i.e., geometric mean are then computed and shown in the table below:

Calculation of chain indices, using the geometric means of the relatives.

Year	Link Relatives of Commodity				$\sum \log x$	$\frac{1}{4} \sum \log x$	G.M.	Chain Index Numbers
	A	B	C	D				
1991	81	77	119	55	7.6109	1.9027	79.9	79.9
	log	1.9085	1.8865	2.0755				
1992	76	70	108	59	7.9325	1.9831	96.2	79.9×96.2 100 -76.9
	log	1.8808	1.8451	2.0334				
1993	168	161	87	122	8.4580	2.1145	130.2	16.9×130.2 100 = 100.1
	log	2.2253	2.2068	1.9395				
1994	89	86	139	96	8.0923	2.0023	100.6	100.1×100.6 100 = 100.7
	log	1.9494	1.9345	2.1430				
1995	65	57	107	93	7.5667	1.8917	77.9	100.7×77.9 100 -78.4
	log	1.8129	1.7559	2.0294				

Hence the chain indices by geometric means are 79.9, 76.9, 100.1, 100.7 and 78.4.

5.4 WEIGHTED INDEX NUMBERS

An index number that measures the change in the prices of a group of commodities when the relative importance of the commodities (i.e. weight) has been taken into account, is called a weighted price index number. Weighted indices are generally divided into Weighted Aggregative indices or Weighted Average of relatives indices.

5.4.1 Weighted Aggregative Price Index Numbers. An index is called a weighted aggregate index when it is constructed for an aggregate of items (prices) that have been weighted in some way corresponding quantities produced, consumed or sold so as to reflect their importance. There are many kinds of weighted aggregative index numbers, some of them are discussed below:

(a) Laspeyres' Price Index, advocated by the German economist E. Laspeyres (1834-1913) in 1864, is defined as

$$P_{0e} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

This is the percentage ratio of the aggregate of the given period prices weighted by the quantities produced, consumed or sold in the base period to the aggregate of base period prices weighted by the base period quantities. The index represents the relative cost in different years of purchasing the base year quantities of various commodities at the given year price. The advantage of Laspeyres formula is that the weight remains unchanged for the subsequent periods and only inflation in latest prices need be taken into account. It has, however, a few limitations. The index value obtained by this formula gets somewhat altered as the series move away from the base period. That is why an index of Laspeyres type is said to have an *asymptotic bias*. It does not satisfy the time-reversal test or the factor-reversal test (discussed later).

(b) Paasche's Price Index, proposed in 1874 by the German economist Herman Paasche (1815-1925) is given by the relation

$$P_{0n} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$$

This is the percentage ratio of the aggregate of given period prices weighed by the quantities produced, consumed or sold in the given period to the aggregate of base period prices weighted by the given period quantities. It represents the relative cost in different years of purchasing the given year quantities of various commodities at the given year price. The computation of this type of index needs base data on quantities (i.e. weights) and to obtain such information on quantities for each given year, an enquiry or a statistical survey, which normally involves considerable time and finances, would be required every year and this is a difficult task. This index has a *moreward bias*, i.e. it deflates the index over long periods of time. It also does not obey the time-reversal test or the factor-reversal test.

(c) Marshall-Edgeworth Price Index. This index was proposed independently by the two English economists Alfred Marshall (1842-1924) and F.Y. Edgeworth (1845-1926). Here the weights are taken as average of the respective quantities in the base period and in the given period. This is, so to say, a 'middle' solution, although it is the index which has no general bias in either direction. Since π_n lies between π_0 and π_n , the Marshall-Edgeworth's index number lies between the Laspeyres' and Paasche's index numbers. The formula for this index is

$$P_{0n} = \frac{\sum p_n (\pi_0 + \pi_n)}{\sum p_0 (\pi_0 + \pi_n)} \times 100$$

(d) Fisher's "Ideal" Index. Fisher's ideal index number, named after its inventor, Irving Fisher (1872), is the geometric mean of the Laspeyres and Paasche type of index numbers. Symbolically, it is given by

$$P_{0n} (\text{Fisher}) = \sqrt{P_{0n} (\text{Laspeyres}) \times P_{0n} (\text{Paasche})}$$

$$= \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}} \times 100$$

Fisher called it "ideal" index because it meets certain theoretical tests of quality which he considered appropriate for a good index number. It is sometimes known as *crossed-weight formula* as the result of geometrically crossing (averaging) two index numbers with different systems of

Fisher's ideal index has a theoretical advantage over other index numbers as it is the only one that obeys both the time-reversal test and the factor-reversal test. It suffers, however from the following disadvantages:

- (i) The Fisher index number being the geometric mean of the Laspeyres index that has an upward bias and the Paasche index which suffers from a downward bias, is considered to give a better result. But we are not sure about the elimination of the biases as "the average of two wrong answers does not necessarily give one right answer".
 - (ii) It is difficult to say specifically what the Fisher index number measures because of its being a hybrid of two index numbers.
 - (iii) The computation of the Fisher index is relatively difficult and laborious.
 - (iv) Its computation needs information on quantities (consumed, produced or sold) for each given period. To get such information, a fresh survey which may be too costly or time-consuming would be needed every year.
- (c) **Walsh Index.** Walsh advocated that the weights should be the geometric mean of the base-period quantities instead of taking their simple average. Symmetrically, it is given as

$$P_{0q} = \frac{\sum p_0 \sqrt{q_0 q_p}}{\sum p_0 \sqrt{q_0}} \times 100$$

There is another formula known as the Lowe price index number in which average weights are used. It gives as

$$P_{0q} = \frac{\sum p_0 q}{\sum p_0 \bar{q}} \times 100$$

where \bar{q} is obtained by averaging quantities of several years.

Example 5.7 From the following table, compute the weighted aggregative price index for the year 1999 on the basis of the year 1995.

Commodity	Unit	Average monthly consumption per family	Average prices in rupees	
			1995	1999
Milk	Litre	30	1.50	4.00
Flour	Kilogram	25	1.25	1.75
Cloth	Meter	12	4.00	8.00
Tea	Kilogram	1	12.00	18.00
Vegetable ghee	Kilogram	5	7.50	4.00
Eggs	Dozen	4	5.00	7.25

To compute the weighted aggregative price index for the year 1999 with 1995 as base, we need $\sum p_0 q_1$ and $\sum p_0 q_0$. The calculations appear below:

Commodity	1995 (p_0)	1999 (p_1)	1999 (p_1)	$p_1 q_1$	$p_1 q_0$
Milk	30	3.50	4.00	105.00	120.00
Flour	25	1.25	1.75	31.25	43.75
Cloth	12	4.00	8.00	48.00	96.00
Tea	1	12.00	18.00	12.00	18.00
Vegetable ghee	5	7.50	4.00	37.50	20.00
Eggs	4	3.00	7.25	20.00	29.00
Σ	—	—	—	253.75	326.75

$$\text{Index } P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{326.75}{253.75} \times 100 = 128.77$$

The weighted aggregative price index for 1999 on the basis of 1995 = 128.77. This means that the prices have increased by 28.77%.

Example 5.8 Construct the following weighted aggregative price index numbers for 2000 and 2001 from the given data.

Laspeyres' index, (ii) Paasche's index and (iii) Fisher's "ideal" index.

Commodity	Prices (Rs per kg)			Quantities (tons)		
	1996 (base)	2000	2001	1996	2000	2001
A	64	75	80	270	276	290
B	48	45	41	124	118	144
C	18	21	20	130	121	137
D	18	68	56	185	267	355

We calculate the necessary products in the following table:

	Price			Quantity								
	P_0	P_1	P_1	q_0	q_1	q_1	$P_0 q_0$ (1996)	$P_1 q_0$ (2000)	$P_1 q_1$ (2001)	$P_0 q_1$ (1996)	$P_1 q_1$ (2001)	
A	64	75	80	270	238	290	17280	20230	23600	17064	20560	23200
B	48	45	41	124	118	144	5792	5380	5984	4720	5160	5904
C	18	21	20	130	121	137	2340	2730	2680	2178	2460	2340
D	18	68	56	185	267	355	3270	12590	18040	15488	20792	16158
Total	—	—	—	—	—	—	35110	41140	39644	43048	47238	51714

Laspeyres' index numbers:

$$\text{Index for 2000 } (P_{01}) = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{41140}{35310} \times 100 = 116.5$$

$$\text{Index for 2001 } (P_{02}) = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{39644}{35310} \times 100 = 112.3$$

(ii) Paasche's index numbers:

$$\text{Index for 2000 } (P_{a_1}) = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{46707}{40048} \times 100 = 116.6$$

$$\text{Index for 2001 } (P_{a_2}) = \frac{\sum p_2 q_2}{\sum p_1 q_2} \times 100 = \frac{51724}{47376} \times 100 = 109.2$$

(iii) Fisher's "Ideal" index numbers:

$$\begin{aligned}\text{Index for 2000 } (P_{a_1}) &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_1} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 \\ &= \sqrt{\frac{41140}{35310} \times \frac{46707}{40048}} \times 100 \\ &\approx \sqrt{1.165 \times 1.166} \times 100 = 116.5\end{aligned}$$

$$\begin{aligned}\text{Index for 2001 } (P_{a_2}) &= \sqrt{\frac{\sum p_2 q_1}{\sum p_1 q_2} \times \frac{\sum p_2 q_2}{\sum p_1 q_2}} \times 100 \\ &= \sqrt{\frac{51724}{47376} \times \frac{51724}{47376}} \times 100 \\ &= \sqrt{1.123 \times 1.092} \times 100 = 110.7\end{aligned}$$

Example 5.9 The prices and quantities of three commodities during 1990 and 1994 are given below:

Commodity	Prices Rs. per maund		Quantities Produced (maunds)	
	1990 (p_0)	1994 (p_1)	1990 (q_0)	1994 (q_1)
A	3.93	6.25	8,673	10,436
B	34.80	38.94	78	83
C	61.56	59.70	118	116

Compute the Marshall-Edgeworth and the Walsh's price index numbers for 1994, using 1990 as the base period.

First we calculate the necessary products with 1990 as base year. These calculations are shown in the table below:

Construction of Weighted Aggregative Price Indices

Com.	$(q_0 + q_1)$	$p_1 q_0 + q_1$	$p_0 (q_0 + q_1)$	$\sqrt{q_0 q_1}$	$p_1 \sqrt{q_0 q_1}$	$p_0 \sqrt{q_0 q_1}$
A	29,111	85,471.75	79,438.45	10,048.3	42,705.28	39,690.78
B	161	6,269.34	5,602.80	80.5	3,124.67	2,801.40
C	234	13,969.80	14,405.04	117.0	6,984.90	7,262.52
Total	—	105,710.89	99,446.29	—	52,824.85	49,694.70

thus the Marshall-Edgeworth price index = $\frac{\sum p_i (q_0 + q_1)}{\sum p_i (q_0 + q_1)} \times 100$

$$= \frac{105710.89}{99446.29} \times 100 = 106.2$$

and the Walsh price index = $\frac{\sum p_i \sqrt{q_0 q_1}}{\sum p_i \sqrt{q_0 q_1}} \times 100$

$$= \frac{52824.85}{49694.7} \times 100 = 106.3.$$

5.4.2. Weighted Average of Relatives (the Index Number). It is computed by multiplying each relative by its weight, summing these products and dividing by the sum of the weights. The weights are always the total values of the commodities. The important types of the weighted average of relatives indices are given below:

(a) Laspeyres' index number is

$$P_{la} = \frac{\sum (p_0 / p_1) p_1 q_0}{\sum p_0 q_0} \times 100$$

thus the price relatives are weighted by the total value of commodities in the base year. This is equal to Laspeyres weighted aggregative price index, i.e. $\frac{\sum p_0 q_0}{\sum p_1 q_0} \times 100$. In other words, these indices are alternative ways of getting the same result.

(b) Paasche's index number is

$$P_{pa} = \frac{\sum \left(\frac{p_1}{p_0} \right) p_0 q_1}{\sum p_1 q_1} \times 100$$

where the price relatives are weighted by the total value of commodities in the given year at base year prices. This is also identical with Paasche's weighted aggregative price index, i.e. $\frac{\sum p_t q_t}{\sum p_0 q_t} \times 100$.

(c) Palgrave's index number is

$$P_{ps} = \frac{\sum (p_s / p_t) p_t q_s}{\sum p_t q_s} \times 100,$$

where the price relatives are weighted by the total value of commodities in the given year.

In all these index numbers, the computational labour is reduced if the weights are made to unity. For example, if $w_s = \frac{p_t q_s}{\sum p_t q_s}$, then the Laspeyres formula becomes:

$$P_{ls} = \sum \left[\left(\frac{p_s}{p_t} \right) w_s \right] \times 100, \quad (\sum w_s = 1)$$

The advantages of this procedure is that it indicates how many points each commodity contributes to the index number each year.

Example 5.10 Using the data of Example 5.1, compute the weighted average of relative price index numbers for 1994 by (i) Laspeyres' method, (ii) Paasche's method, and (iii) Palgrave's method.

Computation of Weighted Average of Relative Price Indices

Com.	Date		Quantity	Price	Weights			Weighted Price Ratios		
	(In 1990)	(In 1994)			(In 1990)	(In 1994)	(In 1990)	(In 1994)	(In 1990)	(In 1994)
A	2.10	4.22	9.872	104.00	1.04	102.04	0.122249	122.00	4121.00	4121.00
B	14.80	18.94	19	100	1.19	104.40	0.198405	102.02	3000.61	3292.11
C	61.35	58.76	110	100	0.978	104.00	0.692530	104.16	3836.71	4010.00
Total	—	—	—	—	—	41584.71	52515.56	34510.32	51286.21	54511.54

Thus (i) Laspeyres' index for 1994 = $\frac{\sum (p_1 / p_0) p_0 q_s}{\sum p_0 q_s} \times 100$

$$= \frac{51,204.25}{48,194.73} \times 100 = 106.2,$$

(ii) Paasche's index for 1994 = $\frac{\sum (p_1 / p_0) p_s q_1}{\sum p_s q_1} \times 100$

$$= \frac{54,513.94}{51,251.56} \times 100 = 106.4, \text{ and}$$

$$(iii) \text{ Palgrave's index for 1994} = \frac{\sum (p_1 / p_0) p_1 q_1}{\sum p_1 q_1} \times 100 \\ = \frac{58,057.90}{54,510.22} \times 100 = 106.5.$$

4.5 QUANTITY INDEX NUMBERS

They are intended to measure the changes in the physical volume or quantity produced, consumed or sold of certain goods or services with respect to time. Like a price relative, we define a quantity relative by the ratio:

$$\text{quantity relative} = \frac{q_r}{q_b} \times 100,$$

where q_r denotes the quantity of a commodity in the given period and q_b denotes the corresponding quantity in the base period and which measures the proportionate change in quantity. The quantity index number formulas are obtained by interchanging p 's and q 's in the weighted price index number formulas. For instance, the Laspeyres' Weighted Aggregative Quantity Index is given by

$$Q_{Lw} = \frac{\sum q_r p_b}{\sum q_b p_b} \times 100$$

where prices of the base year are kept fixed as weights. The Paasche's Weighted Aggregative Quantity Index is given as

$$Q_{Pw} = \frac{\sum q_r p_r}{\sum q_b p_r} \times 100,$$

where prices of the given year are kept fixed as weights. Similarly, we define the Fisher's Quantity Index by the relation

$$Q_f = \sqrt{\frac{\sum q_b p_b}{\sum q_r p_b} \times \frac{\sum q_r p_r}{\sum q_b p_r}} \times 100.$$

Other formulas are also defined and interpreted in a similar way. As the prices are held constant, the differences are attributed to changes in quantity. These index numbers are also called the volume numbers or Quantum index numbers.

When the price p of a commodity during a period is multiplied by its quantity q , produced, consumed or sold during the period, we get the total value, by pq or V . A value relative is then defined as

$$\text{value relative} = \frac{\text{Total value during given year}}{\text{Total value during base year}} \times 100$$

$$\text{or } V_{Lw} = \frac{p_r q_r}{p_b q_b} \times 100 = \frac{V_r}{V_b} \times 100.$$

It is interesting to note that (omitting the factor 100) a value relative is equal to the product of the price relative by the quantity relative. Symbolically, this may be written as

$$\begin{aligned}\text{value relative} &= \frac{P_1 Q_1}{P_0 Q_0} \\ &= \left(\frac{P_1}{P_0} \right) \left(\frac{Q_1}{Q_0} \right) \\ &= \text{Price relative} \times \text{Quantity relative}\end{aligned}$$

In a similar way, a weighted relative quantity-index is defined as:

$$Q_w = \frac{\sum \left(\frac{q_1}{q_0} \right) w}{\sum w} \times 100,$$

where (q_1/q_0) are the quantity relatives and w denotes the weights.

Example 5.11 Construct (i) a Laspeyres' type (ii) a Paasche's type and (iii) Fisher's Ideal quantity index numbers from the following data:

Commodity	Quantity			Price		
	1958 (q ₀)	1959 (q ₁)	1960 (q ₂)	1958 (p ₀)	1959 (p ₁)	1960 (p ₂)
Chair	200	150	350	15	16	20
Desk	100	75	340	18	20	35
Radio	30	45	50	100	120	150

(IU., M.A., Econ. 1989-90)

The necessary calculations are given below:

Commodity	q ₀ p ₀	q ₁ p ₀	q ₂ p ₀	q ₀ p ₁	q ₁ p ₁	q ₂ p ₁
Chair	3000	3220	5250	3200	4000	5000
Desk	1800	3060	6120	2000	2500	4400
Radio	3000	4500	5000	3600	4500	5400
Total	7800	13710	16370	8800	11000	15400

(i) **Laspeyres' type:**

$$\text{Quantity index for 1959} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100 = \frac{13710}{7800} \times 100 = 175.8$$

$$\text{Quantity index for 1960} = \frac{\sum q_2 p_0}{\sum q_0 p_0} \times 100 = \frac{16370}{7800} \times 100 = 209.9$$

(ii) **Paasche's type:**

$$\text{Quantity index for 1959} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100 = \frac{8800}{5500} \times 100 = 175.0$$

$$\text{Quantity index for 1960} = \frac{\sum q_2 p_1}{\sum q_0 p_2} \times 100 = \frac{11000}{7800} \times 100 = 209.1$$

Q10 Fisher's Ideal type:

$$Q_{10} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_0 p_1}{\sum q_1 p_1}} \times 100$$

$$= \sqrt{\frac{13710}{7800} \times \frac{15400}{8800}} \times 100 = 175.4$$

$$Q_{11} = \sqrt{\frac{\sum q_1 p_1}{\sum q_0 p_0} \times \frac{\sum q_0 p_1}{\sum q_1 p_1}} \times 100$$

$$= \sqrt{\frac{16370}{7800} \times \frac{23600}{11000}} \times 100 = 209.5$$

Example 5.12 The following series shows for U.K. total imports (a) the declared value, and (b) value on the basis of average values in 1930.

U.K. Total Imports

Year	Declared value [£ million]	Value on basis of 1930 values (£ million)
1930	1044	1044
1931	991	1069
1932	792	939
1933	674	946
1934	751	991
1935	846	1012
1936	848	1077

Taking 1930 as base year, construct index numbers (i) of average values and (ii) of volume for the 1931-1936. (P.U., M.A., 1961; B.A./B.Sc., 1969; P.C.S. 1971)

It is easy to obtain (i) an index of average values for each year and (ii) a volume index, by writing in columns 2 and 3 in symbols first as below:

Year (1)	Declared value (2)	Value on basis of 1930 values (3)
1930	$\sum p_0 q_0$	$\sum p_0 q_1$
1931	$\sum p_1 q_1$	$\sum p_1 q_1$
1932	$\sum p_2 q_2$	$\sum p_2 q_2$
1933	$\sum p_3 q_3$	$\sum p_3 q_3$
1934	$\sum p_4 q_4$	$\sum p_4 q_4$
1935	$\sum p_5 q_5$	$\sum p_5 q_5$
1936	$\sum p_6 q_6$	$\sum p_6 q_6$

- (i) Index numbers of average values are thus obtained by dividing the entry in column 2 for year by corresponding entry in column 3, i.e., by Paasche's formula. Thus

$$\text{the average value (price) index for 1930} = \frac{\sum p_0 q_0}{\sum p_0 q_0} \times 100 \\ = \frac{1044}{1044} \times 100 = 100,$$

$$\text{and average value index for 1931} = \frac{\sum p_0 q_1}{\sum p_0 q_1} \times 100 \\ = \frac{861}{1069} \times 100 = 80.7.$$

Similarly, the average value indices for 1932-36 are obtained as 74.8, 71.4, 73.8, 74.7 and 78.7.

- (ii) Index numbers of volumes are obtained by dividing the values in column 3 by $\sum p_0 q_0$ by Laspeyres' method.

$$\text{Thus the volume index for 1930} = \frac{\sum p_0 q_0}{\sum p_0 q_0} \times 100 \\ = \frac{1044}{1044} \times 100 = 100,$$

$$\text{and the volume index for 1931} = \frac{\sum p_0 q_1}{\sum p_0 q_1} \times 100 \\ = \frac{1069}{1044} \times 100 = 102.4$$

Proceeding in a similar way, we obtain the volume index numbers for 1932-36 as 89.9, 90.6, 96.9 and 103.2.

Example 5.13 Given the following data:

Commodity	Base Year		Current Year	
	Price (p_0)	Quantity	Price (p_1)	Quantity
A	80.65	276	85.03	290
B	155.00	18	154.75	44
C	32.50	121	30.50	137
D	75.00	267	60.95	315

- Calculate (i) the value index number;
(ii) Laspeyres' weighted aggregative price index;
(iii) Paasche's weighted average of relatives volume index.

We first calculate the necessary products in the following table:

Commodity	$P_0 q_0$	$P_0 q_1$	$P_1 q_0$	$P_1 q_1$	q_1/q_0	$(q_1/q_0) \times q_0 p_1$
A	22259.4	24450.0	23460.0	23388.5	1.0507	24649.42
B	2790.0	6809.0	2785.5	6820.0	2.4444	6808.88
C	3932.5	4175.1	3690.5	4452.5	1.1323	4178.75
D	20025.0	21637.25	16273.65	26625.0	1.3295	21637.45
Σ	49006.9	57274.75	46209.75	61286.0	—	57274.50

- (i) The value index number is:

$$V_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0} \times 100 = \frac{57274.75}{49006.9} \times 100 = 116.87$$

- (ii) The Laspeyres' weighted aggregative price index is:

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 = \frac{46209.65}{49006.9} \times 100 = 94.29;$$

- (iii) The Passche's weighted average of relatives volume index is:

$$P_{01} = \frac{\sum (q_1/q_0) q_0 P_1}{\sum q_0 P_1} \times 100 = \frac{57274.50}{46209.65} \times 100 = 123.94.$$

TESTS FOR INDEX NUMBER FORMULAE

From a theoretical view point, a "good" index number formula is required to satisfy the following test proposed by Irving Fisher (1867-1947).

- 5.6.1 Time Reversal Test. This may be stated as follows:

If the time subscripts 0 and 1 are interchanged in a price (or quantity) index number formula wherever they appear, then the resulting price (or quantity) formula should be the reciprocal of the original index formula, ignoring the factor 100. Symbolically, the test requires that

$$P_{01} = \frac{1}{P_{10}} \text{ or } P_{01} \times P_{10} = 1$$

Since the index number is designed to measure changing values of prices or production, it is reasonable to expect that the formula should give the same result regardless of which the two periods is taken as the base. The base period and the given period are only relative terms and hence should be interchanged. An index number satisfying this test gives consistent results. This is a property we shall desire our formulae to obey.

For the purposes of illustration, let us consider some of the important formulae and see whether they do not satisfy this test.

The Laspeyres' price index expressed in ratio rather than percentage form (*i.e.* omitting the factor 100), is given by

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_1}$$

Interchanging the time subscripts, we get

$$P_{st} = \frac{\sum p_s q_t}{\sum p_t q_s}$$

$$\text{But } P_{st} \neq P_{ts}, \text{ i.e., } \frac{\sum p_s q_t}{\sum p_t q_s} \neq \frac{\sum p_t q_s}{\sum p_s q_t} = 1.$$

Hence it does not satisfy the time reversal test.

- (ii) Again, taking the Paasche's price index formula, i.e. $P_{st} = \frac{\sum p_s q_s}{\sum p_t q_s}$ and interchanging time subscripts, we get

$$P_{st} = \frac{\sum p_t q_s}{\sum p_s q_t}$$

$$\text{But again, } P_{st} \neq P_{ts} = \frac{\sum p_s q_s}{\sum p_t q_s} \neq \frac{\sum p_t q_s}{\sum p_s q_t} \neq 1, \text{ which shows that this formula too does not obey the time reversal test.}$$

- (iii) Interchanging the time subscripts in the Marshall-Eggerth price index, i.e. in the formula $P_{ts} = \frac{\sum p_s (q_0 + q_s)}{\sum p_0 (q_0 + q_s)}$, we get

$$P_{st} = \frac{\sum p_t (q_0 + q_t)}{\sum p_0 (q_0 + q_t)}$$

Multiplying together, we obtain

$$P_{ts} \times P_{st} = \frac{\sum p_s (q_0 + q_s)}{\sum p_0 (q_0 + q_s)} \times \frac{\sum p_t (q_0 + q_t)}{\sum p_0 (q_0 + q_t)} = 1.$$

Hence the time reversal test is satisfied.

- (iv) Fisher's Ideal index number is

$$P_{ts} = \sqrt{\frac{\sum p_s q_0}{\sum p_0 q_s} \times \frac{\sum p_t q_0}{\sum p_0 q_t}}$$

Interchanging the time subscripts, we have

$$P_{st} = \sqrt{\frac{\sum p_t q_0}{\sum p_0 q_t} \times \frac{\sum p_s q_0}{\sum p_0 q_s}}$$

$$\text{Thus } P_{ts} \times P_{st} = \sqrt{\frac{\sum p_s q_0}{\sum p_0 q_s} \times \frac{\sum p_t q_0}{\sum p_0 q_t}} \times \sqrt{\frac{\sum p_t q_0}{\sum p_0 q_t} \times \frac{\sum p_s q_0}{\sum p_0 q_s}} = 1$$

Hence it conforms to the time reversal test.

5.6.2 Factor Reversal Test. It may be stated in this way:

"If the factors p 's (prices) and q 's (quantities) occurring in a price (or quantity) index formula be changed (or reversed) so that a quantity (or price) index formula is obtained, then the product of the index numbers should equal the value index number, i.e. $\frac{\sum p_0 q_0}{\sum p_0 q_0} = \frac{\sum p_0 q_0}{\sum p_0 q_0}$ ". In other words, the factor reversal test requires that

$$(\text{Price index}) \cdot (\text{Quantity index}) = \text{Value index}.$$

We give the following illustrations.

Interchanging the factors p 's and q 's in the formula

$$P_{0t} = \frac{\sum p_t q_0}{\sum p_0 q_t} \quad (\text{Laspeyres' price index}), \text{ we obtain}$$

$$Q_{0t} = \frac{\sum q_t p_0}{\sum q_0 p_t} \quad (\text{Laspeyres' quantity index}).$$

The product of these two is $\frac{\sum p_t q_0}{\sum p_0 q_t} \times \frac{\sum q_t p_0}{\sum q_0 p_t}$, which is not equal to $\frac{\sum p_t q_t}{\sum p_0 q_0}$, the value index.

Hence the factor-reversal test is not obeyed.

Again interchanging the factors p 's and q 's, the formula

$$P_{t0} = \frac{\sum p_0 q_t}{\sum p_t q_0} \quad (\text{Paasche's price index})$$

transforms into $Q_{t0} = \frac{\sum p_t q_0}{\sum q_0 p_t} \quad (\text{Paasche's quantity index})$

$$\text{Thus } P_{t0} \times Q_{t0} = \frac{\sum p_0 q_t}{\sum p_t q_0} \times \frac{\sum q_0 p_t}{\sum q_t p_0} = \frac{\sum p_0 q_t}{\sum p_t q_0} \quad (\text{Value index}).$$

Hence Paasche's price index does not satisfy the factor reversal test.

The Marshall-Edgeworth's price index is

$$P_{0t} = \frac{\sum p_0 (q_0 + q_t)}{\sum p_t (q_0 + q_t)}$$

Interchanging the factors p 's and q 's, the quantity index is obtained as

$$Q_{0t} = \frac{\sum q_0 (p_0 + p_t)}{\sum q_t (p_0 + p_t)}$$

$$\text{But } \frac{\sum p_t(q_0 + q_s)}{\sum p_t(q_0 + q_s)} \times \frac{\sum q_t(p_0 + p_s)}{\sum q_t(p_0 + p_s)} \neq \frac{\sum p_t q_t}{\sum p_t q_t},$$

i.e. (price index) (quantity index) \neq value index.

Thus the factor-reversal test is not satisfied.

(iv) By definition, Fisher's price index is given by

$$P_{3t} = \sqrt{\frac{\sum p_t q_t}{\sum p_0 q_t} \times \frac{\sum p_t q_t}{\sum p_0 q_t}}$$

Fisher's quantity index is obtained by interchanging the factors p 's and q 's as

$$Q_{3t} = \sqrt{\frac{\sum q_t p_t}{\sum q_t p_0} \times \frac{\sum q_t p_t}{\sum q_t p_t}}$$

$$\begin{aligned} \text{Therefore } P_{3t} \times Q_{3t} &= \sqrt{\frac{\sum p_t q_t}{\sum p_0 q_t} \times \frac{\sum p_t q_t}{\sum p_0 q_t}} \times \sqrt{\frac{\sum q_t p_t}{\sum q_0 p_t} \times \frac{\sum q_t p_t}{\sum q_0 p_t}} \\ &= \frac{\sum p_t q_t}{\sum p_0 q_t} = \text{value index.} \end{aligned}$$

Hence, we see that Fisher's Ideal index satisfies the factor reversal test.

Example 5.14 Show with the help of the following data that the factor-reversal and time-tests are satisfied by Fisher's Ideal formula for index number construction.

Commodity	New Year		Current Year	
	Price (in Rs.)	Quantity (units of 40 kg)	Price (in Rs.)	Quantity (units of 40 kg)
A	6	50	10	56
B	4	100	2	120
C	2	60	6	80
D	10	30	12	24
E	8	40	12	36

(P.I) = B.A/B.S.

We calculate the necessary products in the following table:

Commodity	Price		Quantity		$p_t q_t$	$p_0 q_t$	$p_0 q_0$	$p_t q_0$
	p_0	p_t	q_0	q_t				
A	6	10	50	56	300	500	336	560
B	4	2	100	120	400	200	480	240
C	2	6	60	60	120	360	120	360
D	10	12	30	24	300	360	240	288
E	8	12	40	36	320	480	228	432
Total	-	-	-	-	1440	1900	1464	1880
					$\sum p_t q_t$	$= \sum p_0 q_t$	$= \sum p_0 q_0$	$= \sum p_t q_0$

Now Fisher's price index for the current year (omitting the factor 100) is given by

$$P_{II} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_1}} \times \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_0}} = \sqrt{\frac{1900}{1440}} \times \sqrt{\frac{1880}{1464}},$$

∴ Fisher's quantity index for the current year (omitting the factor 100) is given by

$$Q_{II} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_1}} \times \sqrt{\frac{\sum q_1 p_1}{\sum q_0 p_0}} = \sqrt{\frac{1464}{1440}} \times \sqrt{\frac{1880}{1900}},$$

$$\begin{aligned} P_{II} \times Q_{II} &= \sqrt{\frac{1900}{1440}} \times \sqrt{\frac{1880}{1464}} \times \sqrt{\frac{1464}{1440}} \times \sqrt{\frac{1880}{1900}} \\ &= \sqrt{\frac{1880}{1440}} \times \sqrt{\frac{1880}{1440}} = \frac{1880}{1440} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \end{aligned}$$

= Value index.

Hence we see that Fisher's ideal formula for index numbers satisfies the factor-reversal test. Again interchanging the time subscripts in Fisher's price index (omitting the factor 100), we get

$$P_{II} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_0}} \times \sqrt{\frac{\sum p_0 q_0}{\sum p_1 q_1}}$$

$$P_{II} \times P_{II} = \sqrt{\frac{1900}{1440}} \times \sqrt{\frac{1880}{1464}} \times \sqrt{\frac{1440}{1880}} \times \sqrt{\frac{1464}{1900}} = 1$$

Since Fisher's ideal formula for index numbers also satisfies the time-reversal test.

10.3 Circular Test. The circular test is an extension of the time reversal test, which is applicable even, i.e., the given year and/or base year. But in circular test, three or more years are taken into account. This test may be stated as:

If P_{ab} is the price index for the year 'b' with year 'a' as base, P_{bc} is the price index for the year 'c' with year 'b' as base and P_{ac} is the price index for the year 'a' with year 'c' as base, then the product of these three numbers should equal 1. In other words, the circular test requires

$$P_{ab} \times P_{bc} \times P_{ac} = 1$$

In general, the circular test is said to be satisfied, if

$$P_{ij} \times P_{ji} \times P_{ik} \times \dots \times P_{kj} \times P_{ki} = 1,$$

where P_{ij} indicates the price index (without the factor 100) for the year 'j' with the year 'i' as the base.

This test is not obeyed by any of the weighted index numbers unless the weights are constant. For illustration, we consider the weighted aggregate price number with fixed weights, say,

Let there be three years denoted by 'a', 'b', and 'c'.

Then the weighted price index numbers are

$$P_{ab} = \frac{\sum p_b q_a}{\sum p_a q_a}, P_{bc} = \frac{\sum p_c q_b}{\sum p_b q_b} \text{ and } P_{ca} = \frac{\sum p_a q_c}{\sum p_c q_c}.$$

Multiplying them together, we get

$$P_{ab} \times P_{bc} \times P_{ca} = \frac{\sum p_b q_a}{\sum p_a q_a} \times \frac{\sum p_c q_b}{\sum p_b q_b} \times \frac{\sum p_a q_c}{\sum p_c q_c} = 1,$$

which shows that the circular test is satisfied.

5.7 CONSUMER PRICE INDEX NUMBER

5.7.1 Meaning. A *consumer price index* (CPI) is designed to measure the changes in composite price of a specified "basket" of goods and services during the given period as compared with the base period. The so-called "basket" would comprise various commodities consumed and services received in the base period or the given period, grouped under the main headings: (i) Food and beverages; (ii) Clothing and footwear; (iii) Fuel and lighting; (iv) Housing; (v) Services; (vi) Miscellaneous. It is customary to exclude the durable goods and non-consumption monetary transactions such as contribution to Provident Fund, Savings Certificates, etc. The quantities consumed or the expenditure incurred on various groups are used as weights for the average retail prices prevailing in the localities concerned during the base and given periods.

Some countries still retain the name of the cost of living index number as "it measures the change in the cost of living of a person or of a group of persons, having identical tastes for goods." A consumer price index is also called a household-budget price index or a retail price index.

5.7.2 Construction of Consumer Price Index Numbers. The following steps are involved in the compilation of the consumer price index numbers:

- Scope.** The first step is to clearly specify the category of people and the locality where they reside as a consumer price index number relates to a particular segment of population such as low-salaried employees, school teachers, industrial workers, etc. residing in a particular well-defined area such as a city or an industrial town. As far as possible, a homogeneous group of persons, i.e., persons who have identical patterns of living, is considered.
- Household Budget Inquiry and Allocation of weights.** The next step is to conduct a household budget inquiry of the category of people concerned in order to determine the goods and services to be included in the construction and to derive weights to be attached to them. This step has many practical problems as no two households have the same income per head and the same purchasing or consuming patterns. The inquiry or the house hold consumption survey should, therefore, include questions on family size, income, number of taxes, quality and quantities of goods and services consumed and the money spent on them in various headings such as food and beverages, clothing and footwear, fuel and lighting, and miscellaneous. The miscellaneous group includes items such as clothing, education, medical care, recreation, gifts, newspaper, barber, laundry and other charges.

An appropriate sampling technique is to be employed to collect consumption data of the households. The factory payrolls in case of industrial workers or the ration register in other cases may be used as frame. The households, after having been stratified into different income groups, are drawn into the sample by the method of either area sampling or systematic sampling with random start. The consumption data collected are then analysed according to the nature, quality and quantities of the consumed commodities, the proportion which expenditure on each group bears to the total expenditure of all groups. The weights are then attached to the various consumption groups in proportion to the money spent on them so as to truly reflect their relative importance.

Price Data. The third step is to collect data on consumer prices of goods and services included in the basket. These prices (retail prices) should be obtained both for the base period and the given period from the locality in which the people concerned reside or from which they make their purchases.

Computation of the Index. The last step is the computation of the consumer price index number with the help of an appropriate formula. For this purpose, one of the following two methods is employed with the same result.

The Aggregate Expenditure Method. Here the quantities consumed by households in the base year are taken as the weights. The quantity $\sum p_0 q_0$ represents the aggregate expenditure incurred on the various items in the base year and the quantity $\sum p_s q_s$ the in the given year with base year quantities. De Lapeyres' formula, namely

$$P_{Cw} = \frac{\sum p_s q_s}{\sum p_0 q_0} \times 100$$

This applied. The given year quantities are not used as weights because they change from year to year and a fresh sample budget enquiry, which involves considerable expense, labour and time, would be needed every year.

Household Budget Method. In this method, the price relatives are weighted by either the money spent by the households on various items or fixed weights derived from the sample household budgets conducted in the category of people concerned. This method is called the Household Budget Method, because the amounts of money spent by the households concerned are obtained from a household consumption survey. This method is also known as the Weighted Average of Relatives. The formula followed is

$$P_{Ch} = \frac{\sum W \left(\frac{p_s}{p_0} \right)}{\sum W} \times 100 = \frac{\sum W \times 1}{\sum W}$$

— W may be $p_0 q_s$ and $1 = \frac{p_s}{p_0} \times 100$.

The procedure of computation is illustrated by the following example.

Example 5.15 The following table gives average annual prices of ten commodities for the years 1990 and 1994. Calculate the consumer price index number for 1994 on the basis of 1990.

Commodity	Quantity consumed	Unit of Price	Price in 1990	Price in 1994
Wheat	20 kgs	Rs. per kg	1.25	1.50
Rice	10 kgs	Rs. per kg	3.00	3.75
Pulse	12 kgs	Rs. per kg	2.50	3.00
Sugar	4 kgs	Rs. per kg	2.00	3.25
Ghee	3 kgs	Rs. per kg	3.75	4.00
Milk	30 litres	Rs. per litre	0.50	0.75
Vegetables	35 kgs	Rs. per kg	0.25	0.40
Fuel	200 kgs	Rs. per kg	0.50	0.75
Cloth	22 meters	Rs. per meter	1.50	2.10
House Rent	1 unit	Rs. per unit	30.00	60.00

The necessary calculations are shown in the following table.

(i) Consumer price index number by Aggregate Expenditure Method.

Commodity	Unit	Quantity consumed (q_0)	Price per unit		Aggregate Exp	
			p_0	p_1	$p_0 q_0$	$p_1 q_0$
Wheat	kg	25 kgs	1.25	1.50	31.25	37.50
Rice	kg	10 kgs	3.00	3.75	30.00	37.50
Pulse	kg	12 kgs	2.50	3.00	30.00	36.00
Sugar	kg	4 kgs	2.00	3.25	8.00	13.00
Ghee	kg	3 kgs	3.75	4.00	11.25	12.00
Milk	litre	30 litres	0.50	0.75	15.00	22.50
Vegetables	kg	35 kgs	0.25	0.40	8.75	14.00
Fuel	kg	200 kgs	0.50	0.75	100.00	150.00
Cloth	meter	22 meters	1.50	2.10	33.00	46.20
House Rent	unit	1 unit	30.00	60.00	30.00	60.00
Total	-	-	-	-	297.25	428.70

Thus the consumer price index number for 1994 is

$$P_{1994} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{428.70}{297.25} \times 100 = 144.2$$

This result indicates that the prices of consumption goods have increased by 44.2% for 1994 in comparison with 1990.

(iii) Consumer price index number by the Household Budget Method.

Commodity	Unit	Commodity consumed (x_1)	Price per unit		Price Relatives $\frac{P_1}{P_0} \times 100 + 1$	Weights $W = P_0 x_1$	Weights x Price Relatives - $W + 1$
			1990 P_0	1994 P_1			
Wheat	kg	25 kg	1.25	1.50	120.00	31.25	3750.00
Rice	kg	10 kg	3.00	3.75	125.00	30.00	3750.00
Pulse	kg	12 kg	2.50	3.00	120.00	30.00	3600.00
Sugar	kg	4 kg	2.00	3.25	162.50	8.00	1300.00
Ghee	kg	3 kg	3.75	4.00	106.67	11.25	1200.04
Milk	liter	10 liter	0.50	0.75	150.00	15.00	2250.00
Vegetables	kg	35 kg	0.25	0.40	160.00	8.75	1400.00
Fuel	kg	200 kg	0.50	0.75	150.00	100.00	15000.00
Cloth	metres	22 metres	1.50	2.10	140.00	33.00	4620.00
Haus-Rent	unit	1 unit	30.00	60.00	200.00	30.00	6000.00
Total	-	--	-	-	--	297.25	42870.04

Hence the consumer price index number for 1994 is

$$P_{01} = \frac{\sum W x 1}{\sum W}$$

$$= \frac{42870.04}{297.25} = 144.2$$

We see that the consumer price index numbers constructed by both methods are the same except, the process of rounding off the figures may result in a small difference between the results.

5.3.3 Shortcomings or Drawbacks of Consumer Price Index Numbers. Some of the shortcomings of the consumer price index numbers are given below:

- 1. It is practically difficult to clearly demarcate one category of people from another.
- 2. As the construction of consumer price indices involves the sampling of goods and services, the sampling errors and biases may affect indices and render them to suspect. Moreover, the frames used for household consumption inquiry may be incomplete and outdated.
- 3. In case of certain goods, it is difficult to collect prices actually needed. For example, the prices for clothing usually relate to cloth and not to tailored clothes.
- 4. It is also difficult to eliminate the effect of changes in quality and grade of goods and services purchased by households.
- 5. During the course of household budget inquiry, the price of goods and services and their demand may change; some commodities may change in quality, others may disappear and some new goods may enter the market.
- 6. The consumer price indices cannot be used for comparing the price changes in consumption goods and services in two localities or in two households in the same locality as no two households can be homogeneous, i.e. they can neither have precisely the same pattern of consumption nor precisely the same basket of goods and services.

It is therefore relevant to point out that a consumer price index should not be wholly relied upon as it is an imperfect measure.

5.8 USES OF INDEX NUMBERS

A few uses of index numbers are given below:

- i) The price index numbers are used to measure changes in a particular group of prices and help us in comparing the movement in prices of one commodity with another. They are also designed to measure the changes in the purchasing power of money.
- ii) Index numbers of industrial production provide a measure of change in the level of industrial production in a country.
- iii) The quantity index numbers show the rise or fall in the volume of production, volume of exports and imports, etc.
- iv) The import and export price indices are used to measure the changes in the terms of trade of a country. By the terms of trade is meant the ratio of import to export prices.
- v) Index numbers are also used to forecast business conditions of a country and to discuss seasonal fluctuations and business cycles.
- vi) The consumer price indices indicate the movements in retail prices of consumption goods and services. These movements in prices help government in formulating its policies and also appropriate economic measures. They can be used to re-adjust the wages and to take measures of relief by granting dearness allowance and bonus to their employees to meet increased costs by the industrial and commercial establishments as well as mills. They are also used to deflate the gross national product and wages to arrive at the *real* values of the same—product and *real* wages.
- vii) Index numbers are also used to measure enrolment changes, intelligence quotients and performance of students.

5.9 LIMITATIONS OF INDEX NUMBERS

Some of the limitations are described below:

- i) It is not practicable to price all the goods and services as well as to take into account changes in quantity or product.
- ii) Since the construction of almost all index numbers is based on sampling of items, therefore the sampling errors creep into their calculations.
- iii) In price indices the choice of a normal period is difficult as few periods can be regarded as normal for all segments of the economy.
- iv) The results obtained by different methods of construction may not quite agree.
- v) Comparisons of changes in variables over long periods are not reliable.
- vi) All index numbers are not suitable for all purposes. The users are therefore strongly advised to essentially understand the purpose for which the index has been constructed.

EXERCISES

OBJECTIVE

- a) Answer "True" or "False". If the statement is not true then replace the underlined words with words that make the statement true.
- Weighted aggregate index is the simplest form of an index number.
 - When the base year values are used as weights, the weighted average of relatives price index is the same as Paasche's.
 - There are three methods of construction of CPI numbers.
 - Fisher's ideal index is the mean of Laspeyre's and Paasche's index numbers.
 - The most suitable average for index numbers is median.
 - A Laspeyres price index is a CPI.
 - If a price index increased from 150 to 200 over a certain period, the increase in prices was 50% from the beginning to the end of that period.
 - If the current year and base year index numbers are 170 and 160 respectively, then the value of Fisher's ideal index number is 185.
 - Prices are appropriate weights in a weighted aggregates quantity index.
 - If price and quantity of a commodity in a year are multiplied, we get value.

MULTIPLE CHOICE QUESTIONS

- a) If the price of a kg of meat was Rs.40/- in 2000 and Rs.50/- in 2002, the simple price relative in 2002 is
- 125
 - 100
 - 80
 - 50
- b) An un-weighted aggregates price index has a limitation that
- It is difficult to calculate.
 - It is unduly influenced by the price variations of high priced commodities.
 - It is unduly influenced by the price variations of low priced commodities.
 - None of the above.

- iii) The best weights to be used in a quantity index calculated by the weighted average of relatives methods are:
- Base period price weights.
 - Current period price weights.
 - Base period quantity weights.
 - Base period value weights.
- iv) The CPI is basically
- A fixed-weight index.
 - A Laspeyres index.
 - Both of the above.
 - None of the above.
- v) The Laspeyres price index is a weighted aggregate index in which the weights are based on
- Current quantities.
 - Base period quantities.
 - Mean of base and current period quantities.
 - None of the above.
- vi) The Paasche's price index is a weighted aggregate index in which the weights are based on
- Current quantities.
 - Base period quantities.
 - Mean of base and current period quantities.
 - None of the above.
- vii) The Laspeyres price index is:
- Upward biased.
 - Downward biased.
 - No bias.
 - None of the above.
- viii) The following is a price index number series: 1995, 100; 1997, 120; 2002, 150; which of the following statement is incorrect?
- Prices increased by 50% from 1995 to 2002.
 - Prices increased by 30% from 1997 to 2002.
 - Prices in 1995 were $13\frac{1}{3}\%$ lower than in 2002.
 - Prices increased by 25% from 1997 to 2002.

- (a) The following is a price index series for Lahore based on 1990 = 100, 1995 = 120, 2000 = 125. Which of the following statement is correct?
- Prices have increased by 5% from 1995 to 2000.
 - Prices in 1990 were 25% lower than in 2000.
 - Prices in 2000 were 1.2% higher than in 1995.
 - None of the above.
- (b) If wages of a group of workers increased from 1995 to 2000 by 10% and a relevant price index increased by 5%; Real wages have increased over this period by:
- 4.8%.
 - 10%.
 - 6%.
 - None of the above.
- (c) Which of the statement is the for Laspeyres's index number.
- It meets time reversal test.
 - It meets factor reversal test.
 - It meets both time reversal as well as factor reversal tests.
 - None of the above.
- (d) Which of the statement is true for Paasche's index?
- It meets time reversal test.
 - It meets factor reversal test.
 - It meets both time reversal as well as factor reversal tests.
 - None of the above.
- (e) A Laspeyres price index:
- A cost of living index.
 - A weighted index.
 - Both of the above.
 - None of the above.
- In 2000 the price for a certain type of fish was Rs.120/- per kg, and 450 tons were consumed. In 2001 the price for this type of fish was Rs.100/- per kg, and 350 tons of fish were consumed. If the simple price relative in 2000 is Rs.100/- then in 2001 simple price relative would be
- 130
 - 100
 - 90
 - 83

- iv) The index number for a base year is always
 a) Zero.
 b) Greater than 100.
 c) Less than 100.
 d) None of the above.

SUBJECTIVE

- 5.1 Explain the concept of an index number. What is the procedure followed in the preparation of an index number?
- 5.2 Define an index number. Discuss the main steps involved in the construction of index numbers of wholesale prices. Indicate their uses and limitations.
 (C.S.S. 1964; P.U., B.A./B.Sc. 1986, 1993)
- 5.3 What is an index number? Describe the important problems involved in the preparation of an index number. What considerations would you weigh while constructing a wholesale price index number, in connection with the selection of commodities and the base year?
 (P.U., B.A./B.Sc. 1986)
- 5.4 It has been stated that the technique of index number construction involves four major factors:
 (a) Choice of items (b) Base period
 (c) Form of averages (d) Weighting system
 Do you agree with this? If so, explain these four factors and discuss the problems to which they give rise. If you do not agree, give your views on the main problems involved in the construction.
 (P.U., M.A. Econ. 1993)
- 5.5 a) What is an index number? Distinguish between fixed base and chain base methods of constructing index numbers. What are their respective merits and demerits?
 (P.U., B.A./B.Sc. 1993)
- b) Describe the various methods of averaging that can be used in constructing an index number, and point out their merits and demerits.
 (P.U., M.A. Econ. 1993)
- 5.6 a) What is a weighted index number? Describe the various methods of weighting the index numbers of prices. What are their advantages?
 (P.U., B.A./B.Sc. 1993)
- b) Present an interpretation of the (i) Laspeyres and (ii) Paasche price index numbers in terms of the total value of commodities.
- 5.7 Compare the following concepts
 i) Simple index and Composite index.
 ii) Fixed base index and Chain base index.
 iii) Laspeyres' price index and Paasche's price index.
 iv) Weighted aggregative price index and Weighted average of relatives price index.
 (P.U., B.A./B.Sc. 1993)

- 5.8 a) Define Laspeyres, Paasche and Marshall-Edgeworth types of index numbers. Show that the Paasche type is the reciprocal of the Laspeyres type with time subscripts reversed, given that the two have the same value.
 b) Define Fisher's ideal index number. Describe its advantages and disadvantages.
 (C.S.S., 1960; P.U., B.A./B.Sc., 1973)
- 5.9 Define and discuss the following index numbers.
 i) Quantity Index numbers.
 ii) Value Index numbers.
- 5.10 Explain the time and factor reversal tests.
 Which of the following formulae satisfy these tests and which do not?
 i) $P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0}$ ii) $P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$ iii) $P_{01} = \frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}}$
 iv) $P_{01} = \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)}$ v) $P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_1} \times \frac{\sum p_1 q_1}{\sum p_0 q_0}}$
 (P.U., B.A./B.Sc., 1994; B.Z.U. M.A., Exam., 1994)
- 5.11 a) Explain theoretical tests which a good index is expected to satisfy.
 b) Show that the Marshall-Edgeworth index $P = \frac{\sum p_{ra}(q_{ra} + q_{rb})}{\sum p_{rb}(q_{ra} + q_{rb})}$ satisfies the time reversal test but not the circular test, unless the weights in the three years a, b, c, are equal.
 (P.U., B.A./B.Sc., 1977, 1978-S)
- 5.12 a) Prove that the simple aggregate value index numbers (i.e. $\frac{\sum p_n q_n}{\sum p_0 q_0}$) satisfy the time reversal and circular tests but do not satisfy the factor reversal test.
 b) Show that weighted aggregate price index numbers with fixed (quantity) weights satisfy the circular test.
 c) Show that the (i) Laspeyres and (ii) Paasche index do not satisfy the time reversal or factor reversal tests.
 d) Show that the Marshall-Edgeworth index satisfies the time-reversal test but not the factor-reversal test.
 (P.U., B.A./B.Sc., 1963; M.A., Exam., 1980)
- 5.13 a) Prove that Fisher's ideal index satisfies both the time reversal and the factor reversal test but does not conform to the circular test.
 b) Describe the methods for testing the consistency of index numbers. Explain Fisher's formula, giving an example.
 (C.S.S. 1960, P.U., B.A./B.Sc.)

- b) How would you calculate the Consumer Price Index for factory workers in Pakistan?
 (P.U., M.A. Econ., 1997)

- 5.15 Explain the meaning of the consumer price index number. Describe the method of construction adopted. Explain the uses of consumer price index numbers.
 (P.U., B.A./B.Sc., 1962, 63, 71, 81, 83; C.S.S., 1960; M.A. Econ., 1997)

- 5.16 You are required to prepare the consumer price index for industrial workers in Lahore. Describe how you will proceed. Prepare a short questionnaire for the inquiry.
 (P.U., B.A./B.Sc., 1997)

- 5.17 a) Explain the construction of the index of retail prices.
 b) Consumer price index for Rawalpindi stood at 137 in July and 140 in August. During the same months, the index number at Abbottabad stood at 150 and 151. Does this mean that Abbottabad is costlier than Rawalpindi? Give your reasons in detail.
 (P.U., B.A./B.Sc., 1997)

- 5.18 a) Discuss the problems which arise in constructing consumer price index numbers.
 b) Show that the Fisher's Ideal index satisfies both the time-reversal and factor-reversal properties. Discuss the other properties of this index number.
 (P.U., B.A./B.Sc., 1997)

- 5.19 Discuss the nature and method of construction of an index number of wages. Examine the possibilities of constructing such an index number for Pakistan.
 (C.S.S., 1961, 65; P.U., M.A. Econ., 1997)

- 5.20 Find the price relatives for each year from the following average retail prices of wheat:
 (i) 1995 as a base (ii) 1996 as a base.

Year	1995	1996	1997	1998	1999	2000
Retail prices	Rs. 14.95	14.54	15.10	15.65	16.28	16.53

- 5.21 The following table gives the production of cotton cloth (in million meters) for Pakistan during 1954 to 1963:

Year	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963
Production	285	319	436	470	511	555	564	636	642	681

Find out index numbers by taking (i) 1954 as base year, (ii) average of 1958, 59, 60 as the period.
 (P.U., M.A. Econ., 1997)

- 5.22 The following table gives the average wholesale prices in rupees per unit of gold, wheat and cotton during the years 2001–2006.

Commodity	Average price in Rs. Per unit					
	2001	2002	2003	2004	2005	2006
Gold	25.3	30.1	33.4	35.5	35.3	36.0
Wheat	17.3	14.5	4.9	5.7	17.1	11.6
Cotton	7.8	5.4	6.7	5.6	7.2	10.2

Using 2001 as the base period, compute the simple aggregative price indices and the average of relatives price indices for the years 2002 to 2006.

- 1.23 The Prices of four commodities quoted at Multan for May 2001, April 2002 and May 2002 are given below:

Commodity	May 2001	April 2002	May 2002
Wheat	17.30	18.27	17.58
Barley	14.58	14.58	16.50
Jawar	14.67	13.94	15.25
Bajra	17.50	13.75	13.42

Compute price index numbers for April 2002 and May 2002 with May 2001 as base, using (i) simple aggregative method, (ii) simple average (mean) of price relatives and (iii) geometric mean of price relatives.

- 1.24 Describe the chain base method used for the construction of index numbers from the following table and calculate such index numbers. Discuss its merits against the fixed base method.

Commodities	Average Prices in Rs. per 40 kg			
	1928	1929	1930	1931
Rice	7.3	7.7	5.8	4.6
Wheat	7.5	5.5	3.6	2.7
Linseed	7.0	8.0	6.5	4.2
Gur	6.3	7.3	-	4.2
Cotton	34.1	29.8	27.3	13.3
Tobacco	17.3	17.0	14.5	11.6

(P.U., B.A./B.Sc. 1976)

Construct the Chain Indices for the following data for price relatives for 1941 to 1944.

Year	Sugar	Gar	Tea	Coffee
1941	98	75	82	99
1942	100	82	74	100
1943	114	83	78	104
1944	100	84	89	95

(P.U., M.A. Econ., 1975)

The price relatives of three commodities are given below:

Year	Price Relatives of Commodities		
	A	B	C
2001	100	100	100
2002	105	97	121
2003	110	94	125
2004	113	100	130
2005	116	99	128
2006	120	105	130

- a) Compute the link relatives, i.e. the price relatives in each year with reference to the previous year as 100, and calculate the index for three commodities for each year.
- b) Chain the above indices to a 2001 base.

- 5.27 The following table gives the price relatives of four commodities for the years 2002-2005 inclusive, the price of each commodity in 2001 being stated as 100.

Commodity	Price Relatives: 2001 = 100				
	2001	2002	2003	2004	2005
A	100	125	112	123	131
B	100	120	110	120	127
C	100	87	92	108	122
D	100	75	125	150	140

Calculate:

- An index number for each year, 2002-2005, using the simple A.M. of price relatives.
- Index number for each year, using the chain base method.
- Explain why, in general, the indices of (i) and the chained indices are not in agreement.

- 5.28 A firm divides its material into four main groups and tries to estimate the overall effect of price changes by producing an index number weighted according to the tonnages used in 2003. Calculate this index for 2007 from the following information, and comment on the result.

Group	2003		2007	
	Tons	Price (£)	Tons	Price (£)
A	30	85	45	116
B	120	14	185	42
C	35		68	15
D	210	48	250	50

- 5.29 The prices and quantities of three commodities are shown below:

Commodity	Price (Rs.)		Quantity	
	1998	2005	1998	2005
Rice	3.50	3.15	75	80
Barley	2.00	1.80	107	138
Maize	2.60	1.75	62	57

- Using 1998 as the base period and the base period quantities as weights, compute weighted-aggregative price index and the weighted average of relatives price index for 2005.
- Using 1998 as the base period and 2005 quantities as weights, compute the weighted-aggregative price index and the weighted average of relatives price index for 2005.

- 5.30 The prices and quantities of three commodities during 1997 and 2007 are given below:

Commodity	Price		Quantity	
	1997	2007	1997	2007
A	12	10	301	300
B	38	50	100	104
C	40	40	56	76

Compute weighted-aggregative price index for 2007 with 1997 = 100 by (i) Laspeyres' method. If (ii) Paasche's method.

- 5.21 Compute the following index numbers for the data in question 5.20 by Peasche's method:
- Weighted aggregative price index.
 - Weighted average of relatives price index.
- 5.22 Construct the following index numbers of prices for 2004 and 2005 from the given data.
- Base year weighted.
 - Current year weighted.

Commodity	Prices (Rs. per 40 kg)			Quantities (tons)		
	2000 (base)	2004	2005	2000	2004	2005
A	70.50	80.65	85.00	270	276	290
B	146.95	155.00	154.75	24	18	44
C	25.50	32.50	30.50	130	121	137
D	64.75	75.00	60.95	185	267	355

Compute the index numbers of Marshall-Edgeworth and Fisher's "ideal" type for the following data.

Commodity	Base Year		Current Year	
	Price P_0	Quantity Q_0	Price P_1	Quantity Q_1
A	7	20	9	49
B	5	27	7	28
C	10	35	9	29
D	9	50	4	42
E	3	4	10	25

(P.U., B.A./B.Sc., 1973)

Given the following, construct Fisher's ideal index number for

- 1964, taking 1957 as base.
- 1957, taking 1964 as base year.

Commodity	1957		1964	
	Price	Quantity	Price	Quantity
Rice	9.3	100	4.5	90
Wheat	6.4	11	3.7	10
Jawar	5.1	5	2.7	3

(P.U., M.A., Econ., 1968)

Using 2003 as base, construct price index numbers from the following foodgrain data.

Commodity	Prices per 40 kg (Rs.)			Production (000 tons)		
	2003	2004	2005	2003	2004	2005
Wheat	9.35	8.12	8.78	3,974	3,862	3,930
Rice	11.25	11.73	12.08	973	852	722
Gram	7.00	7.68	8.23	755	601	744

- 5.36 Calculate (i) Laspeyres' index, (ii) Paasche's index, (iii) Fisher's Ideal index, (iv) Marshall-Edgeworth index, (v) Walsh index and (vi) Palgrave index for the following dairy products.

Dairy Products	Prices			Quantities Produced		
	1998 (base)	1999	2007	1998	1999	2007
Milk	3.95	3.89	4.13	9,675	9,717	10,436
Butter	61.50	62.20	59.70	118	116	116
Cheese	34.80	35.40	38.90	78	78	83

- 5.37 a) Obtain (i) a volume index and (ii) an index of average values for each year from the following data:

Remained Imports (£ millions)		
Year	Declared value	Value on basis of 1938 values
1938	858	858
1939	840	832
1940	1126	807
1941	1132	704

- b) The following table shows the average prices in rupees per quarter.

Cereals	1960	1961	1962
Wheat	110.70	121.85	129.00
Barley	99.50	124.55	120.50
Maize	—	74.90	76.25

Taking 1960 as base, construct a price index number for the three cereal together; (i) for 1961 and for 1962, using the weights 12 for wheat, 8 for barley and 5 for maize.

(B.Z.U., M.A., Econ.)

- 5.38 Compute Fisher's price index number for 1976 with 1961 as base from the following data:

Commodity	Quantity (units)		Value (Rs.)	
	1961	1976	1961	1976
A	100	150	600	1200
B	80	100	400	700
C	60	72	180	432
D	30	33	450	363

(P.U., B.A./B.Sc.)

- 5.39 Compute the quantity and price index numbers from the following data. Describe how the index number could be computed from data such as those in this question.

Commodity	Quantity (units)		Value (£)	
	1997 (q_0)	2007 (q_1)	1997	2007
1	100	150	500	900
2	80	100	320	500
3	60	72	150	360
4	30	33	360	297

The prices and quantities of three commodities during 1955 and 1965 are given below:

Commodity	1955		1965	
	P	q	P	q
A	10	501	12	600
B	40	100	38	194
C	50	76	40	56

Using 1955 as the base period and base period quantities as the weights, compute the weighted aggregative price index and the weighted average of relative price index for 1965.

(P.U., B.A./B.Sc. 1988)

Compute a quantity index number for (i) 1943 on 1938 as base, using 1938 values as weights; and another for (ii) 1938 on 1943 as base, using 1943 values as weights.

Articles	Export of Cotton Yarns and Manufactures			
	Quantity		Value (in '000 £)	
	1938	1943	1938	1943
Cotton Yarn	A	10.0	5.4	399
	B	2.9	2.8	315
Cotton	A	35.3	69.3	841
	B	21.6	83.2	776
Manufactures	C	81.9	4	1,452
	D	68.8	90.2	5,805
				6,381

(P.U., B.A., (Hons.), 1961)

An inquiry into the budgets of middle class families in a city in England gave the following information:

Expenses on	Fuel 5%	Rent 15%	Clothing 20%	Fuel 10%	Misc. 20%
Prices (1929)	£130	£30	£75	£25	£40
Prices (1938)	£145	£30	£65	£23	£45

What changes in cost of living figures of 1929 as compared with that of 1938 are seen?

(P.U., B.A./B.Sc. 1974, 1981)

Compute the consumer price index number for 1940 on the basis of 1939 from the following data, using (i) Aggregative Expenditure Method and (ii) the Household Budget Method.

Commodity	Quantity Consumed in 1939	Unit of Price	Price in 1939	Price in 1940
Rice	240 kg	Rs. per kg	5.75	6.00
Wheat	240 kg	Rs. per kg	5.00	8.00
Gram	40 kg	Rs. per kg	6.00	9.00
Sugar	40 kg	Rs. per kg	20.00	15.00
Arhar	240 kg	Rs. per kg	8.00	10.00
Ghee	4 kg	Rs. per kg	2.00	1.50

(P.U., B.A./B.Sc. 1973)

- 3.44 Compute the consumer price index number for the following data for 2007 with 2000 as the base year. Use as weights (i) the quantities consumed in the base year, (ii) the value in the base year.

Article	Quantity 2000	Price (Rs.) in	
		2000	2007
Food	5 pounds	18.00	26.50
Cloth	30 metres	2.60	2.80
Electricity	75 units	0.25	0.30
Rent	3 rooms	30.00	27.50
Miscellaneous	34 units	0.50	0.60

- 3.45 Calculate an index number of retail prices for the following selection of foodstuffs for 1 Sept. 2003 (July 2000 = 100).

Item	Unit	Purchases (units)	Price (Rs.)	
			July 2000	1 Sept. 2003
Flour	Kilogram	18	1.90	2.30
Milk	Kilogram	2	22.00	28.00
Bread	200 gram	2	1.00	1.50
Tea	450 gram	4	8.20	10.35
Sugar	Kilogram	3	7.50	7.75
Milk	Litre	2.5	2.50	4.00
Butter	450 gram	2	12.30	15.00
Eggs	Dozen	1.5	6.50	10.50
Potatoes	Kilogram	10	2.60	3.20

- 3.46 a) Define an index number. Discuss the main steps involved in the construction of wholesale price index numbers.
 b) Given:

Commodity	Quantity (units)		Value	
	2001	2006	2001	2006
A	100	150	600	1200
B	50	100	400	700
C	60	72	180	432
D	30	33	450	363

Compute the following:

- i) Fisher's quantity index number for 2006,
 ii) Simple aggregative value index for 2006.

(P.U. B.A./B.Sc. 2006)

- 3.47 Obtain i) A simple aggregative value index
 ii) An index of average values, for each year from the following data:

Year	Retained Imports (Billion Rs.)	
	Declared value	Value on the basis of 2002 values
2002	860	860
2003	950	832
2004	1300	857
2005	1450	764

(P.U. B.A./B.Sc. 2006)

CHAPTER 6

PROBABILITY

PROBABILITY

6.1 INTRODUCTION

The word probability has two basic meanings: (i) a quantitative measure of uncertainty and (ii) a measure of degree of belief in a particular statement or problem.

Probability and statistics are fundamentally interrelated. Probability is often called the vehicle of statistics. The area of inferential statistics in which we are mainly concerned with drawing inferences from experiments or situations involving an element of uncertainty, leans heavily upon probability theory. Uncertainty is also an inherent part of statistical inference as inferences are based on a sample, and a sample being a small part of the larger population, contains incomplete information. A similar type of uncertainty occurs when we toss a coin, draw a card or throw dice, etc. The uncertainty in all these cases is measured in terms of probability.

It is always clear what we mean when we make statements of the type that it is very likely to rain today or I have a fair chance of passing the annual examination or A will probably win a prize, etc. In such of these statements, the natural state of uncertainty is expressed, but on the basis of past evidence, we have some degree of personal belief in the truth of each statement.

The foundations of probability were laid by two French mathematicians of the seventeenth century—Blaise Pascal (1623–1662) and Pierre De Fermat (1601–1665)—in connection with gambling problems. Later on it was developed by Jakob Bernoulli (1654–1705), Abraham De Moivre (1667–1754) and Pierre Simon Laplace (1749–1827). The modern treatment of probability theory which consists of a few axioms and rules resulting from these, was developed during the twenties and thirties of twentieth century.

Today the probability theory has a wide field of application and is used to make intelligent decisions in Economics, Management, Operations Research, Sociology, Psychology, Astronomy, Engineering and Genetics where chance and uncertainty are involved.

As the probability theory these days is best understood through the application of the modern set theory, brief descriptions of the basic concepts, notation and operations of set theory that are relevant to probability, are given below.

6.2 ASIDE—SETS

A set is any well-defined collection or list of distinct objects, e.g. a group of students, the books in a library, the integers between 1 and 100, all human beings on the Earth, etc. The term well-defined here means that any object must be classified as either belonging or not belonging to the set under consideration, and the term distinct implies that each object must appear only once. The objects that are in a set are called members or elements of that set. Sets are usually denoted by capital letters such as A , B , C , etc., while their elements are represented by small letters such as a , b , c , y , etc. Elements are enclosed in braces to represent a set, e.g.

$$A = \{a, b, x, y\} \text{ or } B = \{1, 2, 3, 7\}$$

If x is an element of a set, A , we write $x \in A$, which is read as "x belongs to A" or x is in A.

If x does not belong to A , i.e. x is not an element of A , we write $x \notin A$. The number of a set A , denoted by $n(A)$, is defined as the number of elements in A .

A set that has no elements is called an *empty* or a *null set* and is denoted by the symbol \emptyset . It must be noted that $\{0\}$ is not an empty set as it contains an element 0. If a set contains only one element, it is called a *unit set* or a *singleton set*. It is also important to note the difference between an element "x" and a unit set $\{x\}$. The elements of a set may be sets themselves.

A set may be specified in two ways. We may either give a list of all the elements of a set (the "*Roster*" method), e.g.

$$A = \{1, 2, 5, 7, 9, 11\}; B = \{\text{a book, a city, a clock, a teacher}\}.$$

or we may state a rule that enables us to determine whether or not a given object is a member of this (the "*rule*" method or "*set builder*" method), e.g.

$$A = \{x \mid x \text{ is an odd number and } x < 12\}$$

meaning that A is a set of all elements x such that x is an odd number and x is less than 12. The vertical line is read as "such that". The repetition or the order in which the elements of a set occur does not change the nature of the set. The size of a set is given by the number of elements present in it. This number may be finite or infinite. Thus a set is *finite* when it contains a finite number of elements, otherwise it is an *infinite set*. The empty set is regarded as a finite set. Examples of finite sets are:

- i) $A = \{1, 2, 3, \dots, 99, 100\}$;
- ii) $B = \{x \mid x \text{ is a month of the year}\}$;
- iii) $C = \{x \mid x \text{ is a printing mistake in a book}\}$;
- iv) $D = \{x \mid x \text{ is a living citizen of Pakistan}\}$; etc.

and the sets:

- i) $A = \{x \mid x \text{ is an even integer}\}$;
- ii) $B = \{x \mid x \text{ is a real number between 0 and 1 inclusive}\}, \text{i.e., } B = \{x \mid 0 \leq x \leq 1\}$;
- iii) $C = \{x \mid x \text{ is a point on a line}\}$;
- iv) $D = \{x \mid x \text{ is a sentence in the English language}\}$; etc. see the examples of an infinite set.

A set A is said to be in *one-to-one correspondence* with a set B when every element of set A can be made to correspond to one and only one element of set B and conversely. For example, if

$$A = \{1, 2, 3, 4\} \text{ and } B = \{a, b, c, d\}$$

then the sets A and B are in one-to-one correspondence.

A set is called *countably infinite* or *denumerable* when its elements can be put into a one-to-one correspondence with the sequence of positive integers. A set is said to be *non-denumerable* whose elements cannot be enumerated.

6.2.1 Subsets. A set that consists of some elements or members of another set, is called a *subset* of that set. For example, if B is a subset of A , then every member of set B is also a member of set A and we write

$$B \subset A \text{ or equivalently } A \supset B$$

which is read as 'B is a subset of A' or 'B is contained in A', or 'A contains B'. For example, if

$$A = \{1, 2, 3, 4, 5, 10\} \text{ and } B = \{1, 3, 5\}$$

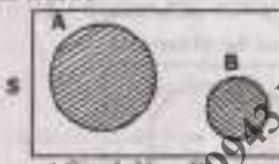
then $B \subset A$, i.e. B is contained in A.

It should be noted that a set A is always regarded a subset of itself and an empty set \emptyset is considered to be (or accepted as) a subset of every set. Two sets A and B are equal or identical, if and only if they contain exactly the same elements. That is $A = B$ if and only if $A \subset B$ and $B \subset A$. If a set B contains some but not all of the elements of another set A while A contains each element of B, i.e. if

$$B \subset A \text{ and } B \neq A$$

then the set B is defined to be a proper subset of A. The large or the original set of which all the sets we talk about, are subsets is called the universal set or the space and is generally denoted by S or Ω . The universal set thus contains all possible elements under consideration. It is also regarded a subset of itself. $S = S$ with a element will produce 2^n subsets, including S and \emptyset .

4.2.2 Venn Diagram. A diagram that is used to represent sets by circular regions, parts of circular regions or their complements with respect to a rectangle representing the space S is called a Venn diagram, named after the English logician John Venn (1834-1923). The Venn diagrams are used to represent sets and subsets in a pictorial way and to verify the relationship among sets and subsets. An example of a Venn diagram follows:



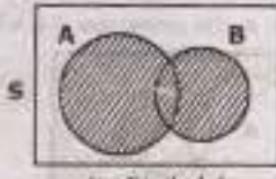
A Simple Venn Diagram

4.2.3 Operations on Sets. Let the sets A and B be the subsets of some universal set S. Then these can be combined and operated on in various ways to form new sets which are also subsets of S. The operations are union, intersection, difference and complementation.

The union or sum of two sets A and B, denoted by $A \cup B$, and read as "A union B" or "A or B", means the set of all elements that belong to at least one of the sets A and B, that is

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}$$

By means of a Venn diagram, $A \cup B$ is shown by the shaded area as below:

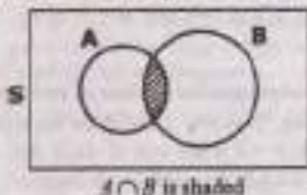


$A \cup B$ is shaded

The intersection of two sets A and B, denoted by $A \cap B$ or by AB , and read as "A intersection B" or "A cap B", means the set of all elements that belong to both A and B; that is

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

Diagrammatically $A \cap B$ is shown by the shaded area as below:



The operations of union and intersection have been defined for two sets only. They can conveniently be extended to any finite number of sets.

Two sets A and B are defined to be *disjoint* or *mutually exclusive* or *non-overlapping* when they have no elements in common, i.e. when their intersection is an empty set or $A \cap B = \emptyset$. On the other hand, two sets A and B are said to be *conjoint* when they have at least one element in common.

- iii) The difference of two sets A and B , denoted by $A - B$ or by $A - (A \cap B)$, is the set of elements of A which do not belong to B . Symbolically,

$$A - B = \{x \mid x \in A \text{ and } x \notin B\}$$

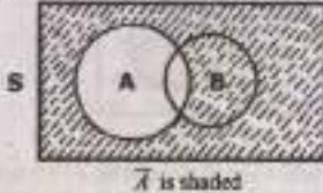
It is to be pointed out that in general $A - B \neq B - A$, because $B - A = \{x \mid x \in B \text{ and } x \notin A\}$. The shaded area of the following Venn diagram shows the difference $A - B$.



It is to be noted that $A - B$ and B are disjoint sets. If A and B are disjoint, then the difference coincides with the set A .

- iv) Complementation: The particular difference $S - A$, that is, the set of all those elements which do not belong to A , is called the *Complement* of A and is denoted by \bar{A} or by A' . In symbols, $\bar{A} = \{x \mid x \in S \text{ and } x \notin A\}$

The complement of S is the empty set \emptyset . The complement of A is shown by the shaded portion in the following Venn diagram.



It should be noted that $A - B$ and $A \cap \bar{B}$, where \bar{B} is the complement of set B , are the same.

6.2.4 The Algebra of Sets. The algebra of sets provides us with laws which can be used to solve many problems in probability calculations. Let A , B and C be any subsets of the universal set S . Then we have

- Commutative laws

$$A \cup B = B \cup A \text{ and } A \cap B = B \cap A$$

- Associative laws

$$(A \cup B) \cup C = A \cup (B \cup C) \text{ and } (A \cap B) \cap C = A \cap (B \cap C)$$

- Distributive laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \text{ and } A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

- Idempotent laws

$$A \cup A = A \text{ and } A \cap A = A$$

- Identity laws

$$A \cup \emptyset = A, A \cap \emptyset = \emptyset, A \cup S = S, \text{ and } A \cap S = A.$$

- Complementation laws

$$A \cup \bar{A} = S, A \cap \bar{A} = \emptyset, (\bar{A}) = A, \text{ and } \bar{\bar{A}} = A.$$

- De Morgan's laws

$$\overline{(A \cup B)} = \bar{A} \cap \bar{B}, \text{ and } \overline{(A \cap B)} = \bar{A} \cup \bar{B}.$$

These laws can be visualized by means of the Venn diagrams.

6.2.5 Partition of Sets. A partition of a set S is a sub-division of the set into non-empty subsets which are disjoint and exhaustive, i.e. their union is the set S itself. This implies that

- $A_i \cap A_j = \emptyset$, where $i \neq j$.

- $A_1 \cup A_2 \cup \dots \cup A_n = S$.

The subsets in a partition are called cells. Let us consider a set $S = \{a, b, c, d, e\}$. Then $\{a, b\}$ and $\{c, d, e\}$ is a partition of S as each element of S belongs to exactly one cell.

6.2.6 Class of Sets. A set of sets is called a class, e.g. in a set of lines, each line is a set of points. The class of all subsets of a set A is called the power set of A and is denoted by $P(A)$. For example, if $A = \{I, II, III\}$, then $P(A) = \{\emptyset, \{I\}, \{II\}, \{III\}, \{I, II\}, \{I, III\}, \{II, III\}, \{I, II, III\}\}$.

6.2.7 Cartesian Product Sets. The Cartesian product of sets A and B , denoted by $A \times B$, (read as A cross B), is a set that contains all ordered pairs (x, y) , where x belongs to A and y belongs to B . Practically, we write

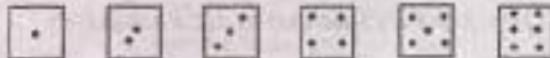
$$A \times B = \{(x, y) \mid x \in A \text{ and } y \in B\}.$$

This set is also called the Cartesian set of A and B , named after the French mathematician René Descartes (1596–1650). The product of a set A by itself is denoted by A^2 . This concept of product may be extended to any finite number of sets.

Let $A = \{H, T\}$ and $B = \{1, 2, 3, 4, 5, 6\}$. Then the Cartesian product set is the collection of the following twelve (2×6) ordered pairs:

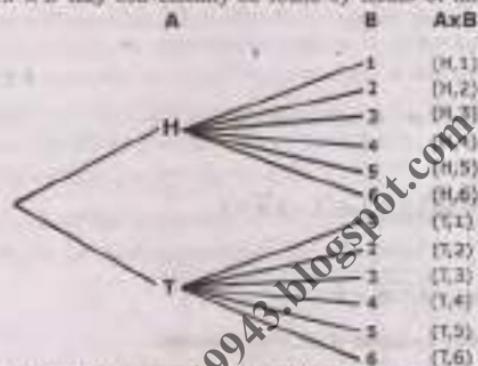
$$A \times B = \{(H, 1); (H, 2); (H, 3); (H, 4); (H, 5); (H, 6);\\ (T, 1); (T, 2); (T, 3); (T, 4); (T, 5); (T, 6)\}$$

Clearly, these twelve elements together make up the universal set S when a coin and a die are tossed together. A die (plural, dice) is a cube of wood or ivory whose six faces are marked with dots, as shown below:



It is relevant to note that, in general, $A \times B \neq B \times A$.

The product $A \times B$ may conveniently be found by means of the so-called tree diagram shown below:



The "tree" is constructed from left to right. A "tree diagram" is a useful device for enumerating all the possible outcomes of two or more sequential events; the possible outcomes are represented by the individual paths, or branches of the tree. This diagram is also used when we apply multiplication rules to compute probabilities.

6.2.8 Relation and function. A relation from a set A to a set B is a subset of the Cartesian product of $A \times B$. Such a relation is usually called a binary relation. That is, a relation is an association between two or more objects. The set of the first elements of a binary relation is called the domain of the relation, while the set of the second elements of the relation is called the range. For example, if a relation is $F = \{(1,4), (2,7), (3,12)\}$, then its domain is $D = \{1, 2, 3\}$ and its range is $R = \{4, 7, 12\}$.

A function is a rule that assigns values in some manner from a set A to a set B such that for each element x of A , there is a unique element y of B . Such an assignment is usually written as $f: A \rightarrow B$ and is a function from A to B . In other words, a function is a special kind of binary relation that associates each element of the domain to a unique element of the range. It is to be emphasized that a function is a binary relation in which no first element is repeated. The value of the function f at $x \in A$ is denoted by $y = f(x) \in B$.

The variable x that represents the elements of the domain, is called the independent variable, while the variable $y (= f(x))$ representing elements of the range is referred to as the dependent variable. Functions are also called single-valued functions. A function, whose range consists of numbers, is called a numerical function. A function whose domain and range consist of sets of real numbers, is said to be

~~an~~ *even function* of a real variable. A function $f(x)$ is defined to be an even function, if for every x in a ~~range~~ $f(-x) = f(x)$, e.g., $f(x) = x^2$. A function $f(x)$ having the property that $f(-x) = -f(x)$, is said to be *odd function*, e.g., $f(x) = x^{2n+1}$.

4.2 RANDOM EXPERIMENT

The term *experiment* means a planned activity or process whose results yield a set of data. A single ~~instance~~ ~~outcome~~ of an experiment is called a *trial*. The result obtained from an experiment or a trial is called ~~outcomes~~.

An experiment which produces different results even though it is repeated a large number of times under essentially similar conditions, is called a *random experiment*. The tossing of a fair coin, the rolling of a balanced die, drawing of a card from a well-shuffled deck of 52 playing cards, selecting a sample, etc. are examples of random experiments. A random experiment has three properties:

- (i) The experiment can be repeated, practically or theoretically, any number of times.
- (ii) The experiment always has two or more possible outcomes. An experiment that has one possible outcome, is not a random experiment.
- (iii) The outcome of each repetition is unpredictable, i.e. it has some degree of uncertainty.

It is to be remembered that an ordinary deck of playing cards contains 52 cards arranged in 4 suits of 13 each. The four suits are called diamonds, hearts, clubs and spades; the first two are red and the last two black. The face values called denominations, of the 13 cards in each suit are ace, 2, 3, ..., 10, jack, queen and king. The term *honour card* refers to the denominations ace, 10, jack, queen and king. The jokers are king, queen and jack. These cards are used for various games such as whist, bridge, etc.

4.3.1 Sample Space. A set consisting of all possible outcomes that can result from a random experiment (real or conceptual), is defined to be a *sample space* for the experiment and is denoted by the symbol S . Each possible outcome is a member of the sample space and is called a *sample point* in that space. For instance, the experiment of tossing a coin results in either of the two possible outcomes: a head (H) or a tail (T); landing on its edge or rolling away is not considered. The sample space for this experiment may be expressed in set notation as $S = \{H, T\}$. The sample space for tossing two coins once (or tossing a coin twice) will contain four possible outcomes denoted by $S = \{HH, HT, TH, TT\}$. Clearly, the Cartesian product $A \times A$, where $A = \{H, T\}$. Similarly, the sample space S for the random experiment of throwing two six-sided dice can be described by the Cartesian product $A \times A$, where $A = \{1, 2, 3, 4, 5, 6\}$. In other words,

$$S = A \times A = \{(x, y) \mid x \in A \text{ and } y \in A\}$$

x denotes the number of dots on the upper face of the first die and y , the number of dots on the upper face of the second die; and S contains 36 outcomes or sample points, which may also be conveniently represented in the following manner:

$$\begin{aligned} S = & \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ & (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ & (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\ & (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ & (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\} \end{aligned}$$

This sample space may more briefly be expressed as

$$S = \{(i, j) \mid i = 1, 2, 3, 4, 5, 6; \quad j = 1, 2, 3, 4, 5, 6\}$$

A sample space that contains a finite number of sample points is said to be a *finite sample space*. It is defined to be a *discrete sample space* if the sample points can be placed in a one-to-one correspondence with the positive integers or if it is a finite. If it satisfies neither of these criteria, it is called *continuous*. It should be noted that our sample space S will be finite unless otherwise stated.

6.3.2 Events. An event is an individual outcome or any number of outcomes (sample points) of a random experiment or a trial. In set terminology, any subset of a sample space S of the experiment is called an event. An event that contains exactly one sample point, is defined a *simple event*. A *compound event* contains more than one sample point and is produced by the union of simple events. For instance, the occurrence of a 6 when a die is thrown, is a simple event, while the occurrence of a sum of 10 with a pair of dice, is a compound event, as it can be decomposed into three simple events (4, 6), (5, 5) and (6, 4).

An event A is said to occur if and only if the outcome of the experiment corresponds to an element of A . The event "not- A " is denoted by \bar{A} or A' and is called the *negation* (or complement) event of A . For example, the complement of "heads" is "tails" for tossing of one coin, the complement "at least one head" on 4 tosses of a coin is "no heads". A sample space consisting of n sample points produce 2^n different subsets of simple and compound events of which sample point can either be included or excluded in forming a subset.

To illustrate, let us consider a sample space S containing three sample points, i.e. $S = \{a, b, c\}$.

Then the eight possible subsets are:

$$\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}$$

Each of these subsets is an event. The subset $\{a, b, c\}$ is the sample space itself and is also an event, always occurs and is known as the *certain* or *true* event. The empty set \emptyset is also an event, sometimes known as *impossible* event, because it can never occur.

This class of $2^3 = 8$ subsets (events) can be thought of as a *field* which is denoted by \mathcal{F} . All events have the following characteristics:

- The union of any number of events will result in a set that belongs to \mathcal{F} .
- The intersection of any number of events will result in a set that belongs to \mathcal{F} .
- The difference of any two events belongs to \mathcal{F} .
- The complement of any event belongs to \mathcal{F} .

Mutually Exclusive Events. Two events A and B of a single experiment are said to be *mutually exclusive* or *disjoint* if and only if they cannot both occur at the same time. That is they have no common. For instance, when we toss a coin, we get either a head or a tail, but not both, the two events head and tail are therefore mutually exclusive; when a die is rolled, the outcomes are mutually exclusive as we get one and only one of six possible outcomes 1, 2, 3, 4, 5 or 6. Similarly, a student either qualifies or fails, a single birth must be either a boy or a girl, it cannot be both, etc. Three or more events originating from the same experiment are *mutually exclusive* if pairwise they are mutually exclusive, the two events can occur at the same time, they are not *mutually exclusive*, e.g., if we draw a card from

any deck of 52 playing cards, it can be both a king and a diamond. Therefore kings and diamonds are not mutually exclusive. Similarly, inflation and recession are not mutually exclusive events.

Exhaustive Events. Events are said to be collectively exhaustive, when the union of mutually exclusive events is the entire sample space S . Thus, in our coin-tossing experiments, head and tail are collectively exhaustive set of events. A group of mutually exclusive and exhaustive events is called a partition of the sample space. For instance, events A and A' form a partition as they are mutually exclusive and their union is the entire sample space.

Equally Likely Events. Two events A and B are said to be equally likely, when one event is as likely to occur as the other. In other words, each event should occur in equal number in repeated trials. For example, when a fair coin is tossed, the head is as likely to appear as the tail, and the proportion of times each side is expected to appear is $\frac{1}{2}$.

6.3.3. Events and Symbolic Representations. For convenience, the verbal statements of some events and their corresponding symbolic representations in sets are listed below:

Verbal statement:

Set Notation:

Event A :

$A \subset S$

Event A is impossible:

$A = \emptyset$

Event A is sure (certain):

$A = S$

non- A (Event A does not occur):

$A^c = S - A$

Event A or event \bar{A} :

$A \cup \bar{A} = S$

Event A or event B :

$A \cup B$

Event A and event B :

$A \cap B$

Events A and B are mutually exclusive:

$A \cap B = \emptyset$

Events A and B are exhaustive:

$A \cup B = S$

Event B occurs when A occurs:

$A \subseteq B$

Event A occurs but B does not occur:

$A \cap \bar{B}$

Event B occurs given that A has occurred:

$B | A$

6.4. Counting Sample Points. When the number of sample points in a sample space S is very large, it becomes very inconvenient and difficult to list them all and to count the number of points in the sample space S and in the subsets of S . We then need some methods or rules which help us to count the number of all sample points without actually listing them. A few of the basic rules frequently used in probability theory are briefly described here.

Rule of Multiplication. If a compound experiment consists of two experiments such that the first experiment has exactly m distinct outcomes and, if corresponding to each outcome of the first experiment there can be n distinct outcomes of the second experiment, then the compound experiment has mn outcomes.

For example, the compound experiment of tossing a coin and throwing a die together consists of the coin-tossing with two distinct outcomes (H, T), and the die-throwing with six outcomes ($1, 2, 3, 4, 5, 6$). The total number of possible distinct outcomes of the compound experiment is $2 \times 6 = 12$.

experiment is therefore $2 \times 6 = 12$, as each of the two outcomes of the coin-tossing experiment can be combined with each of the six outcomes of die-throwing experiment (see Cartesian product on page 178). A tree diagram (page 178) provides a good illustration of this rule.

The rule of multiplication can be readily extended to compound experiments consisting of a number of experiments performed in a given sequence.

(b) Rule of Permutation. A permutation is any ordered subset from a set of n distinct objects. The number of permutations of r objects, selected in a definite order from n distinct objects is denoted by the symbol ${}^n P_r$.

To derive the computational formula for ${}^n P_r$ ($r < n$), we proceed as below:

The first object may be chosen in n ways, and corresponding to each way of first selection, the second object may be chosen in $(n - 1)$ ways. Similarly, once selections have been made for both these second objects, there are $(n - 2)$ objects left and the third object may be selected in $(n - 2)$ ways and so on. The r th object may be chosen in $(n - r + 1)$ ways. Thus the first r objects may be chosen in $n \times (n - 1) \times (n - 2) \times \dots \times (n - r + 1)$ ways.

Hence

$${}^n P_r = n(n-1)(n-2)\dots(n-r+1)$$

In particular, when $r = n$, ${}^n P_n = n(n-1)(n-2)\dots(n-2) \times 1$

$$= n! \text{ (read } n \text{ factorial)}$$

It is relevant to note that $1! = 1$ and that we define $0! = 1$.

The expression for ${}^n P_r$ on multiplication by $\frac{(n-r)!}{(n-r)!}$, may be written as ${}^n P_r = \frac{n!}{(n-r)!}$

The number of permutations of n objects, selected all at a time, when n objects consist of n_1 of one kind, n_2 of a second kind, ..., n_r of r th kind, ($\sum n_i = n$) is $P = \frac{n!}{n_1! n_2! \dots n_r!}$

(c) Rule of Combination. A combination is any subset of r objects, selected without regard to their order, from a set of n distinct objects. The total number of such combinations is denoted by the symbol ${}^n C_r$ or $\binom{n}{r}$, (read " n above r "), where $r \leq n$.

The computational formula for ${}^n C_r$ is derived as below:

Let there be $\binom{n}{r}$ possible combinations of r objects, selected from n distinct objects. Then each combination of r objects, if order is not disregarded, can be arranged in $r!$ different orders and thus makes $r!$ different permutations. The number of permutations or ordered subsets that correspond to

combinations, each having r objects, is therefore $r! \binom{n}{r}$. But the total number of permutations of r objects, selected from n objects is ${}^n P_r$. Thus we conclude that

$$r! \binom{n}{r} = {}^n P_r$$

$$\text{or } \binom{n}{r} = \frac{{}^n P_r}{r!} = \frac{n(n-1)(n-2)\dots(n-r+1)}{r(r-1)(r-2)\dots3 \times 2 \times 1} \\ = \frac{n!}{r!(n-r)!}.$$

The quantity $\binom{n}{r}$ or ${}^n C_r$ is also called a binomial co-efficient because of its appearance in the binomial

expansion $(a+b)^n = \sum_{r=0}^n \binom{n}{r} a^{n-r} b^r$. The binomial coefficient has two important properties:

$$\text{(i) } \binom{n}{r} = \binom{n}{n-r}, \text{ and (ii) } \binom{n}{r-1} + \binom{n}{r} = \binom{n+1}{r}.$$

Large factorials may be conveniently evaluated by using an approximation known as Stirling's

$$n! \approx n^{n+1/2} e^{-n} \sqrt{2\pi},$$

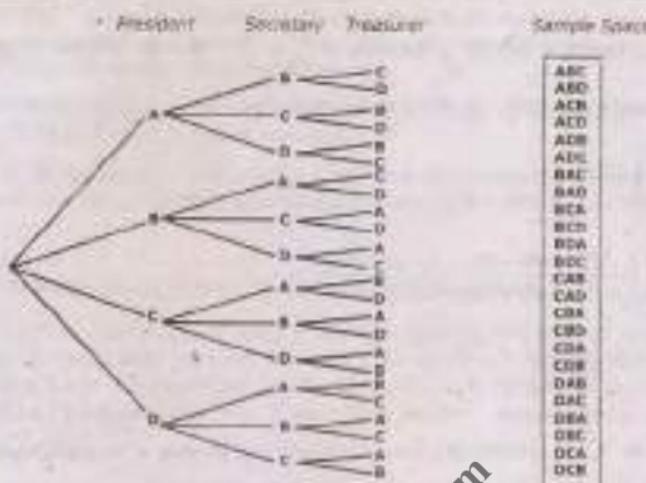
$\pi \approx 3.1416$ and $e \approx 2.7183$

Example 6.1 A club consists of four members. How many sample points are in the sample space if three officers: president, secretary and treasurer, are to be chosen?

It is evident that the order in which 3 officers are to be chosen, is of significance. Thus there are 4 choices for the first office, 3 choices for the second office and 2 choices for the third office. Hence the number of sample points is $4 \times 3 \times 2 = 24$. In other words, the number of permutations is

$${}^4 P_3 = \frac{4!}{(4-3)!} = 4 \times 3 \times 2 = 24.$$

Let the four members be A, B, C and D. Then a tree diagram which provides an organized way of possible arrangements, for this example, is given on the next page.



Example 6.2 A three-person committee is to be formed from a list of four persons. How many sample points are associated with the experiment?

Since the order in which the three persons of the committee are chosen, is unimportant, therefore an example of a problem involving combinations. Thus the desired number of combinations is

$$\frac{n!}{r!(n-r)!} = \frac{4!}{3!(4-3)!} = 4$$

In other words, the sample space associated with the experiment contains 4 sample points.

These two examples serve to illustrate the difference between a permutation and a combination.

Example 6.3 How many sample points are in the sample space when a person draws a hand of 5 cards from a well-shuffled ordinary deck of 52 cards?

The total number of sample points is given by:

$$\binom{n}{r} = \binom{52}{5} = \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1} = 2,598,960$$

6.4 DEFINITIONS OF PROBABILITY

Probability can be discussed from two points of view: the objective and the subjective. Objective probability can be classified into the following categories, each of which is briefly discussed as follows:

(a) The Classical or *A Priori* Definition of Probability is given as follows:

If a random experiment can produce a mutually exclusive and equally likely outcomes and if one of these outcomes are considered favourable to the occurrence of a certain event A , then the probability

the event A , denoted by $P(A)$, is defined as the ratio $\frac{m}{n}$. Symbolically, we write

$$P(A) = \frac{m}{n} = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

This definition was formulated by the French mathematician P.S. Laplace (1749–1827) and can be conveniently used in experiments where the total number of possible outcomes and the number of outcomes favourable to an event can be determined.

The classical definition has the following shortcomings:

- (i) This definition is said to avoid circular reasoning as the term *equally likely* really means *equally probable*. Thus probability is defined by introducing concepts that presume a prior knowledge of the meaning of probability.
- (ii) This definition is not applicable when the assumption of equally likely does not hold.
- (iii) This definition becomes vague when the number of possible outcomes may be infinite.

The Relative Frequency or Aumann's Definition of Probability. If a random experiment is repeated a large number of times, say n , under identical conditions and if an event A is observed to occur m times, then the probability of the event A is defined as the limit of the relative frequency $\frac{m}{n}$ as n tends to infinity. Symbolically, we write

$$P(A) = \lim_{n \rightarrow \infty} \frac{m}{n}$$

This definition assumes that as n increases indefinitely, the value $\frac{m}{n}$ tends to become stable at the real value $P(A)$. To investigate the stand of the long-run stability, several coin-tossing experiments were performed. The results of some well known experiments are shown below:

Experiment	Number of times coins tossed (n)	Number of heads obtained (m)	Ratio $\frac{m}{n}$
Buffon	4,040	2,048	0.5069
K. Pearson	12,000	6,019	0.5016
K. Pearson	24,000	12,012	0.5005

It is quite obvious that the value of the ratio $\frac{m}{n}$ fluctuates about the number 0.5 and becomes near to 0.5 as the number of throws increases. This sort of a long-run frequency property provides a basis of the theory of probability.

This definition is also called the statistical or empirical definition of probability as it is based on empirical data. It is more useful for practical purposes.

This definition too has certain limitations as the conditions under which an experiment is performed may change from trial to trial and it is not possible, in practice, to repeat the experiment an infinite number of times and hence the ratio $\frac{m}{n}$ may not be unique.

(c) **The Axiomatic Definition of Probability.** This definition, introduced in 1933 by the Russian mathematician Andrei N. Kolmogorov (1903–1987), is based on a set of axioms, where an axiom is a statement that is assumed to be true. Let S be a sample space with the sample points E_1, E_2, \dots, E_n . To each sample point, we assign a real number, denoted by the symbol $P(E_i)$, and called the probability of E_i , that must satisfy the following basic axioms:

- Axiom (i).** For any event E_i , $0 \leq P(E_i) \leq 1$.
- Axiom (ii).** $P(S) = 1$ for the sure event S .
- Axiom (iii).** If A and B are mutually exclusive events (outcomes), then $P(A \cup B) = P(A) + P(B)$.

It is to be emphasized that the axiomatic theory of probability assumes that some probability, defined as a non-negative real number is to be attached to each sample point E_i such that the sum of such numbers must equal one. The assignment of probabilities may be based on past evidence or on other underlying conditions.

Probability of an event. If an event A is defined in a sample space S , then its probability will be equal to the sum of the probabilities of all sample points that are included in A , i.e. $P(A) = \sum P(E_i)$. If all the n possible outcomes of a random experiment are equally likely to occur, then each outcome is assigned the same probability $\frac{1}{n}$, e.g. in throwing a fair die once, $P(1) = \frac{1}{6}, P(2) = \frac{1}{6}, \dots, P(6) = \frac{1}{6}$.

It follows from axiom (iii) that for any event A containing m equally likely outcomes (sample points) we have

$$P(A) = \frac{\text{Number of sample points in } A}{\text{Number of sample points in } S} = \frac{n(A)}{n(S)}$$

It is interesting to note that in the classical definition of probability,

- i) $P(A) = \frac{m}{n} \geq 0;$
- ii) $P(S) = \frac{n}{n} = 1;$
- iii) If A and B are mutually exclusive events, and m_1 and m_2 outcomes are favourable to A and B respectively, then

$$P(A \cup B) = \frac{m_1 + m_2}{n} = P(A) + P(B)$$

Thus the classical definition of probability also satisfies Kolmogorov's axioms. It is reasonable to conclude that the probability is always a number between zero and one (inclusive).

6.4.1 Subjective or Personalistic Probability. As its name suggests, the subjective or personalistic probability is a measure of the strength of a person's belief regarding the occurrence of an event A . Probability in this sense is purely subjective and is based on whatever evidence is available to the individual. This definition being flexible, may be applied to those real-world situations where neither equally likely nor a relative frequency approach is possible. The subjective probability has a disadvantage that two or more persons faced with the same evidence may arrive at different probabilities.

Example 6.4 If a card is drawn from an ordinary deck of 52 playing cards, find the probability that the card is a red card, (ii) the card is a diamond, (iii) the card is a 10.

The total number of possible outcomes is 52, and we assume that all possible outcomes are equally likely.

- (i) Let A represent the event that the card drawn is a red card. Then the number of outcomes favourable to the event A is 26 since there are 26 red cards.

$$\text{Hence } P(A) = \frac{n(A)}{n} = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

$$= \frac{26}{52} = \frac{1}{2}$$

- (ii) Let B denote the event that the card drawn is a diamond. Then the number of outcomes favourable to the event B is 13 since there are 13 diamonds.

$$\text{Hence } P(B) = \frac{13}{52} = \frac{1}{4}$$

- (iii) Let C denote the event that the card drawn is a 10. Then the number of outcomes favourable to C is 4 as there are four 10's.

$$\text{Thus } P(C) = \frac{4}{52} = \frac{1}{13}$$

Example 6.5 A fair coin is tossed three times. What is the probability that at least one head appears?

The sample space for the experiment is

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

$$n(S) = 8.$$

As the coin is fair, we assume that each of these outcomes is equally likely to appear. Therefore we assign probability of $\frac{1}{8}$ to each outcome.

Let A denote the event that at least one head appears. Then

$$A = \{HHH, HHT, HTH, THH, HTT, THT, TTH\}$$

$$\therefore n(A) = 7.$$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{7}{8}$$

Example 6.6 If two fair dice are thrown, what is the probability of getting (i) a double six? (ii) sum of 8 or more dots?

The sample space S is represented by the following 36 outcomes:

$$\begin{aligned} S = & \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6) \\ & (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6) \\ & (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) \\ & (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) \\ & (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\} \end{aligned}$$

As the dice are fair, therefore each of these 36 outcomes is equally likely and a probability is attached with each outcome.

- (i) Let A represent the event that a double six occurs.

$$\text{Then } A = \{(6, 6)\} \text{ and thus } P(A) = \frac{1}{36}.$$

- (ii) Let B denote the event that a sum of 8 or more dots occurs.

$$\begin{aligned} \text{Then } B = & \{(6, 2), (5, 3), (4, 4), (3, 5), (2, 6), (6, 1), (5, 4), (4, 5), \\ & (3, 6), (6, 4), (5, 5), (4, 6), (6, 3), (5, 6), (6, 5), \\ & \text{i.e., } n(B) = 15. \end{aligned}$$

$$\text{Hence } P(B) = \frac{15}{36} = \frac{5}{12}.$$

Example 6.7 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 of their being three white and three black.

The total number of positive equally likely outcomes in S is

$$\binom{10}{6} = \frac{10!}{6!(10-6)!} = 210$$

Let A represent the event that three white and three black balls are taken. Then the number of outcomes that correspond to the event A is $\binom{6}{3} \times \binom{4}{3} = 80$.

$$\text{Therefore } P(A) = \frac{n(A)}{n(S)} = \frac{80}{210} = \frac{8}{21}.$$

Example 6.8 An employer wishes to hire three people from a group of 15 applicants, 8 men and 7 women, all of whom are equally qualified to fill the position. If he selects the three at random, what is the probability that (i) all three will be men, (ii) at least one will be a woman? (P.U., M.A., Econ. 2010)

The sample space S contains $\binom{15}{3} = 455$ sample points, the number of ways in which 3 people can be selected from 15 applicants.

- (i) Let A represent the event that the three selected will be men. Then A contains $\binom{8}{3} = 56$ sample points, the number of ways in which 3 men can be selected from 8 men;

$$\text{Therefore } P(A) = \frac{n(A)}{n(S)} = \frac{56}{455} = \frac{8}{65}$$

- (ii) At least one woman means one, two or three women. Let B denote the event that at least one woman is selected.

$$\begin{aligned} \text{Then } n(B) &= \binom{7}{1} \binom{8}{2} + \binom{7}{2} \binom{8}{1} + \binom{7}{3} \\ &= 196 + 168 + 35 = 399 \text{ sample points.} \end{aligned}$$

$$\text{Hence } P(B) = \frac{n(B)}{n(S)} = \frac{399}{455} = \frac{57}{65} = 0.877.$$

Example 6.9 Four items are taken at random from a box of 12 items and inspected. The box is rejected if more than 1 item is found to be faulty. If there are 3 faulty items in the box, find the probability that the box is accepted.

The sample space S contains $\binom{12}{4} = 495$ sample points.

The box contains 3 faulty and 9 good items. The box is accepted if there is (i) no faulty item, or (ii) exactly 1 item in the sample of 4 items selected.

Let A denote the event the number of faulty items chosen is 0 or 1. Then

$$\begin{aligned} n(A) &= \binom{3}{0} \binom{9}{4} + \binom{3}{1} \binom{9}{3} \\ &= 126 + 252 = 378 \text{ sample points.} \end{aligned}$$

$$P(A) = \frac{378}{495} = 0.76$$

The probability that the box is accepted is 0.76.

LAWS OF PROBABILITY

Following are some of the basic rules for the calculations of probability. These rules have many applications.

Theorem 6.1 If ϕ is the impossible event, then $P(\phi) = 0$.

Proof. The sure event S and the impossible event \emptyset are mutually exclusive and their union is S .

$$S \cup \emptyset = S$$

Then $P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset)$

Subtracting $P(S)$ from both sides, we get

$$P(\emptyset) = 0$$

Thus the probability of the impossible event is zero. It is to be kept in mind that the converse of this rule is not generally true. Moreover, a theorem is a statement derived either from axioms or previously proved theorems.

Theorem 6.2 Law of Complementation. If \bar{A} is the complement of an event A relative to the sample space S , then

$$P(\bar{A}) = 1 - P(A).$$

Proof. Since the events A and \bar{A} are mutually exclusive and collectively exhaustive, they together make up the entire sample space S , therefore, we have

$$A \cup \bar{A} = S$$

Thus $P(A \cup \bar{A}) = P(S)$

or $P(A) + P(\bar{A}) = 1$ [since $P(S) = 1$ by axiom (ii)]

or $P(\bar{A}) = 1 - P(A)$.

Hence the probability of the complement of an event is equal to one minus the probability of the event. Complementary probabilities are very useful when the question asks for the probability of "at least one".

* **Example 6.10** A coin is tossed 4 times in succession. What is the probability that at least one head occurs?

The sample space S for this experiment consists of $2^4 = 16$ sample points, as each toss can result in 2 outcomes, and we assume that each outcome is equally likely.

Let A represent the event that at least one head occurs. Then A consists of many sample points. On the other hand, \bar{A} is the event that no head occurs and \bar{A} has the single sample point {TTTT}. Thus

$$P(\bar{A}) = \frac{1}{16}.$$

Hence by the law of complementation, we have

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{1}{16} = \frac{15}{16}$$

Example 6.11 A coin is biased so that the probability that it falls showing tails is $\frac{2}{3}$.

- (a) Find the probability of obtaining at least one head when the coin is tossed five times.
 (b) How many times must the coin be tossed so that the probability of obtaining at least one head is greater than 0.96?

Since $P(\text{a head appears}) = \frac{1}{4}$ and

$$P(\text{no head or tail appears}) = \frac{3}{4}$$

- (c) Let A be the event that at least one head is obtained when the coin is tossed 5 times, and \bar{A} is the event that no head is obtained. Thus by the law of complementation, we have

$$\begin{aligned} P(A) &= 1 - P(\bar{A}) \\ &= 1 - \left(\frac{3}{4}\right)^5 = 1 - 0.237 = 0.763 \end{aligned}$$

Let the coin be tossed n times to obtain the probability of at least one head greater than 0.98.

$$\text{Then } 1 - \left(\frac{3}{4}\right)^n \geq 0.98 \text{ i.e. } \left(\frac{3}{4}\right)^n \leq 0.02$$

Taking loga. we have

$$n \log\left(\frac{3}{4}\right) \leq \log 0.02$$

Dividing both sides by $\log\left(\frac{3}{4}\right)$ and reversing the inequality sign as $\log\left(\frac{3}{4}\right)$ is negative, we have

$$n \geq \frac{\log 0.02}{\log(3/4)} = \frac{-0.019}{-0.8751} = \frac{-1.6990}{-0.1249} \geq 13.6$$

$\therefore n = 14$.

The coin should be tossed 14 times so that the probability of obtaining at least one head is greater than 0.98.

Theorem 6.3 Probability of Subevent. If A and B are two events such that $A \subset B$, then $P(B) = P(A) + P(B \cap \bar{A})$

Proof. For $A \subset B$, the event B may be written as the union of two mutually exclusive events $B \cap A$ and $B \cap \bar{A}$.

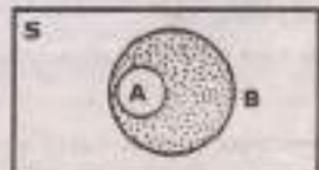
$$B = (B \cap A) \cup (B \cap \bar{A})$$

$$B \cap A = A \text{ as } B = A \cup (B \cap \bar{A})$$

$$\therefore P(B) = P(A) + P(B \cap \bar{A})$$

$$\therefore P(B \cap \bar{A}) \geq 0,$$

$$\therefore P(A) \leq P(B).$$



$B \cap \bar{A}$ is shaded

It is to be noted that an event such as $A \cap \bar{B}$ is called a joint event and probability associated with a joint event is called a joint probability.

Theorem 6.4 If A and B are any two events defined in a sample space S , then

$$P(A \cap \bar{B}) = P(A) - P(A \cap B).$$

Proof. The events $A \cap \bar{B}$ and $A \cap B$ are mutually exclusive and their union is A (see the Venn diagram). That is

$$A = (A \cap \bar{B}) \cup (A \cap B).$$

$$\therefore P(A) = P(A \cap \bar{B}) + P(A \cap B)$$

Hence $P(A \cap \bar{B}) = P(A) - P(A \cap B)$.



Theorem 6.5 Addition Law: If A and B are any two events defined in a sample space S , then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof. The event $A \cup B$ may be written as the union of two mutually exclusive events A and $\bar{B} \cap \bar{A}$ (see the diagram). That is

$$A \cup B = A \cup (\bar{B} \cap \bar{A})$$

Then $P(A \cup B) = P(A) + P(\bar{B} \cap \bar{A})$

Again the event B may also be decomposed into two mutually exclusive events as

$$B = (A \cap B) \cup (\bar{A} \cap B)$$

Thus $P(B) = P(A \cap B) + P(\bar{A} \cap B)$

Subtracting this result from the former, we get

$$P(A \cup B) - P(B) = P(A) - P(A \cap B)$$

Hence $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

This law, often called the *General Rule of Addition* for probabilities may be stated as below:

"If two events A and B are not mutually exclusive, then the probability that at least one occurs, is given by the sum of the separate probabilities of events A and B minus the probability of joint event $A \cap B$."

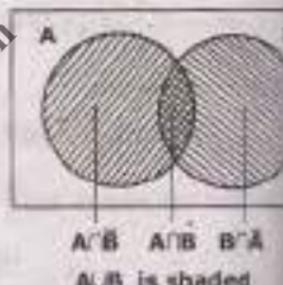
Corollary 1. If A and B are mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

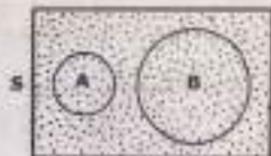
Proof: Since the events A and B are mutually exclusive, therefore

$$A \cap B = \emptyset \text{ and } P(A \cap B) = P(\emptyset) = 0$$

Hence $P(A \cup B) = P(A) + P(B)$, which is just a restatement of axiom (iii).



alternatively, Let n be the total number of sample points, let A and B be two events; A consisting of m_1 sample points and B of m_2 sample points. The occurrence of $A \cup B$ consists of all the sample points belonging to either A or B .



Since A and B are mutually exclusive events, therefore they have no sample points in common. Obviously, the number of points contained in $A \cup B$ is $m_1 + m_2$.

$$\text{Hence } P(A \cup B) = \frac{\text{number of sample points in } A \cup B}{\text{number of sample points in } S}$$

$$= \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B).$$

Corollary 1. If A_1, A_2, \dots, A_k are k mutually exclusive events, then the probability that one of them occurs is the sum of the probabilities of the separate events, i.e.

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k).$$

It is to be noted that if A_1, A_2, \dots, A_k are mutually exclusive and collectively exhaustive, then $P(A_1) + \dots + P(A_k) = 1$.

Corollary 2. If A and B are any two events, then

$$P(A \cup B) = P(A) + P(B)$$

The addition law for any two events A and B may be written as

$$P(A \cup B) + P(A \cap B) = P(A) + P(B);$$

$$P(A \cup B) \leq P(A) + P(B)$$

and, for any k events A_1, A_2, \dots, A_k , the relation

$$P(A_1 \cup A_2 \cup \dots \cup A_k) \leq P(A_1) + P(A_2) + \dots + P(A_k).$$

This result is known as Boole's inequality.

Example 6.12. If one card is selected at random from a deck of 52 playing cards, what is the probability that the card is a club or a face card or both?

Let A represent the event that the card selected is a club, B , the event that the card selected is a face card, $A \cap B$, the event that the card selected is both a club and a face card. Then we need $P(A \cup B)$.

Now $P(A) = \frac{13}{52}$, as there are 13 clubs.

$P(B) = \frac{12}{52}$, as there are 12 face cards.

and $P(A \cap B) = \frac{3}{52}$, since 3 of clubs are also face cards.

Therefore the desired probability is

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52} \end{aligned}$$

Example 6.13 An integer is chosen at random from the first 200 positive integers. What is the probability that the integer chosen is divisible by 6 or by 8?

The sample space for this experiment is

$$S = \{1, 2, 3, \dots, 199, 200\}, \text{ and therefore } n(S) = 200$$

Let A represent the event that the integer chosen is divisible by 6, B, the event that the integer chosen is divisible by 8, and $A \cap B$, the event that the integer chosen is divisible by both 6 and 8, i.e., 24.

Then we need $P(A \cup B)$.

$$\text{Now } n(A) = \left[\frac{200}{6} \right] = 33, \quad n(B) = \left[\frac{200}{8} \right] = 25,$$

$$n(A \cap B) = \left[\frac{200}{24} \right] = 8, \text{ where } [x] \text{ stands for the greatest integer in } x.$$

$$\begin{aligned} \text{Hence } P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{33}{200} + \frac{25}{200} - \frac{8}{200} = \frac{50}{200} = \frac{1}{4} \end{aligned}$$

Example 6.14 A pair of dice are thrown. Find the probability of getting a total of either 5 or 11.

The sample space has 36 outcomes when two dice are thrown. (see example 6.6 on page 118)

Let A be the event that a total of 5 occurs and B be the event that a total of 11 occurs. Then the events are

$$A = \{(1, 4), (2, 3), (3, 2), (4, 1)\}, \text{ and}$$

$$B = \{(5, 6), (6, 5)\}.$$

The events A and B are mutually exclusive as a total of 5 and 11 cannot both occur together. Therefore

$$P(A \cup B) = P(A) + P(B) = \frac{4}{36} + \frac{2}{36} = \frac{1}{6}.$$

Example 6.15 Three horses A, B and C are in a race; A is twice as likely to win as B and it is three times as likely to win as C. What is the probability that A or B wins?

Let $P(C) = p$ as the events are not equally likely.

Then $P(B) = 2p$, as B is twice as likely to win as C .

Similarly $P(A) = 2P(B) = 4p$.

Since A , B and C are mutually exclusive and collectively exhaustive, therefore the sum of their probabilities must be equal to 1. Thus,

$$p + 2p + 4p = 1 \quad \text{or} \quad p = \frac{1}{7}$$

$$P(A) = \frac{4}{7}, P(B) = \frac{2}{7} \quad \text{and} \quad P(C) = \frac{1}{7}$$

$$\text{Hence } P(A \cup B) = P(A) + P(B) = \frac{4}{7} + \frac{2}{7} = \frac{6}{7}.$$

Theorem 6.6: If A , B and C are any three events in a sample space Ω , then the probability of at least one of them occurring is given by

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C).$$

$$\text{Let us write } A \cup B \cup C = A \cup (B \cup C)$$

$$= A \cup D, \text{ where } D = B \cup C.$$

$$\begin{aligned} \text{Now } P(A \cup B \cup C) &= P(A \cup D) \\ &= P(A) + P(D) - P(A \cap D) \\ &= P(A) + P(B \cup C) - P(A \cap (B \cup C)) \\ &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap (B \cup C)) \end{aligned}$$

By distributive law of sets gives

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

Since the two sets are disjoint, their intersection is given by $A \cap B \cap C$.

$$\begin{aligned} P(A \cap (B \cup C)) &= P((A \cap B) \cup (A \cap C)) \\ &= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \end{aligned}$$

Substitution of this result gives

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

This result may be written as

$$P(A_1 \cup A_2 \cup A_3) = \sum_{i=1}^3 P(A_i) - \sum_{i,j} P(A_i \cap A_j) + P(A_1 \cap A_2 \cap A_3).$$

In general, the formula for k event is

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = \sum_i P(A_i) - \sum_{i,j} P(A_i \cap A_j) + \sum_{i,j,l} P(A_i \cap A_j \cap A_l) - \dots + (-1)^{k+1} P(A_1 \cap A_2 \cap \dots \cap A_k).$$

Example 6.16 A card is drawn at random from a deck of ordinary playing cards. What is the probability that it is a diamond, a face card or a king?

Let A represent the event that the card drawn is a diamond, B , the event that the card drawn is a face card, C , the event that the card drawn is a king, $A \cap B$, the event that the card drawn is both a diamond and face card, and so on. Then we need

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

Now $P(A) = \frac{n(A)}{n(S)} = \frac{13}{52}$, (there are 13 diamonds)

$$P(B) = \frac{n(B)}{n(S)} = \frac{12}{52}$$
, (there are 12 face cards)

$$P(C) = \frac{n(C)}{n(S)} = \frac{4}{52}$$
, (there are 4 kings)

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{3}{52}$$
, (diamonds and face card)

$$P(B \cap C) = \frac{n(B \cap C)}{n(S)} = \frac{4}{52}$$
, (face cards and king)

$$P(A \cap C) = \frac{n(A \cap C)}{n(S)} = \frac{1}{52}$$
, (diamond and king)

$$P(A \cap B \cap C) = \frac{n(A \cap B \cap C)}{n(S)} = \frac{1}{52}$$
, (diamond and face card and king)

Hence, we get

$$P(A \cup B \cup C) = \frac{13}{52} + \frac{12}{52} + \frac{4}{52} - \frac{3}{52} - \frac{4}{52} - \frac{1}{52} + \frac{1}{52} = \frac{22}{52} = 0.423$$

6.6 CONDITIONAL PROBABILITY

The sample space for an experiment must often be changed when some additional information pertaining to the outcome of the experiment is received. The effect of such information is to reduce the sample space by excluding some outcomes as being impossible which before receiving the information were believed possible. The probabilities associated with such a reduced sample space are called *conditional probabilities*. The following example illustrates the concept of conditional probability.

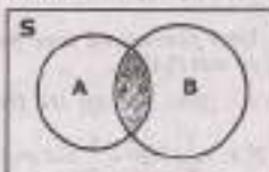
Let us consider the die-throwing experiment with sample space $S = \{1, 2, 3, 4, 5, 6\}$. Suppose we wish to know the probability of the outcome that the die shows 6, say event A . If before seeing the outcome, we are told that the die shows an even number of dots, say event B , then the information that the die shows an even number excludes the outcomes 1, 3 and 5, and thereby reduces the original sample space to a sample space that consists of 3 outcomes 2, 4 and 6, i.e. the reduced sample space is $B = \{2, 4, 6\}$. Then the desired probability in the reduced sample space B is $\frac{1}{3}$, since each outcome in the reduced

sample space is equally likely. We call the $\frac{1}{3}$ as the conditional probability of the event A because it is computed under the condition that the die has shown even number of dots. In other words,

$$P(\text{dice shows 6} | \text{dice shows even numbers}) = \frac{1}{3},$$

where the vertical line is read as given that and the information following the vertical line describes the remaining event. This is in fact the probability of getting a 6 in the reduced sample space B , and is denoted as $P(A|B)$. Thus we have the following:

$$\begin{aligned} P(A|B) &= \frac{\text{number of sample points in } A \cap B}{\text{number of sample points in } B} \\ &= \frac{n(A \cap B)}{n(B)} \end{aligned}$$



Dividing the numerator and the denominator by the number of sample points in original sample space $n(S)$, we get

$$\frac{n(A \cap B)}{n(S)} \cdot \frac{n(S)}{n(B)} = \frac{P(A \cap B)}{P(B)}$$

Reduced Sample Space B

This illustration leads us to make the following definition of conditional probability.

If A and B are two events in a sample space S and if $P(B)$ is not equal to zero, then the conditional probability of the event A given that event B has occurred, written as $P(A|B)$, is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

If $P(B) = 0$, the conditional probability $P(A|B)$ remains undefined.

Similarly, $P(B|A) = \frac{P(A \cap B)}{P(A)}$, where $P(A) > 0$.

It should be noted that $P(A|B)$ satisfies all the basic axioms of probability, namely

$$0 \leq P(A|B) \leq 1,$$

$$P(S|B) = 1 \quad [\because S \cap B = B \therefore P(S|B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1]$$

$$P(A_1 \cup A_2 | B) = P(A_1 | B) + P(A_2 | B), \text{ provided the events } A_1 \text{ and } A_2 \text{ are mutually exclusive}$$

Thus to determine the conditional probability $P(A|B)$, either we directly calculate the probability of A relative to the reduced sample space B or we use $P(A \cap B)$ and $P(B)$, the probabilities of the events in the original sample space S .

Example 6.17 Two coins are tossed. What is the conditional probability that two heads appear given that there is at least one head?

The sample space S for this experiment is

$$S = \{HH, HT, TH, TT\}$$

Let A represent the event that two heads appear, and B , the event that there is at least one head. Then we need $P(A|B)$.

Since $A = \{HH\}$, $B = \{HH, HT, TH\}$ and $A \cap B = \{HH\}$,

$$P(A) = \frac{1}{4}, P(B) = \frac{3}{4} \text{ and } P(A \cap B) = \frac{1}{4}$$

$$\text{Hence } P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Example 6.18 A man tosses two fair dice. What is the conditional probability that the sum of the two dice will be 7, given that (i) the sum is odd, (ii) the sum is greater than 6, (iii) the two dice have the same outcome?

The sample space S for this experiment consists of the following 36 equally likely outcomes:

$$\begin{aligned} S \sim & \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ & (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ & (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\ & (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ & (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\} \end{aligned}$$

Let $A = \{\text{the sum is 7}\}$, $B = \{\text{the sum is odd}\}$,

$C = \{\text{the sum is greater than 6}\}$, and

$D = \{\text{the two dice had the same outcome}\}$. Then

$$A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\},$$

$$B = \{(1, 2), (1, 4), (1, 6), (2, 1), (2, 3), (2, 5), (3, 4), \dots, (6, 5)\},$$

$$C = \{(1, 6), (2, 5), (2, 6), (3, 4), (3, 5), (3, 6), (4, 3), (4, 4), \dots, (6, 6)\},$$

$$D = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\},$$

$$A \cap B = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\},$$

$$A \cap C = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\},$$

$$A \cap D = \emptyset$$

$$\therefore P(A) = \frac{6}{36}, P(B) = \frac{18}{36}, P(C) = \frac{21}{36}, P(D) = \frac{6}{36}$$

$$P(A \cap B) = \frac{6}{36}, P(A \cap C) = \frac{6}{36} \text{ and } P(A \cap D) = 0$$

now using the definition of conditional probability, we get

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{6}{36} \times \frac{36}{18} = \frac{1}{3}$$

$$P(A/C) = \frac{P(A \cap C)}{P(C)} = \frac{6}{36} \times \frac{36}{21} = \frac{2}{7}$$

$$P(A/D) = \frac{P(A \cap D)}{P(D)} = 0 \times \frac{36}{6} = 0.$$

Example 6.19 What is the probability that a randomly selected poker hand, contains exactly 3 aces given that it contains at least 2 aces?

Let A represent the event that exactly 3 aces are selected and B , the event that at least 2 aces are selected. Then we need $P(A/B)$.

Now a poker hand consists of 5 cards, therefore the sample space S contains $\binom{52}{5} = 2,598,960$

$$n(A) = \binom{4}{3} \binom{48}{2} \text{ outcomes;}$$

$$n(B) = \binom{4}{2} \binom{48}{3} + \binom{4}{3} \binom{48}{2} + \binom{4}{4} \binom{48}{1}$$

at least 2 aces means 2 or 3 or 4 aces; and

$$n(A \cap B) = \binom{4}{3} \binom{48}{2} = A = B.$$

$$\therefore P(A \cap B) = \frac{\binom{4}{3} \binom{48}{2}}{\binom{52}{5}}, \text{ and}$$

$$P(B) = \frac{\binom{4}{2} \binom{48}{3} + \binom{4}{3} \binom{48}{2} + \binom{4}{4} \binom{48}{1}}{\binom{52}{5}}$$

Hence: $P(A \cap B) = \frac{P(A \cap B)}{P(B)}$

$$= \frac{\binom{4}{3} \binom{48}{2}}{\binom{4}{2} \binom{48}{3} + \binom{4}{3} \binom{48}{2} + \binom{4}{4} \binom{48}{1}}$$

$$= \frac{4,312}{108,336} = 0.0416.$$

Theorem 6.7 Multiplication Law. If A and B are any two events defined in a sample space, then

$$P(A \cap B) = P(A) P(B/A), \text{ provided } P(A) \neq 0,$$

$$= P(B) P(A/B), \text{ provided } P(B) \neq 0.$$

The conditional probability of B given that A has occurred is

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, \text{ where } P(A) \neq 0$$

Multiplying both sides by $P(A)$, we get

$$P(A \cap B) = P(A) P(B/A).$$

The second form is easily obtained by interchanging A and B .

This is called the *general rule of multiplication for probabilities*.

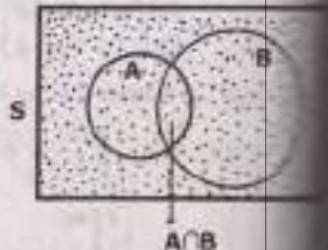
Alternative Proof. Let S represent a sample space of an experiment having n equally likely outcomes (sample points). Let m_1 be the number of sample points contained in A (including those common to B); m_2 be the number of sample points in B and m_3 be the number of sample points belonging both to A and B . Then (assuming $m_1 > 0, m_2 > 0$),

$$P(A \cap B) = \frac{m_3}{n}.$$

The fraction $\frac{m_3}{n}$ may be written as $\frac{m_3}{n} = \frac{m_3}{m_1} \cdot \frac{m_1}{n}$.

But $\frac{m_3}{m_1} = P(A)$ and $\frac{m_1}{n} = \text{conditional probability of } B, \text{ given that } A \text{ has occurred, i.e. } P(B/A)$.

Hence $P(A \cap B) = P(A) P(B/A)$.



Since the joint event $A \cap B$ involves A and B symmetrically, therefore interchanging A and B , we

$$P(A \cap B) = P(B) P(A|B).$$

The rule may be stated as below:

"The probability that two events A and B will both occur is equal to the probability that one of the events will occur multiplied by the conditional probability that the other event will occur given that the first event has occurred."

Generalization. This rule may be extended to several events. In case of three events A , B and C , we have

$$\begin{aligned} P(A \cap B \cap C) &= P(D \cap C), \text{ where } D = A \cap B \\ &= P(D) P(C|D) \\ &= P(A \cap B) P(C|A \cap B) \\ &= P(A) P(B|A) P(C|A \cap B); \end{aligned}$$

Similarly, for more than three events, the formula may be proved by mathematical induction.

Example 6.20. A box contains 15 items, 4 of which are defective and 11 are good. Two items are selected at random. What is the probability that the first is good and the second defective?

Let A represent the event that the first item selected is good and B , the event that the second item is defective.

Then we need to calculate the probability of the joint event $A \cap B$ by the rule $P(A \cap B) = P(A) P(B|A)$.

$$\text{Now } P(A) = \frac{11}{15}$$

Given the event A has occurred, there remains 14 items of which 4 are defective. Therefore the probability of selecting a defective after a good has been selected, i.e. $P(B|A) = \frac{4}{14}$.

$$\text{Hence } P(A \cap B) = P(A) P(B|A) = \frac{11}{15} \times \frac{4}{14} = \frac{44}{210} = 0.16.$$

Example 6.21. Two cards are dealt from a pack of ordinary playing cards. Find the probability that the second card dealt is a heart.

Let H_1 represent the event that the first card dealt is a heart, and H_2 , the event that the second card is a heart. Then

$P(\text{second card is a heart}) = P(\text{first card is a heart and second card is a heart}) + P(\text{first card is not a heart and second card is a heart}).$

$$P(H_2) = P(H_1 \cap H_2) + P(\bar{H}_1 \cap H_2)$$

$$= P(H_1) P(H_2|H_1) + P(\bar{H}_1) P(H_2|\bar{H}_1)$$

$$= \left(\frac{12}{52} \times \frac{12}{51} \right) + \left(\frac{39}{52} \times \frac{13}{51} \right)$$

$$= \frac{1}{17} + \frac{13}{68} = \frac{17}{68} = \frac{1}{4}$$

Example 6.22. Box A contains 5 green and 7 red balls. Box B contains 3 green, 3 red and 6 yellow balls. A box is selected at random and a ball is drawn at random from it. What is the probability that the ball drawn is green?

Let E represent the event that the green ball is drawn. Then E can occur in one of the following mutually exclusive ways:

- Box A is selected and a green ball is drawn, i.e. $A \cap E$, or
- Box B is selected and a green ball is drawn, i.e. $B \cap E$.

Therefore $P(\text{green ball}) = P(\text{box } A \text{ and green ball}) + P(\text{box } B \text{ and green ball})$

$$= P(\text{box } A) \cdot P(\text{green ball}/\text{box } A) + P(\text{box } B) \cdot P(\text{green ball}/\text{box } B)$$

In symbols, $P(E) = P(A \cap E) + P(B \cap E)$

$$= P(A) \cdot P(E/A) + P(B) \cdot P(E/B)$$

$$= \frac{1}{2} \cdot \frac{5}{12} + \frac{1}{2} \cdot \frac{3}{12} = \frac{1}{3}$$

Example 6.23. An urn contains 10 white and 3 black balls. Another urn contains 3 white and 2 black balls. Two balls are transferred from the first urn and placed in the second and then one ball is drawn from the latter. What is the probability that it is a white ball? (P.U., B.A./B.Sc.)

Let A represent the event that 2 balls are drawn from the first urn and transferred to the second. Then A can occur in the following three mutually exclusive ways:

- $$A_1 = 2 \text{ white balls}$$
- $$A_2 = 1 \text{ white ball and 1 black ball}$$
- $$A_3 = 2 \text{ black balls}$$

$$\text{Thus } P(A_1) = \binom{10}{2} + \binom{13}{2} = \frac{45}{78},$$

$$P(A_2) = \binom{10}{1} \binom{3}{1} + \binom{13}{2} = \frac{36}{78}, \text{ and}$$

$$P(A_3) = \binom{3}{2} + \binom{13}{2} = \frac{3}{78}.$$

The second urn after having transferred 2 balls from the first urn, contains

- 5 white and 5 black balls (2 white balls transferred)
- 4 white and 6 black balls (1 white and 1 black ball transferred)
- 3 white and 7 black balls (2 black balls transferred)

Let W' represent the event that a white ball is drawn from the second urn after having transferred 2 balls from the first urn. Then

$$P(W') = P(W' \cap A_1) + P(W' \cap A_2) + P(W' \cap A_3)$$

$$\text{Now } P(W' \cap A_1) = \frac{5}{10} \times \frac{45}{78} = \frac{15}{52}$$

$$P(W' \cap A_2) = \frac{4}{10} \times \frac{30}{78} = \frac{2}{13}, \text{ and}$$

$$P(W' \cap A_3) = \frac{3}{10} \times \frac{3}{78} = \frac{3}{260}$$

Hence the required probability is

$$P(W') = \frac{15}{52} + \frac{2}{13} + \frac{3}{260} = \frac{59}{130} = 0.4538$$

Example 6.24 A card is drawn at random from a deck of ordinary playing cards. What is the probability that it is a diamond, a face card or a king? (P.U., B.A./B.Sc. 1992)

There are two ways of solving this problem. One is given in Example 6.16. A second approach is given below:

Let A = the card drawn is a diamond,

B = the card drawn is a face card, and

C = the card drawn is a king.

So we need

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

$$P(A) = \frac{13}{52}, P(B) = \frac{12}{52}, P(C) = \frac{4}{52},$$

$$P(A \cap B) = P(A)P(B/A) = \frac{13}{52} \times \frac{3}{13} = \frac{3}{52},$$

$$P(B \cap C) = P(B)P(C/B) = \frac{12}{52} \times \frac{4}{12} = \frac{4}{52},$$

$$P(A \cap C) = P(A)P(C/A) = \frac{13}{52} \times \frac{1}{13} = \frac{1}{52}, \quad \left(\text{or } P(C \cap A) = \frac{4}{52} \times \frac{1}{4}\right)$$

$$P(A \cap B \cap C) = P(A)P(B/A)P(C/A \cap B)$$

$$= \frac{13}{52} \times \frac{3}{13} \times \frac{1}{3} = \frac{1}{52}$$

So we get

$$\begin{aligned} P(A \cup B \cup C) &= \frac{13}{52} + \frac{12}{52} + \frac{4}{52} - \frac{3}{52} - \frac{4}{52} - \frac{1}{52} + \frac{1}{52} \\ &= \frac{22}{52} = 0.423 \end{aligned}$$

Example 6.25 Three urns of the same appearance are given as follows:

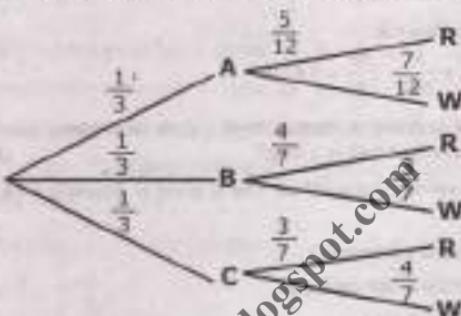
Urn A contains 5 red and 7 white balls.

Urn B contains 4 red and 3 white balls.

Urn C contains 3 red and 4 white balls.

An urn is selected at random and a ball is drawn from the urn.

- What is the probability that the ball drawn is red?
- If the ball drawn is red, what is probability that it came from urn A?



Here we first select one of the three urns, and then we draw a ball which is either red (R) or white (W). In other words, we perform a sequence of two experiments. This process is described by a probability tree diagram, (see page 28), in which each branch of the tree gives the respective probability.

Now the probability of selecting urn A, for instance, and then a red ball (R) is $\frac{1}{3} \cdot \frac{5}{12} = \frac{5}{36}$, and the probability that any particular path of the tree occurs is, by the multiplication law, the product of the probability of each branch in the path.

- Now the probability of drawing a red ball is given by the relation

$$P(R) = P(A)P(R/A) + P(B)P(R/B) + P(C)P(R/C)$$

as there are three mutually exclusive paths leading to the drawing of a red ball.

$$\begin{aligned}\text{Hence } P(R) &= \frac{1}{3} \cdot \frac{5}{12} + \frac{1}{3} \cdot \frac{4}{7} + \frac{1}{3} \cdot \frac{3}{7} \\ &= \frac{119}{252} = 0.4723\end{aligned}$$

- Here we need the probability that urn A is selected, given that the ball drawn is red.

$$\text{By definition, } P(A/R) = \frac{P(A \cap R)}{P(R)}$$

But $P(A \cap R) = \text{Probability that urn } A \text{ is selected and a red ball is drawn}$
 $= \frac{1}{3} \times \frac{5}{12} = \frac{5}{36}$

Hence $P(A|R) = \frac{5/36}{119/252}$
 $= \frac{35}{119} = 0.294$

6.7 INDEPENDENT AND DEPENDENT EVENTS

Two events A and B in the same sample space S , are defined to be *independent* (or *statistically independent*) if the probability that one event occurs, is not affected by whether the other event has or has not occurred, that is

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B).$$

This follows that two events A and B are independent if and only if

$$P(A \cap B) = P(A) P(B)$$

The events A and B are defined to be *dependent* if $P(A \cap B) \neq P(A) P(B)$. This means that the occurrence of one of the events in some way affects the probability of the occurrence of the other event.

It is to be emphasized that two mutually exclusive events A and B are independent if and only if $P(B) = 0$, which is true when either $P(A) = 0$ or $P(B) = 0$. If both events A and B have non-zero probabilities, they must have a sample point in common. Thus two events that are independent, can never be mutually exclusive. Moreover, two events that are mutually exclusive, are also dependent events, and two events that are non-mutually exclusive, may either be independent or dependent events.

Three events A , B and C , all defined on the same sample space, are said to be *mutually independent* if they satisfy the following conditions:

- i) They are pairwise independent, i.e. $P(A \cap B) = P(A) P(B)$.

$$P(A \cap C) = P(A) P(C); P(B \cap C) = P(B) P(C).$$

- ii) They are mutually independent, i.e.

$$P(A \cap B \cap C) = P(A) P(B) P(C).$$

In general, the k events A_1, A_2, \dots, A_k are defined to be *mutually independent* if and only if the probability of the intersection of any 2, 3, ..., or k of them equals the product of their respective probabilities.

Example 6.26 Two events A and B are such that $P(A) = \frac{1}{4}$, $P(A|B) = \frac{1}{2}$, and $P(B|A) = \frac{2}{3}$.

- i) Are A and B independent events?
- ii) Are A and B mutually exclusive events?
- iii) Find $P(A \cap B)$ and $P(B)$.

- i) If A and B are independent events, then $P(A \cap B) = P(A)P(B)$

$$\text{Now } P(A) = \frac{1}{4} \text{ and } P(A \cap B) = \frac{1}{2} \text{ i.e. } P(A \cap B) \neq P(A)$$

Hence A and B are not independent events.

- ii) If A and B are mutually exclusive events, then $P(A \cap B) = 0$.

$$\text{But it is given that } P(A \cap B) = \frac{1}{2}$$

Hence A and B are not mutually exclusive events.

- iii) Now $P(A \cap B) = P(A)P(B|A)$

$$= \frac{1}{4} \times \frac{2}{3} = \frac{1}{6}$$

By definition, we have

$$P(B)P(A|B) = P(A)P(B|A)$$

$$\text{or } P(B)\left(\frac{1}{2}\right) = \left(\frac{1}{4}\right)\left(\frac{2}{3}\right) \text{ so that } P(B) = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3}$$

Example 6.27 Two fair dice, one red and one green, are thrown. Let A denote the event that red die shows an even number and B , the event that the green die shows a 5 or a 6. Show that the events A and B are independent.

The sample space S contains 36 equally likely outcomes when two dice are thrown.

Given A = event that red die shows an even number.

B = event that green die shows a 5 or a 6, and therefore

$A \cap B$ = event that red die shows an even number and green die shows a 5 or a 6.

Then A contains 18 outcomes, B contains 12 and $A \cap B$ contains only 6 outcomes.

Associating with each outcome a probability of $\frac{1}{36}$, we get

$$P(A) = \frac{18}{36} = \frac{1}{2}, P(B) = \frac{12}{36} = \frac{1}{3} \text{ and } P(A \cap B) = \frac{6}{36} = \frac{1}{6}$$

Since $P(A \cap B) = \frac{1}{6} = \frac{1}{2} \times \frac{1}{3} = P(A)P(B)$, therefore the events A and B are independent.

Alternatively,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/3} = \frac{1}{2} = P(A), \text{ and}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{1/2} = \frac{1}{3} = P(B).$$

Hence the events A and B are independent.

Example 6.28 Let A be the event that a family has children of both sexes and B be the event that a family has at most one boy. If a family is known to have (i) three children, then show that A and B are independent events, (ii) five children, then show that A and B are dependent events.

(P.U., B.A. (Hons.) Part-I, 1970)

Let b denote a boy and g a girl. Then

- (i) the equiprobable sample space S would be

$$S = \{bbb, bbg, bgb, gbb, bgg, gbg, ggb, ggg\}$$

The two events are

- $A = \{\text{children of both sexes}\},$
= $\{bbg, bgb, gbb, bgg, gbg, ggb\}$, and
- $B = \{\text{at most one boy}\},$
= $\{bgg, gbg, ggb, ggg\}$

The event $A \cap B$ is

$$A \cap B = \{bgg, gbg, ggb\}$$

Thus their respective probabilities are

$$P(A) = \frac{3}{8} = \frac{3}{4}, \quad P(B) = \frac{4}{8} = \frac{1}{2}, \quad \text{and} \quad P(A \cap B) = \frac{3}{8}$$

$$\Rightarrow P(A) \cdot P(B) = \frac{3}{4} \times \frac{1}{2} = \frac{3}{8} = P(A \cap B).$$

Hence A and B are independent.

- the sample space S may be represented by the following 16 equally likely outcomes:

$$S = \{bbbb, bbhg, bhbg, bhhg, hbpg, hgbg, ghbg, ghhg, hhgb, ggbb, gggb, gggg, ghgg, gghg, gggh, gggh\}$$

The events are:

- $A = \{bbhg, hhgb, hgbb, ghhg, hhgb, hgbg, ghbg, ghhg, hhgb, ggbb, gggb, gggg, ghgg, gghg, gggh, gggh\}$
- $B = \{gggg, ghgg, gghg, gggh, gggb, gggh\}$, and
- $A \cap B = \{gggg, ghgg, gghg, gggh\}$

Their probabilities are

$$P(A) = \frac{14}{16} = \frac{7}{8}, \quad P(B) = \frac{5}{16} \quad \text{and} \quad P(A \cap B) = \frac{4}{16} = \frac{1}{4}.$$

$$\Rightarrow P(A) \cdot P(B) = \frac{7}{8} \cdot \frac{5}{16} \neq P(A \cap B).$$

Hence A and B are dependent events.

Theorem 6.8 If A and B are two independent events, then

$$P(A \cap B) = P(A) \cdot P(B).$$

Proof: Since A and B are independent events, therefore

$$P(A) = P(A \cap B) \text{ and } P(B) = P(B \cap A)$$

Substituting these results in the general rule of multiplication, we obtain

$$P(A \cap B) = P(A) \cdot P(B).$$

Theorem 6.9 If A and B are two independent events in a sample space S , then (i) A and \bar{B} are independent, (ii) \bar{A} and B are independent, and (iii) \bar{A} and \bar{B} are independent.

Proof. (i) The events $A \cap B$ and $A \cap \bar{B}$ are mutually exclusive and their union is A .
 $A = (A \cap B) \cup (A \cap \bar{B})$.

Therefore $P(A) = P(A \cap B) + P(A \cap \bar{B})$

$$\begin{aligned} \text{or } P(A \cap \bar{B}) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \quad [\because A \text{ and } B \text{ are independent}] \\ &= P(A)[1 - P(B)] = P(A)P(\bar{B}) \end{aligned}$$

Hence A and \bar{B} are independent.

(ii) Similarly,

$$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap \bar{A}) \\ \text{or } P(B \cap \bar{A}) &= P(B) - P(B \cap A) \\ &= P(B) - P(B)P(A) \quad [\because A \text{ and } B \text{ are independent}] \\ &= P(B)[1 - P(A)] = P(B)P(\bar{A}) \end{aligned}$$

Therefore \bar{A} and B are independent.

(iii) Using De Morgan's law, $\bar{A} \cap \bar{B} = \overline{A \cup B}$, we have

$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= P(\overline{A \cup B}) \\ &= 1 - P(A \cup B) \\ &= 1 - P(A) - P(B) + P(A \cap B) \\ &= 1 - P(A) - P(B) + P(A)P(B) \\ &= [1 - P(A)][1 - P(B)] = P(\bar{A})P(\bar{B}) \end{aligned}$$

which shows that \bar{A} and \bar{B} are independent.

Example 6.29 Two cards are drawn from a well-shuffled ordinary deck of 52 cards. Find the probability that they are both aces if the first card is (i) replaced, (ii) not replaced.

(P.U., B.A./B.Sc., 1967; M.A. Econ. 1969)

Let A denote the event *ace* on first draw and B denote the event *ace* on the second draw.

In case of replacement, event A and B are independent.

$$\text{Thus } P(\text{both cards are aces}) = P(A \cap B) = P(A)P(B)$$

$$= \frac{4}{52} \times \frac{4}{52} = \frac{1}{169}$$

If the first card is not replaced, then A and B are dependent events and therefore

$P(\text{both cards are aces}) = P(\text{first card is an ace}) \times P(\text{second card is an ace given that the first card is an ace})$

$$\text{i.e. } P(A \cap B) = P(A)P(B/A) = \frac{4}{52} \times \frac{3}{51} = \frac{1}{221}$$

Example 6.30 A pair of fair dice is thrown twice. What is the probability of getting totals of 5 and (P.U., B.A./B.Sc. 1978)

Let A represent the event of getting a total of 5 and B the event of getting a total of 11, when two dice are thrown.

Then A can occur in the following two ways:

$$A_1 = \{\text{a total of 5 occurs on the first throw}\},$$

$$A_2 = \{\text{a total of 5 occurs on the second throw}\},$$

and B can occur in the following two ways:

$$B_1 = \{\text{a total of 11 occurs on the first throw}\},$$

$$B_2 = \{\text{a total of 11 occurs on the second throw}\}.$$

The joint event $A \cap B$ can occur in two mutually exclusive ways $A_1 \cap B_2$ or $B_1 \cap A_2$, i.e. $A \cap B = (A_1 \cap B_2) \cup (B_1 \cap A_2)$.

$$P(A \cap B) = P(A_1 \cap B_2) + P(B_1 \cap A_2)$$

$$= P(A_1)P(B_2) + P(B_1)P(A_2)$$

(\because events are independent)

$$= \frac{4}{36} \times \frac{2}{36} + \frac{2}{36} \times \frac{4}{36} = \frac{1}{162} + \frac{1}{162} = \frac{1}{81}$$

Example 6.31 The probability that a man will be alive in 25 years is $\frac{3}{5}$, and the probability that his wife will be alive in 25 years is $\frac{2}{3}$. Find the probability that (i) both will be alive, (ii) only the man will be alive, (iii) only the wife will be alive, (iv) at least one will be alive and (v) neither will be alive in 25 years.

(P.U., B.A./B.Sc. 1978)

Let A be the event that the man will be alive and B be the event that his wife will be alive in 25 years. Then

$$P(A) = \frac{3}{5}, \quad \text{and} \quad P(B) = \frac{2}{3}.$$

- i) We need the probability that both will be alive, i.e. $P(A \cap B)$.

Since A and B are independent, therefore

$$P(A \cap B) = P(A) \cdot P(B) = \frac{3}{5} \times \frac{2}{3} = \frac{2}{5}.$$

- ii) We need the probability that only the man will be alive, i.e. $P(A \cap \bar{B})$.

Since A and \bar{B} are independent and $P(\bar{B}) = 1 - P(B)$, therefore

$$P(A \cap \bar{B}) = P(A) \cdot P(\bar{B}) = \frac{3}{5} \times \left(1 - \frac{2}{3}\right) = \frac{1}{5}.$$

- iii) We require the probability that only the wife will be alive, i.e. $P(\bar{A} \cap B)$. Thus

$$P(\bar{A} \cap B) = P(\bar{A}) \cdot P(B) = \frac{2}{5} \times \frac{2}{3} = \frac{4}{15} \quad \text{as the events } \bar{A} \text{ and } B \text{ are independent}$$

$$P(\bar{A}) = 1 - P(A).$$

- iv) We require the probability that at least one will be alive, i.e. $P(A \cup B)$.

Since the events A and B are independent and not mutually exclusive, therefore

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{3}{5} + \frac{2}{3} - \frac{2}{5} = \frac{13}{15}. \end{aligned}$$

- v) We need the probability that neither will be alive, i.e. $P(\bar{A} \cap \bar{B})$.

Since \bar{A} and \bar{B} are independent, therefore

$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= P(\bar{A}) \cdot P(\bar{B}) \\ &= [1 - P(A)] [1 - P(B)] \\ &= \frac{2}{5} \times \frac{1}{3} = \frac{2}{15}. \end{aligned}$$

Theorem 6.10. Independent Repeated Trials with two Outcomes. If the probability of an event A occurring in a single trial is p , then the probability of its occurring k times in n independent trials is given by $P(A_k) = \binom{n}{k} p^k q^{n-k}$, where $q = 1 - p$.

Proof: If the event A occurs k times in n independent trials, then the event \bar{A} (not- A) will occur in the remaining $(n-k)$ trials, that is

$$\underbrace{AAA \dots A}_{k \text{ times}} \quad \underbrace{\bar{A} \bar{A} \bar{A} \dots \bar{A}}_{n-k \text{ times}}$$

The probability of this sequence would be

$$\underbrace{PPP \dots P}_{k \text{ times}} \quad \underbrace{QQQ \dots Q}_{n-k \text{ times}} \\ = p^k q^{n-k}.$$

But this is one sequence in which the event A occurs k times and fails to occur $(n-k)$ times. The number of possible sequences is $\binom{n}{k}$.

Hence the required probability is

$$P(A_k) = \binom{n}{k} p^k q^{n-k}$$

It is interesting to note that this is a special case of a very general result, called the *binomial law*, which is discussed in a bit detail in chapter 8.

Example 6.32. Five coins are tossed once (or one coin is tossed five times). What is the probability of getting precisely 3 heads?

We know that the probability of getting a head is $\frac{1}{2}$,

$$P(3 \text{ heads on a single toss of 5 coins}) = \binom{3}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{5-3} \\ = 10 \times \left(\frac{1}{2}\right)^5 = \frac{5}{16}.$$

Example 6.33. If 60 percent of the voters in the City of Lahore prefer candidate X , what is the probability that in a sample of 12 voters exactly 7 will prefer X ?

Here: $n = 12$, $k = 7$, $p = 0.60$ and $q = 0.40$

$$P(7 \text{ out of 12 prefer } X) = \binom{12}{7} (0.60)^7 (0.40)^{12-7} \\ = (792) (0.02799) (0.01024) = 0.227$$

Theorem 6.11 Bayes' theorem. If the events A_1, A_2, \dots, A_k form a partition of sample space S , that is, the events A_i are mutually exclusive and their union is S , and if B is any other event of S such that it can occur only if one of the A_i occurs, then for any i ,

$$P(A_i / B) = \frac{P(A_i)P(B / A_i)}{\sum_{j=1}^k P(A_j)P(B / A_j)}, \text{ for } i = 1, 2, \dots, k.$$

Proof: By the multiplicative law of probabilities, we have

$$\begin{aligned} P(B \cap A_i) &= P(B)P(A_i / B) \\ &= P(A_i)P(B / A_i). \end{aligned}$$

Equating the equivalent relations of $P(B \cap A_i)$ and solving for $P(A_i / B)$, we get

$$P(A_i / B) = \frac{P(A_i)P(B / A_i)}{P(B)}.$$

We may write the event B as $B = S \cap B$ (see the Venn diagram)

$$\begin{aligned} &= (A_1 \cup A_2 \cup \dots \cup A_k) \cap B \\ &= (A_1 \cap B) \cup (A_2 \cap B) \cup \dots \cup (A_k \cap B), \end{aligned}$$

where the $A_i \cap B$ are also mutually exclusive.

$$\text{Therefore } P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \dots + P(A_k \cap B).$$

Using the multiplicative law of probabilities, we may express each term $P(A_i \cap B) = P(A_i)P(B / A_i)$. Then

$$\begin{aligned} P(B) &= P(A_1)P(B / A_1) + P(A_2)P(B / A_2) + \dots + P(A_k)P(B / A_k) \\ &= \sum_{i=1}^k P(A_i)P(B / A_i). \end{aligned}$$



B is shaded

This result is generally known as the *theorem on total probability*. Replacing $P(B)$ by the probability formula for the event B , we obtain *Bayes' formula* as

$$P(A_i / B) = \frac{P(A_i)P(B / A_i)}{\sum_{j=1}^k P(A_j)P(B / A_j)}.$$

This result is known as Bayes' theorem after an English clergymen Thomas Bayes (1702–1761) who derived it and first used in a paper that was published posthumously in 1763. It should be noted that the original probabilities $P(A_i)$ are known as the *a priori* probabilities and the conditional probabilities

are called the *a posteriori* or inverse probabilities because probabilities are revised after some information has been obtained. Bayes' formula is also called the formula for probabilities of *events* on account of the reason that the events A_1, A_2, A_3 may be thought of as hypotheses to account for occurrence of the event B .

Example 6.34 In a bolt factory, machines A , B and C manufacture 25, 35 and 40 percent of the output, respectively. Of their outputs, 5, 4, 2 percent, respectively, are defective bolts. A bolt is selected at random and found to be defective. What is the probability that the bolt came from machine A ?

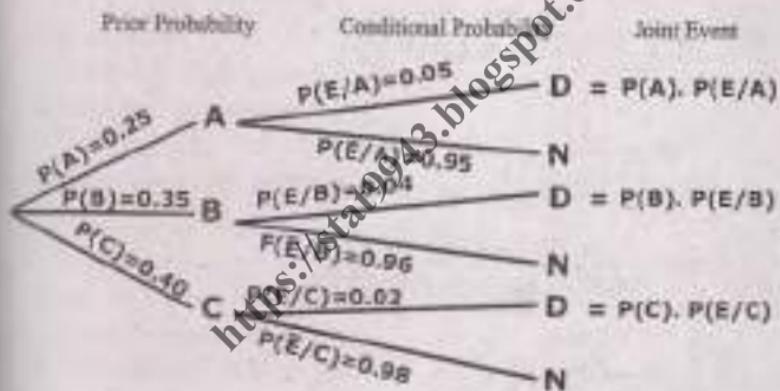
The *a priori* probabilities (before the information that the bolt is defective) are $P(A) = 0.25$, $P(B) = 0.35$, and $P(C) = 0.40$.

Let E represent the event that a bolt is defective (D).

Then the conditional probabilities are

$$P(E/A) = 0.05, P(E/B) = 0.04, \text{ and } P(E/C) = 0.02.$$

The outcomes with their respective probabilities may be shown by tree diagram as below:



$P(E)$ is the *a posteriori* probability that the selected defective bolt came from machine A . By Bayes' theorem, we get

$$\begin{aligned}
 P(E) &= \frac{P(A) \cdot P(E/A)}{P(A) \cdot P(E/A) + P(B) \cdot P(E/B) + P(C) \cdot P(E/C)} \\
 &= \frac{(0.25)(0.05)}{(0.25)(0.05) + (0.35)(0.04) + (0.40)(0.02)} \\
 &= \frac{0.0125}{0.0545} = 0.362
 \end{aligned}$$

Similarly, the posterior probabilities of *measures B* and *C* are

$$P(B/E) = 0.400, \text{ and } P(C/E) = 0.232$$

Example 4.35. An urn contains four balls which are known to be either (i) all white or (ii) one white and two black. A ball is drawn at random and is found to be white. What is the probability that the balls are white? (P.U., B.A./B.Sc. (Hons.) Part-III, 1996)

Let A_1 be the hypothesis that all the balls are white and A_2 be the hypothesis that two are white and two black. Then the *a priori* probabilities must be

$$P(A_1) = P(A_2) = \frac{1}{2}, \text{ as the selection of a hypothesis is random.}$$

Again let B the event that the ball drawn is white. Then the conditional probabilities are

$$P(B/A_1) = \frac{4C_1}{4C_1} = 1 \quad \text{and} \quad P(B/A_2) = \frac{2C_1}{4C_1} = \frac{1}{2}$$

Therefore by Bayes' theorem, we get the *a posteriori* probabilities

$$P(A_1/B) = \frac{P(A_1)P(B/A_1)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2)}$$

$$= \frac{\left(\frac{1}{2}\right)(1)}{\left(\frac{1}{2}\right)(1) + \frac{1}{2}\left[\frac{1}{2}\right]} = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{4}} = \frac{2}{3},$$

$$P(A_2/B) = \frac{P(A_2)P(B/A_2)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2)}$$

$$= \frac{\frac{1}{2}\left(\frac{1}{2}\right)}{\frac{1}{2}\left(\frac{1}{2}\right) + \frac{1}{2}\left(\frac{1}{3}\right)} = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}.$$

Hence the first hypothesis, i.e. all the balls are white, is preferred as it has larger probability.

EXERCISES

OBJECTIVE

Answer 'True' or 'False'. If the statement is not true then replace the underlined words with words to make the statement true.

- (i) The probability of an event is a whole number.

If two events are mutually exclusive, they are also independent.

If A and B are mutually exclusive events, the sum of their probabilities equal to one.

The sample points of a sample space are equally likely events.

If the sets of sample points belonging to two different events do not intersect, the events are independent.

Probability of head on tossing a coin is $\frac{1}{2}$.

If $P(A|B) = P(A)$ and $P(B|A) = P(B)$ then the two events A and B are independent.

It is always true that $P(A) = 1 - P(\bar{A})$.

The probabilities of complementary events always are equal.

If events A and B are statistically independent then $P(A \cap B) = P(A) \cdot P(B)$.

MULTIPLE CHOICE QUESTIONS

A simple event is

- a collection of exactly two outcomes.
- does not include any outcome.
 includes one and only one outcome.- includes more than one events.

A compound event includes

- at least four outcomes.
- one and only one outcome.
 at least two outcomes.- all the outcomes of an experiment.

The probability of an event is always

- greater than zero
- less than 1
 in the range zero to 1- greater than 1

The classical probability method is applied to an experiment that

- cannot be repeated.
 has equally likely outcomes.- has all independent outcomes.
- does not have more than two outcomes.

- v) The relative frequency method is applied to an experiment that
- a) does not have equally likely outcomes but can be repeated.
 - b) does not have equally likely outcomes and cannot be repeated.
 - c) has equally likely outcomes and cannot be repeated.
 - d) has all independent outcomes.
- vi) Which of the following values cannot be the probability of an event?
- a) .82
 - b) 0
 - c) 1.78
 - d) 0.36
- vii) In a group of 400 families, 300 own houses. If one family is randomly selected from the group the probability that this family owns a house is:
- a) .75
 - b) .25
 - c) .80
 - d) .40
- viii) Two mutually exclusive events
- a) always occur together.
 - b) cannot occur together.
 - c) can sometimes occur together.
 - d) can never occur together.
- ix) The two events A and B are mutually exclusive. Which one of the following statements can be true?
- a) $P(A \cap B) = 0$.
 - b) $P(A \cap B) = 1$.
 - c) $P(A \cup B) = 0$.
 - d) $P(A \cup B) = 1$.
- x) $P(A) = 0.6$ and $P(B) = 0.5$. Which of the following statements is true?
- a) A and B are mutually exclusive.
 - b) A and B are not mutually exclusive.
 - c) A and B are independent.
 - d) A and B are dependent.

The conditional probability of event A given that the event B has already occurred is written as:

- $P(A \cup B)$
- $P(B|A)$
- $P(A \cap B)$
- $P(A|B)$

Two complementary events:

- have no common outcomes
- have common outcomes
- contain the same outcomes
- can have common outcomes

The union of two events A and B is written as:

- $(A \cup B)$
- $(A \text{ and } B)$
- (B/A)
- (A/B)

The intersection of two events A and B is written as:

- $(A \text{ or } B)$
- $(A \text{ and } B)$
- (AB)
- (A/B)

The joint probability of two independent events A and B is:

- $P(A) + P(B)$
- $P(A) - P(B) = P(A \cap B)$
- $P(A)P(B)$
- $P(A)P(A/B)$

SUMMATIVE

1. List all the proper subsets of the universal set.

$$S = \{\text{chair, student, pen}\}$$

2. Construct a Venn diagram to illustrate the following subsets of

$$S = \{\text{ball, pen, table, coin, die, card, book}\},$$

$$A = \{\text{ball, pen, book}\}, B = \{\text{pen, table, coin}\}, C = \{\text{card}\}.$$

- 6.3 a) Let $A = \{1, 2\}$, $B = \{2, 3\}$ and $C = \{3\}$ be subsets of the universal set $S = \{1, 2, 3, 4, 5\}$. Determine the elements of the following sets:
 i) $A \times A$, ii) $A \times B$, iii) $B \times A$,
 iv) $(A \times B) \cup (B \times C)$, v) $(A \times B) \cap (B \times C)$
- b) Let $A = \{2, 3\}$, $B = \{1, 3, 5\}$ and $C = \{4, 5\}$. Construct the "tree diagram" of $A \times B \times C$.
- 6.4 a) Explain what is meant by a Random Experiment, a Sample Space and an Event.
 b) Let A, B and C be events (subsets) in a sample space S defined by
 $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
 $A = \{2, 3, 4\}$, $B = \{3, 4, 5\}$, $C = \{5, 6, 7\}$
- List the members of the following events:
- $\overline{A} \cap B$, ii) $\overline{A} \cup B$, iii) $\overline{A} \cap \overline{B}$, iv) $\overline{A} \cap (\overline{B} \cup C)$
- 6.5 If $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $A = \{0, 2, 4, 6, 8\}$,
 $B = \{1, 3, 5, 7, 9\}$, $C = \{2, 3, 4, 5\}$ and $D = \{1, 6, 7\}$, list the elements in the following sets:
 i) $A \cup C$, ii) $A \cap B$, iii) \overline{C} , iv) $(\overline{C} \cap D) \cup B$,
 v) $(\overline{S} \cap \overline{C})$, vi) $A \cap C \cap \overline{D}$
- 6.6 a) A pair of dice is rolled. List the elements of the sample space S . Let A denote the event "the sum is less than 5" and B the event "a 6 occurs on either die". List the elements corresponding to event A and to event B .
- b) Two dice are rolled. Let A be the event that the sum of dots on the faces shown is 8 and B the event that there is at least one 3 shown. Describe $A \cup B$; $A \cap B$; $A - B$; $(A \cap \overline{B}) \cup \overline{A}$.
- 6.7 a) Enumerate all the (i) combinations and (ii) permutations of 3 letters chosen from the four letters A, B, C and D.
 b) A club consists of 15 members. In how many ways can:
 i) the three officers: president, vice-president, and secretary-treasurer, be chosen?
 ii) a committee of three members be selected?
- 6.8 How many different bridge hands can be selected from an ordinary deck of 52 playing cards?
- 6.9 We have $n = 16$ persons and we wish to divide them at random into 5 groups consisting of 3 persons and 2 persons respectively. In how many ways is this possible?
- 6.10 a) Define the terms: Experiment, Outcome, Event, Sample Space, Simple and Compound Events, Mutually Exclusive Events.
 b) Show that a sample space consisting of 4 elements has 2^4 different events.
 c) Prove that $\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{n} = 2^n$,
 using the fact that a sample space with n elements has 2^n subsets.

- 6.11 Distinguish among classical or *a priori* probability, relative frequency or a posterior probability, axiomatic probability and subjective or personalistic probability. What is the disadvantage of each? Why do we study probability theory?
- 6.12 Explain what is wrong with each of the following statements.
- An investment counselor claims that the probability that a stock's price will go up is 0.60, remain unchanged is 0.38, or go down is 0.25.
 - If two coins are tossed, there are 3 possible outcomes: 2 heads, one head and one tail, and 2 tails. Hence probability of each of these outcomes is $\frac{1}{3}$.
 - The probabilities that a certain truck driver would have no, one and two or more accidents during the year are 0.90, 0.02, and 0.08.
 - $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$, $P(C) = \frac{1}{6}$ for the probabilities of three mutually exclusive events A, B and C.

Find the probability for each of the following events:

- An odd number appears in a single toss of a fair die.
- The sum 8 appears in a single toss of a pair of fair dice.
- At least one head appears in three tosses of a fair coin.
- A king, ace, jack of clubs or queen of diamonds appears in drawing a single card from a well-shuffled ordinary deck of 52 cards. (P.U., B.A./B.Sc. 1970)
- Describe the classical, relative frequency and subjective concepts of probability.
- A marble is drawn at random from a box containing 10 red, 10 white, 20 blue and 12 orange marbles. Find the probability that it is (i) orange or red, (ii) not 'red or blue', (iii) not blue, (iv) white, (v) red, white or blue. (P.U., B.A./B.Sc. 1970)
- If two dice are thrown, what are the various total number of dots that may turn up? What are the probabilities of each of them? What is the probability that the number of dots will total at least four?
- Show that in a single throw of two dice, the probability of throwing more than 7 is equal to that of throwing less than 7, and hence find the probability of throwing exactly 7. State clearly what assumptions you are making. (P.U., B.A./B.Sc. 1983)
- Two dice are thrown. Let A be the event that the sum of the upper face numbers is odd, and B the event of at least one ace. Assuming a sample space of 36 points, list the sample points which belong to the events $A \cap B$, $A \cup B$ and $A \cap \bar{B}$. Find the probabilities of these events, assuming equally likely events.
- Two good dice are rolled simultaneously. Let A denote the event "the sum shown is 8" and B the event "the two show the same number." Find $P(A)$, $P(B)$, $P(A \cap B)$, and $P(A \cup B)$.

A box contains six discs numbered 1 to 6. Find for each integer k from 3 to 11, the probability that the numbers on two discs drawn without replacement have a sum equal to k.

(P.U., B.A./B.Sc. 1977)

- 6.18 In a single throw of two fair dice, find the probability that the product of the numbers on the dice is (i) between 8 and 16 (both inclusive), (ii) divisible by 4. (P.U., B.A./B.Sc.)
- 6.19 a) Elaborate the statement that "two mutually exclusive events need not be equally likely".
 b) Compare the probabilities of at least one 6 in 4 tosses of a fair die with the probability of at least one double 6 in 24 tosses of two fair dice.
 c) Compare the probability of a total of 9 with that of a total of 10 when three fair dice are tossed once.
- Hint:** To find the sample points in different events, combine the faces of the third die with the relevant sums of the first two dice.
- 6.20 a) From a pack of 52 cards, two are drawn at random. Find the probability that one is a king and the other a queen. (P.U., B.A./B.Sc.)
- b) A set of eight cards contains one joker. A and B are two players and A chooses 5 cards at random, B taking the remaining 3 cards. What is the probability that A has the joker?
- 6.21 a) Of 12 eggs in a refrigerator, 2 are bad. From these 12 eggs are chosen at random to make a cake. What are the probabilities that (i) exactly one is bad? (ii) at least one is bad?
 b) A certain carton of eggs has 3 bad eggs and 12 good eggs. An omelette is made of 3 randomly chosen from the carton. What is the probability that there are (i) no bad eggs, (ii) at least 1 bad egg, (iii) exactly 2 bad eggs in the omelette?
- 6.22 a) An integer between 3 and 12 inclusive is chosen at random. What is the probability that it is an even number? That it is an even and divisible by 3?
 b) Three distinct integers are chosen at random from the first 20 positive integers. Find the probability that (i) their sum is even, (ii) their product is even.
 c) A box contains 3 red, 4 white and 5 green balls. These balls are drawn from the box together. Find the probability that they may be (i) all of different colours, (ii) all of the same colour.
- 6.23 a) Find the probability of obtaining at least one 6 when (i) 5 dice are thrown, (ii) a pair of dice are thrown.
 b) How many dice must be thrown so that the probability of obtaining at least one 6 is at least 0.99?
 c) A missile is fired at a target and the probability that the target is hit is 0.7. Find how many missiles should be fired so that the probability that the target is hit at least once is more than 0.995.
- 6.24 A bag contains 14 identical balls, 4 of which are red, 5 black and 5 white. Six balls are drawn from the bag. Find the probability that (i) 3 are red, (ii) at least two are white. (P.U., B.A., Hons., M.Tech.)
- 6.25 a) Three applicants are to be selected at random out of 4 boys and 6 girls. Find the probability of selecting (i) all girls, (ii) all boys, (iii) at least one boy?

- b) From a group of 6 men and 8 women, 3 people are chosen at random. Find the probability that there are more men chosen than women.
- 6.26 A firm buys 3 shipments of parts each month. The purchasing agent selects at random from among four in-state suppliers and six out-of-state suppliers. What is the probability that the orders are placed with
- the in-state suppliers only?
 - the out-of-state suppliers only?
 - at least one in-state supplier?
- (P.U., M.A. Econ. 1979)
- 6.27 Q. a) Using a sample space of otherwise, show that for any two events A and B,
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
- b) A class contains 10 men and 20 women of which half the men and half the women have brown eyes. Find the probability that a person chosen at random is a man or has brown eyes.
(P.U., B.A./B.Sc. 1974)
- c) In a group of 20 adults, 4 out of the 7 women and 2 out of the 13 men wear glasses. What is the probability that a person chosen at random from the group is a woman or someone who wears glasses?
- 6.28 a) The events E_1 and E_2 are neither independent nor mutually exclusive. Denote by p_{ij} the probability that E_1 and E_2 both happen. Prove that the probability that at least one of E_1 and E_2 happens, is $p_1 + p_2 - p_{12}$.
(P.U., B.A./B.Sc., 1975)
- b) One integer is chosen at random from the numbers 1, 2, 3, ..., 50. What is the probability that the chosen number is divisible by 6 or by 8?
- c) Define the probability of an event.
- d) State and prove the addition law of probability for any two events A and B.
- e) A drawer contains 100 bolts and 150 nuts. Half of the bolts and half of the nuts are rusted. If one item is taken at random, what is the probability that it is rusted or is a bolt?
- f) Define Mutually Exclusive Events. State and prove the theorem of addition of probabilities concerning mutually exclusive events.
- g) What is the probability of throwing either 7 or 11 with two dice?
- h) If A and B are mutually exclusive events and $P(A) = 0.4$ and $P(B) = 0.5$, find (i) $P(A \cup B)$, (ii) $P(\bar{A})$.
- i) If A and B are any two events defined on a sample space S, show that
$$P[(A \cap \bar{B}) \cup (\bar{A} \cap B)] = P(A) + P(B) - 2P(A \cap B)$$

(P.U., B.A./B.Sc., 1989)
- j) Let A and B be events with $P(A \cup B) = \frac{3}{4}$, $P(\bar{A}) = \frac{2}{3}$ and $P(A \cap B) = \frac{1}{4}$. Find (i) $P(A)$, (ii) $P(B)$, (iii) $P(A \cap \bar{B})$.

- c) If $P(A) = \frac{1}{2}$, $P(A \cup B) = \frac{3}{4}$ and $P(\bar{B}) = \frac{5}{8}$, then find (i) $P(A \cap B)$, (ii) $P(A \cap \bar{B})$,
 (iii) $P(A \cup \bar{B})$ and, (iv) $P(B \cap \bar{A})$.

6.32 Using the Venn diagram, show that

- i) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 ii) $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$
 (P.U., B.A./B.Sc., 1976, 78)

6.33 a) Explain what is meant by conditional probability.

- b) A pair of fair dice is thrown. If the two numbers appearing are different, find the probability that (i) the sum is 8, (ii) the sum is four or less. (P.U., M.A. Econ., 1977)

- 6.34 a) A box contains 4 bad and 6 good tubes. Two are drawn out together. One of them is tested and found to be good. What is the probability that the other one is also good?

- b) If two balanced dice are rolled, find the conditional probability that sum of dice will be given that it is odd.

- c) In a firm, 20 percent of the employees have accounting background, while 5 percent of the employees are executives and have accounting background. If an employee has accounting background, what is the probability that the employee is an executive?

6.35 A box contains 10 red and 12 white roses. Flowers are picked up at random one by one without replacement. What is the probability that;

- i) the first 3 flowers are red?
 ii) there are 2 red and 2 white flowers in the first four picked up?
 iii) the third one is red given that the first 2 are white? (P.U., B.A./B.Sc., 1976, 78)

6.36 a) State and prove the multiplication law of probabilities for any two events A and B.

- b) Define Independent Events and Dependent Events. Give examples.

- c) Given $P(A) = 0.60$, $P(B) = 0.40$, $P(A \cap B) = 0.24$. Find $P(A|B)$, $P(A \cap B)$, $P(B|A)$, $P(\bar{B}|A)$. What is the relation between A and B?

6.37 d) Differentiate between independent and mutually exclusive events. Are independent events always disjoint?

- b) Let A and B be two events associated with an experiment. Suppose that $P(A) = 0.4$, $P(A \cup B) = 0.7$. Let $P(B) = p$.

- i) For what choice of p are A and B mutually exclusive?
 ii) For what choice of p are A and B independent? (P.U., M.A. Stat., 1977)

- c) Given $P(A) = 0.5$ and $P(A \cup B) = 0.6$. find $P(B) if$

- i) A and B are mutually exclusive.
 ii) A and B are independent.
 iii) $P(A|B) = 0.4$

- Q.3E Indicate whether each of the following statements is true or false. If false, indicate why.
- If $P(A \cap B) = 0$, then A and B are mutually exclusive.
 - If $P(A \cup B) = 0$, then A and B are independent.
 - If $P(A|B) = P(B|A)$, then $P(A) = P(B)$.
 - If A and B are independent, then $P(A) = P(B)$.
- Q.3F a) State and prove the multiplication law of probabilities for independent and dependent events.
- b) Let A and B be two independent events such that the probability is $\frac{1}{6}$ that they will occur simultaneously and $\frac{3}{8}$ that neither of them will occur. Find $P(A)$ and $P(B)$.
(B.Z.U. & P.I.U., B.A/B.Sc. 1976)
- Q.3G Two dice are cast: E_1 is the event that a 6 appears on at least one die, E_2 is the event that a 5 appears on exactly one die and E_3 is the event that same number appears on both dice.
- Are E_1 and E_2 independent?
 - Are E_2 and E_3 independent?
 - Are E_3 and E_1 independent?
(P.I.U., B.A/B.Sc. 1980)
- Q.3H a) A can solve 75% of the problems in this book and B can solve 70%. What is the probability that either A or B can solve a problem chosen at random?
- b) Three cards are drawn in succession without replacement from an ordinary deck of playing cards. Find the probability that the first card is a red ace, the second card is a ten or jack, and the third card is greater than 3 but less than 7.
- The probability that A will be alive after 10 years to come is 5/7 and for B it is 7/9. Find out the probability that (i) both of them will die, (ii) A will be alive and B dead, (iii) B will be alive and A dead, (iv) both of them will be alive, in 10 years to come.
- A bag contains 3 red and 5 black balls and another 4 red and 7 black balls. A ball is drawn from a bag selected at random. Find the probability that it is red.
 - One urn contains 3 white and 2 black balls, another contains 5 white and 3 black balls. If an urn is chosen at random and a ball is taken from it, what is the probability that it is white?
- Two drawings each of three balls are made from a bag containing 5 white and 8 black balls; the balls are not being replaced before the next trial. What are the probabilities that the first drawing will give 2 white balls and the second 2 black balls?
- c) Show that the multiplication law $P(A \cap B \cap C) = P(A|B) P(B) P(C)$, established for two events, may be generalized to three events as follows:
- $$P(A \cap B \cap C) = P(A|B \cap C) P(B|C) P(C)$$

- b) A farmer has a box containing 30 eggs, 5 of which have blood spots. He checks eggs by taking them at random one after another from the box. What is the probability that the first two eggs have spots and the third will be clear?

Ques. 6.45 There groups of children contain respectively 3 girls and 1 boy; 2 girls and 2 boys; 1 girl and 2 boys. One child is selected at random from each group. Show that the probability that the selected consists of 1 girl and 2 boys is 13/32.

6.47 There are three families, each having four children, 2 boys and 2 girls; 2 boys and 1 girl, and 1 boy and 3 girls. A child from each family is invited to a party. Find the probability (i) that only girls turn up for the party, (ii) that two girls and one boy turn up for the party.

- Ques.* 6.48 The odds that a book will be favorably reviewed by three independent critics are 3 to 2, 3 and 2 to 3 respectively. What is the probability that of three reviews a majority will be favorable?

Hint: If we are given the odds that an event A will occur as a to b ,

$$p = \frac{a}{a+b} \text{ and } q = \frac{b}{a+b}$$

- 6.49 a) A can hit a target four times in 5 shots; B three times in 4 shots; C twice in 3 shots fire a volley. What is the probability that two hits at least by?
- b) A committee of three - A, B, and C, - make a decision on the basis of a vote. What is the probability of a wrong decision by the committee if the probabilities of a wrong decision by each member are $P(A) = 0.05$, $P(B) = 0.02$, and $P(C) = 0.10$?

(P.U., B.A.B.Sc.)

- 6.50 The probability that three men hit a target are respectively $\frac{1}{6}$, $\frac{1}{4}$ and $\frac{1}{3}$. Each shoots once at the target.

- i) Find the probability that exactly one of them hits the target.
ii) If only one hits the target, what is the probability that it was the first man?

(P.U., B.A.B.Sc.)

- 6.51 Three missiles are fired at a target. If the probabilities of hitting the target are 0.4, 0.3 and 0.2 respectively, and if the missiles are fired independently, what is the probability?
- that all the missiles hit the target?
 - that at least one of the three hits the target?
 - that exactly one hits the target?
 - that exactly 2 hit the target?

- 6.52 A and B play 12 games of chess, of which 6 are won by A, 4 are won by B, and 2 end in a draw. They agree to play a tournament consisting of 5 games. Find the probability that (a) A wins all three games, (b) two games end in a tie, (c) A and B win alternately, (d) B wins at least one game.

(P.U., B.A.B.Sc.)

- 6.53 The contents of two urns are as follows:

Urn A contains 3 red and 2 white balls.

Urn B contains 2 red and 3 white balls.

An urn is selected at random; a ball is drawn and put into the other urn, then a ball is drawn from the second urn. Find the probability that both balls drawn are of the same colour.

Hint: Construct the tree diagram.

- 6.54 A man invited 5 friends. He was born in April as also all the invited friends. What is the probability that none of the friends was born on the same day of the month as the host?

(C.S.S., 1962)

- 6.55 a) What are respective chances of winning of A and B who toss a coin alternately on the understanding that the first to obtain heads, wins the toss?

- b) Three men toss in succession for a prize to be given to the one who first obtains heads. Show that their respective chances of winning are $\frac{4}{7}$, $\frac{2}{7}$ and $\frac{1}{7}$.

- 6.56 a) Find the probability of getting exactly 4 heads when 6 coins are tossed.

(B.Z.U., B.A./B.Sc. 1976)

- b) Sixteen coins are tossed once. What is the probability of obtaining (i) exactly 8 heads, (ii) exactly 11 heads?

The national pass rate for an examination is 70%. A school enters 6 candidates. Calculate the probability that (i) 2 candidates will pass and (ii) 5 candidates will pass. Explain why the probability of all passing is not equal to the probability of all failing.

- a) State and prove Bayes' theorem.

- b) Three urns of the same appearance have the following proportions of white and black balls:

Urn A: 1 white, 2 black balls.

Urn B: 2 white, 1 black ball.

Urn C: 2 white, 2 black balls.

One of the urns is selected and a ball is drawn from it. It turns out to be white. What is the probability that urn C was chosen? (P.U., B.A., (Hons.) – Part III, 1965, B.A./B.Sc. 2007)

There are three coins, identical in appearance, one of which is ideal and the other two biased with probabilities $\frac{1}{3}$ and $\frac{2}{3}$ respectively for a head. One coin is taken at random and tossed twice. If a head appears both the times, what is the probability that the ideal coin was chosen?

In a certain college, 4% of the men and 1% of the women are taller than 6 feet. Further more, 60% of the students are women. Now if a student is selected at random and is taller than 6 feet, what is the probability that the student is a woman? (P.U., B.A. Hons. 1974)

Three cooks A, B and C, bake a special kind of cake, and with respective probabilities 0.02, 0.03 and 0.05 it fails to rise. In the restaurant where they work, A bakes 50 percent of these cakes, B 30 percent, and C 20 percent. What proportion of "failures" is caused by A?

- 6.62 The stock of a warehouse consists of boxes of high, medium and low quality lightbulbs in respective proportions 1:2:2. The probabilities of bulbs of three types being unsatisfactory are 0.0, 0.1 and 0.2 respectively. If a box is chosen at random and two bulbs in it are tested are found to be satisfactory, what is the probability that it contains bulbs (i) of high quality, (ii) of medium quality, (iii) of low quality?
- 6.63 A patient is thought to have one of three diseases A_1 , A_2 and A_3 whose probabilities under the given conditions are $\frac{1}{2}$, $\frac{1}{6}$ and $\frac{1}{3}$ respectively. A test is carried out to help the diagnosis and yields a positive result with a probability of 0.1 for disease A_1 , a probability of 0.2 for disease A_2 and a probability of 0.9 for disease A_3 . The test is conducted 5 times and the results are positive 4 times and negative once. What is the probability of each disease after testing?
[P.U., B.A. (Hons.) - Part-III, 1987]

<https://stat9943.blogspot.com>

CHAPTER 7

**RANDOM
VARIABLES**

INTRODUCTION

Usually, we are not interested in a particular outcome of a random experiment but our interest is in some numerical description of the outcome. For example, when two coins are tossed, we may be interested only in the number of heads which appear and not in the actual sequence of heads and tails which is not a numerical quantity. To express the outcomes in numbers, we assign to each non-numerical outcome of the sample space $S = \{HH, HT, TH, TT\}$, one of the numbers i ($i = 0, 1, 2$) corresponding to the number of heads appearing. That is, we express the outcomes in terms of numerical values as

Domain or Sample Space (E):

(HH), (HT), (TH), (TT)

Range or Corresponding Value $X = f(E)$:

2, 1, 1, 0

Again in the experiment of throwing a pair of dice, if we are interested only in the sum of the upper faces of the two dice and not in the particular dots, we assign to each possible outcome of the experiment, one of numbers i ($i = 2, 3, 4, \dots, 12$) corresponding to the sum of dots appearing on their faces.

It is clear that the numbers 0, 1, 2 and 2, 3, 4, ..., 12 in the above cited examples, are random variables determined by the outcomes of the random experiments. Such a numerical quantity whose value is determined by the outcome of a random experiment, is called a random variable. Specifically, we assign a single real number to each outcome of the sample space, and hence we state that a random variable is a real-valued function defined on a sample space. Thus the number of heads obtained in tossing of two coins and the sum of the dots obtained with a pair of dice in the above examples are the values of random variables.

It should be noted that in the above definition, a function has been named a random variable. This nomenclature, though inappropriate and somewhat unfortunate, is universally accepted and used.

A random variable is also called a chance variable, a stochastic variable or simply a variate and is denoted as $r.v.$. The random variables are usually denoted by capital Latin letters such as X, Y, Z whose values taken by them are represented by the corresponding small letters such as x, y, z . It is to be understood that more than one $r.v.$ can be defined on the same sample space. There are two types of $r.v.$: discrete and the continuous.

DISTRIBUTION FUNCTION

The distribution function of a random variable X , denoted by $F(x)$, is defined by $F(x) = P(X \leq x)$. Hence $F(x)$ gives the probability of the event that X takes a value less than or equal to a specified value x . The distribution function is abbreviated to d.f. and is also called the cumulative distribution (c.d.) or cumulative probability function of the X from the smallest upto specific values of x .

Since $F(x)$ is a probability, it is quite obvious that

$$F(-\infty) = P(\emptyset) = 0 \quad \text{and} \quad F(+\infty) = P(S) = 1.$$

Let a and b be two real numbers such that $a < b$. Then the probability of the interval (a, b) is

$$\begin{aligned} F(b) - F(a) &= P(X \leq b) - P(X \leq a) \\ &= P(a < X \leq b), \end{aligned}$$

which is non-negative and hence $F(x)$ is a non-decreasing function of x .

Again $\lim_{b \rightarrow a^+} F(x+a) = F(x)$, i.e. the function $F(x)$ is continuous on the right at each value of x .

A d.f. $F(x)$ thus has the following properties:

- $F(-\infty) = 0, F(+\infty) = 1$.
- $F(x)$ is a non-decreasing function of x , i.e. $F(x_1) \leq F(x_2)$ if $x_1 \leq x_2$.
- $F(x)$ is continuous at least on the right of each x .

All random variables have distribution functions. Distribution functions for different r.v.s are distinguished by using the notation F_x, F_y , etc.

7.3 DISCRETE RANDOM VARIABLE AND ITS PROBABILITY DISTRIBUTION

A random variable X is defined to be discrete if it can assume values which are finite or countably infinite. When X takes on a finite number of values, they may be listed as x_1, x_2, \dots, x_n . In the countably infinite case, the values may be listed as $x_1, x_2, \dots, x_n, \dots$. The number of heads in coin tossing experiments, the number of defective items observed in a consignment, the number of accidents, the number of cactuses in 1cc of soil, etc. are the examples of discrete r.v.s.

Let X be a discrete r.v. taking n distinct values $x_1, x_2, \dots, x_n, \dots$. Then the function, denoted by $f(x)$, and defined by

$$\begin{aligned} f(x_i) &= P(X = x_i), \quad i = 1, 2, \dots, n, \\ &= 0, \quad \text{for } x \neq x_i. \end{aligned}$$

is called the probability function (pf) of the r.v. X , and the values $x_1, x_2, \dots, x_n, \dots$ among which the probability 1 is distributed, are called the probability points or jump points. $P(X = x_i)$ is the probability that the discrete r.v. X takes the value x_i .

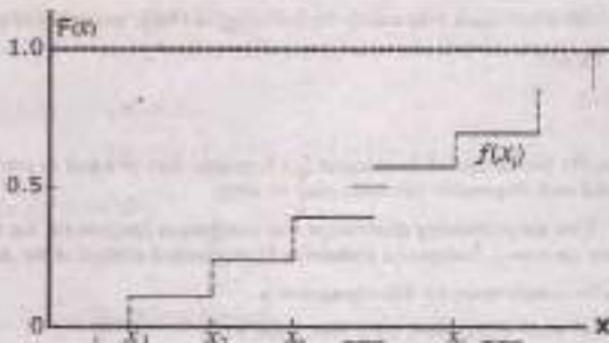
The distribution function for a discrete r.v. is

$$F(x) = \sum_i f(x_i),$$

where sum is taken over all x_i that are less than or equal to x .

Letting x tend to $+\infty$, we have $F(+\infty) = \sum_i f(x_i) = 1$.

$F(x)$ in case of a discrete r.v. is a step function. That is, its graph consists of horizontal segments between any two successive values and has a step or jump of height $f(x_i)$ at each value x_i (see figure on page 229). It should be noted that $F(x)$ is continuous but between jumps, it is constant.



A discrete r.v. may also be defined as a r.v. whose d.f. jumps at the possible values of X and is constant between adjacent jump points. The height of a jump at each point x is the probability of event $X = x$.

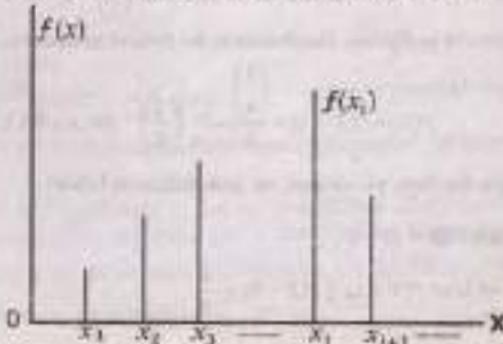
$$f(x_i) = [\text{Jump at } x_i] = P(X = x_i).$$

The set whose elements are the ordered pairs $\{x_i, f(x_i)\}, i = 1, 2, \dots$ defines the probability function. Some writers do not make any distinction between the terms probability function and probability distribution but they use them interchangeably. The probability distribution of a r.v. may easily be expressed either in a tabular form by giving all the possible values of X and the probabilities $f(x_i) = P(X = x_i)$ as

Values	(x)	x_1	\dots	x_n	\dots
Probability	$f(x)$	$f(x_1)$	$f(x_2)$	\dots	$f(x_n)$

in the form of an equation for $f(x)$ or a list of the possible values of X .

The graph of a probability distribution is obtained by locating the values $x_1, x_2, \dots, x_n, \dots$ along the x -axis and drawing vertical lines of heights equal to $f(x_1), f(x_2), \dots, f(x_n), \dots$ above them. A probability distribution can also be graphically displayed by a probability histogram.



A probability distribution must satisfy the following two basic properties of probability.

- $f(x_i) \geq 0$, for all i .
- $\sum_i f(x_i) = 1$.

In other words, the probability of an outcome (x_i) is greater than or equal to zero and the sum of probabilities associated with all possible outcomes must be unity.

Example 7.1 Find the probability distribution and distribution function for the number of heads when 3 balanced coins are tossed. Construct a probability histogram and a graph of the distribution.

The equiprobable sample space for this experiment is

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

Let X be the r.v. that denotes the number of heads. Then the values of x are 0, 1, 2 and 3, and their probabilities are:

$$f(0) = P(X = 0) = P\{\{TTT\}\} = \frac{1}{8},$$

$$f(1) = P(X = 1) = P\{\{HTT, THT, TTH\}\} = \frac{3}{8}$$

$$f(2) = P(X = 2) = P\{\{HHT, HTH, THH\}\} = \frac{3}{8}$$

$$f(3) = P(X = 3) = P\{\{HHH\}\} = \frac{1}{8}$$

Putting this information in the following form, we obtain the desired probability distribution as follows:

Number of Heads (x_i)	$f(x_i)$	0	1	2	3
Probability $f(x_i)$		$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

To obtain a formula, we need expressions for x heads to be selected out of 3 heads and also the denominators for all values. Now, x heads can be selected in $\binom{3}{x}$ ways and the total number of outcomes is 8. Therefore the desired probability distribution in the form of an equation is

$$f(x) = P(X = x) = \frac{\binom{3}{x}}{8} = \binom{3}{x} \left(\frac{1}{2}\right)^3 \quad \text{for } x = 0, 1, 2, 3.$$

For distribution function, we compute the probabilities as below:

If $x < 0$, we have $P(X < x) = 0$

If $0 \leq x < 1$, we have $P(X < x) = P(X = 0) = \frac{1}{8}$

For $1 \leq x < 2$, we have

$$P(X < x) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8}$$

Similarly, for $2 \leq x < 3$, we have

$$P(X < x) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{1}{8} + \frac{3}{8} + \frac{4}{8} = \frac{7}{8}$$

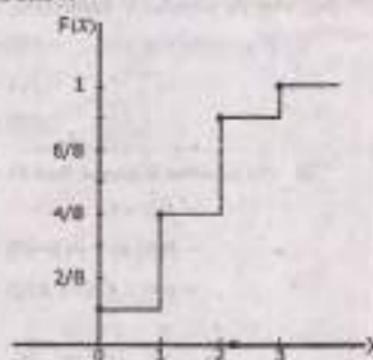
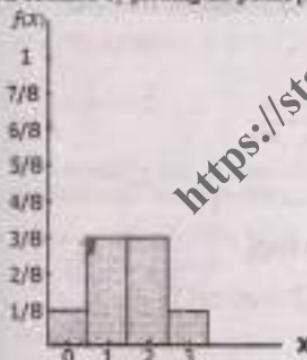
Finally for $x \geq 3$, we have

$$P(X < x) = \sum_{i=0}^x P(X = i) = 1.$$

Hence the desired distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{8} & \text{for } 0 \leq x < 1 \\ \frac{4}{8} & \text{for } 1 \leq x < 2 \\ \frac{7}{8} & \text{for } 2 \leq x < 3 \\ 1 & \text{for } x \geq 3 \end{cases}$$

The probability histogram is obtained by plotting the points $(x, f(x))$, while the graph of distribution function is obtained by plotting the points $(x, F(x))$ as shown below:



Examples 7.2 (a) Find the probability distribution of the sum of the dots when two fair dice are thrown.

- (i) Use the probability distribution to find the probabilities of obtaining (i) a sum of 8 or 11, that is greater than 8, (ii) a sum that is greater than 5 but less than or equal to 10.
- (ii) The sample space S for the experiment of throwing two dice contains 36 sample points, which are equally likely, i.e. each point has probability $\frac{1}{36}$.

Let X be the random variable representing the sum of dots which appear on the dice. The values of the r.v. are $2, 3, 4, \dots, 12$. The probabilities of these values are computed as below;

$$f(2) = P(X=2) = P\{(\{1,1\})\} = \frac{1}{36}, \text{ as there is only one point resulting in a sum of 2,}$$

$$f(3) = P(X=3) = P\{(\{1,2\}, (2,1)\}) = \frac{2}{36},$$

$$f(4) = P(X=4) = P\{(\{1,3\}, (2,2), (3,1)\}) = \frac{3}{36},$$

$$\text{Similarly, } f(5) = \frac{4}{36}, f(6) = \frac{5}{36}, f(7) = \frac{6}{36}, f(8) = \frac{5}{36}, f(9) = \frac{4}{36},$$

$$f(10) = \frac{3}{36}, f(11) = \frac{2}{36} \text{ and } f(12) = \frac{1}{36}.$$

Therefore the desired probability distribution of the r.v. X is

x_i	2	3	4	5	6	7	8	9	10	11	12
$f(x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

It is interesting to note that this result may also be expressed by the equation as

$$f(x) = \frac{6 - |7 - x|}{36}, \text{ for } x = 2, 3, 4,$$

(b) Using the probability distribution, we get the required probabilities as follows:

$$\text{i)} P(\text{a sum of 8 or 11}) = P(X=8 \text{ or } X=11)]$$

$$= P(X=8) + P(X=11)$$

$$= f(8) + f(11) = \frac{1}{36} + \frac{2}{36} = \frac{3}{36}$$

$$\text{ii)} P(\text{a sum that is greater than 8})$$

$$= P(X > 8)$$

$$= P(X=9) + P(X=10) + P(X=11) + P(X=12)$$

$$= f(9) + f(10) + f(11) + f(12)$$

$$= \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} = \frac{10}{36}$$

$$\text{iii)} P(\text{a sum that is greater than 5 but less than or equal to 10})$$

$$= P(5 < X \leq 10)$$

$$= P(X=6) + P(X=7) + P(X=8) + P(X=9) + P(X=10)$$

$$= f(6) + f(7) + f(8) + f(9) + f(10)$$

$$= \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} = \frac{23}{36}$$

CONTINUOUS RANDOM VARIABLE AND ITS PROBABILITY DENSITY FUNCTION

A random variable X is defined to be continuous if it can assume every possible value in an interval $a < b$, where a and b may be $-\infty$ and $+\infty$ respectively. The height of a person, the temperature at time, the amount of rainfall, time to failure for an electronic system, etc. are examples of continuous random variable.

A r.v. X may also be defined as continuous if its d.f. $F(x)$ is continuous and is differentiable except at isolated points in the given range. The graph of $F(x)$ has no jumps or steps but is a continuous function for all x .

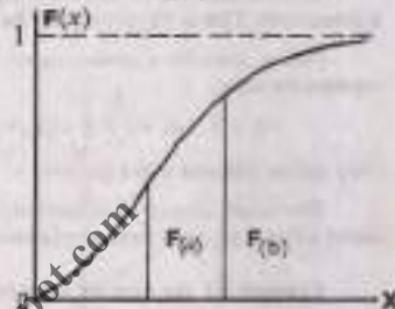
Let the derivative of $F(x)$ be denoted by $f(x)$, i.e.

$$\frac{dF(x)}{dx} = f(x)$$

Since $F(x)$ is a non-decreasing function of x , we have

- (i) $f(x) \geq 0$,

- (ii) $F(x) = \int_{-\infty}^x f(x) dx$, for all x .



The function $f(x)$ is called the *probability density function*, abbreviated to p.d.f., or simply *density* of the r.v. X .

A p.d.f. has the following properties:

- (i) $f(x) \geq 0$, for all x

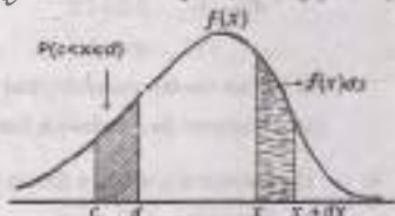
- (ii) $\int_{-\infty}^{\infty} f(x) dx = 1$

- (iii) The probability that X takes on a value in the interval $[c, d]$, $c < d$, is given by

$$P(c < x \leq d) = F(d) - F(c)$$

$$= \int_c^d f(x) dx - \int_c^d f(x) dx$$

$$= \int_c^d f(x) dx$$



∴ the area under the curve $y = f(x)$ between $X = c$ and $X = d$, (see figure above),

In other words, $f(x)$ is a non-negative function, the integration takes place over all possible values of X between the specified limits and the probabilities are given by appropriate areas under the

$$\text{Also } P(x < X \leq z + dx) = F(z + dx) - F(z) \\ \approx f(z)dx$$

The quantity $f(z)dx$ is called the probability differential or probability element (p.e.) of X .

$$\text{Since } P(X = z) = \int_z^z f(z) dz = 0.$$

it should therefore be noted that the probability of a continuous r.v. X taking any particular value is always zero. That is why probability for a continuous r.v. is measurable only over a given interval.

Further, since for a continuous r.v. X , $P(X=x) = 0$ for every x , the following four probabilities are regarded the same:

$$P(x < X \leq d), P(x < X \leq d), P(x \leq X < d) \text{ and } P(x < X < d).$$

They may be different with a discrete r.v.

The values (expressed as intervals) of a continuous r.v. and their associated probabilities are shown either in a tabular form or expressed by means of a formula.

Example 7.3 (a) Find the value of k so that the function $f(x)$ defined as follows, may be a density function

$$f(x) = kx, \quad 0 \leq x \leq 2 \\ = 0, \quad \text{elsewhere}$$

- (b) Find also the probability that sum of two sample values will exceed 1.
- (c) Compute the distribution function $F(x)$.
- (d) The function $f(x)$ will be a density function, if

i) $f(x) \geq 0$ for every x and

ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

The first condition is satisfied when $k \geq 0$. The second condition will be satisfied, if

i.e. if $1 = \int_{-\infty}^0 f(x)dx + \int_0^2 f(x)dx + \int_2^{\infty} f(x)dx$

i.e. if $1 = \int_{-\infty}^0 0 dx + \int_0^2 kx dx + \int_2^{\infty} 0 dx$

$$\text{i.e. if } 1 = 0 + \left[k \frac{x^2}{2} \right]_0^2 + 0 = 2k$$

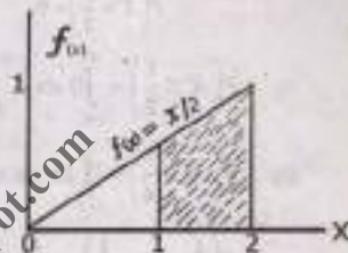
This gives $k = \frac{1}{2}$

$$\text{Hence } f(x) = \begin{cases} \frac{x}{2}, & \text{for } 0 \leq x \leq 2 \\ 0, & \text{elsewhere} \end{cases}$$

$P(X > 1)$ = area of shaded region

$$\begin{aligned} &= \int_1^2 f(x) dx \\ &= \int_1^2 \frac{x}{2} dx = \left[\frac{x^2}{4} \right]_1^2 - \frac{1}{4} \end{aligned}$$

$$P(\text{two sample values exceeding one}) = \frac{3}{4} \times \frac{3}{4} = \frac{9}{16}$$



To compute the distribution function, we find

$$F(x) = P(X < x) = \int_{-\infty}^x f(t) dt$$

for x such that $-\infty < x \leq 0$, $F(x) = \int_{-\infty}^0 0 dt = 0$,

$$\text{for } x \leq 2, \text{ we have } F(x) = \int_{-\infty}^x 0 dt + \int_0^x \frac{t}{2} dt = \left[\frac{t^2}{4} \right]_0^x = \frac{x^2}{4},$$

$$\text{and for } x > 2, \text{ we have } F(x) = \int_{-\infty}^0 0 dt + \int_0^2 \frac{t}{2} dt + \int_2^x 0 dt = 1$$

$$\therefore F(x) = 0, \quad \text{for } x < 0$$

$$= \frac{x^2}{4}, \quad \text{for } 0 \leq x \leq 2$$

$$= 1, \quad \text{for } x > 2$$

Example 7.4 A r.v. X is of continuous type with p.d.f.

$$\begin{aligned} f(x) &= 2x, \quad 0 < x < 1, \\ &= 0, \quad \text{elsewhere.} \end{aligned}$$

Find (i) $P\left(X = \frac{1}{2}\right)$, (ii) $P\left(X \leq \frac{1}{2}\right)$, (iii) $P\left(X > \frac{1}{4}\right)$, (iv) $P\left(\frac{1}{4} \leq X < \frac{5}{2}\right)$

(v) $P\left(X \leq \frac{1}{2} \mid \frac{1}{3} \leq X \leq \frac{2}{3}\right)$

Clearly $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x) dx = \int_0^1 2x dx = 1$.

i) Since $f(x)$ is a continuous probability function, therefore

$$P\left(X = \frac{1}{2}\right) = 0.$$

ii) $P\left(X \leq \frac{1}{2}\right) = \int_0^{1/2} 0 dx + \int_0^{1/2} 2x dx = 0 + [x^2]_0^{1/2} = \frac{1}{4}$

iii) $P\left(X > \frac{1}{4}\right) = \int_{1/4}^1 2x dx + \int_1^1 0 dx = [x^2]_{1/4}^1 + 0 = \frac{15}{16}$

iv) $P\left(\frac{1}{4} \leq X < \frac{1}{2}\right) = \int_{1/4}^{1/2} 2x dx = [x^2]_{1/4}^{1/2} = \frac{3}{16}$

v) Applying the definition of conditional probability, we get

$$\begin{aligned} P\left(X \leq \frac{1}{2} \mid \frac{1}{3} \leq X \leq \frac{2}{3}\right) &= \frac{P\left(\frac{1}{3} \leq X \leq \frac{1}{2}\right)}{P\left(\frac{1}{3} \leq X \leq \frac{2}{3}\right)} = \frac{\int_{1/3}^{1/2} 2x dx}{\int_{1/3}^{2/3} 2x dx} \\ &= \left[x^2 \right]_{1/3}^{1/2} + \left[x^2 \right]_{1/3}^{2/3} \\ &= \frac{5}{36} \times \frac{9}{4} = \frac{5}{12}. \end{aligned}$$

Example 7.5 A continuous r.v. X has the d.f. $F(x)$ as follows:

$$F(x) = 0, \quad \text{for } x < 0,$$

$$= \frac{2x^2}{5}, \quad \text{for } 0 < x \leq 1,$$

$$= -\frac{3}{5} + \frac{2}{5} \left(3x - \frac{x^2}{2} \right), \quad \text{for } 1 < x \leq 2,$$

$$= 1 \quad \text{for } x > 2.$$

Find the p.d. and $P(|X| < 1.5)$.

By definition, we have $f(x) = \frac{d}{dx} F(x)$

$$\begin{aligned}\text{Therefore } f(x) &= \frac{4x}{5} && \text{for } 0 < x \leq 1 \\ &= \frac{2}{5}(3-x) && \text{for } 1 < x \leq 2 \\ &= 0 && \text{elsewhere.}\end{aligned}$$

$$\therefore P(|X| < 1.5) = P(-1.5 < X < 1.5)$$

$$\begin{aligned}&= \int_{-1.5}^{1.5} dx + \int_{-1.5}^0 dx + \int_0^1 \frac{4x}{5} dx + \int_1^{1.5} \frac{2(3-x)}{5} dx \\ &= 0 + 0 + \left[\frac{2x^2}{5} \right]_0^1 + \left[\frac{2}{5} \left(3x - \frac{x^2}{2} \right) \right]_1^{1.5} \\ &= \frac{2}{5} + \frac{2}{5} \left[\left(\frac{4.5}{2} - \frac{2.25}{2} \right) - \left(3 - \frac{1}{2} \right) \right] \\ &= 0.40 + 0.35 = 0.75.\end{aligned}$$

JOINT DISTRIBUTIONS

The distribution of two or more random variables which are observed simultaneously when an experiment is performed, is called their *joint distribution*. It is customary to call the distribution of a single r.v. as *univariate*. Likewise, a distribution involving two, three or many r.v.'s simultaneously is referred to as *bivariate*, *trivariate* or *multivariate*.

7.3.1 Bivariate Distribution Function. Let X and Y be two r.v.'s defined on the same sample space. Then the function $F(x, y)$ defined by $F(x, y) = P(X \leq x \text{ and } Y \leq y)$, where $F(x, y)$ gives the probability that X will take on a value less than or equal to x and, at the same time, Y will take on a value less than or equal to y , is called a *bivariate or joint distribution function* of X and Y .

A bivariate d.f. $F(x, y)$ possesses the following properties:

- $F(x, -\infty) = F(-\infty, y) = 0$, $F(+\infty, +\infty) = 1$
 - $F(x, y)$ is a non-decreasing function of x and y , and is continuous on the right.
 - If $x_1 < x_2$ and $y_1 < y_2$, then
- $$\begin{aligned}P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) &= P(X < x_2, Y < y_2) - P(X < x_2, Y < y_1) - P(X < x_1, Y < y_2) + P(X < x_1, Y < y_1) \\ &= F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) > 0\end{aligned}$$

The probability that a random point (X, Y) falls in the interval $(x_1 \leq X < x_2; y_1 \leq Y < y_2)$ is shown graphically.

A bivariate distribution may be discrete when the possible values of (X, Y) are finite or countably infinite. It is continuous if (X, Y) can assume all values in some non-countable set of the plane. A bivariate distribution is said mixed when one r.v. is discrete and the other is continuous.

7.5.2 Bivariate Probability Function. Let X and Y be two discrete r.v.'s defined on the sample space S , X taking the values x_1, x_2, \dots, x_m and Y taking the values y_1, y_2, \dots, y_n . The probability that X takes on the value x_i and, at the same time, Y takes on the value y_j , denoted by f_{ij} or p_{ij} , is defined to be the joint probability function or simply the joint distribution of X and Y . The joint probability function, also called the bivariate probability function $f(x, y)$ is a function whose value at the point (x_i, y_j) is given by

$$f(x_i, y_j) = P(X = x_i \text{ and } Y = y_j), \quad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, n.$$

The joint or bivariate probability distribution consists of all pairs of values (x_i, y_j) and their associated probabilities $f(x_i, y_j)$ (i.e. the set of triples $(x_i, y_j, f(x_i, y_j))$) can either be shown in a two-way table as follows:

Joint Probability Distribution of X and Y

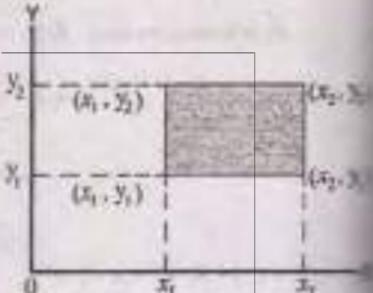
$X \setminus Y$	y_1	y_2	\dots	y_j	\dots	y_n	$P(X=x_i)$
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	\dots	$f(x_1, y_j)$	\dots	$f(x_1, y_n)$	$g(x_1)$
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	\dots	$f(x_2, y_j)$	\dots	$f(x_2, y_n)$	$g(x_2)$
\vdots	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
x_i	$f(x_i, y_1)$	$f(x_i, y_2)$	\dots	$f(x_i, y_j)$	\dots	$f(x_i, y_n)$	$g(x_i)$
\vdots	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
x_m	$f(x_m, y_1)$	$f(x_m, y_2)$	\dots	$f(x_m, y_j)$	\dots	$f(x_m, y_n)$	$g(x_m)$
$P(Y=y_j)$	$h(y_1)$	$h(y_2)$	\dots	$h(y_j)$	\dots	$h(y_n)$	1

or be expressed by means of a formula for $f(x, y)$. The probabilities $f(x, y)$ can be obtained by appropriate values of x and y in the table or formula.

A joint probability function has the following properties,

- $f(x_i, y_j) \geq 0$, for all (x_i, y_j) , i.e. for $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n$.
- $\sum_{i=1}^m \sum_{j=1}^n f(x_i, y_j) = 1$

7.5.3 Marginal Probability Functions. From the joint probability function for (X, Y) we obtain the individual probability function of X and Y . Such individual probability function are called marginal probability function.



Let $f(x, y)$ be the joint probability function of two discrete r.v.'s X and Y . Then the marginal probability function of X is defined as

$$\begin{aligned} g(x_i) &= \sum_{j=1}^n f(x_i, y_j) \\ &= f(x_i, y_1) + f(x_i, y_2) + \dots + f(x_i, y_n) \text{ as } x_i \text{ must occur either with } y_1 \text{ or } y_2 \text{ or } \dots \text{ or } y_n \\ &= P(X = x_i). \end{aligned}$$

Thus, the individual probability function of X is found by adding over the rows of the two-way table.

Similarly, the marginal probability function for Y is obtained by adding over the columns as

$$h(y_j) = \sum_{i=1}^m f(x_i, y_j) = P(Y = y_j).$$

The values of *marginal probabilities* are often written in the margins of the joint table as they are row and column totals in the table. The probabilities in each marginal probability function add to 1.

7.5.4 Conditional Probability Functions. Let X and Y be two discrete r.v.'s with joint probability function $f(x, y)$. Then the conditional probability function for X given Y , denoted as $f(x|y)$, is defined by

$$\begin{aligned} f(x_i|y_j) &= P(X = x_i | Y = y_j) \\ &= \frac{P(X = x_i \text{ and } Y = y_j)}{P(Y = y_j)} \\ &= \frac{f(x_i, y_j)}{h(y_j)}, \quad \text{for } i = 1, \dots, j = 1, 2, \dots \end{aligned}$$

$f(x_i|y_j)$ is the marginal probability and $f(x_i|y_j) \geq 0$. It gives the probability that X takes on the values x_i when Y has taken on the values y_j . The conditional probability $f(x_i|y_j)$ is non-negative and (for a fixed y_j) adds to 1 on i and hence is a probability function.

Similarly, the conditional probability function for Y given $X = x_i$ is

$$\begin{aligned} f(y_j|x_i) &= P(Y = y_j | X = x_i) \\ &= \frac{P(Y = y_j \text{ and } X = x_i)}{P(X = x_i)} \\ &= \frac{f(x_i, y_j)}{g(x_i)}, \quad \text{where } g(x_i) > 0. \end{aligned}$$

7.5.5 Independence. Two discrete r.v.'s X and Y are said to be *statistically independent*, if and only if for all possible pairs of values (x_i, y_j) the joint probability function $f(x, y)$ can be expressed as the product of the two marginal probability functions. That is, X and Y are independent, if

$$\begin{aligned} f(x, y) &= P(X = x_i \text{ and } Y = y_j) \\ &= P(X = x_i) \cdot P(Y = y_j) \quad \text{for all } i \text{ and } j \\ &= g(x_i) h(y_j). \end{aligned}$$

It should be noted that the joint p.f. of X and Y when they are independent, can be obtained by multiplying together their marginal probability functions.

Example 7.6 An urn contains 3 black, 2 red and 3 green balls and 2 balls are selected at random from it. If X is the number of black balls and Y is the number of red balls selected, then find

- the joint probability function $f(x, y)$;
 - $P(X + Y \leq 1)$;
 - the marginal p.d. $g(x)$ and $h(y)$;
 - the conditional p.d. $f_{X|Y}(x|y)$;
 - $P(X=0 | Y=1)$, and
 - Are X and Y independent?
- i) The sample space Ω for this experiment contains $\binom{8}{2} = 28$ sample points. The possible values of X are 0, 1 and 2, and those for Y are 0, 1 and 2. The values that (X, Y) can take on are $(0, 0), (0, 1), (1, 0), (1, 1), (0, 2)$ and $(2, 0)$. We desire to find $f(x, y)$ for each value (x, y) . Now $f(0, 0) = P(X=0 \text{ and } Y=0)$, where the event $(X=0 \text{ and } Y=0)$ represents that neither a black nor red ball is selected, implying that the 2 selected are green balls. This event Ω contains $\binom{3}{2} \binom{2}{0} = 3$ sample points, and
- $$f(0, 0) = P(X=0 \text{ and } Y=0) = \frac{3}{28}.$$

$$\begin{aligned} \text{Again } f(0, 1) &= P(X=0 \text{ and } Y=1) \\ &= P(\text{none is black, 1 is red and 1 is green}) \\ &= \frac{\binom{3}{2} \binom{2}{1}}{28} = \frac{6}{28} \end{aligned}$$

$$\text{Similarly, } f(1, 1) = P(X=1 \text{ and } Y=1)$$

$$= P(1 \text{ is black, 1 is red and none is green})$$

$$= \frac{\binom{3}{1} \binom{2}{1} \binom{3}{0}}{28} = \frac{6}{28}$$

Similar calculations give the probabilities of other values and the joint p.f. of X and Y is given below:

(x, y)	$(0, 0)$	$(0, 1)$	$(1, 0)$	$(1, 1)$	$(0, 2)$	$(2, 0)$
$f(x, y)$	$\frac{3}{28}$	$\frac{6}{28}$	$\frac{9}{28}$	$\frac{6}{28}$	$\frac{1}{28}$	$\frac{3}{28}$

These probabilities can also be represented in another tabular form as follows:

Joint Probability Distribution

X \ Y	0	1	2	$P(X=x_i)$ $g(x)$	
0	$\frac{3}{28}$	$\frac{6}{28}$	$\frac{1}{28}$	10/28	
1	$\frac{9}{28}$	$\frac{6}{28}$	0	15/28	
2	$\frac{3}{28}$	0	0	3/28	
	$P(Y=y_j)$ $h(y)$	15/28	12/28	1/28	1

Thus, this joint p.d. of the two r.v.'s (X, Y) can be represented by the formula:

$$f(x, y) = \frac{\binom{3}{x} \binom{2}{y} \binom{3}{2-x-y}}{28}, \quad \begin{aligned} &x = 0, 1, 2 \\ &y = 0, 1, 2 \\ &0 \leq x + y \leq 2 \end{aligned}$$

- To compute $P(X + Y \leq 1)$, we see that $x + y \leq 1$ for the cells (0, 0), (0, 1) and (1, 0). Therefore

$$\begin{aligned} P(X + Y \leq 1) &= f(0, 0) + f(0, 1) + f(1, 0) \\ &= \frac{3}{28} + \frac{6}{28} = \frac{18}{28} = \frac{9}{14} \end{aligned}$$

- The marginal p.d.'s are

x	0	1	2
$g(x)$	10/28	12/28	3/28

y	0	1	2
$h(y)$	15/28	12/28	1/28

- By definition the conditional p.d. $f(x | 1)$ is

$$\begin{aligned} f(x | 1) &= P(X=x | Y=1) \\ &= \frac{P(X=x \text{ and } Y=1)}{P(Y=1)} = \frac{f(x, 1)}{h(1)} \end{aligned}$$

$$h(1) = \sum_{x=0}^2 f(x, 1) = \frac{0}{28} + \frac{6}{28} + 0 = \frac{12}{28} = \frac{3}{7}$$

$$f(x | 1) = \frac{f(x, 1)}{h(1)} = \frac{7}{3} f(x, 1), \quad x = 0, 1, 2$$

$$f(0 | 1) = \frac{7}{3} f(0, 1) = \left(\frac{7}{3} \right) \left(\frac{6}{28} \right) = \frac{1}{2}$$

$$f(1|1) = \frac{7}{3} f(1,1) = \left(\frac{7}{3}\right)\left(\frac{6}{28}\right) = \frac{1}{2}$$

$$f(2|1) = \frac{7}{3} f(2,1) = \left(\frac{7}{3}\right)(0) = 0$$

Hence the conditional p.d. of X given the $Y=1$, is

x	0	1	2
$f(x 1)$	1/2	1/2	0

v) Finally, $P(X=0 | Y=1) = f(0|1) = \frac{1}{2}$

vi) We find that $f(0, 1) = \frac{6}{28}$.

$$g(0) = \sum_{y=0}^2 f(0, y) = \frac{3}{28} + \frac{6}{28} + \frac{1}{28} = \frac{10}{28}$$

$$h(1) = \sum_{x=0}^2 f(x, 1) = \frac{6}{28} + \frac{6}{28} + 0 = \frac{12}{28}$$

$$\text{Now } \frac{6}{28} \neq \frac{10}{28} \cdot \frac{12}{28},$$

$$\text{i.e. } f(0, 1) \neq g(0) h(1),$$

and therefore X and Y are not statistically independent.

Example 7.7 The joint p.d. of two discrete r.v.'s X and Y is given by

$$f(x, y) = \begin{cases} \frac{1}{30} & \text{for } x = 1, 2, 3 \text{ and } y = 1, 2, \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

The marginal p.d. for X is

$$\begin{aligned} g(x) &= \sum_y f(x, y) \\ &= \sum_{y=1}^2 \frac{xy^2}{30} = \frac{x(1)^2}{30} + \frac{x(2)^2}{30} = \frac{x}{6}, \text{ for } x = 1, 2, 3; \end{aligned}$$

and the marginal p.d. for Y is

$$\begin{aligned} h(y) &= \sum_x f(x, y) \\ &= \sum_{x=1}^3 \frac{xy^2}{30} = \frac{1y^2}{30} + \frac{2y^2}{30} + \frac{3y^2}{30} = \frac{y^2}{5}, \text{ for } y = 1, 2, 3. \end{aligned}$$

$$\text{Ans} \quad \frac{xy^2}{30} = \frac{x}{6} \times \frac{y^2}{5}, \text{ for } x = 1, 2, 3 \text{ and } y = 1, 2,$$

$$\therefore f(x, y) = g(x) \cdot h(y)$$

X and Y are independent.

5.6 Continuous Bivariate Distributions. The bivariate probability density function of two r.v.'s X and Y is an integrable function $f(x, y)$ satisfying the following properties:

- (i) $f(x, y) \geq 0$ for all (x, y) .
- (ii) $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$, and
- (iii) $P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) dx dy$.

The distribution function (d.f.) of the bivariate r.v. (X, Y) is defined by

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du$$

It should be noted that analogous to the relationship $\frac{d}{dx} F(x) = f(x)$, we have $\frac{d}{dy} F(x, y) = f(x, y)$, wherever F is differentiable.

The marginal p.d.f. of the continuous r.v.

$$g(x) = \int_{-\infty}^{+\infty} f(x, y) dy$$

of the r.v. X is

$$h(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

This is the marginal p.d.f. of any of the variables is obtained by integrating out the other variable p.d.f. between the limits $-\infty$ and $+\infty$.

The conditional p.d.f. of the continuous r.v. X given that Y takes the value y , is defined to be

$$f(x|y) = \frac{f(x, y)}{h(y)}$$

and $h(y)$ are respectively the joint p.d.f. of X and Y , and the marginal p.d.f. of Y and $h(y) > 0$. Similarly, the conditional p.d.f. of the continuous r.v. Y given that $X = x$, is

$$f(y|x) = \frac{f(x, y)}{g(x)}, \text{ provided that } g(x) > 0.$$

It is worth noting that the conditional p.d.f.'s satisfy all the requirements for a univariate probability density function.

Finally, two continuous r.v.'s X and Y are said to be statistically independent, if and only if their joint density $f(x, y)$ can be factored in the form $f(x, y) = g(x) \cdot h(y)$ for all possible values of X and Y .

Example 7.8 Given the following joint p.d.f.

$$f(x, y) = \begin{cases} \frac{1}{8}(6-x-y), & 0 \leq x \leq 2, 2 \leq y \leq 4, \\ 0, & \text{elsewhere.} \end{cases}$$

- a) Verify that $f(x, y)$ is a joint density function.
- b) Calculate (i) $P\left(X \geq \frac{3}{2}, Y \geq \frac{5}{2}\right)$, (ii) $P(X + Y < 3)$.
- c) Find the marginal p.d.f. $g(x)$ and $h(y)$.
- d) Find the conditional p.d.f. $f(x|y)$ and $f(y|x)$.
- e) The joint density $f(x, y)$ will be a p.d.f. if
 - i) $f(x, y) \geq 0$ and
 - ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$.

P.U., B.A. Stat.

Now $f(x, y)$ is clearly ≥ 0 for all x and y in the given region, and

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= \frac{1}{8} \int_0^2 \int_2^4 (6-x-y) dy dx \\ &= \frac{1}{8} \int_0^2 \left[6y - xy - \frac{y^2}{2} \right]_2^4 dx \\ &= \frac{1}{8} \int_0^2 (6-2x) dx = \frac{1}{8} [6x - x^2]_0^2 \\ &= \frac{1}{8} [12 - 4] = 1. \end{aligned}$$

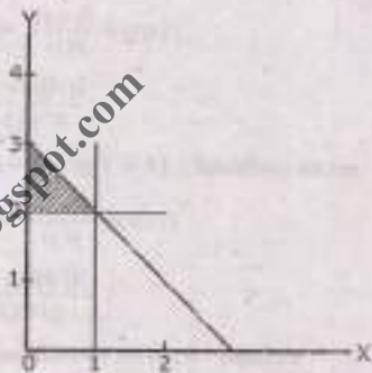
Thus $f(x, y)$ has the properties of a joint p.d.f.

- b) (i) To determine the probability of a value of the r.v. (X, Y) falling in the region we find

$$\begin{aligned}
 P(X \leq \frac{3}{2}, Y \leq \frac{5}{2}) &= \int_{x=0}^{\frac{3}{2}} \int_{y=2}^{\frac{5}{2}} \frac{1}{8}(6-x-y)dy dx \\
 &= \frac{1}{8} \int_0^{\frac{3}{2}} \left[6y - xy - \frac{y^2}{2} \right]_2^{\frac{5}{2}} dx \\
 &= \frac{1}{8} \int_0^{\frac{3}{2}} \left(\frac{15}{2} - \frac{x}{2} \right) dx = \frac{1}{64} \left[15x - 2x^2 \right]_0^{\frac{3}{2}} = \frac{5}{32}
 \end{aligned}$$

$$(ii) P(X+Y \leq 3) = \frac{1}{8} \int_0^1 \int_0^{3-x} (6-x-y)dy dx \quad (\because x+y \leq 3 \Rightarrow y = 3-x)$$

$$\begin{aligned}
 &= \frac{1}{8} \int_0^1 \left[6y - xy - \frac{y^2}{2} \right]_0^{3-x} dx \\
 &= \frac{1}{8} \int_0^1 \left(\frac{x^2}{2} - 4x + \frac{7}{2} \right) dx \\
 &= \frac{1}{8} \left[\frac{x^3}{6} - 2x^2 + \frac{7x}{2} \right]_0^1 \\
 &= \frac{1}{8} \cdot \frac{10}{6} = \frac{5}{24}
 \end{aligned}$$



Event "X + Y ≤ 3" is shaded.

The marginal p.d.f. of X is

$$\begin{aligned}
 f(x) &= \int_{-\infty}^{\infty} f(x,y) dy, \quad -\infty < x < \infty \\
 &\Rightarrow \frac{1}{8} \int_2^5 (6-x-y) dy, \quad 0 < x < 2 \\
 &= \frac{1}{8} \left[6y - xy - \frac{y^2}{2} \right]_2^5, \quad 0 \leq x \leq 2 \\
 &= \frac{1}{8} (3-x) \\
 &= 0 \quad x < 0 \text{ or } x \geq 2
 \end{aligned}$$

Similarly, the marginal p.d.f. of Y is

$$\begin{aligned} h(y) &= \frac{1}{3} \int_0^y (6-x-y) dx, & 2 \leq y \leq 4 \\ &= \frac{1}{4}(5-y) & 2 \leq y \leq 4 \\ &= 0, \text{ elsewhere.} \end{aligned}$$

- d) The conditional p.d.f. of X given $Y = y$, is

$$\begin{aligned} f(x|y) &= \frac{f(x,y)}{h(y)}, \text{ where } h(y) > 0. \\ &= \frac{(1/3)(6-x-y)}{(1/4)(5-y)} = \frac{6-x-y}{2(5-y)} \end{aligned}$$

and the conditional p.d.f. of Y given $X = x$, is

$$\begin{aligned} f(y|x) &= \frac{f(x,y)}{g(x)}, \text{ where } g(x) \\ &= \frac{(1/3)(6-x)}{(1/4)(3-x)} = \frac{6-x-y}{2(3-x)}. \end{aligned}$$

Example 7.9 Let the bivariate continuous random variable (X, Y) have the joint probability function given by

$$\begin{aligned} f(x,y) &= \frac{xy}{3}, & 0 \leq x \leq 1, 0 \leq y \leq 2, \\ &= 0, \text{ elsewhere.} \end{aligned}$$

- a) Check that $f(x, y)$ is a p.d.f.
 b) Find the marginal p.d.f.'s.
 c) Find the conditional p.d.f.'s and verify that $f(x|y)$ is a p.d.f.
 d) The function $f(x, y)$ will be a p.d.f. if
 i. $f(x, y) \geq 0$ and
 ii. $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1.$

RANDOM VARIABLES

Now $f(x, y)$ is clearly ≥ 0 for all x and y in the given interval and

$$\begin{aligned} \int_{-x}^x \int_{-y}^{+y} f(x, y) dx dy &= \int_0^1 \int_0^2 \left(x^2 + \frac{xy}{3} \right) dx dy \\ &= \int_0^1 \left[\frac{x^3}{3} + \frac{x^2 y}{6} \right]_0^2 dy = \int_0^1 \left(\frac{1}{3} + \frac{y}{6} \right) dy \\ &= \left[\frac{y}{3} + \frac{y^2}{12} \right]_0^1 = \frac{2}{3} + \frac{1}{12} = 1. \end{aligned}$$

The marginal p.d.f.'s are

$$\begin{aligned} f(x) &= \int_{-x}^x f(x, y) dy = \int_0^2 \left(x^2 + \frac{xy}{3} \right) dy \\ &= \left[x^2 y + \frac{xy^2}{6} \right]_0^2 = 2x^2 + \frac{2}{3}x = \frac{2}{3}x(1+3x), \quad 0 \leq x \leq 1 \\ f(y) &= \int_{-y}^y f(x, y) dx = \int_0^2 \left(x^2 + \frac{xy}{3} \right) dx \\ &= \left[\frac{x^3}{3} + \frac{x^2 y}{6} \right]_0^2 = \frac{1}{3} + \frac{y}{6} = \frac{1}{6}(2+y), \quad 0 \leq y \leq 2. \end{aligned}$$

The conditional p.d.f. of X for given $Y = y$ is

$$f(x|y) = \frac{f(x, y)}{h(y)} = \frac{\frac{y^2}{3}}{\frac{1}{6}(2+y)} = \frac{6x^2 + 2xy}{2+y}, \quad 0 \leq x \leq 1, 0 \leq y \leq 2.$$

Conditional p.d.f. of Y for given $X = x$ is

$$f(y|x) = \frac{f(x, y)}{g(x)} = \frac{\frac{y^2}{3}}{\frac{1}{3}(3+3x)} = \frac{3x^2 + xy}{6x^2 + 2x} = \frac{3x + y}{6x + 2}, \quad 0 \leq y \leq 2, \quad 0 \leq x \leq 1.$$

To verify that the conditional p.d.f. $f(x|y)$ is a p.d.f., we have

$$\int_0^1 \frac{6x^2 + 2xy}{2+y} dx = \left[\frac{1}{2+y} (2x^3 + x^2 y) \right]_0^1$$

$$= \frac{1}{2+y} (2+y) = 1, \text{ for all } y.$$

Hence the conditional p.d.f. $f(y|x_i)$ satisfies the requirements for a univariate density function.

7.6 MATHEMATICAL EXPECTATION OF A RANDOM VARIABLE

Let a discrete r.v. X have possible values $x_1, x_2, \dots, x_n, \dots$ with corresponding probabilities $f(x_1), f(x_2), \dots, f(x_n)$, such that $\sum f(x) = 1$. Then the mathematical expectation or the expectation or the expected value of X , denoted by $E(X)$, is defined as

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \dots + x_n f(x_n) + \dots$$

$$= \sum_{i=1}^{\infty} x_i f(x_i), \text{ provided it converges absolutely.}$$

The sum converges absolutely if and only if $\sum (x_i f(x_i))$ is finite.

When X takes on only a finite number of values, we have $E(X) = \sum_{i=1}^n x_i f(x_i)$, which we regard as a weighted mean of the variable's possible values x_1, x_2, \dots, x_n , each being weighted by its respective probability. In case the values are equally likely, $E(X) = \frac{1}{n} \sum x_i$, which represents ordinary arithmetic mean of the n possible values. $E(X)$ is also called the mean of X and is denoted by the letter μ . It should be noted that $E(X)$ is the average value of the r.v. X over a large number of trials.

If the r.v. X is continuous, i.e. p.d.f. $f(x)$, then,

$$E(X) = \int x f(x) dx, \text{ provided the integral converges absolutely, i.e. } \int |x| f(x) dx \text{ is finite.}$$

Clearly the definition of mathematical expectation in the case of a continuous r.v. is the same with summation being replaced by integral.

Example 7.10 (a) What is the mathematical expectation of the number of heads when 3 coins are tossed?

(b) What is the expectation of the number of failure preceding the first success in an infinite sequence of independent trials with constant probability of success? (P.II, 2008)

(c) Let the r.v. X represent the number of heads when three fair coins are tossed. Then find the following p.d.

x	0	1	2	3
$f(x)$	1/8	3/8	3/8	1/8

Hence the mathematical expectation of X is

$$E(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{12}{8} = 1.5$$

It should be noted that $E(X)$ is 1.5, which is not an integer and not any value X could actually have. This means that, if the experiment of tossing 3 coins is repeated a very large number of times, we expect on the average to get 1.5 heads.

- Let the r.v. X denote the number of failures preceding the first success. Then as X takes the value 0, 1, 2, 3, ..., the respective probabilities are p, qp, q^2p, q^3p, \dots where $q = 1 - p$.

$$\begin{aligned} \text{Hence } E(X) &= x_0f(x_0) + x_1f(x_1) + x_2f(x_2) + \dots \\ &= 0 \cdot p + 1 \cdot qp + 2 \cdot q^2p + 3 \cdot q^3p + \dots \\ &= qp(1 + 2q + 3q^2 + \dots) \\ &= qp(1 - q)^{-2} = qp(p)^{-1} = \frac{q}{p} \end{aligned}$$

Example 7.11 (a) If it rains, an umbrella salesman can earn \$30 per day. If it is fair, he can lose \$6. What is his expectation if the probability of rain is 0.3? (P.U., B.A./B.Sc., 1982)

(b) A man draws 2 balls from a bag containing 3 white and 5 black balls. If he receives Rs. 70 for one ball he draws and Rs. 7 for every black ball, find his expectation. (P.U., B.A./B.Sc., 1987)

- Let X represent the number of dollars the salesman earns. Then X is a r.v. with possible values 30 and -6, where -6 corresponds to the event that salesman loses, and the corresponding probabilities are 0.3 and 0.7 respectively.

$$\begin{aligned} \text{Hence } E(X) &= 30 \times 0.3 + (-6) \times 0.7 \\ &= \$ (9.00 - 4.20) = \$ 4.80 \text{ per day.} \end{aligned}$$

- Two balls from a bag containing 3 white and 5 black balls can be drawn in the following three mutually exclusive ways:

- 2 white balls,
- 1 white and 1 black ball,
- 2 black balls.

Let p denote the probability of drawing 2 balls.

$$\text{Then } p_1 = \binom{3}{2} + \binom{8}{2} = \frac{3}{28}$$

$$p_2 = \binom{3}{1} \binom{5}{1} + \binom{8}{2} = \frac{15}{28} \text{ and}$$

$$p_3 = \binom{5}{2} + \binom{8}{2} = \frac{10}{28}$$

Let X denote the amount to be received. Then

$$x_1 = 2 \times \text{Rs. } 70 + 0 \times \text{Rs. } 7 = \text{Rs. } 140,$$

$$x_2 = 1 \times \text{Rs. } 70 + 1 \times \text{Rs. } 7 = \text{Rs. } 77,$$

$$x_3 = 0 \times \text{Rs. } 70 + 2 \times \text{Rs. } 7 = \text{Rs. } 14.$$

Hence the required expectation $= x_1 p_1 + x_2 p_2 + x_3 p_3$

$$= \frac{3}{28} \times 140 + \frac{15}{28} \times 77 + \frac{10}{28} \times 14$$

$$= 15 + 41.25 + 5 = \text{Rs. } 61.25.$$

Example 7.12 Find the expected value of the r.v. X having the p.d.f.

$$f(x) = 2(1-x), \quad 0 < x < 1$$

$$= 0, \quad \text{elsewhere.}$$

$$\begin{aligned} \text{Now } E(X) &= \int_{-\infty}^{\infty} x f(x) dx \\ &= 2 \int_0^1 x(1-x) dx \\ &= 2 \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_0^1 = 2 \left[\frac{1}{2} - \frac{1}{3} \right] = \frac{1}{3} \end{aligned}$$

7.6.1 Expectation of a Function of a Random variable. Let $H(X)$ be a function of the r.v. X . Then $H(X)$ is also a r.v. and also has an expected value, as any function of a r.v. is also a r.v. If X is a discrete r.v. with p.d.f. $f(x)$, then since $H(X)$ takes the value $H(x_i)$ when $X = x_i$, the expected value of function $H(X)$ is

$$E[H(X)] = H(x_1) f(x_1) + H(x_2) f(x_2) + \dots + H(x_n) f(x_n)$$

$$= \sum_i H(x_i) f(x_i), \text{ provided the series converges absolutely.}$$

Similarly, if X is a continuous r.v. with p.d.f. $f(x)$, then

$$E[H(X)] = \int_{-\infty}^{\infty} H(x) f(x) dx, \text{ provide the integral exists.}$$

In particular, if $H(X) = X^2$, then $E[X^2] = \sum x_i^2 f(x_i)$

It is relevant to note that $E(X^2)$ is not the same as $[E(X)]^2$.

Again if $H(X) = (X - \mu)^2$, where μ is the population mean, then

$$E(X - \mu)^2 = \sum (x_i - \mu)^2 f(x_i)$$

We call this expected value the variance and denote it by $\text{Var}(X)$ or σ^2 . That is $\sigma^2 = E(X - \mu)^2 = E(X^2) - [E(X)]^2$. The positive square root of the variance, as before, is called the standard deviation.

It is useful to note the following important results about variance.

- $\text{Var}(X)$ cannot be negative.
- $\text{Var}(\alpha) = 0$, where α is a constant.
- $\text{Var}(\alpha X) = \alpha^2 \text{Var}(X)$, where α is a constant.
- $\text{Var}(\alpha X + b) = \alpha^2 \text{Var}(X)$, where α and b are constants.

More generally, if $H(X) = X^k$, $k = 1, 2, 3, \dots$, then

$$E(X^k) = \sum x_i^k f(x_i)$$

which we call the k th moment about the origin of the r.v. X and we denote it by μ'_k .

Similarly, if $H(X) = (X - \mu)^k$, $k = 1, 2, 3, \dots$, then we get an expected value, called the k th moment about the mean of the r.v. X , which we denote by μ_k . That is

$$\mu_k = E(X - \mu)^k = \sum (x_i - \mu)^k f(x_i)$$

In case of a continuous r.v., the summations are replaced by integrals.

The skewness is often measured by

$$\gamma_1 = \frac{E(X - \mu)^3}{\sigma^3} \quad \text{and} \quad \text{Kurtosis } \beta_2 = \frac{\mu_4}{\mu_2^2} \quad \text{as discussed earlier.}$$

Example 7.13 Let X have the following probability distribution:

x_i	1	2	3	4	5
$f(x_i)$	0.2	0.3	0.2	0.2	0.1

Find the probability functions of $3X - 1$, X^2 and $X^2 + 2$; and find $E(3X - 1)$, $E(X^2)$ and $E(X^2 + 2)$.

The probability distribution of the r.v. $H(X) = 3X - 1$, is

Values of X ,	x_i	1	2	3	4	5
Probabilities,	$f(x_i)$	0.2	0.3	0.2	0.2	0.1
Values of $3X - 1$,	$(3x_i - 1)$	2	5	8	11	14

$$E(3X - 1) = \sum H(x_i) f(x_i) = \sum (3x_i - 1) f(x_i)$$

$$= 2 \times 0.2 + 5 \times 0.3 + 8 \times 0.2 + 11 \times 0.2 + 14 \times 0.1$$

$$= 0.4 + 1.5 + 1.6 + 2.2 + 1.4 = 7.1$$

The p.d. of $H(X) = X^2$ is

x_i	1	2	3	4	5
$f(x_i)$	0.2	0.3	0.2	0.2	0.1
x_i^2	1	4	9	16	25

and $E(X^2) = \sum x_i^2 f(x_i)$
 $= 1 \times 0.2 + 4 \times 0.3 + 9 \times 0.2 + 16 \times 0.2 + 25 \times 0.1$
 $= 0.2 + 1.2 + 1.8 + 3.2 + 2.5 = 8.9$

Similarly, $E(X^2 - 2) = \sum (x_i^2 - 2)f(x_i)$
 $= 3 \times 0.2 + 6 \times 0.3 + 11 \times 0.2 + 18 \times 0.2 + 27 \times 0.1$
 $= 0.6 + 1.8 + 2.2 + 3.6 + 2.7 = 10.9$

Example 7.14 Let X be a r.v. with p.d.f.

$$f(x) = 2(x-1), \quad 1 < x < 2 \\ = 0 \quad \text{elsewhere}$$

Find the expected values of $E(X) = 2X - 1$ and $E(X^2)$.

$$\text{Now } E(2X - 1) = \int_{-\infty}^{\infty} (2x - 1)f(x)dx = 2 \int_1^2 (2x - 1)(x-1)dx \\ = 2 \int_1^2 (2x^2 - 3x + 1)dx = 2 \left[\frac{2x^3}{3} - \frac{3x^2}{2} + x \right]_1^2 \\ = 2 \left[\left(\frac{2}{3} - \frac{3}{2} + 2 \right) - \left(\frac{2}{3} - \frac{3}{2} + 1 \right) \right] \\ = 2 \left[\frac{4}{3} - \frac{1}{6} \right] = \frac{17}{6}$$

$$\text{and } E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx \\ = 2 \int_1^2 x^2(x-1)dx = 2 \left[\frac{x^3}{4} - \frac{x^2}{3} \right]_1^2 \\ = 2 \left[\left(4 - \frac{8}{3} \right) - \left(\frac{1}{4} - \frac{1}{3} \right) \right] \\ = 2 \left[\frac{4}{3} + \frac{1}{12} \right] = \frac{17}{6}$$

Example 7.15 If the continuous r.v. X has p.d.f.

$$f(x) = \begin{cases} \frac{3}{4}(3-x)(x-5), & 3 \leq x \leq 5 \\ 0, & \text{elsewhere.} \end{cases}$$

Find the arithmetic mean, variance and standard deviation of X .

$$\begin{aligned} \text{Now } E(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \frac{3}{4} \int_3^5 x(3-x)(x-5)dx = \frac{3}{4} \left[-\frac{x^3}{4} + \frac{8x^2}{3} - \frac{15x^1}{2} \right] \\ &= \frac{3}{4} \left[\left(-\frac{625}{4} + \frac{1000}{3} - \frac{375}{2} \right) - \left(-\frac{81}{4} + \frac{216}{3} - \frac{45}{2} \right) \right] \\ &= \frac{3}{4} \left[\left(-\frac{125}{12} + \frac{63}{4} \right) \right] = \frac{3}{4} \left(\frac{64}{12} \right) = \end{aligned}$$

$$\begin{aligned} \text{Again } E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)dx \\ &= \frac{3}{4} \int_3^5 x^2 (3-x)(x-5)dx \\ &= \frac{3}{4} \int_3^5 (-x^3 + 8x^2 - 15x^1)dx = \frac{3}{4} \left[-\frac{x^4}{5} + \frac{8x^3}{4} - \frac{15x^2}{3} \right] \\ &= \frac{3}{4} \left[\left(-\frac{1}{5}(3125) + 2(243) - 5(25) \right) - \left(-\frac{1}{5}(243) + 2(81) - 5(25) \right) \right] \\ &= \frac{3}{4} \left[0 + \frac{243}{5} - 162 + 135 \right] = \frac{3}{4} \left[\frac{168}{5} \right] = \frac{81}{5} \end{aligned}$$

$$\therefore \text{Var}(X) = E(X^2) - [E(X)]^2$$

$$= \frac{81}{5} - (4)^2 = \frac{1}{5} = 0.2, \text{ and}$$

$$S.D.(X) = \sqrt{0.2} = 0.447$$

The mean = 4, variance = 0.2 and standard deviation = 0.447.

Example 7.16 The continuous r.v. X has p.d.f. $f(x)$, where $f(x) = \frac{3}{4}(1+x^2)^{-\frac{3}{2}}$ for $0 \leq x \leq 1$.

$E(X) = \mu$ and $\text{Var}(X) = \sigma^2$, find $P(|X - \mu| < \sigma)$.

$$\text{Now } E(X) = \int_0^{\infty} xf(x)dx$$

$$\begin{aligned} &= \frac{3}{4} \int_0^1 x(1+x^2)^{-\frac{3}{2}} dx = \frac{3}{4} \int_0^1 (x+x^3)^{-\frac{1}{2}} dx \\ &= \frac{3}{4} \left[\frac{x^2}{2} + \frac{x^4}{4} \right]_0^1 = \frac{3}{4} \left[\frac{1}{2} + \frac{1}{4} \right] = \frac{9}{16} = 0.5625 \end{aligned}$$

$$\text{And } E(X^2) = \int_0^{\infty} x^2 f(x)dx$$

$$\begin{aligned} &= \frac{3}{4} \int_0^1 x^2(1+x^2)^{-\frac{3}{2}} dx = \frac{3}{4} \int_0^1 (x^2+x^4)^{-\frac{1}{2}} dx \\ &= \frac{3}{4} \left[\frac{x^3}{3} + \frac{x^5}{5} \right]_0^1 = \frac{3}{4} \left[\frac{1}{3} + \frac{1}{5} \right] \\ &= \frac{3}{4} \left(\frac{8}{15} \right) = \frac{2}{5} = 0.4 \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E(X^2) - [E(X)]^2 = \frac{2}{5} - \left(\frac{9}{16} \right)^2 \\ &= \frac{107}{1280} = 0.0836, \text{ so that} \end{aligned}$$

$$\text{S.D. (X)} \text{ or } \sigma = \sqrt{0.0836} = 0.289$$

$$\begin{aligned} \text{Now } P(|X - \mu| < \sigma) &= P(-\sigma < X - \mu < \sigma) \\ &= P(\mu - \sigma < X < \mu + \sigma) \\ &= P(0.5625 - 0.289 < X < 0.5625 + 0.289) \\ &= P(0.2735 < X < 0.8515), \text{ and} \end{aligned}$$

$$P(0.2735 < X < 0.8515) = \frac{3}{4} \int_{0.2735}^{0.8515} (1+x^2)^{-\frac{3}{2}} dx$$

$$\begin{aligned}
 &= \frac{1}{4} \left[x + \frac{x^3}{3} \right]_{-0.2735}^{0.8515} \\
 &= \frac{3}{4} \left[0.8515 + \frac{(0.8515)^3}{3} - (-0.2735 + \frac{(0.2735)^3}{3}) \right] \\
 &= \frac{3}{4} [0.85729 - 0.28032] = 0.8527
 \end{aligned}$$

$$\therefore P(X - \mu < \sigma) = 0.8527.$$

Example 7.17 A continuous r.v. X has the p.d.f.

$$\begin{aligned}
 f(x) &= \frac{3}{4}x(2-x), \quad 0 \leq x \leq 2 \\
 &= 0, \quad \text{elsewhere}
 \end{aligned}$$

Find the first four moments about the mean and the coefficient of skewness.

We first calculate the moments about origin as:

$$\begin{aligned}
 \mu'_1 &= E(X) = \int_0^2 x f(x) dx \\
 &= \frac{3}{4} \int_0^2 x(2x - x^2) dx = \frac{3}{4} \left[\frac{2x^3}{3} - \frac{x^4}{4} \right]_0^2 \\
 &= \frac{3}{4} \left[\frac{16}{3} - \frac{16}{4} \right] = \frac{1}{4} \left[\frac{16}{12} \right] = 1;
 \end{aligned}$$

$$\begin{aligned}
 \mu'_2 &= E(X^2) = \int_0^2 x^2 f(x) dx \\
 &= \frac{3}{4} \int_0^2 x^2 (2x - x^2) dx = \frac{3}{4} \left[\frac{2x^4}{4} - \frac{x^5}{5} \right]_0^2 \\
 &= \frac{3}{4} \left[8 - \frac{32}{5} \right] = \frac{3}{4} \left[\frac{8}{5} \right] = \frac{6}{5};
 \end{aligned}$$

$$\begin{aligned}
 \mu'_3 &= E(X^3) = \int_0^2 x^3 f(x) dx \\
 &= \frac{3}{4} \int_0^2 x^3 (2x - x^2) dx = \frac{3}{4} \left[\frac{2x^5}{5} - \frac{x^6}{6} \right]_0^2
 \end{aligned}$$

$$= \frac{3}{4} \left[\frac{64}{5} - \frac{64}{6} \right] = \frac{3}{4} \left[\frac{64}{30} \right] = \frac{8}{5}$$

$$\mu'_2 = E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx$$

$$= \frac{3}{4} \int_0^7 x^2 (2x + x^2) dx = \frac{3}{4} \left[\frac{2x^3}{3} + \frac{x^4}{4} \right]_0^7$$

$$= \frac{3}{4} \left[\frac{64}{3} - \frac{128}{7} \right] = \frac{3}{4} \left[\frac{64}{21} \right] = \frac{16}{7}$$

Then we find the moments about the mean as;

$$\mu_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = \frac{8}{5} - (0)^2 = \frac{8}{5}$$

$$\begin{aligned}\mu_3 &= \mu'_3 = 3\mu'_1 \mu'_2 + 2(\mu'_1)^3 \\ &= \frac{8}{5} - 3(0)\left(\frac{8}{5}\right) + 2(0)^3 = \frac{8}{5} - \frac{16}{5} + 0 = 0\end{aligned}$$

$$\begin{aligned}\mu_4 &= \mu'_4 = 4\mu'_1 \mu'_3 + 6(\mu'_1)^2 \mu'_2 + 3(\mu'_1)^4 \\ &= \frac{16}{7} - 4(0)\left(\frac{8}{5}\right) + 6(0)^2 \left(-\frac{16}{5}\right) \\ &= \frac{16}{7} - \frac{32}{5} + \frac{36}{35} = \frac{3}{5}\end{aligned}$$

The coefficient of skewness is

$$\gamma_1 = \frac{E(X - \mu)^3}{\sigma^3} = 0 \quad (\because E(X - \mu)^3 = \mu_3)$$

7.6.2 Properties of Expected Values. The important properties of the expected values are given below.

1. If a is a constant, then $E(a) = a$.

Proof: Let X be a discrete r.v. with p.d. $\{x_i, f(x_i)\}, i = 1, 2, \dots, n$.

Then $E(X)$ for $X = a$ is given by

$$E(a) = \sum_{i=1}^n a f(x_i)$$

$$\begin{aligned}
 &= a f(x_1) + a f(x_2) + \dots + a f(x_n) \\
 &= a [f(x_1) + f(x_2) + \dots + f(x_n)] \\
 &= a \sum_i f(x_i) = a \quad (\because \sum_i f(x_i) = 1)
 \end{aligned}$$

Thus the expected value of a constant is constant itself.

2. If X is a discrete r.v. and if a and b are constants, then

$$E(aX + b) = a E(X) + b.$$

Proof. Let the p.d. of the r.v. X be $\{x_i, f(x_i)\}$, $i = 1, 2, \dots, n$.

Then by definition of expected value, we have:

$$\begin{aligned}
 E(aX + b) &= \sum_{i=1}^n (ax_i + b)f(x_i) \\
 &= (ax_1 + b)f(x_1) + (ax_2 + b)f(x_2) + \dots + (ax_n + b)f(x_n) \\
 &= a[x_1f(x_1) + x_2f(x_2) + \dots + x_nf(x_n)] + b[x_1f(x_1) + \dots + x_nf(x_n)] \\
 &= a \sum_i x_i f(x_i) + b \sum_i f(x_i) = a E(X) + b
 \end{aligned}$$

This result is often called the expected value of linear transformation of the r.v. X .

If $b = 0$, then $E(aX) = a E(X)$.

If $a = 1$, $b = -\mu$, then $E(X - \mu) = E(X) - \mu = \mu - \mu = 0$.

That is, the expected value of the deviation of any r.v. from its mean is always equal to zero.

The expected value of the sum of any two random variables is equal to the sum of their expected values, i.e.

$$E(X + Y) = E(X) + E(Y)$$

Let X and Y be two discrete r.v.'s defined on the same sample space S . Let X assume m values x_1, x_2, \dots, x_m with probabilities $g(x_1), g(x_2), \dots, g(x_m)$ and Y assume n values y_1, y_2, \dots, y_n with probabilities $h(y_1), \dots, h(y_n)$. The sum $Z = X + Y$ is a r.v. taking the values $x_i + y_j$ with probabilities $f(x_i, y_j)$ for all combinations of values of i and j . Hence by definition, we have

$$\begin{aligned}
 E(X+Y) &= \sum_i \sum_j (x_i + y_j) f(x_i, y_j) \\
 &= \sum_i x_i f(x_i, y_1) + \sum_i \sum_j y_j f(x_i, y_j)
 \end{aligned}$$

$$\text{But } \sum_i \sum_j x_i f(x_i, y_j) = \sum_i x_i \sum_j f(x_i, y_j)$$

$$\begin{aligned}
 &= \sum_i x_i [f(x_1, y_1) + f(x_1, y_2) + \dots + f(x_1, y_n)] \\
 &= \sum_i x_i g(x_i) \quad (\because \text{all possible values of } y \text{ are included in } \Sigma) \\
 &= E(X)
 \end{aligned}$$

$$\text{Similarly, } \sum_j \sum_i y_j f(x_i, y_j) = \sum_j y_j \sum_i f(x_i, y_j)$$

$$\begin{aligned}
 &= \sum_j y_j [f(x_1, y_1) + f(x_2, y_1) + \dots + f(x_n, y_1)] \\
 &= \sum_j y_j h(y_j) = E(Y)
 \end{aligned}$$

$$\text{Hence } E(X + Y) = E(X) + E(Y)$$

It is interesting to note that this result holds in general, that is, the expected value of any number of r.v.'s is the sum of their expected values. In other words,

$$\begin{aligned}
 E(Y_1 + Y_2 + \dots + Y_n) &= E(Y_1) + E(Y_2) + \dots + E(Y_n) \\
 \text{or} \quad E(\sum Y_i) &= \sum E(Y_i)
 \end{aligned}$$

The result also holds for the difference of r.v.'s, i.e.

$$E(X - Y) = E(X) - E(Y).$$

4. The expected value of the product of two independent r.v.'s is equal to the product of their expected values, i.e.

$$E(XY) = E(X) E(Y).$$

Proof. Let the r.v. X assume n values x_1, x_2, \dots, x_n and the r.v. Y , the m values y_1, y_2, \dots, y_m . Let the probability of X assuming the value x_i be $g(x_i)$ and the probability of Y assuming the values y_j be $h(y_j)$. The product XY is a r.v. taking the value $x_i y_j$ with probabilities $f(x_i, y_j)$. As X and Y are independent, joint p.d. $f(x_i, y_j)$ can be factored ($f(x_i, y_j) = g(x_i) h(y_j)$).

$$\begin{aligned}
 \text{Hence, } E(XY) &= \sum_i \sum_j x_i y_j f(x_i, y_j) \\
 &= \sum_i \sum_j x_i y_j g(x_i) h(y_j) \\
 &= \sum_i x_i g(x_i) \sum_j y_j h(y_j) \\
 &= E(X) E(Y).
 \end{aligned}$$

This result can easily be extended to several independent r.v.'s.

It should be noted that these properties are valid for continuous r.v.'s in which case the sums are replaced by integrals.

COM VARIABLES

Example 7.18 Let X and Y be two discrete r.v.'s with the following joint p.d.

		2	4
		1	0.10
y	3	0.20	0.30
	5	0.10	0.15

Find $E(X)$, $E(Y)$, $E(X+Y)$, $E(2X-3Y)$ and $E(XY)$.

To determine the expected values of X and Y , we first find the marginal p.d. $g(x)$ and $h(y)$ by summing over the columns and rows of the two-way table as below:

		2	4	$h(y)$
		1	0.10	0.15
y	3	0.20	0.30	0.50
	5	0.10	0.15	0.25
$g(x)$		0.40	0.60	0.00

$$E(X) = \sum x_i g(x_i) = 2 \times 0.40 + 4 \times 0.60 = 0.80 + 2.40 = 3.2$$

$$\begin{aligned} E(Y) &= \sum y_j h(y_j) = 1 \times 0.25 + 3 \times 0.50 + 5 \times 0.25 \\ &= 0.25 + 1.50 + 1.25 = 3.0 \end{aligned}$$

$$\begin{aligned} E(X+Y) &= \sum \sum (x_i + y_j) f(x_i, y_j) \\ &= (2+1)(0.10) + (2+3)(0.20) + (2+5)(0.10) + (4+1)(0.15) \\ &\quad + (4+3)(0.30) + (4+5)(0.15) \\ &= 0.30 + 1.20 + 0.70 + 0.75 + 2.10 + 1.35 = 6.20 \\ &= E(X) + E(Y) \end{aligned}$$

$$\begin{aligned} E(2X-3Y) &= 2E(X) - 3E(Y) \\ &= 2(3.2) - 3(3.0) = -2.8 \end{aligned}$$

X and Y are independent, therefore

$$E(XY) = E(X)E(Y) = (3.2)(3.0) = 9.6$$

Example 7.19 X and Y are two independent r.v.'s such that

$$g(x) = \frac{1}{3} \text{ for } x = 1, 2, 3 \text{ and } h(y) = \frac{1}{2} \text{ for } y = 0, 1.$$

$Z = 2X - Y$, then verify that $E(Z) = 2E(X) - E(Y)$.

The joint distribution of the two independent r.v.'s X and Y is

Y	X			$A(Y)$
	1	2	3	
0	1/6	1/6	1/6	1/2
1	1/6	1/6	1/6	1/2
$g(x)$	1/3	1/3	1/3	1

Now

$$E(X) = \sum x g(x) = \left(1 \times \frac{1}{3}\right) + \left(2 \times \frac{1}{3}\right) + \left(3 \times \frac{1}{3}\right) = 2,$$

$$E(Y) = \sum y A(y) = \left(0 \times \frac{1}{2}\right) + \left(1 \times \frac{1}{2}\right) = \frac{1}{2}, \text{ and}$$

$$E(Z) = E[2X - Y] = 2E(X) - E(Y) = 2 \times 2 - \frac{1}{2} = 3\frac{1}{2}$$

For verification, we find the value of $E[2X - Y]$ directly as below:

$$\begin{aligned} E[2X - Y] &= \sum_{i,j} E[(2x_i - y_j)f(x_i, y_j)] \\ &= (2 \times 1 - 0)\frac{1}{6} + (2 \times 1 - 1)\frac{1}{6} + (2 \times 2 - 0)\frac{1}{6} \\ &\quad + (2 \times 2 - 1)\frac{1}{6} + (2 \times 3 - 0)\frac{1}{6} + (2 \times 3 - 1)\frac{1}{6} \\ &= \frac{2}{6} + \frac{3}{6} + \frac{6}{6} + \frac{3}{6} \\ &= \frac{1}{6} = 3\frac{1}{2} \end{aligned}$$

Hence the result.

Example 7.20 Three dice, each numbered in the usual way from one to six, are coloured red and blue respectively. After casting them, a boy 'scores' in the following way. To the white he adds twice the red number and then subtracts the blue number. Thus a white three, a red five and blue two would score $3 + 8 - 2 = 9$. Assume that the boy casts the dice a large number of times. Find the mean and the variance of the scores.

Let X_w , X_r and X_b denote the numbers when the white, red and blue dice are cast respectively. Let Y represent the score a boy gets by casting the dice. Then his score according to the condition $Y = X_w + 2X_r - X_b$, and we need $E(Y)$ and $\text{Var}(Y)$.

Now $E(Y) = E[X_w + 2X_r - X_b]$

$$= E(X_w) + 2E(X_r) - E(X_b), \text{ where}$$

$$E[X_i, i = w, r, b] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{21}{6}$$

$$\text{Therefore } E(Y) = \frac{21}{6} + 2\left(\frac{21}{6}\right) - \frac{21}{6} = \frac{21}{3} = 7$$

COM VARIABLES

By definition, $\text{Var}(Y) = E(Y^2) - [E(Y)]^2$.

$$\begin{aligned} E(Y^2) &= E[X_1 + 2X_2 - X_3]^2 \\ &= E(X_1^2) + 4 E(X_2^2) + E(X_3^2) + 4 E(X_1 X_2) + 2 E(X_2 X_3) - 4 E(X_1 X_3) \\ &= E(X_1^2) + 4 E(X_2^2) + E(X_3^2) + 4 E(X_1) E(X_2) + 2 E(X_2) E(X_3) - 4 E(X_1) E(X_3) \end{aligned}$$

$$E[X_i^2, i=w, r, b] = 1 \times \frac{1}{6} + 4 \times \frac{1}{6} + 9 \times \frac{1}{6} = 15 \times \frac{1}{6} + 25 \times \frac{1}{6} + 36 \times \frac{1}{6} = \frac{91}{6}$$

$$\begin{aligned} E(Y^2) &= \frac{91}{6} + 4 \left(\frac{91}{6} \right) + \frac{91}{6} + 4 \left(\frac{21}{6} \right) \left(\frac{21}{6} \right) - 2 \left(\frac{21}{6} \right) \left(\frac{21}{6} \right) - 4 \left(\frac{21}{6} \right) \left(\frac{21}{6} \right) \\ &= \frac{91}{6} + \frac{364}{6} + \frac{91}{6} + 49 - \frac{49}{2} - 49 - \frac{133}{2} \end{aligned}$$

$$E(Y) = \frac{133}{2} - (7)^2 = 17.5.$$

Example 7.21 Let X and Y be independent r.v's with joint p.d.f.

$$f(x, y) = \frac{x(1+3y^2)}{4}, \quad 0 < x < 2, 0 < y < 1 \\ = 0, \quad \text{elsewhere}$$

$E(X)$, $E(Y)$, $E(X + Y)$ and $E(XY)$.

To determine $E(X)$ and $E(Y)$, we first find the marginal p.d.f. $g(x)$ and $h(y)$ as below:

$$\begin{aligned} g(x) &= \int f(x, y) dy = \int \frac{x(1+3y^2)}{4} dy \\ &= \frac{1}{4} \left[xy + xy^3 \right]_0^1 = \frac{x}{2} \quad \text{for } 0 < x < 2. \end{aligned}$$

$$\begin{aligned} h(y) &= \int f(x, y) dx = \int \frac{x(1+3y^2)}{4} dx = \frac{1}{4} \left[\frac{x^2}{2} + 3xy^2 \right]_0^1 \\ &= \frac{1}{2}(1+3y^2), \quad \text{for } 0 < y < 1. \end{aligned}$$

$$\begin{aligned} E(X) &= \int x g(x) dx \\ &= \int x \left(\frac{x}{2} \right) dx = \frac{1}{2} \left[\frac{x^2}{3} \right]_0^2 = \frac{4}{3}, \text{ and} \end{aligned}$$

$$E(Y) = \int y h(y) dy = \frac{1}{2} \int y(1+3y^2) dy$$

$$= \frac{1}{2} \left[\frac{x^2}{2} + \frac{3y^2}{4} \right]_0^1 = \frac{1}{2} \left[\frac{1}{2} + \frac{3}{4} \right] = \frac{5}{8}.$$

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y) f(x,y) dx dy \\ &= \int_0^1 \int_0^1 (x+y) \cdot \frac{6(1+3y^2)}{4} dy dx \\ &= \int_0^1 \int_0^1 \frac{x^2 + 3xy^2}{4} dy dx + \int_0^1 \int_0^1 \frac{xy + 3y^3}{4} dy dx \\ &= \int_0^1 \frac{1}{4} \left[x^2 y + x^2 y^3 \right]_0^1 dx + \int_0^1 \frac{1}{4} \left[\frac{xy^2}{2} + \frac{3y^4}{4} \right]_0^1 dx \\ &= \int_0^1 \frac{1}{4} (2x^2) dx + \int_0^1 \frac{1}{4} \left(\frac{x}{2} + \frac{27}{16} \right) dx \\ &= \frac{1}{2} \left[\frac{x^2}{3} \right]_0^1 + \frac{1}{2} \left[\frac{x^2}{8} + \frac{27}{16} x \right]_0^1 \\ &= \frac{4}{3} + \frac{5}{24}, \text{ and} \end{aligned}$$

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x,y) dx dy \\ &= \int_0^1 \int_0^1 xy \cdot \frac{6(1+3y^2)}{4} dy dx = \int_0^1 \int_0^1 \frac{x^2 y + 3x^2 y^3}{4} dy dx \\ &= \int_0^1 \frac{1}{4} \left[\frac{x^2 y^2}{2} + \frac{3x^2 y^4}{4} \right]_0^1 dx = \int_0^1 \frac{1}{4} \left[\frac{5x^2}{12} \right]_0^1 dx = \frac{5}{6}. \end{aligned}$$

It should be noted that

i) $E(X) + E(Y) = \frac{4}{3} + \frac{5}{8} = \frac{47}{24} = E(X+Y)$, and

ii) $E(X) \cdot E(Y) = \left(\frac{4}{3}\right)\left(\frac{5}{8}\right) = \frac{5}{6} = E(XY)$

7.4.3 Covariance of Two Random Variables. The covariance of two r.v.'s X and Y is a measure of the extent to which their values tend to increase or decrease together. It is denoted as $\text{Cov}(X, Y)$, and is defined as the expected value of the product $[X - E(X)][Y - E(Y)]$. That is

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E[XY - XE(Y) - YE(X) + E(X)E(Y)] \\ &= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y).\end{aligned}$$

If X and Y are independent, then $E(XY) = E(X)E(Y)$, and

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0.$$

It is very important to note that covariance is zero when the r.v.'s X and Y are independent but it is not generally true. The covariance of a r.v. with itself is obviously its variance.

7.4.4 Variance of the Sum or Difference of two Random Variables. Let X and Y be two discrete r.v.'s. The variance of the r.v. $X + Y$ is defined as

$$\begin{aligned}\text{Var}(X + Y) &= E[(X + Y - E(X + Y))^2] \\ &= E[(X - E(X)) + (Y - E(Y))]^2 \\ &= E[X - E(X)]^2 + E[Y - E(Y)]^2 + 2E[X - E(X)](Y - E(Y)) \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)\end{aligned}$$

If X and Y are independent, then $\text{Cov}(X, Y) = 0$ and we get

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Similarly, when X and Y are independent, the result for the difference $X - Y$ is

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y).$$

7.4.5 Correlation Co-efficient of Random Variables. Let X and Y be two r.v.'s with non-zero σ_x^2 and σ_y^2 . Then the correlation coefficient which is a measure of linear relationship between X and Y (the Greek letter rho) or $\text{Corr}(X, Y)$, is defined as

$$\rho_{xy} = \frac{E[X - E(X)][Y - E(Y)]}{\sigma_x \sigma_y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

If X and Y are independent r.v.'s, then ρ_{xy} will be zero but zero correlation does not necessarily indicate independence. The correlation coefficient has the following properties:

Correlation co-efficient is unitless and symmetric in X and Y , i.e., $\rho_{xy} = \rho_{yx}$.

Correlation co-efficient remains unchanged if constants are added to the r.v.'s or if the r.v.'s are multiplied by constants having the same sign.

Correlation co-efficient is a number between -1 and $+1$ inclusive. To prove this property, we standardize the r.v.'s X and Y as

$$Z_x = \frac{X - E(X)}{\sqrt{\text{Var}(X)}} \text{ and } Z_y = \frac{Y - E(Y)}{\sqrt{\text{Var}(Y)}}$$

It is obvious that $E(Z_1) = E(Z_2) = 0$ and $\text{Var}(Z_1) = \text{Var}(Z_2) = 1$.

$$\text{Now } \text{Var}(Z_1 + Z_2) = \text{Var}(Z_1) + \text{Var}(Z_2) + 2 \text{Cov}(Z_1, Z_2)$$

$$\text{But } \text{Cov}(Z_1, Z_2) = E(Z_1 Z_2) - E(Z_1) E(Z_2)$$

$$= E(Z_1 Z_2) \quad [\because E(Z_1) = E(Z_2) = 0]$$

$$= E\left[\frac{[X - E(X)][Y - E(Y)]}{\sqrt{\text{Var}(X)\text{Var}(Y)}}\right] = \rho$$

$$\text{Thus } \text{Var}(Z_1 + Z_2) = 1 + 1 + 2\rho \quad [\because \text{Var}(Z_1) = \text{Var}(Z_2) = 1]$$

$$= 2(1 + \rho)$$

Since $\text{Var}(Z_1 + Z_2)$ must be non-negative, therefore it follows that

$$2(1 + \rho) \geq 0 \text{ which implies that } \rho \geq -1.$$

Similarly, $\text{Var}(Z_1 - Z_2) = 2(1 - \rho)$, which implies that $\rho \leq 1$.

Hence from these two results, we get

$$-1 \leq \rho \leq 1.$$

Example 7.22: From the following joint PMF of X and Y , find $\text{Var}(X)$, $\text{Var}(Y)$, $\text{Cov}(X, Y)$ and

$X \backslash Y$	0	1	2	3	$g(x)$
0	0.25	0.05	0.10	0	0.20
1	0.05	0.10	0.25	0.10	0.30
2	0	0.15	0.30	0.05	0.30
$g(y)$	0.10	0.30	0.45	0.15	1.00

$$\text{Now } E(X) = \sum x_i g(x_i) = 0 \times 0.20 + 1 \times 0.30 + 2 \times 0.30$$

$$= 0 + 0.30 + 0.60 = 1.10$$

$$E(Y) = \sum y_j g(y_j) = 0 \times 0.10 + 1 \times 0.30 + 2 \times 0.45 + 3 \times 0.15$$

$$= 0 + 0.30 + 0.90 + 0.45 = 1.65$$

$$E(X^2) = \sum x_i^2 g(x_i) = 0 \times 0.20 + 1 \times 0.50 + 4 \times 0.30 = 1.70$$

$$E(Y^2) = \sum y_j^2 g(y_j) = 0 \times 0.10 + 1 \times 0.30 + 4 \times 0.45 + 9 \times 0.15 = 3.45$$

$$\text{Thus } \text{Var}(X) = E(X^2) - [E(X)]^2 = 1.70 - (1.10)^2 = 0.49, \text{ and}$$

$$\text{Var}(Y) = E(Y^2) - [E(Y)]^2 = 3.45 - (1.65)^2 = 0.7275$$

$$\begin{aligned} \Rightarrow E(XY) &= \sum_{i,j} (x_i y_j) f(x_i, y_j) \\ &= 1 \times 0.10 + 2 \times 0.15 + 2 \times 0.25 + 4 \times 0.10 + 3 \times 0.10 + 6 \times 0.05 \\ &= 0.10 + 0.30 + 0.50 + 0.40 + 0.30 + 0.30 = 1.90 \end{aligned}$$

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 1.90 - 1.10 \times 1.65 = 0.085, \text{ and}$$

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{0.085}{\sqrt{(0.49)(0.7275)}} = \frac{0.085}{0.535} = 0.14$$

Example 7.23 If $f(x, y) = x^2 + \frac{xy}{3}$, $0 \leq x \leq 1, 0 < y \leq 2$
 $= 0, \quad \text{elsewhere.}$

Find $\text{Var}(X)$, $\text{Var}(Y)$ and $\text{Cov}(X, Y)$.

The marginal p.d.f.'s are

$$g(x) = \int_0^2 \left(x^2 + \frac{xy}{3} \right) dy = 2x^2 + \frac{x}{3}, \text{ and}$$

$$h(y) = \int_0^1 \left(x^2 + \frac{xy}{3} \right) dx = \frac{1}{3} + \frac{y}{6}$$

$$\text{Now } E(X) = \int_0^1 x g(x) dx = \int_0^1 x \left(2x^2 + \frac{x}{3} \right) dx = \frac{13}{18},$$

$$E(Y) = \int_0^2 y h(y) dy = \int_0^2 y \left(\frac{1}{3} + \frac{y}{6} \right) dy = \frac{10}{9}.$$

Thus $\text{Var}(X) = E[X - E(X)]^2 = \int_0^1 [x - \mu_x]^2 g(x) dx$:

$$= \int_0^1 \left(x - \frac{13}{18} \right)^2 \left(2x^2 + \frac{x}{3} \right) dx = \frac{73}{1620}$$

$$\text{Var}(Y) = E[Y - E(Y)]^2 = \int_0^2 [y - \mu_y]^2 h(y) dy$$

$$= \int_0^2 \left(y - \frac{10}{9} \right)^2 \left(\frac{1}{3} + \frac{y}{6} \right) dy = \frac{26}{81}, \text{ and}$$

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

$$= \int_0^2 \int_0^1 \left(x - \frac{13}{18} \right) \left(y - \frac{10}{9} \right) \left(x^2 + \frac{xy}{3} \right) dy dx$$

$$= \left[\left(-\frac{2}{9}x^2 + \frac{25}{81}x^2 - \frac{26}{162}x \right) dx = -\frac{1}{162} \right]$$

$$\text{Hence } \text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{-1/162}{\sqrt{(73/1620)(26/81)}} \\ = -0.05$$

7.7 MEDIAN AND MODES OF CONTINUOUS RANDOM VARIABLES

If X is a discrete r.v., then a value ' a ' that satisfies the inequalities

$$P(X \leq a) \geq \frac{1}{2}, P(X \geq a) \geq \frac{1}{2}$$

is called a median. If the r.v. X is continuous, then median is a value of X that satisfies the condition $F(x) = \frac{1}{2}$. In other words, the median ' a ' is given by

$$\int_{-\infty}^a f(x) dx = \int_a^\infty f(x) dx = \frac{1}{2} \int_{-\infty}^\infty f(x) dx = \frac{1}{2}$$

Similarly, the two quantiles Q_1 and Q_3 are defined by the relations

$$\int_{-\infty}^{Q_1} f(x) dx = \frac{1}{4} \text{ and } \int_{Q_3}^\infty f(x) dx = \frac{1}{4}$$

The mode in case of a continuous r.v. X is such a stationary value of $f(x)$ for which the probability density function $f(x)$ attains its maximum. That is we get a maximum value when

$$(i) \quad \frac{d}{dx} f(x) = 0, \text{ i.e. } f'(x) = 0 \text{ and } (ii) \quad f''(x) < 0,$$

provided that the solution of $f'(x) = 0$ lies within the given range of the r.v. X .

It should also be noted that in case of continuous r.v. X , the geometric mean, G , and the harmonic mean, H , are defined as

$$\ln G = E(\ln X) = \int_{-\infty}^\infty \ln x f(x) dx, \text{ provided the integral exists, and}$$

$$\frac{1}{H} = E\left(\frac{1}{X}\right) = \int_{-\infty}^\infty \frac{1}{x} f(x) dx.$$

Example 7.24 Let X be a r.v. with p.d.f.

$$f(x) = k(x - x^2) \quad 0 \leq x \leq 1, \\ = 0, \quad \text{elsewhere,}$$

where k is a constant. Then find the mean, median, mode, harmonic mean and standard deviation.

First of all, we find the value of k which should be such as to make

$$\int_0^1 k(x - x^2) dx = 1$$

That is $k \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_0^1 = 1$ or $k \left(\frac{1}{2} - \frac{1}{3} \right) = 1$ or $k = 6$.

Now, the mean, μ or $E(X)$ is given by

$$\begin{aligned}\mu = E(X) &= \int_0^1 x \cdot 6(x - x^2) dx \\ &= 6 \int_0^1 (x^2 - x^3) dx = 6 \left[\frac{x^3}{3} - \frac{x^4}{4} \right]_0^1 = \frac{1}{2}\end{aligned}$$

The median, a , is given by $\int_0^a f(x) dx = \frac{1}{2}$. Thus

$$6 \int_0^a (x - x^2) dx = \frac{1}{2} \text{ or } 6 \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_0^a = \frac{1}{2}$$

$$\therefore 4a^3 - 6a^2 + 1 = 0$$

Factorize the equation $4a^3 - 6a^2 + 1 = 0$

$$(2a - 1)(2a^2 - 2a - 1) = 0$$

Now either $2a - 1 = 0$ which gives $a = \frac{1}{2}$, or

$$2a^2 - 2a - 1 = 0, \text{ which gives } a = \frac{1 \pm \sqrt{3}}{2}, \text{ i.e. } a = -0.366 \text{ or } 1.366.$$

But the values -0.366 and 1.366 are unacceptable since a must lie in the interval $(0, 1)$. The median is given by $a = \frac{1}{2}$.

Now what is that value of x for which

$$f'(x) = 0, \text{ and } (ii) f''(x) < 0$$

$$f'(x) = 6(x - x^2), \text{ and } f''(x) = 6(1 - 2x)$$

$$f''(x) = 0 \text{ when } 1 - 2x = 0 \text{ or } x = \frac{1}{2}$$

To check that this is maximum, we find

$$f''(x) = 6(-2) = -12 < 0$$

i.e. $f''(x)$ is negative for all values of x , there is a maximum at $x = \frac{1}{2}$.

Hence the mode $= \frac{1}{2}$.

The harmonic mean, H , is given by

$$\begin{aligned} \frac{1}{H} &= \int_{-\infty}^{\infty} \frac{1}{x} \cdot 6(x-x^2) dx \\ &= 6 \int_{-\infty}^{\infty} \left(\frac{1}{x} - x \right) dx = 6 \left[x - \frac{x^2}{2} \right]_{-\infty}^{\infty} = 3 \end{aligned}$$

$$H = \frac{1}{3}$$

$$\text{Again, } \mu'_2 = E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx$$

$$= \int_{-\infty}^{\infty} x^2 \cdot 6(x-x^2) dx = 6 \left[\frac{x^4}{4} - \frac{x^3}{3} \right]_{-\infty}^{\infty} = \frac{3}{10}$$

$$\mu_2 = E(X^2) - [E(X)]^2 = \mu'_2 - \mu_1^2$$

$$= \frac{3}{10} - \left(\frac{1}{3} \right)^2 = \frac{1}{20} = 0.05$$

$$\text{Hence } \sigma = \sqrt{\mu_2} = \sqrt{0.05} = 0.2234.$$

7.8 CHEBYSHEV'S INEQUALITY

If X is a r.v. having mean μ and variance $\sigma^2 > 0$, and k is any positive constant, the probability that a value of X falls within k standard deviations of the mean is at least $\left(1 - \frac{1}{k^2}\right)$. Then

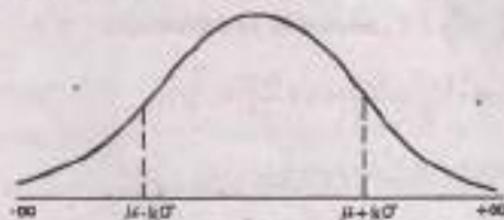
$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2},$$

or equivalently $P(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2}$.

Proof. By definition, we have

$$\sigma^2 = E(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx,$$

Dividing the range into 3 disjoint parts $(-\infty, \mu - k\sigma]$, $(\mu - k\sigma, \mu + k\sigma)$ and $(\mu + k\sigma, \infty)$, we have



$$\sigma^2 = \int_{-\infty}^{\mu-k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu+k\sigma}^{\infty} (x - \mu)^2 f(x) dx$$

Moving the middle term, we get

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu+k\sigma}^{\infty} (x - \mu)^2 f(x) dx$$

$|x - \mu| \geq k\sigma$, we have $(x - \mu)^2 \geq k^2\sigma^2$, whenever $x \geq \mu + k\sigma$

$\leq \mu - k\sigma$. It therefore follows that

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x) dx + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(x) dx$$

$$= \int_{-\infty}^{\mu-k\sigma} f(x) dx + \int_{\mu+k\sigma}^{\infty} f(x) dx$$

$$P(\mu - k\sigma < X < \mu + k\sigma) = \int_{\mu-k\sigma}^{\mu+k\sigma} f(x) dx \geq 1 - \frac{1}{k^2}$$

This inequality is due to Russian mathematician P.L. Chebyshev (1821–1894) and it provides a understanding how the variance measures variability about the mean of a r.v. It holds for all r.v.s with finite mean and variance. In case of a discrete r.v., the proof is the same with integrals replaced by summations.

MOMENT GENERATING FUNCTION

The moment generating function (m.g.f) usually denoted by $M(t)$, of a random variable X about the mean exists, is defined as the expected value of the r.v. e^{tX} , where t is a real variable lying in a neighborhood of zero. That is

$$M_0(t) = E(e^{tX}) = \sum_{i=1}^{\infty} e^{tx_i} f(x_i), \text{ if } X \text{ is discrete,}$$

$$= \int e^{tx} f(x) dx, \text{ if } X \text{ is continuous.}$$

Since $e^{\theta} = 1 + \theta + \frac{\theta^2}{2!} + \frac{\theta^3}{3!} + \dots$, therefore for the discrete case

$$\begin{aligned} M_0(t) &= \sum \left[1 + tx_i + \frac{(tx_i)^2}{2!} + \frac{(tx_i)^3}{3!} + \dots + \frac{(tx_i)^r}{r!} + \dots \right] f(x_i) \\ &= \sum f(x_i) + t \sum x_i f(x_i) + \frac{t^2}{2!} \sum x_i^2 f(x_i) + \frac{t^3}{3!} \sum x_i^3 f(x_i) + \dots + \frac{t^r}{r!} \sum x_i^r f(x_i) + \dots \\ &= 1 + tE(X) + \frac{t^2}{2!} E(X^2) + \frac{t^3}{3!} E(X^3) + \dots + \frac{t^r}{r!} E(X^r) + \dots \\ &= 1 + t\mu'_1 + \frac{t^2}{2!} \mu'_2 + \frac{t^3}{3!} \mu'_3 + \dots + \frac{t^r}{r!} \mu'_r + \dots \\ &\sim \sum_{i=1}^{\infty} \frac{t^r}{r!} \mu'_r. \end{aligned}$$

Thus we find that the coefficient of $\frac{t^r}{r!}$ in the expansion of $M_0(t)$ is just $E(X^r)$ or μ'_r .

moments about zero. We call the function $M_0(t)$ moment generating function because it generates all the moments of the r.v. X . It should be noted that a m.g.f. would exist only if the sum or integral converges for all values of t in a neighbourhood of zero.

It is more convenient to find the moments by differentiating the m.g.f. r times w.r.t. to t and putting $t = 0$. The m.g.f. of X is

$$M_0(t) = 1 + tE(X) + \frac{t^2}{2!} E(X^2) + \dots + \frac{t^r}{r!} E(X^r) + \dots$$

Differentiating w.r.t. to t , we get

$$\frac{dM_0(t)}{dt} = E(X) + tE(X^2) + \frac{t^2}{2!} E(X^3) + \dots + \frac{t^{r-1}}{(r-1)!} E(X^r) + \dots$$

and $\left. \frac{dM_0(t)}{dt} \right|_{t=0} = E(X) = \mu'_1$

Differentiating it again, we obtain

$$\frac{d^2 M_0(t)}{dt^2} = E(X^2) + tE(X^3) + \dots + \frac{t^{r-2}}{(r-2)!} E(X^r) + \dots$$

and $\left. \frac{d^2 M_0(t)}{dt^2} \right|_{t=0} = E(X^2) = \mu'_2$

Differentiating t times and equating $t = 0$, we get

$$\frac{d^r M_x(t)}{dt^r} \Big|_{t=0} = E(X^r) = \mu_r$$

∴ m.g.f. about the value a is defined as the expected value of $e^{(x-a)t}$ and is denoted as

$$M_a(t) = E\left[e^{(x-a)t}\right] = e^{-at} E(e^{xt}) = e^{-at} M_x(t)$$

∴ the m.g.f. about the value a is equal to e^{-at} times the m.g.f. about the origin.

Similarly, we make a change in the scale, and define the m.g.f. by introducing a new variable u as

$$u_i = \frac{x_i - a}{h} \text{ so that } x_i = a + hu_i$$

$$\text{then } M_a(t) = \sum_i e^{itx_i} f(x_i) = \sum_i e^{i(t+a+hu_i)} f(x_i) = e^{it} \sum_i e^{ihu_i} f(x_i)$$

$$\text{--- taking } \frac{t}{h} \text{ for } t \text{ in both sides, we get } M_a(t) = e^{-at/h} M_x\left(\frac{t}{h}\right)$$

The m.g.f. has a very important property, namely the m.g.f. of the sum of independent random variables is the product of the individual m.g.f.'s. Let us consider two independent r.v.'s X and Y . Then

$$E[e^{(X+Y)}] = E[e^X e^Y] = (E[e^X]) (E[e^Y])$$

If the variables have identical probability distributions, then we have

$$\text{m.g.f. of } \sum_{i=1}^n (\text{m.g.f.}(X))^n$$

7.9.1 Cumulant Generating Function. The cumulants are a set of parameters of a probability distribution defined by the following identity in t :

$$\exp \left[\sum_{r=1}^{\infty} \frac{K_r t^r}{r!} \right] = \sum_{r=1}^{\infty} \mu_r \frac{t^r}{r!}$$

K_r is the r th cumulant. In other words, the cumulants are given by the co-efficients in the expansion of a power series from the natural logarithm of the m.g.f. of a random variable, provided such expansion exists. Thus

$$K(t) = \log_e M_x(t)$$

$$= K_1 t + K_2 \frac{t^2}{2!} + K_3 \frac{t^3}{3!} + \dots + K_r \frac{t^r}{r!} + \dots$$

The co-efficient $\kappa_1, \kappa_2, \kappa_3, \dots$ are the first, the second, the third cumulant, etc. and $\kappa(t)$ is called the *cumulant generating function (c.g.f.)*. The c.g.f. possesses a very important property that the c.g.f. of the sum of a number of independent random variables is the sum of their cumulant generating functions.

We differentiate the above relation r times with respect to t and then put $t = 0$ to find the cumulant. That is

$$\kappa_r = \left[\frac{d^r}{dt^r} \log_e M_x(t) \right]_{t=0}$$

7.9.2 Relation between Cumulants and Moments. By definition, we have

$$\kappa(t) = \log_e M_x(t) = \log_e \left[\sum_{r=0}^{\infty} \mu'_r \frac{t^r}{r!} \right]$$

$$\kappa_1 t + \kappa_2 \frac{t^2}{2!} + \dots + \kappa_r \frac{t^r}{r!} + \dots = \log_e \left[1 + \sum_{r=1}^{\infty} \frac{\mu'_r}{r!} t^r \right]$$

$$= \log_e (1 + x), \text{ where } x = \sum_{r=1}^{\infty} \mu'_r \frac{t^r}{r!}$$

$$= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

$$= \left[\sum_{r=1}^{\infty} \mu'_r \frac{t^r}{r!} \right] - \frac{1}{2} \left[\sum_{r=1}^{\infty} \mu'_r \frac{t^r}{r!} \right]^2 + \frac{1}{3} \left[\sum_{r=1}^{\infty} \mu'_r \frac{t^r}{r!} \right]^3 -$$

$$= \left[\mu'_1 t + \mu'_2 \frac{t^2}{2!} + \mu'_3 \frac{t^3}{3!} + \mu'_4 \frac{t^4}{4!} + \dots \right] - \frac{1}{2} \left[\mu'_1 t + \mu'_2 \frac{t^2}{2!} + \mu'_3 \frac{t^3}{3!} + \dots \right]^2 +$$

$$- \frac{1}{3} \left[\mu'_1 t + \mu'_2 \frac{t^2}{2!} + \dots \right]^3 - \frac{1}{4} \left[\mu'_1 t + \dots \right]^4 + \dots$$

$$= \mu'_1 t + (\mu'_2 - \mu'_1)^2 \frac{t^2}{2!} + (\mu'_3 - 3\mu'_2\mu'_1 + 2\mu'_1^3) \frac{t^3}{3!} + \dots$$

Comparing the co-efficients of like powers of t , we get

$$\kappa_1 = \mu'_1,$$

$$\kappa_2 = \mu'_2 - \mu'_1^2 = \mu_2,$$

$$\kappa_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'_1^3 = \mu_3$$

$$\begin{aligned}\kappa_4 &= \mu_4' - 4\mu_3'\mu_1' - 3\mu_2'^2 + 12\mu_2'\mu_1'^2 - 6\mu_1'^3 \\ &= \mu_4 - 3\mu_2^2\end{aligned}$$

19.3 Characteristic Function. The *M.g.f.* does not exist for many probability distributions. We have another function, called the *characteristic function (c.f.)*. The *characteristic function* of a r.v. X , denoted by $\phi(t)$, is defined as the expected value of the r.v. e^{tX} , i.e.

$$\begin{aligned}\phi(t) &= E(e^{tX}) \\ &= \sum e^{tx} f(x) \text{ or } \int e^{tx} f(x) dx\end{aligned}$$

depending as X is discrete or continuous, and where t is a real number and $i = \sqrt{-1}$, the imaginary unit. The characteristic function always exists because $|e^{itx}| = 1$ for all real t , and hence may be defined for any probability distribution. The characteristic function has therefore an advantage over the moment generating function. The c.f. may be written as a series

$$\phi(t) = 1 + it\mu_1' + \frac{(it)^2}{2!} \mu_2' + \frac{(it)^3}{3!} \mu_3' + \dots + \frac{(it)^n}{n!} \mu_n' + \dots$$

Thus the k th moment of X about the origin is the co-efficient of $\frac{(it)^k}{k!}$. Applications of these are given in the chapters that follow.

EXERCISES

TESTIVE

Answer 'True' or 'False'. If the statement is not true then replace the undefined words with words to make the statement true:

A random variable can assume only one value with a given probability.

A random variable that can assume a finite set of possible values is known as a continuous random variable. 0

A random variable that can assume any value in a given interval $[a, b]$ is known as discrete random variable.

The expected value for any discrete random variable is always equal to $\sum(x - \mu)^2 p(x)$.

The variance for any discrete random variable is $\sum xp(x)$.

Discrete random variables may assume only positive values.

The expected value of a constant is zero.

- viii) The expected value of the product of two any random variables is equal to the product of their expected values.
- ix) Sum of probabilities for any probability distribution is equal to zero.
- x) For any random variable having mean μ and variance σ^2 and $k > 1$, then the probability that a value of X falls within k standard deviation of the mean is less than $\left(1 - \frac{1}{k^2}\right)$.

b) MULTIPLE CHOICE QUESTIONS

- i) A random variable is also known as a
- Chance variable.
 - Stochastic variable.
 - Variate.
 - All of above.
- ii) The distribution function of a random variable X , denoted by $F(x)$ is defined as
- $F(x) = P(X \leq x)$
 - $F(x) = P(X > x)$
 - $F(x) = P(X = x)$
 - None of above.
- iii) A discrete probability distribution may be represented by
- A table.
 - A mathematical function.
 - A histogram.
 - All of above.
- iv) A continuous probability distribution is not represented by
- A table.
 - A mathematical function.
 - A graph.
 - A density function.
- v) If X and Y are two independent random variables, then $\text{Var}(X - Y)$ is equal to
- $\text{Var}(X) - \text{Var}(Y)$
 - $\text{Var}(X) + \text{Var}(Y)$
 - $\text{Var}(X) + \text{Var}(Y) - 2 \text{COV}(X, Y)$
 - None of above.

- If X is a random variable and a and b are constants, then $\text{Var}(aX + b)$ is equal to

- a) $a^2 \text{Var}(X)$
- b) $\text{Var}(aX)$
- c) $a^2 \text{Var}(X) + \text{Var}(b)$
- d) $a \text{Var}(X)$

Given the following distribution for a random variable X :

x_i	1	2	3	4	5	6	Total
$f(x_i)$	0.10	0.20	0.20	0.25	0.15	0.10	1.00

the standard deviation of X is

- a) 2.000
- b) 1.4654
- c) 3.5064
- d) 2.1475

Suppose X has a p.d. given by

x_i	-2	-1	0	Total
$f(x_i)$	2α	3α	α	1.00

then α is

- a) 0.1000
- b) 0.1111
- c) 0.2000
- d) None of above.

If X and Y are two random variables, then $E(X + Y)$ is equal to

- a) $E(X) + E(Y)$
- b) $E(X) + Y$
- c) $E(X) - E(Y)$
- d) None of above

Two discrete random variables X and Y are independent, which of the following statements is untrue?

- a) $P(X = 4) = P(X = 4 | Y = 2)$
- b) $P(X = 4 \text{ and } Y = 2) = P(X = 4) P(Y = 2)$
- c) $P(X = 4 \text{ and } Y = 2) \neq P(X = 4) P(Y = 2)$
- d) $P(X = 2) = P(X = 2 | Y = 4)$

SUBJECTIVE

- 7.1 a) Explain the concept of a random variable. What is a distribution function and what are its properties?

- b) Let X be a random variable denoting the number of points appearing in a throw of a die. Determine the distribution function $F(x)$, x is a real number and draw its graph.

- 7.2 a) Define a discrete random variable and its probability distribution. What are the properties of all probability distributions?

- b) Suppose X has a p.d. given by

	-1	0	1
$f(x)$	$3e^{-x}$	$3e^{-x}$	$6e^{-x}$

- (i) Determine c . (ii) What is the p.d. of $Y=2X+1$?

- c) Determine the p.d. of a r.v. X , where X denotes the number of aces in a hand of bridge.

- 7.3 a) Define a discrete r.v. Giving illustrations, explain what it means by a discrete p.d.

- b) A bag contains 4 red and 6 black balls. A sample of 4 balls is selected from the bag with replacement. Let X be the number of red balls. Find the p.d. for X . (P.U., B.A.B.Sc.)

- 7.4 a) A large store places in last 15 clock radios in a clearance sale. Unknown to any one, 3 radios are defective. If a customer buys 4 different clock radios selected at random, find the p.d. of X = number of defective radios in the sample?

- b) 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 p.d. of X .

- 7.5 a) Explain the concept of distribution function. Hence or otherwise, differentiate between discrete and continuous random variables.

- b) Given the following probability function

$$f(x) = \begin{cases} 4x - 2x^2 & \text{for } 0 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$$

Calculate the value of A so as $f(x)$ may be a p.d.f. (P.U., B.A.B.Sc.)

- 7.6 a) Define a continuous r.v. and its probability density function.

- b) A continuous r.v. X has the p.d.f. as follows:

$$f(x) = \begin{cases} x/2 & \text{for } 0 < x \leq 1 \\ \frac{1}{4}(3-x) & \text{for } 1 < x \leq 2 \\ \frac{1}{4} & \text{for } 2 < x \leq 3 \\ \frac{1}{4}(4-x) & \text{for } 3 < x \leq 4 \\ 0, & \text{elsewhere.} \end{cases}$$

Compute $P(X \geq 3)$, $P(X = 2)$, $P(|X| < 1.5)$ and $P(1 < X < 3)$.

CONTINUOUS VARIABLES

- i) Explain the concepts of the Probability Function, Probability Density Function and Distribution Function
 (P.U., B.A./B.Sc. 1990, 92)

- ii) A continuous r.v. X has the p.d.f.

$$f(x) = 4(2-x)(2+x), \quad 0 \leq x \leq 2 \\ = 0, \quad \text{elsewhere.}$$

Find (i) the value of A , (ii) $P(X = \frac{1}{2})$, (iii) $P(X \leq 1)$, (iv) $P(X \geq 2)$, (v) $P(1 \leq X \leq 2)$

(P.U., B.A./B.Sc. 1993)

Let X be a continuous r.v. with p.d.f.

$$f(x) = 6x(1-x), \quad 0 \leq x \leq 1. \\ = 0, \quad \text{otherwise.}$$

- i) Check that $f(x)$ is a p.d.f. and sketch it.

- ii) Obtain an expression for the distribution function of X .

- iii) Compute $P\left(\frac{1}{3} < X < \frac{2}{3}\right)$ and $P\left(X \leq \frac{1}{2} \mid \frac{1}{3} < X < \frac{2}{3}\right)$

- iv) Determine a number b such that $P(X < b) = 2P(X > b)$

(P.U., M.Sc. 1994)

Suppose that the life length (in hours) of a certain type of tube is continuous random variable X with probability density function

$$f(x) = \frac{100}{x^2}, \quad x > 100 \quad \text{and zero elsewhere.}$$

What is the probability that a tube will last less than 200 hours, if it is known that the tube is still functioning after 150 hours of service?

What is the probability that if 3 such tubes are installed in a set exactly one will have to be replaced after 150 hours of service?

What is the maximum number of tubes that may be inserted into a set so that there is a probability of 0.5 that after 150 hours of service all of them are still functioning?

(P.U., B.A. (Hons.) Part-I, 1970)

Explain the terms: a joint distribution function, a joint probability distribution, marginal and conditional distributions.

Given the joint p.d. of two r.v.'s X and Y , whose values $f(x, y)$

$$\text{are: } f(1, 1) = \frac{4}{30}, \quad f(1, 2) = \frac{1}{30}, \quad f(1, 3) = \frac{1}{30}, \quad f(2, 1) = \frac{4}{30},$$

$$f(2, 2) = \frac{5}{30}, \quad f(2, 3) = \frac{1}{30}, \quad f(3, 1) = \frac{2}{30}, \quad f(3, 2) = \frac{6}{30},$$

$f(3, 3) = \frac{6}{30}$, find all the marginal and conditional distributions.

- 7.11 Suppose that the following table represents the joint p.d. of the discrete r.v. (X, Y):

$X \backslash Y$	1	2	3
1	$\frac{1}{12}$	$\frac{1}{6}$	0
2	0	$\frac{1}{9}$	$\frac{1}{5}$
3	$\frac{1}{18}$	$\frac{1}{4}$	$\frac{2}{15}$

- a) Compute $g(x), h(y), f(x|y)$ and $f(y|x)$.
 b) Decide whether X and Y are independent.

- 7.12 Let X and Y have the joint probability functions given by

i) $f(x, y) = \frac{xy^2}{66}, \quad x = 2, 4, 5; y = 1, 2, 3.$

ii) $f(x, y) = \frac{xy^2}{30}, \quad x = 1, 2, 3; y = 1, 2.$

Find the marginal probability functions of X and Y , and find out if X and Y are independent. (P.U., B.A., B.Sc.)

- 7.13 a) What is a joint probability density function? How does a marginal probability function differ from a conditional probability function?
 b) Given the joint p.d.f. $f(x, y) = xy(x+y)$, $0 \leq x \leq 1, 0 \leq y \leq 1$, and 0 elsewhere, find the marginal and conditional p.d.f.s.

- 7.14 Suppose the joint p.d.f. of (X, Y) is given by

$$f(x, y) = x^2 + \frac{1}{2}, \quad 0 \leq x \leq 1, 0 \leq y \leq 2 \text{ and } 0 \text{ elsewhere.}$$

- a) Check that $\int \int f(x, y) dx dy = 1$.

- b) Compute (i) $P\left(X > \frac{1}{2}\right)$, (ii) $P(Y < X)$, (iii) $P(X+Y > 1)$ and (iv) $P\left(Y < \frac{1}{2} \mid X < \frac{1}{2}\right)$.

- 7.15 a) Explain clearly the meaning of marginal and conditional distributions with reference to bivariate density function $f(x, y)$.
 b) If two r.v.s X and Y have the joint p.d.f.

$$f(x, y) = \begin{cases} \frac{2}{3}(x+2y) & \text{for } 0 < x < 1, 0 < y < 1, \\ 0 & \text{elsewhere;} \end{cases}$$

then find the marginal distributions of X and Y , and their conditional distributions:

$$P\left(X < \frac{1}{2} \mid Y = \frac{1}{2}\right)$$

- Given the joint density function of the r.v. (X, Y) as

$$f(x, y) = 3x^2y + 3xy^2, \quad \text{for } 0 \leq x \leq 1, 0 \leq y \leq 1, \\ = 0, \quad \text{otherwise.}$$

Find the marginal and conditional density functions. Also find the conditional probability

$$P\left[\frac{1}{2} \leq X \leq \frac{3}{4} \mid \frac{1}{2} \leq Y \leq \frac{2}{3}\right] \quad (\text{P.U., B.A. (Hons.) Part-III, 1964, 65, 68, 70})$$

Explain the idea of joint probability distribution, conditional distribution and marginal distribution. The random variables X and Y are jointly distributed as

$$f(x, y) = 24x^2y(1-x), \quad 0 \leq x, y \leq 1.$$

Obtain the marginal distribution of X , and the conditional distribution of Y . Are X and Y independent? (P.U., M.Sc. 1970; B.Sc. Hon. Part-II, 1972)

- a) Define $E(X)$, the expected value of a random variable X .

- b) If X is a r.v. and if a and b are constants, then prove that

$$E(aX + b) = aE(X) + b.$$

- c) If X and Y are r.v.'s, then show that $E(X+Y) = E(X) + E(Y)$

- d) Show that, under certain conditions to be specified,

$$E(XY) = E(X)E(Y).$$

- e) Explain random variable and its mathematical expectation.

- f) Prove that $E(cX) = cE(X)$, where c is a constant.

- Two unbiased dice are cast. A payment equal to the sum of the spots on the top sides is given the caster. Compute the expected value of the payment.

- g) A committee of size 3 is to be selected at random from 3 women and 5 men. Find the expected number of women in the committee. (P.U., B.A./B.Sc. 1979)

- h) Let X have the possible values

$$2, -2, \frac{8}{3}, -4, \dots, (-1)^{j+1} \frac{2^j}{j}, \text{ and the corresponding probabilities } \frac{1}{2}, \frac{1}{4}, \frac{1}{3}, \frac{1}{16}, \dots, \left(\frac{1}{2}\right)^j, \dots$$

- Find the expectation of X if it exists.

(P.U., B.A. Hons. 1970)

- i) Define expectation and prove that the expectation of the sum of two random variables is equal to the sum of their expectations. (P.U., B.A./B.Sc. 1979, 81)

- b) Verify that $E(X) + E(Y) = E(X+Y)$ by using the random variable X with the p.d. $f(x)$

$x = 1, 2, 3, 4$; and the r.v. Y with the p.d. $f(y) = \begin{cases} 3 \\ y \end{cases} \left(\frac{1}{2}\right)^y \left(\frac{1}{2}\right)^{3-y}$, $y = 0, 1, 2, 3$.

(Goud, B.A./B.Sc.)

- 7.22 a) The p.d. of a discrete random variable X is

$$f(x) = \begin{cases} 3 \\ x \end{cases} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{3-x}, \quad x = 0, 1, 2, 3$$

Find $E(X)$ and $E(X^2)$.

(P.U., B.A./B.Sc.)

- b) Let X be a random variable with probability distribution

y	-1	0	1	2	3
$f(x)$	0.125	0.50	0.20	0.05	0.125

i) Find $E(X)$ and $\text{Var}(X)$.

ii) Find the p.d. of the r.v. $T = 2X+1$. Using the p.d. of T determine $E(T)$ and $\text{Var}(T)$.

iii) How are $E(T)$ and $E(Y)$, and $\text{Var}(X)$ and $\text{Var}(Y)$ related?

- 7.23 a) For what value of A , the function defined below will be a p.d.f?

$$f(x) = Ax^2(1-x), \quad 0 \leq x \leq 1 \\ = 0, \quad \text{otherwise}$$

- b) Find its mean and variance.

- c) Also find $P\left(\frac{1}{4} < X < \frac{1}{2}\right)$ using its distribution function.

(P.U., B.A./B.Sc.)

- 7.24 a) Show that $E(X^2 - 3X + 5) = E(X^2) - [E(X)]^2$.

- b) Let X_1 and X_2 be two independent r.v.'s having variances 3 and 2 respectively. If $\text{Var}(3X_1 - X_2) = 25$, find k .

- c) The following table shows the distribution function $F(x)$ of the r.v. X :

$x:$	$x \leq 1$	$x \leq 2$	$x \leq 3$	$x \leq 4$
$F(x):$	1/8	3/8	3/4	1

Find (i) probability distribution of the r.v. X , (ii) $E(X)$ and $\text{Var}(X)$.

(P.U., B.A./B.Sc.)

- 7.25 a) If $f(x) = \frac{6-(7-x)}{36}$ for $x = 2, 3, \dots, 12$, then find the mean and variance of the variable X .

- b) If $f(x) = \frac{1}{n}$, ($x = 1, 2, \dots, n$), then find $E(X)$ and $\text{Var}(X)$.

- i) Suppose X can take the values $0, 1, 2, \dots, n$ with frequencies proportional to the binomial co-efficients $\binom{n}{0}, \binom{n}{1}, \dots, \binom{n}{n}$. Show that $E(X) = \frac{n}{2}$ and $V(X) = \frac{n}{4}$

(P.U., M.A., 1970; B.A. (Hons.) Part-II, 1970)

- Practicals
- i) A and B throw with one die for a prize of Rs. 11, which is to be won by the player who first throws 6. If A has the first throw, what are their respective expectations?
- ii) A, B, C and D cut a pack of cards successively in the order mentioned. If the person who cuts a spade first, receives £ 175, what are their expectations?
- iii) A bag contains 2 white and 2 black balls. Three persons A, B and C in the order named above draw a ball and do not replace it. The person who draws a white ball first receives Rs. 18. What are their expectations?

(P.U., B.A./B.Sc., 1986)

- iv) A distribution is given by $f(x) = x^2(1-x)$ between $x=0$ and $=1$. Find the mean and standard deviation.
- v) A continuous r.v. X has p.d.f. given by $f(x) = cx$ for $1 < x < 2$.
- (i) Determine the constant c . (ii) Find the mean, variance and standard deviation of X .

(P.U., B.A./B.Sc., 1990)

A variable X has the p.d.f.

$$\begin{aligned}f(x) &= 0, && \text{for } x < 0 \\&= \frac{3}{4}(x-2)^2, && \text{for } 0 < x < 2 \\&= 0, && \text{for } x \geq 2.\end{aligned}$$

Find the expected value of X and its standard deviation.Find the value of k so that the function $f(x)$ defined as follows, may be a density function:

$$\begin{aligned}f(x) &= kx^2(1-x), && 0 < x < 1, \\&= 0, && \text{otherwise}\end{aligned}$$

Also determine its mean and variance.

(P.U., B.A./B.Sc., 1978)

A r.v. X has the p.d.f. as

$$\begin{aligned}f(x) &= Ax(9-x^2), && 0 \leq x \leq 3 \\&= 0, && \text{otherwise}\end{aligned}$$

Find the value of A , the mean and the standard deviation of X .

(P.U., B.A./B.Sc., 1990)

- 7.29 A r.v. X has the probability density function given by

$$f(x) = \begin{cases} \frac{1}{16}(3+x)^2, & -3 < x \leq -1 \\ \frac{1}{16}(5-2x^2), & -1 < x \leq 1 \\ \frac{1}{16}(3-x)^2, & 1 < x \leq 3 \\ 0, & \text{elsewhere.} \end{cases}$$

Find the mean and the standard deviation of the r.v. X .

(P.U., B.Sc. Hon. Part-I)

- 7.30 Find k , the mode and the mean for the distribution, the equation of whose p.d.f. is $f(x) = kx^2$ for $0 \leq x \leq 3$. Also determine its variance. (P.U., B.A./B.Sc.)

- 7.31 A continuous r.v. X has the p.d.f. given by

$$f(x) = k(2-x)(x-5), \quad 2 \leq x \leq 5 \\ = 0, \quad \text{elsewhere.}$$

Find the value of k , mean and variance. What are the values of the mode and median of the distribution of X ?

- 7.32 Let X be a.r.v. with p.d.f.

$$f(x) = 6(2-x)(x-1), \quad 1 \leq x \leq 2 \\ = 0, \quad \text{elsewhere.}$$

Then show that the geometric mean of x is given by

$$6 \log (16/7) = 19. \quad (\text{P.U., B.A. Hon. Part-I})$$

- 7.33 Explain what is meant by 'skewness of a distribution', and define a suitable measure of skewness. Use this definition to calculate the skewness of the distribution defined by the density function

$$f(x) = kx^2(1-x)^n, \quad 0 \leq x \leq 1 \text{ and } 0, \quad \text{otherwise,}$$

where k is a normalizing constant to be determined. It may be assumed that

$$\int_0^1 x^m (1-x)^n dx = \frac{m! n!}{(m+n+1)!} \quad \left(\text{using } \int_0^1 x^m (1-x)^n dx = \frac{m! n!}{(m+n+1)!} \right)$$

- 7.34 a) The rectangular distribution is given by $y = 1$ between $x = -a$ and $x = a$. Find the mean, variance and standard deviation.

- b) The p.d.f. of a rectangular distribution is

$$f(x) = 1, \quad \text{for } -a \leq x \leq a.$$

Obtain the first four mean moments and obtain also the mean deviation of the distribution function. (P.U., B.A. Hon. Part-I)

- 33 Let X be a continuous random variable with p.d.f. given by

$$f(x) = \begin{cases} \frac{x}{2}, & 0 \leq x \leq 1 \\ \frac{1}{2}, & 1 \leq x \leq 2 \\ \frac{(3-x)}{2}, & 2 \leq x \leq 3 \\ 0, & \text{elsewhere.} \end{cases}$$

Find the mean, the variance and the moment measure of kurtosis of X .

- Find the first four moments about the mean of the distribution $f(x) = x^2(6-x)^2$ between $x=0$ and $x=6$, and the kurtosis of the distribution.

- Let X and Y have the joint p.d.f. described as follows:

(x, y)	(1, 1)	(1, 2)	(1, 3)	(2, 1)	(2, 2)	(2, 3)
$f(x, y)$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{4}{15}$

and $f(x, y)$ is equal to zero elsewhere. Find the two marginal p.d.f.s and the correlation co-efficient ρ .

- Prove that the correlation co-efficient between two r.v.'s must be a number between -1 and +1 inclusive.
- Given $f(x, y) = 2 - x - y$, $0 \leq x \leq 1$, $0 \leq y \leq 1$, and 0 elsewhere. Find co-efficient of correlation between X and Y .
- State and prove the Chebyshev's inequality. Explain the significance of this inequality.
- Show that the variance of the sum (or difference) of two independent r.v.'s is equal to the sum of their variances.
- Define Moment Generating Function and Characteristic Function of a random variable X .
- Show that the m.g.f. of the sum of two independent r.v.'s is the product of their m.g.f.'s.
(P.U., B.A., B.Sc. 1993)

Two fair dice are rolled. Find the probability distribution of minimum of two numbers.

A bag contains 2 white and 3 black balls. Four persons A, B, C and D in the order named above, each draw a ball and does not replace it. The person who draws a white ball gets Rs. 10. What are their respective expectations?

(P.U., B.A., B.Sc. 2016)

Define:

- Probability Function
- Distribution Function

- Probability Density Function

- b) A random variable X has the moment generating function about origin as

$$M_X(t) = (1 - 3t)^{-4}$$

Obtain the mean and variance of random variable X .

- c) The annual gross earnings of a certain pop-singer are a random variable with an expected value of Rs.40,00,000/- and a standard deviation of Rs.8,00,000/-. The singer's manager receives 15 percent of this amount. Determine the expected value and the standard deviation of the amount received by the manager.

(P.U., B.A./B.Sc. 2010)

CHAPTER 8

**DISCRETE
PROBABILITY
DISTRIBUTIONS**

DISCRETE PROBABILITY DISTRIBUTIONS

1.1 INTRODUCTION

As discussed in the previous chapter, a discrete probability distribution gives the probability of every possible value of a discrete random variable. We shall introduce here some important discrete probability distributions which are often used in statistical theory and analysis.

1.2 BINOMIAL PROBABILITY DISTRIBUTION

Many experiments consist of repeated independent trials, each trial having only two possible complementary outcomes. For example, the two possible outcomes of a trial may be head and tail, success and failure, right and wrong, alive and dead, good and defective, infected and not infected and so on. If the probability of each outcome remains the same throughout the trials then such trials are called Bernoulli trials and the experiment having a Bernoulli trials is called binomial experiment. In other words, an experiment is called a binomial probability experiment if it possesses the following four qualities:

- i) The outcomes of each trial may be classified into one of two categories, conventionally called Success (S) and Failure (F). It is to be noted that the outcome of interest is called a success and the other, a failure.
- ii) The probability of success, denoted by p , remains constant for all trials.
- iii) The successive trials are all independent.
- iv) The experiment is repeated a fixed number of times, say n .

When X denotes the number of successes in n trials of a binomial probability experiment, X is a binomial random variable and its p.d. is called the *Binomial Probability Distribution*. The r.v. X can only take on anyone of the $(n+1)$ integer value $0, 1, 2, \dots, n$.

When the binomial r.v. X assumes a value x , the binomial p.d. is given by:

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, 2, \dots, n,$$

where $q = 1 - p$, the probability of failure on each trial. The binomial p.d. has two parameters n and p and is usually denoted by $b(x; n, p)$. The binomial probability distribution is appropriate when a random sample of size n is drawn with replacement from a finite population of size N , or sampling is done from infinite population.

The binomial p.d., which is the most widely used distribution in two-outcome situations, was derived by the Swiss mathematician Jakob Bernoulli (1654–1704) whose main work on probability, *Conjectural* (the art of conjecturing) was published posthumously in Basel in 1713.

1.2.1 Deviation of Binomial Probability Distribution. To derive a formula that gives the probability of successes in n trials for a binomial experiment, we proceed as follows:

The experiment has n trials, each of which may result in S or F . The sample space has 2^n possible points of outcomes, each outcome consisting of a sequence $\{a_1, a_2, \dots, a_n\}$, where each a_i is S or F . We desire to find the probability of these outcomes according to the number of successes.

First, we consider the probability of zero success, i.e. $P(X = 0)$. In case of zero success, every trial is F and the event consists of a sequence of n F 's, i.e. $\{FF \dots F\}$. Because in each trial, $P(F) = p$,

and $P(F) = q$ and trials are independent, so we apply the multiplicative law of probability for independent events and obtain

$$\begin{aligned} P(FF\ldots F) &= P(F)P(F)\ldots P(F) \quad [\text{n times}] \\ &= q^n \end{aligned}$$

Since there is only one sequence of outcomes of n trials resulting in F^n , therefore

$$P(X=0) = q^n$$

Next, we consider the probability of one success, i.e., $P(X=1)$. In this case, one trial results in S and the remaining $(n-1)$ trials result in F 's. The event consisting of one S and $(n-1) F$'s can occur in n different sequences. One such sequence is $\{SFF\ldots F\}$ and the probability for this sequence is pq^{n-1} . Another possible sequence is $\{FFSF\ldots F\}$ and the probability for this sequence is the same as the first sequence. In other words, the probability for any possible sequence consisting of one S and $(n-1) F$'s is pq^{n-1} . But the number of mutually exclusive sequences in which one S and $(n-1) F$'s can occur is n .

Therefore the probability of exactly one success for all possible sequences combined, is

$$P(X=1) = \binom{n}{1} pq^{n-1}$$

The above argument may be repeated for $X = 2, 3, 4, \dots$

Finally, we consider the general case, i.e., when $X=x$. The probability of a sequence that has x successes and $(n-x)$ failures is $p^x q^{n-x}$ and there are $\binom{n}{x}$ different sequences in which x successes and $(n-x)$ failures can occur. Therefore the probability of x success in n trials is

$$P(X=x) = \binom{n}{x} p^x q^{n-x} \quad \text{for } x = 0, 1, 2, \dots, n$$

Thus we have obtained the formula for the binomial probability distribution having a success probability p for success. The binomial probability distribution derives its name from the fact that probabilities $\binom{n}{x} p^x q^{n-x}$, for $x = 0, 1, 2, \dots, n$ are the successive terms of the binomial expansion $(q+p)^n$. That is

$$\begin{aligned} (q+p)^n &= \sum \binom{n}{x} p^x q^{n-x}, \quad \text{where } x = 0, 1, 2, \dots, n \\ &= \binom{n}{0} q^n + \binom{n}{1} q^{n-1} p + \binom{n}{2} q^{n-2} p^2 + \dots + \binom{n}{n} p^n \\ &= b(0; n, p) + b(1; n, p) + b(2; n, p) + \dots + b(n; n, p) \end{aligned}$$

The sum of probabilities, i.e., $\sum_{x=0}^n \binom{n}{x} p^x q^{n-x}$ or $\sum_{x=0}^n b(x; n, p) = 1$ because $q+p=1$. This is a necessary condition for any probability distribution. It is to be noted that the probability

succes is given by $P(X \leq r) = \sum_{x=0}^r \binom{N}{x} p^x q^{N-x}$, where $P(X \leq r)$ is the cumulative binomial distribution of X . There are tables for the cumulative probabilities $P(X \leq r) = \sum_{x=0}^r M(x; n, p)$ for some values of n , p and r .

Example 8.1: A fair coin is tossed 5 times. Find the probabilities of obtaining various numbers of heads.

Let us regard the tossing of a coin as an experiment. Then we observe that

- each toss of a coin (i.e. each trial) has two possible outcomes, heads (success) and tails (failure);
- the probability of a head (success) is $p = \frac{1}{2}$ and remains the same for successive tosses;
- the successive tosses of the coin are independent; and
- the coin is tossed 5 times.

Therefore the r.v. X which denotes the number of heads (successes) has a binomial probability distribution with $p=1/2$ and $n=5$. The possible value of X are 0, 1, 2, 3, 4 and 5. Hence

$$P(\text{no head}) = P(X=0) = \binom{5}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

$$P(1 \text{ head}) = P(X=1) = \binom{5}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^4 = 5 \times \left(\frac{1}{2}\right)^5 = \frac{5}{32}$$

$$P(2 \text{ heads}) = P(X=2) = \binom{5}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = 10 \times \left(\frac{1}{2}\right)^5 = \frac{10}{32}$$

$$P(3 \text{ heads}) = P(X=3) = \binom{5}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = 10 \times \left(\frac{1}{2}\right)^5 = \frac{10}{32}$$

$$P(4 \text{ heads}) = P(X=4) = \binom{5}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^1 = 5 \times \left(\frac{1}{2}\right)^5 = \frac{5}{32}, \text{ and}$$

$$P(5 \text{ heads}) = P(X=5) = \binom{5}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0 = 1 \times \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

These probabilities can also be obtained by expanding the binomial $\left(\frac{1}{2} + \frac{1}{2}\right)^5$. The binomial distribution for the number of heads obtained in 5 tosses of a fair coin is

x	0	1	2	3	4	5
$f(x)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$

Example 8.2 An event has the probability $p = \frac{3}{8}$. Find the complete binomial distribution for $n = 5$ trials.

Hence $p = \frac{3}{8}$ so that $q = 1 - p = \frac{5}{8}$; and $n = 5$.

Hence the desired probabilities are the successive terms in the binomial expansion of $\left(\frac{3}{8} + \frac{5}{8}\right)^5$:

$$\left[\left(\frac{3}{8}\right)^0 + \binom{5}{1} \left(\frac{3}{8}\right)^1 \left(\frac{5}{8}\right)^4 + \binom{5}{2} \left(\frac{3}{8}\right)^2 \left(\frac{5}{8}\right)^3 + \binom{5}{3} \left(\frac{3}{8}\right)^3 \left(\frac{5}{8}\right)^2 + \binom{5}{4} \left(\frac{3}{8}\right)^4 \left(\frac{5}{8}\right)^1 + \left(\frac{3}{8}\right)^5 \right]$$

i.e. $\frac{1}{(8)^5} [1(5)^0 + 5(5)^1(3) + 10(5)^2(3)^2 + 10(5)^3(3)^3 + 5(5)^4(3)^4]$

i.e. $\frac{1}{32768} [3125 + 9375 + 11250 + 6750 + 3025 + 2]$

i.e. $[0.0954 + 0.2861 + 0.3433 + 0.2060 + 0.0618 + 0.0074]$

We can now write these probabilities in the form of a probability table as below:

x	0	1	2	3	4	5
$P(X=x)$	0.0954	0.2861	0.3433	0.2060	0.0618	0.0074

Example 8.3 Let X have a binomial distribution with $n = 4$ and $p = \frac{1}{3}$. Find $P(X=1)$, $P(X=2)$, $P(X=3)$, $P(X=6)$ and $P(X \leq 2)$.

The binomial probability distribution for $n = 4$ and $p = \frac{1}{3}$, is

$$f(x) = \binom{4}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{4-x} \quad \text{for } x = 0, 1, 2, 3, 4$$

Now $P(X=1) = \binom{4}{1} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^{4-1} = \frac{32}{81}$

$P\left(X = \frac{3}{2}\right) = f\left(\frac{3}{2}\right) = 0$; because a r.v. X with a binomial distribution takes only one of the integer values $0, 1, 2, \dots, n$.

$$P(X=3) = f(3) = \binom{4}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^{4-3} = \frac{8}{81}.$$

$P(X=6) = f(6) = 0$, because X can take only values $0, 1, 2, 3, 4$.

$$\begin{aligned} P(X \leq 2) &= \sum_{x=0}^2 f(x) = f(0) + f(1) + f(2) \\ &= \binom{4}{0} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^4 + \binom{4}{1} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^3 + \binom{4}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^2 \\ &= \frac{16}{81} + \frac{32}{81} + \frac{24}{81} = \frac{72}{81} = \frac{8}{9}. \end{aligned}$$

Example 8.4 A and B play a game in which A's probability of winning is $2/3$. In a series of 8 games, what is the probability that A will win (i) exactly 4 games, (ii) at least 4 games, (iii) 6 or more, and (iv) from 3 to 6 games?

We observe that

- (i) there are two possible outcomes, i.e. A will win or will not win the game;
- (ii) the probability of A's winning in each game is $p = 2/3$;
- (iii) the successive games are independently won or lost; and
- (iv) there are 8 games.

Therefore the Binomial probability distribution with $n = 8$ and $p = 2/3$ is appropriate. Let X denote the number of games won by A. Then

$$P(X=4) = \binom{8}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^{8-4} = \frac{1120}{6561} = 0.1707$$

$$P(X \geq 4) = 1 - P(X < 4); \quad (\because \text{at least 4 means 4 or more})$$

$$\begin{aligned} &= 1 - \sum_{x=0}^3 \binom{8}{x} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{8-x} \\ &= 1 - \left[\left(\frac{1}{3}\right)^8 + 8 \left(\frac{2}{3}\right) \left(\frac{1}{3}\right)^7 + 28 \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^6 + 56 \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^5 \right] \\ &= 1 - \frac{1}{6561} [1 + 16 + 112 + 448] \\ &= 1 - \frac{577}{6561} = \frac{5984}{6561} = 0.9121 \end{aligned}$$

$$\text{iii) } P(X \geq 6) = \sum_{x=6}^8 \binom{8}{x} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{8-x}$$

$$= \binom{8}{6} \left(\frac{2}{3}\right)^6 \left(\frac{1}{3}\right)^2 + \binom{8}{7} \left(\frac{2}{3}\right)^7 \left(\frac{1}{3}\right) + \binom{8}{8} \left(\frac{2}{3}\right)^8$$

$$= \frac{64}{6561} [28 + 16 + 4] = \frac{64 \times 48}{6561} = \frac{1024}{2187} \approx 0.4682$$

$$\text{iv) } P(3 \leq X \leq 6) = \sum_{x=3}^6 \binom{8}{x} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{8-x}$$

$$= \binom{8}{3} \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^5 + \binom{8}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^4 + \binom{8}{5} \left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)^3 + \binom{8}{6} \left(\frac{2}{3}\right)^6 \left(\frac{1}{3}\right)^2$$

$$= \frac{2^3}{(3)^3} [56 + 140 + 224 + 224]$$

$$= \frac{8 \times 644}{6561} = \frac{5152}{6561} \approx 0.7852$$

Example 8.5 The experience of a house-agent indicates that he can provide accommodation for 75 percent of the clients who come to him. If on a particular occasion, approach him independently, calculate the probability that (i) less than 4 clients, (ii) exactly 4 (iii) at least 5 clients, will get satisfactory accommodation.

We observe that

- a) there are two possible outcomes, i.e. each client will get or will not get accommodation,
- b) on each occasion, probability of getting accommodation is $p = 3/4$,
- c) clients approach the house-agent independently, and
- d) there are 6 clients.

Therefore the binomial probability distribution with $n=6$ and $p=3/4$ is appropriate to find the desired probabilities.

Let X denote the number of clients who get satisfactory accommodation. Then we need to find

- (i) $P(X < 4)$, (ii) $P(X=4)$ and (iii) $P(X \geq 5)$. Hence,

$$\text{i) } P(X < 4) = \sum_{x=0}^3 \binom{6}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{6-x}$$

$$= \left(\frac{1}{4}\right)^6 + \binom{6}{1} \left(\frac{3}{4}\right) \left(\frac{1}{4}\right)^5 + \binom{6}{2} \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^4 + \binom{6}{3} \left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right)^3$$

$$= \left(\frac{1}{4}\right)^6 [1 + (6)(3) + (15)(9) + (20)(27)]$$

$$= \frac{694}{4096} = 0.169$$

$$(i) P(X=4) = \binom{6}{4} \left(\frac{1}{4}\right)^4 \left(\frac{3}{4}\right)^2 = \frac{15 \times 81}{(4)^6} = \frac{1215}{4096} = 0.297$$

$$(ii) P(X \geq 5) = \sum_{n=5}^6 \binom{6}{n} \left(\frac{1}{4}\right)^n \left(\frac{3}{4}\right)^{6-n}$$

$$= \binom{6}{5} \left(\frac{3}{4}\right)^5 \left(\frac{1}{4}\right) + \binom{6}{6} \left(\frac{3}{4}\right)^6$$

$$= \left(\frac{1}{4}\right)^6 [(6)(3)^5 + (3)^6] = \frac{2187}{4096} = 0.534$$

8.2.2 Binomial Frequency Distribution. If the binomial probability distribution is multiplied by a number of experiments or sets, the resulting distribution is known as the *binomial frequency distribution*. Thus the expected frequency of x successes in N experiments is $N \cdot \binom{n}{x} p^x q^{n-x}$. It should be noted that the n independent trials constitute one experiment or one set.

Example 8.6 Six dice are thrown 729 times. How many times do you expect at least three dice to show a five or a six?

The probability of getting a five or a six with one die is $p = 2/6$. Since 6 dice are thrown and there are two outcomes, the binomial frequency distribution is given by

$$729 \left(\frac{2}{3} + \frac{1}{3}\right)^6$$

Hence the expected number of times at least 3 dice showing 5 or 6

$$= 729 \sum_{x=3}^6 \binom{6}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{6-x}$$

$$= 729 \left[\binom{6}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^3 + \binom{6}{4} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^2 + \binom{6}{5} \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right) + \binom{6}{6} \left(\frac{1}{3}\right)^6 \right]$$

$$= \frac{729}{(3)^6} [160 + 60 + 12 + 1] = 233.$$

Example 8.7 A certain event is believed to follow the binomial distribution. In 1024 samples of size 5, the result was observed once 405 times and twice 270 times. Find p and q.

The first three term in the expansion of the Binomial Frequency Distribution $Nq^n + N$ corresponding to $x = 0, 1$, and 2 are Nq^n , $N\binom{n}{1}q^{n-1}p$ and $N\binom{n}{2}q^{n-2}p^2$.

We are given $N = 1024$, $n = 5$ and the following information:

$$1024 \binom{5}{1} q^{5-1} p = 405,$$

$$1024 \binom{5}{2} q^{5-2} p^2 = 270,$$

Dividing the second equation by the first, we get

$$\frac{10q^4 p^2}{5q^4 p} = \frac{270}{405} \text{ or } \frac{2p}{q} = \frac{2}{3}$$

$$\text{or } 3p = q \text{ or } 3p = 1 - p \text{ or } 4p = 1$$

$$\text{Hence } p = \frac{1}{4} \text{ and } q = \frac{3}{4}$$

4.2.3 Properties of the Binomial Probability Distribution. The properties of the binomial distribution include the mean number of successes, the variance of the number of successes, skewness and kurtosis, etc. and the shape of the distribution. Some of the properties are described below.

- Let X be a random variable with the binomial distribution $B(x; n, p)$. Then its mean and variance are given by $\mu = np$ and $\sigma^2 = npq$ respectively.

Now Mean, $\mu = E(X)$

$$= \sum_{x=0}^n x \binom{n}{x} p^x q^{n-x}, \text{ where } x = 0, 1, 2, \dots, n.$$

$$= 0 \cdot q^n + 1 \left(\binom{n}{1} q^{n-1} p + 2 \binom{n}{2} q^{n-2} p^2 + \dots + np^n \right)$$

$$= np(q^{n-1} + \binom{n-1}{1} q^{n-2} p + \binom{n-1}{2} q^{n-3} p^2 + \dots + p^{n-1})$$

$$= np(q + p)^{n-1}$$

$$= np, \text{ because } q + p = 1.$$

Alternative Method.

$$\text{Mean} = E(X) = \sum_{x=0}^n x \binom{n}{x} p^x q^{n-x}$$

$$\text{But } \binom{n}{x} = \frac{n(n-1)(n-2)\dots(n-x+1)}{x(x-1)(x-2)\dots(2)(1)} = \binom{n-1}{x-1}$$

$$E(X) = n \sum_{x=1}^n \binom{n-1}{x-1} p^x q^{n-x}, \text{ for } x = 1, 2, \dots, n \text{ (since the first term in the summation being zero } (x=0) \text{ is omitted).}$$

$$= np \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} q^{(n-1)-(x-1)}$$

Substituting $y = x - 1$ and $m = n - 1$ in the summation, we get

$$E(X) = np \sum_{y=0}^m \binom{m}{y} p^y q^{m-y} \quad (\text{as } y \text{ ranges from 1 to } n, \text{ so } y = (x-1) \text{ must range from 0 to } n-1 \text{ i.e., } m)$$

$$= np \quad (\because \text{ summation is the expansion of } (q+p)^m)$$

Hence mean = np. In other words, the mean number of successes is np. Similarly the mean number of failures is np.

By definition, the variance σ^2 , is given by

$$\sigma^2 = E[X - \mu]^2 = E[X^2] - [E(X)]^2$$

$$\text{But } E(X^2) = E[X(X-1)] + E[X(X-1)] + E(X)$$

$$= E[X(X-1)] + np$$

$$\begin{aligned} \text{Now } E[X(X-1)] &= \sum_{x=0}^n x(x-1) \binom{n}{x} p^x q^{n-x} \\ &= \sum_{x=0}^n x(x-1) \frac{n(n-1)(n-2)\dots(n-x+1)}{x(x-1)(x-2)\dots(2)(1)} p^x p^{x-2} q^{n-x} \\ &= n(n-1)p^2 \sum_{x=2}^n \frac{(n-2)!}{(x-2)(n-x)!} p^{x-2} q^{n-x}. \end{aligned}$$

(x starts at 2 since x = 0, 1 add nothing to the sum)

The terms $(n-x)$ may be written as $[(n-2) - (x-2)]$.

Substituting $y = x - 2$ and $m = n - 2$ in the summation, we obtain

$$\begin{aligned}
 E[X(X-1)] &= n(n-1)p^2 \sum_{j=0}^{n-2} \frac{n!}{j!(n-1-j)!} p^j q^{n-j} \\
 &= n(n-1)p^2 \sum_{j=0}^{n-2} \binom{n}{j} p^j q^{n-j} \\
 &= n(n-1) p^2 \quad (\text{as summation is 1})
 \end{aligned}$$

$$\begin{aligned}
 \text{Thus } \sigma^2 &= E(X^2) - [E(X)]^2 \\
 &= E[X(X-1)] = E(X) - [E(X)]^2 \\
 &= n(n-1)p^2 + np - (np)^2 \\
 &= np^2 - np^2 + np - np^2 \\
 &= np - np^2 = np(1-p) = npq, \text{ and} \\
 \sigma &= \sqrt{npq}.
 \end{aligned}$$

Hence the variance of the number of successes is npq , and the standard deviation is \sqrt{npq} .

2) Higher Moments of the distribution are found below:

By definition, the moments about the origin are given by the relation

$$\mu_r' = E(X^r)$$

$$\text{Now } \mu_1' = E(X) = np$$

$$\mu_2' = E(X^2) = n(n-1)p^2 + np$$

$$\mu_3' = E(X^3)$$

$$\text{But } X^3 = (X)(X-1)(X-2) + 3X(X-1) + X,$$

$$\therefore E(X^3) = E[X(X-1)(X-2)] + 3E[X(X-1)] + E(X)$$

$$\text{Now } E(X) = \sum_{x=0}^n x \binom{n}{x} p^x q^{n-x} = np,$$

$$\begin{aligned}
 3E[X(X-1)] &= 3 \sum_{x=0}^n x(x-1) \binom{n}{x} p^x q^{n-x} \\
 &= 3n(n-1)p^2, \text{ and}
 \end{aligned}$$

$$E[X(X-1)(X-2)] = \sum_{x=0}^n x(x-1)(x-2) \binom{n}{x} p^x q^{n-x}$$

$$\text{But } x(x-1)(x-2) \binom{n}{x} = x(x-1)(x-2) \frac{x(n-1)(n-2)(n-3)!}{(x-1)(x-2)(x-3)(n-x)!}$$

$$= n(n-1)(n-2) \binom{n-3}{x-3}$$

$$E[X(X-1)(X-2)] = n(n-1)(n-2) \sum_{x=3}^n \binom{n-3}{x-3} p^x q^{n-x} \quad (\text{x starts at 3 since } x=0, 1, 2 \text{ add nothing})$$

to the summation!

Substituting $y = x - 3$ and $m = y + 3$ in the summation, we have

$$E[X(X-1)(X-2)] = n(n-1)(n-2) p^3 \sum_{y=0}^m \binom{n}{y} p^y q^{m-y}$$

$$= n(n-1)(n-2)p^3$$

$$\mu'_3 = n(n-1)(n-2) p^3 + 3n(n-1)p^2 + np$$

$$\mu'_4 = E(\lambda^4)$$

$$= \sum_{x=0}^n x^4 \binom{n}{x} p^x q^{n-x}$$

$x^4 \equiv x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x$ and proceeding as above, we get

$$\mu'_4 = n(n-1)(n-2)(n-3) p^4 + 6n(n-1)(n-2)(n-3)p^3 + 7n(n-1)p^2 + np$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^2$$

$$= [n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np] - 3np(p^2p^2 - np^2 + np) + 2n^2p^3$$

$$= np[1 - 3p + 2p^2] \cdot np(1-p)(1-2p)$$

$$= npq(q-p);$$

$$\mu_4 = \mu'_4 - 4\mu'_2\mu'_3 + 6(\mu'_1)^2(\mu'_2) - 3(\mu'_1)^4$$

Substitution and simplification give

$$\begin{aligned} \mu_4 &= n^4[p^4 - q^4] + n^3[-6p^4 + 6p^3 + 6p^2 - 6p] + \\ &\quad n^2[11p^4 - 18p^3 + 7p^2 - 4p^2 + 12p^3 - 8p^4] + n[-6p^4 + 12p^3 - 7p^2 + p] \\ &= 3p^2p^2(1-p)^2 + np(1-p)(1-6p+6p^2) \\ &= 3p^2p^2q^2 + npq(1-6pq) = npq[1 + 3(n-2)pq]. \end{aligned}$$

$$\text{Now } \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{[(npq)(q-p)]^2}{(npq)^3} = \frac{(q-p)^2}{npq} = \frac{(1-2p)^2}{npq},$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3n^2 p^2 q^2 + npq(1-6pq)}{(npq)^2} = 3 + \frac{1-6pq}{npq}$$

Hence moment co-efficient of skewness and kurtosis are respectively $\frac{1-2p}{\sqrt{npq}}$ and $3 + \frac{1-6pq}{npq}$.

- 3) The shape of the binomial probability distribution depends on the values of the two parameters p and n . The sketches indicate the influence of p and n on the shape of the distribution:

$b(x; 5, 0.2)$



$b(x; 5, 0.4)$



$b(x; 5, 0.6)$



$b(x; 10, 0.3)$



$b(x; 10, 0.5)$



Thus we observe that, when $p < \frac{1}{2}$, the distribution is positively skewed and when $p > \frac{1}{2}$, the distribution becomes negatively skewed. In general, when $p \neq q$, the distribution will be more the difference between p and q , the greater the skewness will be. When $p = \frac{1}{2}$, the distribution is always symmetrical. As n , the number of trials, increases to ∞ , $\beta_1 \rightarrow 0$ and $\beta_2 \rightarrow 3$. Hence as n , the binomial probability distribution is symmetrical and mesokurtic.

Example 8.8 In the B.A. Examination, 24 candidates offered Statistics. If the probability of passing the subject be $\frac{1}{3}$, find the mean and the dispersion of the distribution.

Here $n = 24$ and $p = \frac{1}{3}$ so that $q = 1 - p = \frac{2}{3}$.

Hence mean $= np = 24 \times \frac{1}{3} = 8$; and

$$\sigma^2 = npq = 24 \times \frac{1}{3} \times \frac{2}{3} = \frac{16}{3} = 5.33.$$

8.2.4 The Recurrence Formula for the Binomial Distribution. Beginning with the value of $P(X=0)$, probabilities for other values of X , the number of successes can be computed more easily by the recurrence formula

$$P(X=x) = \frac{n-x+1}{x} \cdot \frac{p}{q} \cdot P(X=x-1)$$

established as follows:

Let X be a random variable with the binomial probability distribution $\delta(x; n, p)$. Then

$$P(X=x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \text{ and}$$

$$P(X=x-1) = \binom{n}{x-1} p^{x-1} q^{n-x+1} = \frac{(x-1)!(n-x+1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x+1}.$$

Dividing $P(X=x)$ by $P(X=x-1)$, we get

$$\begin{aligned} \frac{P(X=x)}{P(X=x-1)} &= \frac{n!}{x!(n-x)!} \cdot \frac{(x-1)!(n-x+1)!}{x!} \cdot \frac{p^x q^{n-x}}{p^{x-1} q^{n-x+1}} \\ &= \frac{n-x+1}{x} \cdot \frac{p}{q} \end{aligned}$$

$$\text{Hence } P(X=x) = \frac{n-x+1}{x} \cdot \frac{p}{q} \cdot P(X=x-1) \quad \text{for } x = 1, 2, \dots, n.$$

Although computations can be carried out more easily by this method, great care must be taken for accuracy.

Example 8.9 If X is binomially distributed with mean 3.20 and variance 1.152, find the complete probability distribution.

If X is a binomial r.v. with parameters n and p , then $E(X) = np$ and $\text{Var}(X) = npq$.

$$\text{Now } E(X) = 3.20 \text{ so } np = 3.20,$$

$$\text{and } \text{Var}(X) = 1.152 \text{ so } npq = 1.152$$

Substituting for np in the second equation, we obtain

$$(3.20)q = 1.152$$

$$\therefore q = \frac{1.152}{3.20} = 0.36 \text{ so } p = 1 - q \approx 0.64$$

$$\text{and } n(0.36) = 3.20 \text{ gives } n = 5.$$

Therefore $n = 5$ and $p = 0.64$ so that X is $\delta(x; 5, 0.64)$.

$$\text{Now } P(X=x) = \binom{5}{x} (0.64)^x (0.36)^{5-x},$$

$$\text{and } P(X=0) = (0.36)^5 = 0.006047$$

Beginning with the value of $P(X = 0)$, we compute the probabilities of other values of X , using a recurrence formula.

$$P(X = x) = \frac{n - x + 1}{x} \cdot \frac{p}{q} P(X = x - 1)$$

$$= \frac{n - x}{x} \left(\frac{0.64}{0.36} \right) P(X = x - 1)$$

$$P(X = 1) = \frac{5}{1} \left(\frac{0.64}{0.36} \right) (0.006047) = 0.053751,$$

$$P(X = 2) = \frac{4}{2} \left(\frac{0.64}{0.36} \right) (0.053751) = 0.191115,$$

$$P(X = 3) = \frac{3}{3} \left(\frac{0.64}{0.36} \right) (0.191115) = 0.339760,$$

$$P(X = 4) = \frac{2}{4} \left(\frac{0.64}{0.36} \right) (0.339760) = 0.302009,$$

$$P(X = 5) = \frac{1}{5} \left(\frac{0.64}{0.36} \right) (0.302009) = 0.141051$$

The sum of probabilities turns out to be 0.99993 instead of 1. Error has been introduced in rounding process.

8.2.5 Fitting a Binomial Distribution to Observed Data. Fitting a binomial distribution given frequency distribution consists of (i) estimating the values of p and n , the two parameters completely determine a binomial distribution, and (ii) calculating the probabilities as well as the expected frequencies of $x = 0, 1, 2, \dots$. Assuming that the given frequency distribution has the characteristics of a fitted theoretical binomial distribution, we calculate the mean of the observed frequency distribution \bar{x} and taking it as the estimate of μ , we equate it to its theoretical value, i.e. np , where n is the number of trials and p is the largest x value given. Having found the value of p , we compute the expected frequencies. This procedure is illustrated by the following example.

Example 8.10 Fit a binomial distribution to the following data, obtained by tossing a biased coin 200 times:

No. of Heads	0	1	2	3	4	5	Total
Frequency	12	50	74	39	18	1	200

(P.U., B.A. (Hons.), 1998)

To fit a binomial distribution, we need to find n and p . Hence $n = 5$, the largest x -value. Now we use the relationship $\bar{x} = np$.

$$\text{Now } \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{0 + 20 + 148 + 117 + 72 + 3}{200} = \frac{398}{200} = 1.99$$

Using the relation $\bar{x} = np$, we get $1.99 = 1.99$ or $p = 0.398$.

Letting r.v. X represent the number of heads, gives the fitted binomial distribution as

$$b(x; 5, 0.398) = \binom{5}{x} (0.398)^x (0.602)^{5-x}$$

Now the probabilities and frequencies are calculated as below:

No. of head (x)	Probability $f(x)$	Expected frequency
0	$\binom{5}{0} q^5 = (0.602)^5$	= 0.07907
1	$\binom{5}{1} q^4 p = 5(0.602)^4 (0.398)$	= 0.23726
2	$\binom{5}{2} q^3 p^2 = 10(0.602)^3 (0.398)^2$	= 0.34559
3	$\binom{5}{3} q^2 p^3 = 10(0.602)^2 (0.398)^3$	= 0.22847
4	$\binom{5}{4} qp^4 = (0.602)(0.398)^4$	= 0.07553
5	$\binom{5}{5} (0.398)^5$	= 0.00998
Total		= 1.00000
		200.0

The expected frequencies are obtained by multiplying each of the probability by 200. These can also be calculated using the binomial recurrence formula.

$$P(X = x) = \frac{x - x + 1}{x} \cdot \frac{p}{q} P(X = x - 1)$$

12.8 Moment Generating and Cumulant Generating Functions of the Binomial Distribution.
The moment generating function of the binomial probability distribution $b(x; n, p)$ is derived as below:

$$M(t) = E(e^{tX})$$

(By definition)

$$= \sum_{x=0}^n x^t \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{n=0}^{\infty} \binom{n}{r} (pe')^r q^{n-r} = (q + pe')^r.$$

The expansion of this binomial is purely algebraic and need not be interpreted in probabilities.

We get the moments by differentiating $M_0(t)$ once, twice, etc. with respect to t and putting $t=0$. Thus

$$\begin{aligned}\mu'_1 &= E(X) = \left[\frac{d}{dt} (q + pe')^r \right]_{t=0} \\&= [nppe' (q + pe')^{r-1}]_{t=0} = np, \text{ and} \\ \mu'_2 &= E(X^2) = \left[\frac{d^2}{dt^2} (q + pe')^r \right]_{t=0} \\&= [nppe' (q + pe')^{r-1}]_{t=0} + [n(n-1)p^2e'^2 (q + pe')^{r-2}]_{t=0} \\&= np + n(n-1)p^2 \\ \therefore \mu'_2 &= \mu'_1 + \mu'_1^2 = npq\end{aligned}$$

In a similar way, the higher moments are obtained.

The cumulant generating function ($\kappa(t)$) is given by

$$\begin{aligned}\kappa(t) &= \log_e M_0(e^t) \\&= \log_e [1 + pe'(e^t)] = n \log_e [q + pe'] \\&= \log_e \left[q + p \left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \dots \right) \right] \\&= n \log_e \left[1 + \left(pt + p \frac{t^2}{2!} + p \frac{t^3}{3!} + \dots \right) \right] (\because q = 1 - p)\end{aligned}$$

Expanding the log and comparing powers in t , we find the first four cumulants to be

$$k_1 = \mu'_1 = np,$$

$$k_2 = \mu'_2 = np(1-p) = npq;$$

$$k_3 = \mu'_3 = np(1-p)(1-2p) = npq(q-p);$$

$$k_4 = np(1-p)(1-6p+6p^2) = npq(1-6pq).$$

Example 8.11 Prove for the binomial distribution, the following relation

$$\mu_{r+1} = pq \left(n\mu_{r-1} + \frac{d\mu_r}{dp} \right)$$

Hence find μ_2 , μ_3 and μ_4 .

(P.U., M.A. (Stats.), 1969)

By definition, the r th moment about the mean is

$$\mu_r = \sum_{j=0}^n (j - np)^r \binom{n}{j} p^j q^{n-j}, \quad \text{where } q = 1 - p$$

$$= -np \sum_{j=1}^n (j - np)^{r-1} \binom{n}{j} p^j q^{n-j} - \sum_{j=1}^n (j - np)^r \binom{n}{j} p^j (n-j) q^{n-j-1} + \sum_{j=1}^n (j - np)^r \binom{n}{j} dp^{j-1} q^{n-j}$$

$$= -np\mu_{r-1} + \sum_{j=1}^n (j - np)^r \binom{n}{j} p^j q^{n-j} \left[-(n-j) \frac{1}{q} + j \frac{1}{p} \right]$$

$$= -np\mu_{r-1} + \frac{1}{qp} \sum_{j=1}^n (j - np)^{r-1} \binom{n}{j} p^j q^{n-j}$$

$$= -np\mu_{r-1} + \frac{1}{qp} \mu_{r+1}.$$

Multiplying both sides by pq and transposing, we get

$$\mu_{r+1} = pq \left(np\mu_{r-1} + \frac{d\mu_r}{dp} \right)$$

Using $r = 1, 2$ and 3 , we get

$$\mu_2 = pq \left(np\mu_0 + \frac{d\mu_1}{dp} \right) = pq [0 + 0] = npq, \quad (\because \mu_0 = 1)$$

$$\mu_3 = pq \left(2np\mu_1 + \frac{d\mu_2}{dp} \right) = pq \left[2n(0) + \frac{d}{dp}(npq) \right] \quad (\because \mu_1 = 0)$$

$$= pq [0 + n(1 - 2p)] = pq(nq - np) = npq(q - p),$$

$$\mu_4 = pq \left[3np\mu_2 + \frac{d\mu_3}{dp} \right]$$

$$= pq \left[3n(npq) + \frac{d}{dp}(np - np^2)(1 - 2p) \right]$$

$$= 3n^2 p^2 q^2 + npq(1 - 6pq).$$

8.3 HYPERGEOMETRIC PROBABILITY DISTRIBUTION

There are many experiments in which the condition of *independence* is violated and the probability of success does not remain constant for all trials. Such experiments are called *hypergeometric experiments*. In other words, a hypergeometric experiment has the following properties:

- The outcomes of each trial may be classified into one of two categories, success and failure.
- The probability of success changes on each trial.
- The successive trials are dependent.
- The experiment is repeated a fixed number of times.

The number of successes, X in a hypergeometric experiment is called a hypergeometric r.v. and its probability distribution is called the *hypergeometric distribution*. When the hypergeometric r.v. X assumes a value x , the hypergeometric p.d. is given by the formula:

$$P(X = x) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}, \text{ for } x \text{ such that } x = 0, 1, 2, \dots, n \text{ and } x = 0, 1, 2, \dots, k.$$

where N = number of units in the set or population (a positive integer)

n = number of units in the subset or sample (a positive integer) and

k = number of successes in the set or population (a non-negative integer less than or equal to N)

The hypergeometric p.d. has three parameters N , n and k , (or N , n and $p = \frac{k}{N}$), and is generally denoted by $H(x; N, n, k)$. The hypergeometric p.d. is appropriate when

- a random sample of size n is drawn without replacement from a finite population of N units.
- k of the units are of one kind (classified as *success*) and the remaining $N - k$ of units are of another kind (classified as *failure*).

8.3.1 Derivation of Hypergeometric Distribution. Suppose a set contains N elements, of which k are classified as *success* and $N - k$ are classified as *failure*; the two classes being mutually exclusive and exhaustive, and we select a subset of n elements ($n \leq N$) from the set without replacement.

Then the total number of ways in which a subset of n elements can be chosen from N is $\binom{N}{n}$.

Let X denote the number of successes and let $X = x$ if and only if we choose x successes and $n - x$ failures from $N - k$ failures in the set. Then the number of ways in which x successes and $n - x$ failures can be chosen, is $\binom{k}{x} \binom{N-k}{n-x}$.

Hence the probability that x of the n elements are successes, is

$$P(X = x) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}, \text{ for } x \text{ such that } 0 \leq x \leq n \text{ and } 0 \leq k \leq n.$$

which is the required formula for the hypergeometric probabilities.

To provide that the sum of the hypergeometric probabilities is one, we use the following useful statistical result (or formula).

$$\sum_{j=0}^m \binom{a}{j} \binom{b}{m-j} = \binom{a+b}{m}$$

by expanding both sides of

$$(1+x)^a (1+x)^b = (1+x)^{a+b}$$

equating the coefficients of x^m .

$$\begin{aligned} \sum_{x=0}^n P(X=x) &= \frac{1}{\binom{N}{n}} \sum_{x=0}^n \binom{k}{x} \binom{N-k}{n-x} \\ &= \frac{1}{\binom{N}{n}} \binom{N-k+k}{n-x+x} = \frac{1}{\binom{N}{n}} \binom{N}{n} = 1. \end{aligned}$$

The hypergeometric distribution derives its name from the fact that probability generating function is given in the form of a hypergeometric series.

Example 8.12 An urn contains 10 red balls and 6 black balls. A sample of 4 balls is selected from it without replacement. Let X be the number of red balls contained in the sample, then find the probability distribution for X .

Here X is a hypergeometric r.v. because

- the results of each draw may be classified as either red (success) or black (failure),
- the probability of success changes on each draw,
- the successive draws are dependent as the selection is made without replacement,
- the drawing is repeated a fixed number of times ($n = 4$).

Now, $N = 4 + 6 = 10$, $k = 4$, $n = 4$ and the possible values of X are 0, 1, 2, 3, and 4. Therefore the probabilities of these possible outcomes are

$$P(X = 0) = h(0; 10, 4, 4) = \frac{\binom{4}{0} \binom{6}{4}}{\binom{10}{4}} = \frac{15}{210}$$

$$P(X=1) = h(1; 10, 4, 4) = \frac{\binom{4}{1} \binom{6}{3}}{\binom{10}{4}} = \frac{80}{210}$$

$$P(X=2) = h(2; 10, 4, 4) = \frac{\binom{4}{2} \binom{6}{2}}{\binom{10}{4}} = \frac{90}{210}$$

$$P(X=3) = h(3; 10, 4, 4) = \frac{\binom{4}{3} \binom{6}{1}}{\binom{10}{4}} = \frac{24}{210}$$

$$P(X=4) = h(4; 10, 4, 4) = \frac{\binom{4}{4} \binom{6}{0}}{\binom{10}{4}} = \frac{1}{210}$$

Hence the hypergeometric p.d. of X is as follows:

X	0	1	2	3	4
$h(x; 10, 4, 4)$	$\frac{15}{210}$	$\frac{80}{210}$	$\frac{90}{210}$	$\frac{24}{210}$	$\frac{1}{210}$

Example 8.13 The names of 5 men and 5 women are written on slips of paper and placed in a box. Four names are drawn. What is the probability that 2 are men and 2 are women?

Let X denote the number of men. Then

$N = 5 + 5 = 10$ names to be drawn from;

$n = 4$, and (here possible values of x are 0, 1, 2, 3, 4, i.e. n)

$K = 5$.

Hence the hypergeometric distribution is

$$h(x; 10, 4, 5) = \frac{\binom{5}{x} \binom{5}{4-x}}{\binom{10}{4}}$$

the required probability i.e. $P(X = 2)$ is

$$h(2; 10, 4, 5) = \frac{\binom{3}{2} \binom{5}{2}}{\binom{10}{4}} = \frac{10}{21}$$

Example 8.14 What is the probability that a poker hand will contain exactly 2 aces?

Let us regard the 4 aces as success and the 48 nonaces as failures. Then we have

$N = 52$, $n = 5$, (number of cards in a poker hand).

$x = 4$ and $s = 2$. (here possible values of x are 0, 1, 2, 3 and 4, i.e. k)

the probability that a poker hand will contain exactly 2 aces is

$$P(X = 2) = h(2; 52, 4, 5) = \frac{\binom{4}{2} \binom{48}{2}}{\binom{52}{5}} = \frac{2162}{54145} = 0.0399.$$

8.3.2 Properties of Hypergeometric Distribution The important properties of the hypergeometric probability distribution are given here.

- i) The mean and variance of the hypergeometric probability distribution are $\mu = xp$ and $\frac{N-x}{N-1}$, where $p = \frac{k}{N}$ and $q = \frac{N-k}{N}$.

$$P(x, k) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}, \text{ for } x \text{ such that } 0 \leq x \leq n \text{ and } 0 \leq x \leq k.$$

The mean, μ , is given by

$$\begin{aligned} \mu &= E(X) \\ &= \sum_{x=0}^n x \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}} \\ &= \sum_{x=0}^n \frac{x \cdot k \binom{k-1}{x-1} \binom{N-k}{n-x}}{x(n-1)(k-x)\binom{N}{n}} = \sum_{x=1}^n \frac{k \binom{k-1}{x-1} \binom{N-k}{n-x}}{\binom{N}{x}} \end{aligned}$$

Let $y = x - 1$, then

$$\begin{aligned} i) \quad \mu &= \frac{k}{N} \sum_{y=0}^{k-1} \binom{k-1}{y} \binom{N-k}{N-1-y} \quad [\because \text{when } x=1, y=0, \text{ and when } x=n, y=n-1] \\ &= \frac{k}{N} \binom{N-1}{n-1} \quad \left[\because \sum_{j=0}^{\infty} \binom{a}{j} \binom{b}{m-j} = \binom{a+b}{m} \right] \\ &= \frac{kn(n-1)(N-n)!}{N(N-1)!} \cdot \frac{(N-1)!}{(n-1)(N-n)!} \\ &= \frac{nk}{N} = np, \text{ where } p = \frac{k}{N}. \end{aligned}$$

Thus the mean of the hypergeometric distribution is the same as that of the binomial distribution.

By definition, the variance, σ^2 , is given by

$$\sigma^2 = E[X^2] - \mu^2 = E[X^2] - \mu^2.$$

$$\begin{aligned} \text{Now } E(X^2) &= E[X(X-1) + X] = E(X) + E[X(X-1)] \\ &= \sum_{x=0}^k x \cdot h(x; N, n, k) + \sum_{x=1}^k (x-1) \cdot h(x; N, n, k) \\ &= \frac{nk}{N} + \sum_{x=1}^k \frac{n_x}{N} \binom{N-1}{x-1} \binom{N-k}{N-x} \\ &= \frac{nk}{N} + \frac{k(k-1)}{N} \sum_{x=2}^k \binom{k-2}{x-2} \binom{N-k}{N-x} \end{aligned}$$

Let $y = x - 2$, then

$$E(X^2) = \frac{nk}{N} + \frac{k(k-1)}{N} \sum_{y=0}^{k-2} \binom{k-2}{y} \binom{N-k}{N-2-y} \quad [\because \text{when } x=2, y=0, \text{ and when } x=n, y=n-2]$$

$$= \frac{nk}{N} + \frac{k(k-1) \binom{N-x}{n-2}}{\binom{N}{n}} = \frac{nk}{N} + \frac{k(k-1).n(n-1)}{N(N-1)}$$

$$\sigma^2 = E(X^2) - \mu^2$$

$$= \frac{nk}{N} + \frac{k(k-1)n(n-1)}{N(N-1)} - \left(\frac{nk}{N}\right)^2$$

$$= \frac{nk(N-k)}{N^2} \cdot \frac{N-n}{N-1}$$

$$= npq \frac{N-n}{N-1}, \text{ where } p = \frac{k}{N} \text{ and } q = \frac{N-k}{N}.$$

Proceeding along the same steps as above, we may obtain higher moments, but they are rather complicated.

ii) If N becomes indefinitely large, the hypergeometric probability distribution tends to the normal probability distribution.

Let $p = \frac{k}{N}$. Then $k = Np$ and $N-k = N(1-p) = Ng$.

Substituting these values in the hypergeometric formula, we get

$$\begin{aligned} h(x; N, n, k) &= \frac{\binom{Np}{x} \binom{Ng}{N-x}}{\binom{N}{n}} \\ &= \frac{n! (Np)!(Ng)!(N-n)!}{n! (Np-x)!(N-x)!(Ng-N+x)!N!} \\ &= \binom{n}{x} \frac{(Np)!(Ng)!(N-n)!}{(Np-x)!(Ng-N+x)!N!} \end{aligned}$$

Using Stirling's approximation ($a! \approx e^{-a} \cdot a^a \sqrt{2\pi a}$) to all factorial terms and simplifying, we get

$$h(x; N, n, k) \approx \binom{n}{x} \frac{p^{Np+1/2} q^{Ng+1/2} \left(1 - \frac{x}{N}\right)^{N-n+1/2}}{\left(p - \frac{x}{N}\right)^{Np-x+1/2} \left(q - \frac{N-x}{N}\right)^{Ng-N+x+1/2}}$$

Now, if N is allowed to become indefinitely large, then $\frac{x}{N}$, $\frac{\delta-x}{N}$, and $\frac{n}{N}$ each approach zero.

Therefore,

$$b(x; N, n, \delta) = \binom{n}{x} p^x q^{n-x} = b(x; n, p).$$

8.4 POISSON DISTRIBUTION

The Poisson distribution, named after the French mathematician Siméon Denis Poisson (1781-1840) who published its derivation in 1837, is used as:

- a limiting approximation of the binomial distribution $b(x; n, p)$, when p , the probability of success is very small but n , the number of trials is so large that the product np is of moderate size;
- a distribution in its own right by considering a Poisson process where events occur over a specified interval of time or space or length. Such random events might be the number of deaths by horse-kicks per year; the number of telephone calls received per minute at a switchboard; the number of taxicab arrivals at an intersection per day; the number born blind per year in a large city; the number of typing errors per page in a book; the number of red blood cells in a specimen of blood; the number of radioactive particles emitted in a given period; the number of flaws per unit length of some material, the number of claims made to a company in a given time.

Generally, most statisticians use Poisson approximation when p is 0.05 or less and $n > 20$, but in fact, the larger n is and the smaller p is, the better will be the approximation. If we assume n goes to infinity and p approaches zero in such a way that $\mu = np$ remains constant, then the sum of the binomial probability distribution

$$\lim_{p \rightarrow 0} b(x; n, p) = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, 2, \dots, \infty$$

where $e = 2.71828$. The Poisson distribution has only one parameter $\mu > 0$, and is denoted $P(\mu)$. The parameter μ may be interpreted as the mean rate of occurrence of events. It is relevant since this is a probability distribution as the function is obviously non-negative, i.e., $b(x; \mu) \geq 0$ for all unity, i.e.

$$\begin{aligned} \sum_{x=0}^{\infty} b(x; \mu) &= \sum_{x=0}^{\infty} P(X=x) \\ &= \sum_{x=0}^{\infty} \frac{\mu^x e^{-\mu}}{x!} = e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^x}{x!} \\ &= e^{-\mu} \left[1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \dots \right] = e^{-\mu} e^{\mu} = 1. \end{aligned}$$

The Poisson probability distribution is also called the *law of small numbers* or the *rare events distribution*. It has found wide application in the field of Biology, Physics, Operation Research and Management Sciences. The Poisson distribution is appropriate when the number of possible occurrences is very large and the number of actual occurrence is very small in a fixed period of time.

8.4.1 Derivation of Poisson Approximation to the Binomial. To derive an approximation formula to the binomial distribution $b(x; n, p)$ when $n \rightarrow \infty, p \rightarrow 0$, and the product np remains constant, we proceed as below:

The binomial distribution $b(x; n, p)$ may be written as

$$\begin{aligned} b(x; n, p) &= \binom{n}{x} p^x q^{n-x}, \quad \text{for } x = 0, 1, \dots, n. \\ &= \frac{n(n-1)(n-2)\dots(n-x+1)}{x!} p^x q^{n-x} \end{aligned}$$

Let $np = \mu$. Then $p = \frac{\mu}{n}$ and $q = 1-p = 1-\frac{\mu}{n}$.

Ignoring all terms involving μ , we get

$$\begin{aligned} b(x; n, p) &\approx \frac{n(n-1)(n-2)\dots(n-x+1)}{x!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \\ &= \frac{\mu^x}{x!} \cdot \frac{n(n-1)(n-2)\dots(n-x+1)}{n^x} \left(1 - \frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{-x} \\ &= \frac{\mu^x}{x!} \left[1 \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{x-1}{n}\right)\right] \left(1 - \frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{-x} \end{aligned}$$

Ignoring $n \rightarrow \infty$ and $\mu \rightarrow 0$, so that $np = \mu$ remains constant, we observe that each of the terms $\left(1 - \frac{1}{n}\right), \left(1 - \frac{2}{n}\right), \dots, \left(1 - \frac{x-1}{n}\right)$ and $\left(1 - \frac{\mu}{n}\right)^{-x}$ approaches unity. The term $\left(1 - \frac{\mu}{n}\right)^x$ may be written as

$$\left(1 - \frac{\mu}{n}\right)^x = \left[\left(1 - \frac{\mu}{n}\right)^{n/\mu}\right]^{\mu}$$

Ignoring $\frac{\mu}{n} = k$, the expression becomes $\left[\left(1 - \frac{1}{k}\right)^k\right]^{\mu}$

If n increases indefinitely, so does k . Therefore $\left(1 - \frac{1}{k}\right)^k$ tends to e^{-1} , where $e = 2.71828$. Then

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\mu}{n}\right)^n = e^{-\mu}$$

Thus the limiting value of $P(X=x)$ is given by the expression

$$\lim_{n \rightarrow \infty} b(x; n, p) = \frac{\mu^x}{x!} \cdot 1 \cdot 1 \cdots 1 \cdot e^{-\mu} = \frac{\mu^x e^{-\mu}}{x!}, \quad \text{for } x = 0, 1, 2, \dots, \infty$$

In other words, if X is a binomial r.v. such that

$$P(X=x) = \binom{n}{x} p^x q^{n-x}, \text{ then}$$

$$\lim_{n \rightarrow \infty} P(X=x) = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, 2, \dots, \infty$$

It is denoted by $p(x; \mu)$. Hence a t.r.v. X having $p(x; \mu)$ is said to have a Poisson distribution with parameter μ .

Example 8.15 If X is a Poisson random variable with parameter $\mu=2$, find the probability $x = 0, 1, 2, 3$ or more. (P.U., B.A.B.Sc.)

Here the Poisson distribution is

$$p(x; 2) = \frac{e^{-2} (2)^x}{x!} \quad (x = 0, 1, 2, \dots)$$

The desired probabilities for $x = 0, 1, 2, 3$ or more are computed as below.

$$P(X=0) = p(0; 2) = e^{-2} = 0.135335$$

$$P(X=1) = p(1; 2) = \frac{e^{-2} \cdot 2}{1!} = 2(0.135335) = 0.27067$$

$$P(X=2) = p(2; 2) = \frac{e^{-2} (2)^2}{2!} = \frac{4}{2}(0.135335) = 0.27067,$$

$$P(X \geq 3) = 1 - P(X < 3)$$

$$= 1 - [P(X=0) + P(X=1) + P(X=2)]$$

$$= 1 - [0.135335 + 0.27067 + 0.27067]$$

$$= 1 - 0.676675 = 0.353325$$

Example 8.16 Two hundred passengers have made reservations for an airplane flight. If the probability that a passenger who has a reservation will not show up is 0.01, what is the probability that exactly three will not show up?

Let us regard a "no show" as success. Then this is essentially a binomial experiment with $n = 200$ and $p = 0.01$. Since p is very small and n is considerably large, we shall apply the Poisson distribution, using $\mu = (200)(0.01) = 2$.

Therefore, if X represents the number of successes (not showing up), we have

$$\begin{aligned} P(X=3) &= p(3; 2) = \frac{(2)^3 e^{-2}}{3!} \\ &= \frac{(8)(0.1353)}{3 \times 2 \times 1} \quad \left(\because e^{-2} = \frac{1}{(2.71828)^2} = 0.1353 \right) \\ &= 0.1804. \end{aligned}$$

Example 8.17 The probability that a man aged 50 years will die within a year is 0.01125. What is the probability that of 12 such men at least 11 will reach their fifty-first birthday?

Here $p = 0.01125$ and $n = 12$. We compute the desire probability by means of Poisson distribution because the probability of death is very small.

Therefore, $\mu = np = 12 \times 0.01125 = 0.135$, and the Poisson distribution is

$$p(x; 0.135) = \frac{e^{-0.135} (0.135)^x}{x!}$$

Now the probability that no person will die, i.e. all the 12 persons will survive, is

$$\begin{aligned} p(0; 0.135) &= e^{-0.135} \\ &= \frac{1}{0.135 + \frac{(0.135)^2}{2!} + \frac{(0.135)^3}{3!} + \dots} \\ &= 0.8737, \end{aligned}$$

The probability that 1 person will die, i.e. 11 persons will survive, is

$$\begin{aligned} p(1; 0.135) &= \frac{e^{-0.135} (0.135)^1}{1!} \\ &= (0.8737)(0.135) = 0.1179 \end{aligned}$$

The probability that at least 11 persons will survive

$$\begin{aligned} &= p(0; 0.135) + p(1; 0.135) \\ &= 0.8737 + 0.1179 = 0.9916. \end{aligned}$$

Example 8.18 A sampling plan calls for taking 100 items from a very large lot which is one defective. Let X be the number of defectives found in the sample of 100. Construct a table of values $P(X=x)$, for $x = 0, 1, 2, 3, 4$, using first the binomial probability distribution and then the approximation.

Let p denote the probability that an item is defective. Then the binomial probabilities with $n = 100$, $p = 0.01$ and $q = 0.99$ are

$$P(X = x) = \binom{100}{x} (0.01)^x (0.99)^{100-x}, \quad x = 0, 1, 2, 3 \text{ and } 4.$$

These probabilities by means of the Poisson approximation, using $\mu = np = (100)(0.01) = 1.00$,

$$P(X = x) \approx p(x; 1) = \frac{e^{-1}}{x!}, \quad x = 0, 1, 2, 3 \text{ and } 4.$$

The desired probabilities, on simplification, are given in the following table:

x	$P(X = x)$	
	Binomial	Poisson
0	0.3660	0.3679
1	0.3697	0.3679
2	0.1849	0.1839
3	0.0610	0.0608
4	0.0149	0.0133

These results indicate that the approximation is very good.

8.4.2 Poisson Frequency Distribution. When the Poisson distribution is multiplied by the number of sets of experiments, each of n trials, the resulting distribution is known as the *frequency distribution*, and is denoted by

$$f(x) = N \cdot \frac{e^{-\mu} \cdot \mu^x}{x!}, \quad \text{where } x = 0, 1, 2, \dots, \infty.$$

Example 8.19 For a machine making parts, there is a small probability of 0.002 for a part being defective. The parts are supplied in bundles of 10. Calculate approximately the number of bundles containing no defective, one defective or two defectives in a consignment of 10,000 bundles, given $e^{-0.02} = 0.9802$.

Let p be the probability of part being defective. Then $p = 0.002$ and $n = 10$.

Since p is extremely small, we apply the Poisson approximation, using $\mu = np = 10 \times 0.002 = 0.02$.

Hence the approximate number of bundles containing no defective, one defective or two defective are as follows for $x = 0, 1$, and 2 is

$$N.p(x; \mu) = 10,000 \cdot \frac{(0.02)^x \cdot e^{-0.02}}{x!}$$

Putting $x = 0$, we get

$$10,000 \times e^{-0.02} = 10,000 \times 0.9802 = 9,802.$$

Putting $x = 1$, we get

$$10,000 \times e^{-0.02} (0.02) = 10,000 \times 0.9802 \times 0.02 = 196.$$

Putting $x = 2$, we obtain

$$10000 \cdot \frac{e^{-0.02} (0.02)^2}{2!} = \frac{10000 \times 0.9802 \times (0.02)^2}{2} = 2 \text{ approx.}$$

8.4.3 Properties of the Poisson Distribution. Some of the main properties of the Poisson distribution are given below:

- If the random variable X has a Poisson distribution with parameter μ , then its mean and variance are given by $E(X) = \mu$ and $\text{Var}(X) = \mu$.

By definition,

$$\text{Mean} = E(X)$$

$$\begin{aligned} &= \sum_{x=0}^{\infty} x \cdot p(x; \mu), \text{ where } p(x; \mu) = \frac{e^{-\mu} \mu^x}{x!} \\ &= 0 \cdot e^{-\mu} + 1 \cdot \mu e^{-\mu} + 2 \cdot \frac{\mu^2}{2!} e^{-\mu} + 3 \cdot \frac{\mu^3}{3!} e^{-\mu} \\ &= \mu e^{-\mu} \left[1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \dots \right] \\ &= \mu e^{-\mu} \cdot e^{\mu} = \mu \end{aligned}$$

Method

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\mu} \mu^x}{x!} \\ &= e^{-\mu} \mu \sum_{x=0}^{\infty} \frac{\mu^x}{(x-1)!} \\ &= \mu \cdot e^{-\mu} \sum_{x=1}^{\infty} \frac{\mu^{x-1}}{(x-1)!} \quad (\text{since the first term in the summation being zero is omitted}) \end{aligned}$$

$y = x - 1$, then

$$\begin{aligned} E(X) &= \mu \cdot e^{-\mu} \sum_{y=0}^{\infty} \frac{\mu^y}{y!} \quad (y = 0, 1, 2, \dots, \infty) \\ &= \mu \cdot e^{-\mu} \cdot e^{\mu} = \mu \end{aligned}$$

μ , the parameter of the distribution.

$$\text{Var}(X) = E(X^2) - [E(X)]^2, \text{ where}$$

$$\begin{aligned}
 E(X^2) &= E[X(X-1) + X] = E(X) + E[X(X-1)] \\
 &= \sum_{x=0}^{\infty} x \cdot \frac{\mu^x \mu^{x-1}}{x!} + \sum_{x=0}^{\infty} x(x-1) \mu^{x-2} \frac{\mu^x}{x!} \\
 &= \mu + e^{-\mu} \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\mu^x}{x(x-1)(x-2)!} \\
 &= \mu + \mu^2 e^{-\mu} \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!} \quad (x \text{ starts at 2, as the first two terms in the sum are zero})
 \end{aligned}$$

Let $y = x - 2$, then

$$\begin{aligned}
 E(X^2) &= \mu + \mu^2 e^{-\mu} \sum_{y=0}^{\infty} \frac{\mu^y}{y!} \quad (y = 0, 1, 2, \dots) \\
 &= \mu + \mu^2 e^{-\mu} e^{-\mu} = \mu + \mu^2
 \end{aligned}$$

Hence $\text{Var}(X) = \mu + \mu^2 - \mu^2 = \mu$

We observe that the Poisson distribution has an interesting property that its mean equals its variance.

2. Higher moments of the distribution are found as below:

By definition, $\mu_3' = E(X^3) = \sum_{x=0}^{\infty} x^3 p(x; \mu)$

Writing x^3 is the factorial form, i.e.

$$x^3 = x(x-1)(x-2) + 3x(x-1) + x, \text{ we have}$$

$$\begin{aligned}
 \mu_3' &= e^{-\mu} \sum_{x=0}^{\infty} [x(x-1)(x-2) + 3x(x-1) + x] \frac{\mu^x}{x!} \\
 &= e^{-\mu} [\mu^3 \sum_{x=0}^{\infty} \frac{\mu^{x-3}}{(x-3)!} + 3\mu^2 \cdot e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^{x-2}}{(x-2)!} + \mu \cdot e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^{x-1}}{(x-1)!}] \\
 &= e^{-\mu} [\mu^3 \sum_{x=0}^{\infty} \frac{\mu^x}{x!} + 3\mu^2 \cdot e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^x}{x!} + \mu \cdot e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^x}{x!}] \\
 &= \mu^3 + 3\mu^2 + \mu, \text{ and}
 \end{aligned}$$

$$\mu_4' = E(X^4) = \sum_{x=0}^{\infty} x^4 p(x; \mu)$$

DISCRETE PROBABILITY DISTRIBUTIONS

writing x^4 in the factorial form as

$$x^4 = x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x \text{ we get}$$

$$\begin{aligned}\mu_4' &= v^{-v} \sum_{x=2}^{\infty} [x(x-1)(x-2) + 6x(x-1)(x-2) + 7x(x-1) + x] \frac{\mu^{x-1}}{x!} \\ &= \mu^4 \cdot v^{-v} \sum_{x=2}^{\infty} \frac{\mu^{x-4}}{(x-4)!} + 6\mu^3 \cdot v^{-v} \sum_{x=2}^{\infty} \frac{\mu^{x-3}}{(x-3)!} + 7\mu^2 \cdot v^{-v} \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!} + \mu \cdot v^{-v} \sum_{x=2}^{\infty} \frac{\mu^{x-1}}{(x-1)!} \\ &= \mu^4 + 6\mu^3 + 7\mu^2 + \mu.\end{aligned}$$

$$\mu_3 = \mu_2' - 3\mu_1'\mu_2^2 + 2(\mu_1')^3$$

$$= (\mu^2 + 3\mu^2 + \mu) - 3\mu(\mu^2 + \mu) + 2\mu^3 = \mu;$$

$$\mu_4 = \mu_4' - 4\mu_2'\mu_2^2 + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4,$$

$$= (\mu^4 + 6\mu^3 + 7\mu^2 + \mu) - 4\mu(\mu^2 + 3\mu^2 + \mu) + 6\mu^3(\mu^2 + \mu) - 3\mu^4$$

$$= 3\mu^2 + \mu.$$

$$\beta_1 = \frac{\mu_2^2}{\mu_2'} = \frac{\mu^2}{\mu^2} = \frac{1}{\mu}, \text{ and}$$

$$\beta_2 = \frac{\mu_3}{\mu_2'} = \frac{3\mu^2 + \mu}{\mu^2} = 3 + \frac{1}{\mu}.$$

3. The shape of the Poisson distribution depends on the value of its parameter μ . As the distribution takes on an infinite number of x values (theoretically), the distribution will be positively skewed. The distribution tends to be symmetrical as μ becomes larger and larger.

4. Reproductive Property. If two independent r.v.'s X and Y have Poisson distributions with parameter μ and v , then their sum $X+Y$ has also a Poisson distribution with parameter $\mu+v$.

Now, Here X is $p(x; \mu)$ and Y is $p(y; v)$; and we desire to find $P(X+Y=k)$ for $k = 0, 1, 2, \dots$

$$k=0: P(X+Y=0) = P(X=0) \cdot P(Y=0) \quad (\because X \text{ and } Y \text{ are independent})$$

$$= e^{-\mu} \cdot \mu^0 / 0! \cdot e^{-v} \cdot v^0 / 0!$$

$$k=1: P(X+Y=1) = P(X=0) \cdot P(Y=1) + P(X=1) \cdot P(Y=0)$$

$$= e^{-\mu} \cdot \mu^0 / 0! + \mu e^{-\mu} \cdot \mu^1 / 1! \cdot e^{-v} \cdot v^0 / 0!$$

$$= e^{-(\mu+v)} (\mu + v)$$

$$\text{Similarly for } k=2, P(X+Y=2) = \frac{e^{-(\mu+\nu)}}{2!}(\mu+\nu)^2.$$

And, in general, for $X+Y=k$ when $X=i$ and $Y=k-i$ for $i=0, 1, 2, \dots, k$, we have

$$P(X+Y=k) = P(X=0)P(Y=k) + P(X=1)P(Y=k-1) + \dots + P(X=k)P(Y=0)$$

$$\begin{aligned} &= \frac{e^{-\mu}}{0!} \frac{\mu^k e^{-\nu}}{k!} + \frac{\mu e^{-\mu}}{1!} \frac{\nu^{k-1} e^{-\nu}}{(k-1)!} + \dots + \frac{\mu^k e^{-\mu}}{k!} \frac{\nu^0}{0!} \\ &= e^{-(\mu+\nu)} \left[\frac{\nu^k}{k!} + \mu \frac{\nu^{k-1}}{(k-1)!} + \dots + \frac{\mu^k}{k!} \right] \end{aligned}$$

Multiplying the r.h.s. by $\frac{k!}{k!}$, we get

$$\begin{aligned} P(X+Y=k) &= \frac{e^{-(\mu+\nu)}}{k!} \left[\binom{k}{0} \nu^k + \binom{k}{1} \nu^{k-1} \mu + \dots + \binom{k}{k} \mu^k \right] \\ &= \frac{e^{-(\mu+\nu)}}{k!} (\mu+\nu)^k, \end{aligned}$$

which is a Poisson distribution with parameter $\mu+\nu$. This result can easily be generalized.

The converse of this result, that is, if X and Y are independent, and $X+Y$ has a Poisson distribution, then each of the r.v.'s X and Y has a Poisson distribution, is also true was proved by Raikov.

8.4.4 The Recurrence Formula for the Poisson Distribution. Let X be a Poisson variable with parameter μ .

$$P(X=x) = e^{-\mu} \frac{\mu^x}{x!}, \text{ and}$$

$$P(X=x-1) = e^{-\mu} \frac{\mu^{x-1}}{(x-1)!}$$

Dividing, we get

$$\frac{P(X=x)}{P(X=x-1)} = \frac{e^{-\mu} \mu^x}{x!} \cdot \frac{(x-1)!}{e^{-\mu} \mu^{x-1}} = \frac{x}{\mu}$$

Therefore $P(X=x) = \frac{\mu^x}{x!} P(X=x-1)$, for $x=1, 2, 3, \dots$ is the recurrence formula for the Poisson distribution. Using this recurrence relationship, the Poisson probabilities can be obtained more easily.

8.4.5 Fitting a Poisson Distribution to Observed Data. The Poisson distribution can be fitted (or derived) if we know the value of its mean which is usually obtained by equating μ to the mean — use from the observed frequency distribution, provided that the probability of occurrence is very small. Using this value of mean, the expected or theoretical frequencies are computed. The following numerical example will illustrate the procedure.

Example 8.20 Bortkiewicz (1868–1931) collected data on the number of deaths from horse-kicks in Prussian Army Corps over a period of 20 years. This distribution of deaths was as follows:

No. of deaths	0	1	2	3	4	5	Total
Frequency	109	65	22	3	1	0	200

Fit a Poisson distribution to these data and compute the theoretical frequencies.

To fit a Poisson distribution, we need to compute (estimate) value of mean of the given distribution and equate it to μ , the mean of the Poisson distribution.

$$\text{Now } \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{0+65+44+9+4}{200}$$

$$= \frac{122}{200} = 0.61, \text{ which is an estimate of } \mu.$$

The fitted Poisson distribution is given by

$$P(X = x) = p(x; 0.61) = \frac{e^{-0.61} (0.61)^x}{x!}, \text{ where } x = 0, 1, 2, 3,$$

The theoretical or expected frequencies of x deaths are computed by multiplying the probabilities which is 200 here. The probabilities are computed by using the Poisson recurrence formula, which has the form

$$P(X = x) = \frac{0.61}{x} P(X = x - 1), \text{ for } x = 0, 1, 2, 3, 4, 5.$$

In case the table values for $e^{-0.61}$ are not available, they are computed by use of logarithms as

$$\text{Let } y = e^{-0.61}$$

$$\text{Then } \log y = -0.61 \log e = -0.61(0.4343)$$

$$= -0.2649 - 1.7351 \text{ so that } y = 0.5434.$$

The probabilities and the frequencies are then computed as below:

No. of death (x)	Probability $p(x; 0.61)$	Expected frequency $200 \times p(x)$
0	$e^{-0.61}$	= 0.5434
1	$e^{-0.61} \frac{(0.61)}{1} = \frac{0.61}{1} (0.5434)$	= 0.3315
2	$e^{-0.61} \frac{(0.61)^2}{2!} = \frac{0.61}{2} (0.3315)$	= 0.1011
3	$e^{-0.61} \frac{(0.61)^3}{3!} = \frac{0.61}{3} (0.1011)$	= 0.0206
4	$e^{-0.61} \frac{(0.61)^4}{4!} = \frac{0.61}{4} (0.0206)$	= 0.0031
5	$e^{-0.61} \frac{(0.61)^5}{5!} = \frac{0.61}{5} (0.0031)$	= 0.0004
Total		= 1.000
		= 200.02

8.4.6 Poisson Process. In an earlier section, the Poisson distribution was obtained as an approximation to the binomial distribution. The terms of a Poisson distribution can also be derived from a Poisson process which may be defined as a physical process governed at least in part by some random mechanism. The occurrence of traffic deaths per month in a city is an example of a Poisson process. A Poisson process has the following properties:

- The probability that an event occurs in a very short time interval Δt , is proportional to the length of the time interval, i.e., approximately $\lambda\Delta t$, where λ is a positive quantity which can be interpreted as the average number of occurrences per unit of time.
- The probability that no or more events occur in such a short interval is so small that it can be neglected.
- Events occurring in non-overlapping intervals of time are statistically independent.

These properties are assumed to hold for events occurring randomly in regions of space. Using these properties, it can be shown that the probability for the number of occurrences of a random event in an interval of stated length t is given by the Poisson distribution with the parameter λt . The Poisson process formula is

$$p(x; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!},$$

where

t = number of units of time,

x = number of occurrences in t units of time, and

λ = average number of occurrences per unit of time.

The derivation of the Poisson process formula is beyond the scope of this book.

Example 8.21 Telephone calls are being placed through a certain exchange at random times or the rate of four per minute. Assuming a Poisson process, determine the probability that in a 15-second interval, there are 3 or more calls. (P.U., B.A./B.Sc. 1979)

Taking a minute as the unit of time, we have

$$\lambda = 4 \text{ calls per minute}$$

$$= 4 \text{ calls per } 60\text{-seconds.}$$

As 15 second is $\frac{15}{60} = \frac{1}{4}$ units of time, so $t = \frac{1}{4}$ and therefore the average number of calls per second interval i.e. $\lambda t = 4 \times \frac{1}{4} = 1$.

Hence, using the Poisson process formula $p(x; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$, we have $P(3 \text{ or more calls in } 15\text{-second interval})$

$$= 1 - P(0, 1 \text{ or } 2 \text{ calls in } 15\text{-second interval})$$

$$= 1 - \sum_{x=0}^2 p(x; \lambda t) = 1 - \sum_{x=0}^2 \frac{e^{-1} (1)^x}{x!}$$

$$= 1 - \sum_{x=0}^2 \frac{(0.3679)(1)^x}{x!} \quad : e^{-1} = 0.3679$$

$$= 1 - (0.9975) = 0.0025$$

Example 8.22 Flaws in a certain type of drapery material appear on the average of one in 150 square feet. If we assume the Poisson distribution, find the probability of at most one flaw in 225 square feet.

Taking 150 square feet as the unit of area, we have

$$\lambda = 1 \text{ flaw per } 150 \text{ square feet.}$$

As 225 square feet are $\frac{225}{150} = 1.5$ units of area, so $t = 1.5$ and therefore the average number of flaws in 225 square feet, i.e. $\lambda t = 1 \times 1.5 = 1.5$. Assuming the flaws are a Poisson process, we have

$$P(\text{at most one flaw in } 225 \text{ square feet}) = \sum_{x=0}^1 p(x; \lambda t)$$

$$= \sum_{x=0}^1 \frac{e^{-1.5} (1.5)^x}{x!}, \text{ where } e^{-1.5} = 0.2231$$

$$= 0.2231 + 0.3347 = 0.5578$$

8.4.7 Moment Generating and Cumulant Generating Functions of the Poisson Distribution
 The m.g.f. for the Poisson distribution with respect to the origin, is found as below:

$$\begin{aligned} M_0(t) &= E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} p(x; \mu) \\ &= \sum_{x=0}^{\infty} \mu^x \frac{\mu^x e^{-\mu}}{x!} = e^{-\mu} \sum_{x=0}^{\infty} \frac{(\mu e^t)^x}{x!} \\ &= e^{-\mu} e^{\mu e^t} = e^{\mu(e^t - 1)}. \end{aligned}$$

The mean and variance are obtained as:

$$\begin{aligned} \mu'_1 &= E(X) = \left[\frac{d}{dt} \left\{ e^{\mu(e^t - 1)} \right\} \right]_{t=0} \\ &= \left[\mu e^t \cdot e^{\mu(e^t - 1)} \right]_{t=0} = \mu, \\ \text{and } \mu'_2 &= E(X^2) = \left[\frac{d^2}{dt^2} \left\{ e^{\mu(e^t - 1)} \right\} \right]_{t=0} \\ &= \left[\mu e^t \cdot e^{\mu(e^t - 1)} + \mu^2 e^t \cdot e^{\mu(e^t - 1)} \right]_{t=0} \\ &= \mu + \mu^2. \end{aligned}$$

Hence the variance is

$$\sigma^2 = \mu'_2 - \mu'_1 = \mu + \mu^2 - \mu^2 = \mu.$$

The cumulant generating function (c.g.f.) of the Poisson distribution with respect to origin, is given by

$$\begin{aligned} \kappa(t) &= \log_e M_0(t) \\ &= \mu (e^t - 1) \\ &= \mu \left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \dots + \frac{t^r}{r!} + \dots \right) \end{aligned}$$

$$\kappa_r = \text{co-efficient of } \frac{t^r}{r!} \text{ in } \kappa(t) = \mu, \text{ for all } r.$$

Hence all the cumulants of the Poisson distribution are equal to μ .

PROBABILITY DISTRIBUTIONS

Example 8.23. Prove the following recurrence formula for a Poisson distribution $p(x; m)$:

$$\mu_{r+1} = rm\mu_{r-1} + m \frac{d\mu_r}{dm}$$

$$\text{or, } \mu_r = \sum_{x=0}^{\infty} p(x; m) (x - m)^r = \sum_{x=0}^{\infty} \frac{e^{-m} m^x}{x!} (x - m)^r$$

Differentiating with respect to m , we get

$$\frac{d\mu_r}{dm} = \sum_{x=0}^{\infty} \left[-x \frac{e^{-m} m^{x-1}}{x!} (x - m)^{r-1} + x \frac{e^{-m} m^{x-1}}{x!} (x - m)^r - \frac{e^{-m} m^x}{x!} (x - m)^r \right]$$

Multiplying both sides by m and simplifying, we get

$$m \frac{d\mu_r}{dm} = -rm\mu_{r-1} + \sum_{x=0}^{\infty} \frac{e^{-m} m^x}{x!} (x - m)^r (x - m)$$

$$= -rm\mu_{r-1} + \mu_{r+1}$$

$$\mu_{r+1} = rm\mu_{r-1} + m \frac{d\mu_r}{dm}$$

NEGATIVE BINOMIAL DISTRIBUTION

In binomial experiments, the number of successes is fixed and the number of trials is fixed. But in experiments in which the number of successes is fixed and the number of trials varies to produce a fixed number of successes. Such experiments are called *negative binomial experiments*. In other words, negative binomial experiment possesses the following four properties:

The outcomes of each trial may be classified into one of two categories: success (S) and failure (F).

The probability of success, denoted by p , remains constant for all trials.

The successive trials are all independent.

The experiment is repeated a variable number of times to obtain a fixed number of successes.

If X denotes the number of trials to produce k successes in a negative binomial experiment, it is a discrete binomial variable, and its p.d., is called the *negative binomial distribution*. When the random c.v. X assumes a value x , on which the k th success occurs, the negative binomial p.d. is given by

$$P(X=x) = \binom{x-1}{k-1} p^k q^{x-k}, \quad x = k, k+1, k+2, \dots$$

The negative binomial p.d. has two parameters: k and $p > 0$, and is generally denoted by $b^k(x; k, p)$. To prove that for the negative binomial distribution, the probability function sums to one, we proceed as follows:

$$b^k(x; k, p) = \sum_{x=k}^{\infty} \binom{x-1}{k-1} p^k q^{x-k}, \quad (x = k, k+1, k+2, \dots)$$

Let $y = \pi - k$. Then

$$\begin{aligned}\text{Sum of prob} &= \sum_{y=0}^{\infty} \binom{y+k-1}{k-1} p^y q^x, \quad (y=0, 1, 2, \dots) \\ &= p^k \sum_{z=1}^{\infty} \binom{y+k-1}{k-1} q^z \\ &= p^k \left[\binom{k-1}{k-1} q^0 + \binom{k}{k-1} q + \binom{k+1}{k-1} q^2 + \dots \right] \\ &= p^k \left[1 + kq + \frac{k(k+1)}{2!} q^2 + \dots \right] \\ &= p^k [1-q]^k = p^k p^{-k} = 1.\end{aligned}$$

The distribution takes its name from the fact that $\binom{y+k-1}{k-1} p^y q^{x-y}$ is a term in the expansion of $p^k (1-q)^k$, a binomial with negative index. Thus the probabilities at k th, $(k+1)$ th, $(k+2)$ th, ... trials are $p^k \left[1, kq, \frac{k(k+1)}{2} q^2, \dots \right]$.

This distribution is sometimes also called the Pascal distribution, after the French mathematician Pascal (1623–1662). The distribution is found to occur in many biological situations and in sampling from a binomial population.

8.5.1. Derivation of the Negative Binomial Distribution. The negative binomial distribution can be derived in various ways. The following derivation is based on the binomial trials.

To find an expression for the probability that x trials are made to achieve k successes under the condition that the last trial must be a success, we proceed as follows:

A sequence containing k successes in exactly x independent trials with the condition that the last trial must be a success, can be obtained as

$$\overbrace{SS\dots S}^{k \text{ times}} \overbrace{FF\dots FS}^{x-k \text{ times}},$$

Thus the probability of a success on the x th trial preceded by $(k-1)$ successes and $(x-k)$ failures is

$$p^{k-1} q^{x-k} p = p^k q^{x-k}$$

Since the last trial must be a success, therefore the total number of mutually exclusive sequences which $(k-1)$ successes and $(x-k)$ failures preceding the last success can occur in any order, a given

$$\binom{x-1}{k-1}$$

Hence the formula for the probability that k th success occurs on the x th trial is

$$P(X=x) = \binom{x-1}{k-1} p^k q^{x-k}, \text{ where } x = k, k+1, k+2, \dots$$

The negative binomial distribution can also be obtained when two or more Poisson r.v.'s are summed term by term.

Example 8.24 A person throws a pair of fair dice. What is the probability that he will get a total of 7 for the second time on the eighth throw?

The probability that he gets a total of 7 is $\frac{6}{36}$, i.e. $p = \frac{1}{6}$.

Since the number of successes is fixed, therefore the negative binomial distribution with $k = 2$ (one success) and $x = 8$ is used.

$$\text{Since } P(X=8) = \binom{8-1}{2-1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{8-2} \\ = 7 \cdot \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^6 \approx 0.0651$$

Example 8.25 Three people each toss a coin and a good man pays for the coffee. If the coins all land on tail or all show tails, they are tossed again. What is the probability that a decision is reached in 3 tosses or fewer?

To reach a decision on any trial, the coins must result in either '2 heads and 1 tail' or '2 tails and 1 head'. The probability of these events is computed by the binomial distribution because each coin has two possible results; a head or a tail.

The probability of getting a head is $p = \frac{1}{2}$ and remains the same for each coin.

Three people toss the coins independently.

Three coins are tossed in each case, (i.e. $n = 3$)

$$P(2 \text{ heads and 1 tail}) = \binom{3}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right) = \frac{3}{8},$$

$$P(1 \text{ head and 2 tails}) = \binom{3}{1} \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)^2 = \frac{3}{8}.$$

Probability of reaching a decision is $\frac{3}{8} + \frac{3}{8} = \frac{6}{8} = 0.75$

Let us find the probability that a decision is reached in 5 tosses or fewer. We observe that we may or may not reach a decision with each set.

The probability of reaching a decision is $p = 0.75$ and remains the same for all sets,

- iii) the results of all sets are independent;
- iv) a variable number of sets is required to produce 1 decision

We therefore compute the required probability by means of the negative binomial distribution where

$$k = 1, p = 0.75 \text{ and } x \leq 5 \text{ (the number of trials).}$$

$$\begin{aligned} \text{Thus } P(X \leq 5) &= \sum_{x=0}^5 \binom{x-1}{k-1} (0.75)^k (0.25)^{x-k}, k=1 \\ &= 0.75 + 0.1875 + 0.0469 + 0.0117 + 0.0029 \\ &= 0.9990 \end{aligned}$$

8.5.2 Properties of the Negative Binomial Distribution. The important properties of the negative binomial distribution are given below:

1. The mean of the negative binomial distribution is less than its variance.

We find the mean and variance by deriving the m.g.f. of the negative binomial distribution.

The m.g.f. about the origin is

$$\begin{aligned} M_0(t) &= E(e^{xt}) \\ &= p^k \sum_{x=0}^{\infty} \binom{x+k-1}{k-1} t^x e^{px} \\ &= p^k \sum_{x=0}^{\infty} \binom{x+k-1}{k-1} (qe^t)^x \\ &= p^k (1-qe^t)^{-k} \end{aligned}$$

$$\begin{aligned} \text{Now Mean} &= \left[\frac{dM_0(t)}{dt} \right]_{t=0} \\ &= [q^k \cdot kqe^t (1-qe^t)^{-k-1}]_{t=0} \\ &= p^k \cdot kq (1-q)^{-k-1} \\ &= kpq^k p^{-k-1} = \frac{kq}{p} \end{aligned}$$

and $\sigma^2 = E(X^2) - [E(X)]^2$, where

$$\begin{aligned} E(X^2) &= \left[\frac{d^2 M_0(t)}{dt^2} \right]_{t=0} \\ &= [p^k kqe^t (1-qe^t)^{-k-2} + p^k \cdot k(k+1)q^2 e^t (1-qe^t)^{-k-3}]_{t=0} \\ &= p^k kq (1-q)^{-k-1} + k(k+1)q^2 p^k (1-q)^{-k-2} \\ &= \frac{kq}{p} + \frac{k(k+1) q^2}{p^2} \end{aligned}$$

$$\begin{aligned}\sigma^2 &= \frac{kq}{p} + \frac{k(k+1)q^2}{p^2} - \left(\frac{kq}{p}\right)^2 \\ &= \frac{kqp + k^2q^2 + kq^2 - k^2q^2}{p^2} = \frac{kq(p+q)}{p^2} = \frac{kq}{p^2}\end{aligned}$$

The variance will be greater than mean, if

$$\frac{kq}{p^2} > \frac{kq}{p} \text{ or if } \frac{1}{p^2} > \frac{1}{p}$$

if $1 > p$, which is obviously true.

Hence we observe that the variance of the negative binomial distribution is greater than its mean, is an important feature of this distribution.

2. The negative binomial distribution is always positively skewed.

GEOMETRIC DISTRIBUTION

When an experiment consists of independent trials with probability p of success and the trials are continued until the first success occurs, it is called a geometric distribution. In other words, a geometric distribution has the following four properties:

- i. The outcomes of each trial may be classified into one of two categories, success and failure.
- ii. The probability of success p remains constant for all trials.
- iii. The successive trials are all independent.
- iv. The experiment is repeated a variable number of times until the first success is obtained.

If X represents the number of trials needed for the first success, then X is called a geometric r.v. and is called the geometric distribution. It has only one parameter p and is denoted by $g(x; p)$. The distribution derives its name from the fact that its successive terms constitute a geometric progression. Since a geometric r.v. represents how long one has to wait for a success, it is also called a waiting r.v. It is interesting to note that a geometric distribution is a special case of a negative binomial distribution when $k = 1$.

4.6.1 Derivation of the Geometric Distribution. Let the random variable X denote the number of trials required upto and including the first success of an event. Then X takes the values $1, 2, 3, \dots, \infty$. $X=x$ if and only if the first $(x-1)$ trials result in failures and the x th trial yields a success in the given sequence of trials, we therefore have the probability distribution of X , as

$$P(X=x) = q^{x-1} p, \quad x = 1, 2, 3, \dots, \infty.$$

This is obviously a probability distribution as (i) $P(X=x) \geq 0$ and (ii) the sum of the probabilities, i.e.

$$\begin{aligned}\sum_{x=1}^{\infty} P(X=x) &= p + qp + q^2p + q^3p + \dots \\ &= p[1 + q + q^2 + q^3 + \dots] \\ &= p[1-q]^{-1} = pp^{-1} = 1.\end{aligned}$$

Example 8.24 If the probability that a person will believe a rumour about the retirement of a certain politician is 0.25, what is the probability that

- a) the sixth person to hear the rumour will be first to believe it;
- b) the twelfth person to hear the rumour will be the fourth to believe it?

Let X denote the number of a person who hears the rumour. Then the number of a person who believes it, will be considered a success.

- a) Since the sixth person is the *first* to believe the rumour, i.e. the first success occurs on the 6th trial, therefore the geometric distribution with $p = 0.25$ and $n = 6$ is appropriate.

Hence, using the geometric distribution $P(X=x) = pq^{x-1}$, we get

$$P(X=6) = (0.25)(0.75)^5 = 0.059.$$

- b) Since the twelfth person hearing the rumour will be the *fourth* to believe implies that the success occurs on the 12th trial, therefore the negative binomial distribution with $p=0.25$ and $n=12$ is appropriate.

Hence, using the negative binomial distribution

$$bN(x; k, p) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \text{ we get}$$

$$\begin{aligned}bN(12; 4, 0.25) &= \binom{12-1}{4-1} 0.25^4 (0.75)^{12-4} \\ &= \binom{11}{3} \left(\frac{1}{4}\right)^4 \left(\frac{3}{4}\right)^8 = \frac{365 \times 6561}{265 \times 65536} \\ &= 0.0645.\end{aligned}$$

8.6.2 Properties of the Geometric Distribution:

1. The mean and variance of the geometric distribution are $\mu = 1/p$ and $\sigma^2 = q/p^2$.

Let the r.v. X have a geometric distribution $g(x; p) = pq^{x-1}$. Then

$$\mu = E(X) = \sum x_i g(x_i; p)$$

$$= \sum x_i q^{x-1} p, \text{ where } x = 1, 2, 3, \dots, \infty$$

$$\begin{aligned}
 &= p + 2qp + 3q^2p + 4q^3p + \dots \\
 &= p[1 + 2q + 3q^2 + 4q^3 + \dots] \\
 &= p[1 - q]^{-2} = p \cdot p^{-2} = \frac{1}{p}; \text{ and}
 \end{aligned}$$

$\sigma^2 = E(X^2) - [E(X)]^2$, where

$$\begin{aligned}
 E(X^2) &= \sum_{x=1}^{\infty} x^2 \cdot q^{x-1} p \\
 &= p + 2^2 qp + 3^2 q^2 p + 4^2 q^3 p + \dots \\
 &= p[1 + 4q + 9q^2 + 16q^3 + \dots] \\
 &= p[(1+3q+6q^2+10q^3+\dots)+(q+3q^2+5q^3+\dots)] \\
 &= p[(1-q)^{-2} + q(1-q)^{-2}] \\
 &= \frac{1}{p^2} + \frac{q}{p^2} \\
 &= \sigma^2 = \frac{1}{p^2} + \frac{q}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{q}{p^2}.
 \end{aligned}$$

The distribution is positively skewed.

2. Moment Generating Function of the Geometric Distribution. The m.g.f. of the geometric distribution is derived as below:

$$\begin{aligned}
 M_0(t) &= E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} p \\
 &= pe^t \sum_{x=0}^{\infty} (qe^t)^{x-1} \\
 &= pe^t [1 + qe^t + (qe^t)^2 + \dots] \\
 &= pe^t [1 - qe^t]^{-1} = \frac{pe^t}{1 - qe^t}, \text{ where } qe^t < 1.
 \end{aligned}$$

To differentiate the m.g.f., we write it as

$$M_0(t) = \frac{p}{e^{-t} - q}$$

Thus $M'_x(t) = pe^{-t} (e^{-t} - q)^{-1}$, and

$$M''_x(t) = 2pe^{-2t} (e^{-t} - q)^{-2} - pe^{-t} (e^{-t} - q)^{-3}$$

Therefore $E(X) = p(1-q)^{-1} = \frac{1}{p}$,

$$E(X^2) = 2p(1-q)^{-2} - p(1-q)^{-3}$$

$$= \frac{2}{p^2} - \frac{1}{p}, \text{ and}$$

$$\sigma^2 = E(X^2) - [E(X)]^2$$

$$= \frac{2}{p^2} - \frac{1}{p} - \left(\frac{1}{p}\right)^2 = \frac{q}{p^2}$$

8.7 MULTINOMIAL DISTRIBUTION

A binomial experiment becomes a *multinomial experiment* when there are more than two outcomes of each trial. For example, manufactured items may be classified as good, average, or defective; or a road accident may result in no injury, minor injuries, severe injuries, or fatal injuries. A multinomial experiment has the following properties.

- The outcomes of each trial may be classified into one of k mutually exclusive categories C_1, \dots, C_k .
- The probability of the i th outcome is p_i , which remains constant and $\sum p_i = 1$.
- The successive trials are independent.
- The experiment is repeated a fixed number of times, say, n .

8.7.1 Derivation of the Multinomial Distribution. If n independent trials be made in a specified order in which C_1 occurs x_1 times, C_2 occurs x_2 times, ..., C_k occurs x_k times, $x_1+x_2+\dots+x_k=n$, is

$$\frac{C_1, C_1}{x_1 \text{ times}} \cdot \frac{C_2, C_2}{x_2 \text{ times}} \cdots \frac{C_k, C_k}{x_k \text{ times}}$$

The probability of this happening by multiplicative law is

$$\frac{p_1 \cdot p_1}{x_1 \text{ times}} \cdot \frac{p_2 \cdot p_2}{x_2 \text{ times}} \cdots \frac{p_k \cdot p_k}{x_k \text{ times}} = p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

But we are interested in events occurring in any order. Therefore the total number of non-exclusive orders in which this can happen, is

$$\binom{n}{x_1, x_2, \dots, x_k} = \frac{n!}{x_1! x_2! \cdots x_k!}$$

Hence the required multinomial probability is

$$\begin{aligned} P(X_1=x_1, X_2=x_2, \dots, X_k=x_k) \\ = \frac{n!}{x_1! x_2! \dots x_k!} (p_1)^{x_1} (p_2)^{x_2} \dots (p_k)^{x_k} \\ = \frac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k (p_i)^{x_i}, \text{ where } \sum_{i=1}^k x_i = n \text{ & } \sum_{i=1}^k p_i = 1 \end{aligned}$$

The multinomial distribution takes its name from the fact that the above probabilities correspond to terms in the multinomial expansion of $(p_1 + p_2 + \dots + p_k)^n$. The parameters of this distribution are n, p_1, p_2, \dots, p_k .

The mean and variance of the multinomial distribution are

$$E(X_i) = np_i \text{ and } \text{Var}(X_i) = np_i q_i$$

When $k = 2$, the multinomial distribution reduces to the binomial probability distribution. Thus binomial distribution is a special case of the multinomial distribution.

Example 8.27 A box contains 5 red, 4 white and 3 blue marbles. A sample of six marbles is drawn with replacement, i.e., each marble is replaced before the next one is drawn. Find the probability that out of marbles selected, 3 are red, 2 are white and one is blue.

Let X_1, X_2 and X_3 denote the red, white and blue marbles. Then

$$p_1 = P(X_1 = 3) = \frac{5}{12}$$

$$p_2 = P(X_2 = 2) = \frac{4}{12}$$

$$p_3 = P(X_3 = 1) = \frac{3}{12}$$

$$P(X_1=3, X_2=2, X_3=1) = \frac{6!}{3!2!1!} \left(\frac{5}{12}\right)^3 \left(\frac{4}{12}\right)^2 \left(\frac{3}{12}\right)^1 = \frac{625}{5184}$$

EXERCISES

DISJUNCTIVE

Answer 'True' or 'False'. If the statement is not true then replace the underlined words with words that make the statement true:

A discrete probability distribution is graphically represented by a curve.

A binomial experiment always has three or more possible outcomes to each trial.

Discrete random variables may assume any values.

In a binomial experiment, the individual trials are independent of each other.

- v) In a binomial distribution the mean is equal to its variance.
- vi) In a hypergeometric probability distribution trials are dependent.
- vii) The mean and variance of the Poisson distribution are not equal.
- viii) The Poisson distribution is defined as a limiting approximation of the binomial distribution when p is large, but n is small so that $\mu = np$ is of moderate size.
- ix) Another name for Poisson probability distribution is the rare events distribution.
- x) Poisson probability distribution has two parameters.
- xi) In a negative binomial distribution, the successive trials are all independent.
- xii) The mean of the negative binomial distribution is greater than its variance.
- xiii) The geometric probability distribution is symmetrical.
- xiv) A binomial distribution is a special case of a multinomial distribution when each trial can assume two possible outcomes.
- xv) The outcomes of each trial for binomial distributions may be classified into one or more mutually exclusive categories.

b) MULTIPLE CHOICE QUESTIONS

- i) Which of the following is not a property of a binomial experiment?
- The successive trials are independent.
 - The experiment is repeated a fixed number of times say n .
 - The probability of success, denoted by p , remains constant for all trials.
 - There are three or more possible outcomes for each trial.
- ii) The standard deviation of the binomial distribution is:
- \sqrt{np}
 - \sqrt{npq}
 - npq
 - pq
- iii) For a binomial distribution, the mean and variance are related by:
- $\mu < \sigma^2$
 - $\mu = \sigma^2$
 - $\mu > \sigma^2$
 - $\mu < \sqrt{\sigma^2}$

For a binomial distribution $P(X = x) = x! / C_x (0.5)^x (0.5)^{12-x}$, $x = 0, 1, 2, \dots, 12$, the mean is:

- a) 3
- b) 6
- c) $\sqrt{6}$
- d) $\sqrt{3}$

For a Poisson distribution, the mean and variance are related by:

- a) $\mu = \sigma^2$
- b) $\mu < \sigma^2$
- c) $\mu > \sigma^2$
- d) None of above

A binomial distribution may be approximated by a Poisson distribution when:

- a) n is large and p is small
- b) n is small and p is large
- c) n is small and p is small
- d) n is large and p is large

Which of the following is not a property of a binomial experiment?

- a) The probability of success changes on each trial.
- b) The successive trials are independent.
- c) The experiment is repeated a fixed number of times.
- d) The outcomes of each trial may be classified into one of two categories, success and failure.

For a negative binomial distribution, the mean and variance are related by:

- a) $\mu = \sigma^2$
- b) $\mu < \sigma^2$
- c) $\mu > \sigma^2$
- d) None of above

Which of the following is not a property of a multinomial experiment?

- a) The successive trials are all independent.
- b) The experiment is repeated a fixed number of times.
- c) The outcomes of each trial may be classified into one of k categories ($k \geq 2$).
- d) The probability of success changes on each trial

SUBJECTIVE

- 8.1 a) What is a binomial experiment and what are its properties?
 b) Derive the binomial distribution and find its mean and variance. (P.U., B.A./B.Sc.)
- 8.2 a) The probability of an event occurring on any one occasion is p . Prove that the probability of its occurring on exactly x of n occasions is $\binom{n}{x} p^x q^{n-x}$, where $q=1-p$. (P.U., B.A./B.Sc. 1984)
- b) Let X have a binomial distribution with $n=3$ and $p=0.4$.
 $P\left(X = \frac{3}{2}\right)$, $P(X = 2)$, $P(X \leq 2)$, $P(X = -2)$ and $P(X \geq 2)$. (B.Z.U., B.A./B.Sc.)
- 8.3 a) A die is rolled five times and a 3 or 6 is considered a success. Find the probability of (i) no success, (ii) at least 2 successes, (iii) at least one but not more than 3 successes. (B.Z.U., B.A./B.Sc.)
- b) Using the binomial distribution, find the probability of
 i) 3 successes in 8 trials when $p = 0.4$,
 ii) 2 failures in 6 trials when $p = 0.6$,
 iii) 2 or fewer successes in 9 trials when $p = 0.4$. (P.U., M.A./M.Sc.)
- 8.4 a) Find the probability of getting (i) exactly 4 heads and (ii) not more than 4 heads when 8 coins are tossed.
 b) Find the probability of (i) 6 or more heads, (ii) fewer than 4 heads in a single toss of 10 coins.
- 8.5 a) If the probability of getting caught copying someone else's exam is 0.2, find the probability of not getting caught in 3 attempts. Assume independence.
 b) If 60% of voters in a large district prefer candidate A, what is the probability that a sample of 12 voters exactly 7 will prefer A?
 c) The probability that a patient recovers from a delicate heart operation is 0.9. What is the probability that exactly five of the next 7 patients having this operation survive? (P.U., B.A./B.Sc.)
- 8.6 a) The incidence of occupational disease in an industry is such that the workers have a chance of suffering from it. What is the probability that out of 6 workers (i) not more than 2, and (ii) 4 or more will catch the disease? (P.U., B.A./B.Sc.)
- b) If on the average rain falls on twelve days in every thirty, find the probability that three days of a given week will be fine and the remaining wet, (i) rain will fall on two days of a given week.
- 8.7 An insurance salesman sells policies to 5 men, all of identical age and in good health. According to the actuarial tables, the probability that a man of this particular age will be alive 30 years is $2/3$. Find the probability that in 30 years (i) all men, (ii) at least 3 men, (iii) only two or not one man will be alive.

- A multiple-choice quiz has 15 questions, each with 4 possible answers of which only 1 is the correct answer. What is the probability that sheer guess work yields from 5 to 10 correct answers?
- A commuter drives to work each morning. The route she takes each day includes ten stoplights. Assume the probability each stoplight is red when she gets to it, is 0.2 and that those stoplights (trials) are independent. What is the distribution for X , the number of times she must stop for a red light on her way to work? Evaluate $P(X=0)$ and $P(X \leq 5)$.
- Find the successive terms of the binomial frequency distribution $600(0.3 + 0.7)^n$.
- Five dice are tossed 96 times. Find the expected frequencies when throwing of a 4, 5 or 6 is regarded as a success.
- A perfect cubic die is thrown a large number of times in sets of 8. The occurrence of a 5 or a 6 is called a success. In what proportion of the sets would you expect 3 successes?
- An irregular six-faced die is thrown and the expectation that in 10 throws it will give five even numbers is twice the expectation that it will give four even numbers. How many times, 10,000 sets of 10 throws would you expect it to give no even number?

(P.U., B.A./B.Sc. (1975))

Four dice are thrown and the number of sixes in each throw are recorded. This is repeated 108 times. Write down the theoretical frequencies of 0, 1, 2, 3 and 4 sixes. Calculate the mean number of sixes in a single throw.

- Find the mean and standard deviation of the binomial distribution $(q+p)^n$.
- Find the mean and variance of the binomial $(q-p)^n$. (P.U., B.A./B.Sc. (1962), 69)
- If $v \sim N$ is binomially distributed with mean 3 and variance 2, compute $P(v=7)$.
- In a binomial distribution, the mean and the standard deviation were found to be 16 and 4.8 respectively. Find p and n . (P.U., B.A./B.Sc. (Hons.), 1966)
- A random variable is binomially distributed with mean 12.38 and variance 8.64. Find n and p . (I.I.T., M.A. Econ., 1989)
- Is it possible to have a binomial distribution with mean = 3 and s.d. = 3?

(P.U., B.A./B.Sc. (Hons.) 1960)

Obtain the first four moments of a binomial distribution.

Show that in a binomial distribution where p is the probability of success, the moments about the mean are given by $\mu_1 = npq$, $\mu_2 = npq(n-p)$, $\mu_3 = npq[1+3(n-2)pq]$. (P.U., B.Sc. 1962)

Show that for the binomial distribution $B(n; x, p)$

$$\beta_1 = \frac{(1-2p)^2}{npq} \text{ and } \beta_2 = 3 + \frac{1-6pq}{npq}$$

Let X be a random variable having a binomial distribution with parameters $n=25$ and $p=0.1$. Evaluate $P[X < \mu - 3\sigma]$.

- b) Given binomial probability function $P(X=x) = \binom{10}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x}$. Find the median of the distribution. (P.U., B.A.B.S., 2008)
- 8.18 A biased coin is tossed 4 times and the number of heads noted. The experiment is performed 47 times in all. The results obtained are shown in the table:

No. of heads	0	1	2	3	4
Frequency	12	50	151	200	47

- a) Find the probability of obtaining a head when the coin is tossed.
 b) Calculate the theoretical frequencies of 0, 1, 2, 3, 4 heads, using the associated binomial distribution.
- 8.19 The incidence of defective items in 200 samples of 6 is shown in the following table:

No. of defectives per sample	0	1	2	3	4	5	6	Total
No. of samples	36	70	61	27	7	1	0	200

Assuming these results follow a binomial distribution, compute the theoretical probabilities and frequencies.

- 8.20 Fit a binomial distribution to the following data

x	0	1	2	3	4
f	30	62	46	10	2

(P.U., B.A.B.S., 2008)

- 8.21 Following data give the number of questions correctly answered out of 10 questions for 100 students.

x	0	1	2	3	4	5	6	7	8	9	10
f	0	1	3	8	36	28	18	13	9	4	0

Examine whether the distribution is binomial and find its mean and standard deviation.

- 8.22 Show that, if two symmetrical binomial distributions ($p = q = \frac{1}{2}$) of degree n (and of the same number of observations) are superposed, the r th term of the one coincides with the $(n-r)$ th term of the other; the distribution formed by adding superposed terms is symmetrical binomial of degree $(n+1)$.

- BQ 8.23 a) If X has a binomial distribution $b(x; n, p)$, then show that $E(X) = np$, $\text{Var}(X) = npq$, $M_2(r) = (q+pe^r)^2$.
 b) If the m.g.f. of X is $M_X(t) = [1/4 + (3/4)e^t]^{1/2}$, find $E(X)$, $\text{Var}(X)$ and $P(X \geq 10)$.

- a) Derive the m.g.f. of the binomial distribution. Use it to find the mean and variance of the binomial distribution. (P.U., I.A./B.Sc. 1992)
- 134 Show that for the binomial distribution $(q+p)^n$, where $p=1-q$,
- $$k_{r+1} = pq \frac{d^r}{dp^r}, \quad r \geq 1; \text{ and}$$
- hence find out the first four cumulants. (P.U., M.Sc. (Stat.) 1969)
- 135 a) What is a hypergeometric experiment and what are its properties?
- b) Derive the hypergeometric probability distribution.
- c) Find the mean and variance of the hypergeometric distribution. (P.U., B.A./B.Sc. 1980, 82, 84, 91, 94)
- d) Determine the probability distribution for the number of white heads among 5 heads drawn at random from a bowl containing 4 white and 7 black beads. Use this to compute the mean and variance and check the results by using the formulae.
- e) A committee of size 3 is selected from 4 men and 2 women. Find the probability distribution for the number of men on the committee.
- f) A homeowner plants 6 bulbs selected at random from a box containing 4 tulip bulbs and 4 daffodil bulbs. What is the probability that he planted 2 daffodil bulbs and 4 tulip bulbs?
- 136 Ten vegetable cans, all the same size, have lost their labels. It is known that 5 contain tomatoes and 5 contain corn. If five are selected at random, what is the probability that all contain tomatoes? What is the probability that 3 or more contain tomatoes? (B.Z.U., B.A./B.Sc. 1991)
- 137 a) Determine the probability that Income Tax Authorities will catch 3 income tax returns with illegitimate deductions if it randomly selects 6 returns from among 20 income tax returns of which 8 contain illegitimate deductions.
- b) To avoid detection by customs, a traveller has placed six narcotic tablets in a bottle containing nine vitamin pills that are similar in appearance. If the customs official selects 3 of the tablets at random for analysis, what is the probability that the traveller will be arrested for illegal possession of narcotics?
- c) Discuss the difference in conditions that must exist in a problem situation for application of the hypergeometric and the binomial distributions.
- 138 Show that the hypergeometric distribution $\delta(r; N, n, k)$ can be approximated by the binomial distribution for large N and k . (P.U., B.A./B.Sc. 1994)
- 139 a) A random sampling of 4 members of a 150 members club has shown that 3 prefer no smoking in the clubhouse dining room. What is the probability that this will occur if in fact only 20% of members prefer no smoking in the dining room. Find this probability assuming that the sample was drawn under
- a) sampling without replacement, and
 - b) sampling with replacement.
- Compare the two answers.

- 8.32 a) Describe a Poisson distribution.
 b) Derive the Poisson distribution as the limiting form of the binomial distribution, clearly the assumptions you make. (P.U., B.A./B.Sc. 1975, 30, 31)
- 8.33 a) If X is a Poisson random variable with $\mu = 1.6$, find $P(X=0)$, $P(X=1)$, $P(X=2)$ and $P(X \geq 2)$.
 b) A Poisson distribution is given by $P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}$. Find the probabilities for $x=0$, 3 and 4. (P.U., B.A./B.Sc. 1975, 31)
 c) Suppose that X has a Poisson distribution if $P(X=1)=0.3$ and $P(X=2)=0.2$, calculate $P(X=0)$ and $P(X=3)$.
 d) A random variable X has a Poisson distribution such that $P(X=2) = 3P(X=4)$. Find $P(X=0)$ and $P(X \leq 3)$ upto three places of decimal. (P.U., B.A./B.Sc. 1975, 31)
- 8.34 a) Prove that, if $n \rightarrow \infty$ and $p \rightarrow 0$ such that $np = \mu$, then
- $$\left(\frac{n}{k} \right) p^k (1-p)^{n-k} \rightarrow \frac{e^{-\mu} \mu^k}{k!}$$
- where $e = \lim_{n \rightarrow \infty} (1 + 1/n)^n = 2.7183$ approximately. (P.U., B.A./B.Sc. 1975, 31)
- b) Past experience in the production of a certain component has shown that the proportion defectives is 0.03. Components leave the factory in boxes of 300. What is the probability
 i) a box contains 3 or more defectives;
 ii) two successive boxes contain 6 or more defectives between them? (P.U., B.A./B.Sc. 1975, 31)
- 8.35 a) Define the Poisson distribution and derive its mean and variance.
 b) Ten percent of the tools produced in a certain manufacturing process turn out to be defective. Find the probability that in a sample of 10 tools chosen at random, exactly two will be defective by using (i) the binomial distribution and (ii) the Poisson approximation to the binomial distribution. (P.U., B.A./B.Sc. 1975, 31)
- 8.36 a) Suppose that the number of insurance claims closely approximates a Poisson distribution with $\mu = 0.05$. Find the probability of (i) no claim and (ii) 1 or fewer claims.
 b) Assume that the probability of being killed in an accident in a coal mine during a year is $\frac{1}{1400}$. Use the Poisson distribution to calculate the probability that in the mine employing 350 miners, there will be at least one fatal accident in a year.
- 8.37 a) A secretary makes 2 errors per page on the average. What is the probability that on the next page she makes (i) 4 or more errors? (ii) no error?

- 1) A car hire firm has 2 cars, which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with parameter 1.5. Calculate the proportion of days on which neither car is used, and the proportion of days on which some demand is refused. ($e^{-1.5} = 0.2231$) (B.Z.U., B.A./B.Sc., 1990)

- 2) A manufacturer of copper pins knows that 5 percent of his product is defective. If he sells copper pins in boxes of 100, and guarantees that not more than 4 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality? ($e^{-1} = 0.067$) (B.Z.U., B.A./B.Sc., 1988)

- 3) Find the mean and the variance of the Poisson distribution. State the relation between binomial and Poisson distributions. (P.U., B.A./B.Sc., 1970)

- 4) Given that X has a Poisson distribution with variance 1, calculate $P(X=2)$.

- 5) Criticise the following statement:

"The mean of a Poisson distribution is 5 while its standard deviation is 4."

- 6) In a Poisson distribution the first two frequencies were 250 and 160. Find the frequencies of the next two values of the variable.

- 7) Show that the mean and the variance of a Poisson distribution are equal.

(P.U., B.A./B.Sc., 1986)

- 8) Find the first four moments of the Poisson distribution $p(x; \mu)$, and hence prove that $\beta_1 = \frac{1}{\mu}$ and $\beta_2 = 3 + \frac{1}{\mu}$. (P.U., B.A./B.Sc., 1983, 88)

- 9) Show that the discontinuous (Poisson) distribution whose probabilities corresponding to the values $0, 1, 2, \dots, j_1, \dots$ are

$$e^{-\lambda}, e^{-\lambda}\lambda, \frac{e^{-\lambda}\lambda^1}{1!}, \dots, \frac{e^{-\lambda}\lambda^j}{j!}, \dots$$

- 10) the second, third, fourth and fifth central moments given by $\mu_2 = \mu_1 = \lambda$, $\mu_4 = 4(1+3\lambda)$, $\mu_3 = \lambda(3+10\lambda)$ (P.U., M.A. Stats., 1969)

- 11) Show that, if X_1 and X_2 are independent Poisson variables with parameters λ_1 and λ_2 respectively, then $T = X_1 + X_2$ is a Poisson variable with parameters $\lambda_1 + \lambda_2$.

- 12) Do the difference of two Poisson variables follow Poisson distribution?

- 13) Fit a Poisson distribution to "Student's" yeast cell data.

Yeast cell count	0	1	2	3	4	5
Frequency	213	128	37	18	3	1

Poisson 44-46

- 8.45 (a) Fit (i) a Poisson (ii) a binomial to the following distribution:

x	0	1	2	3	4	Total
f	511	354	99	15	1	1000

- (b) The frequency of accidents per shift in a factory is shown in the following table:

Accidents per shift	0	1	2	3	4	5
Frequency	300	96	34	9	1	0

Use the Poisson distribution to estimate the probability of

- (i) no accidents in a shift (ii) more than one accident in a shift.

(P.U., B.A., B.Sc., etc.)

- 8.46 A skilled typist, on routine work, kept a record of mistakes made per day during 200 working days.

Mistakes per day	0	1	2	3	4	5	6
No. of days	143	90	42	9	3	1	0

Compute the frequencies of the Poisson distribution which has the same total frequency as the above distribution.

- 8.47 The number of road accidents notified to a certain police station per day is shown in the frequency table relating to a series of 21 successive days.

Accidents per day	1	2	3	4	5	6	7
Frequency	90	113	64	21	7	3	1

Calculate the mean number of accidents a day. Use the Poisson distribution, with this mean, calculate the expected frequencies. Assuming that this distribution continues to apply, calculate the probability of 4 or more accidents being notified on any one day.

- 8.48 (a) Define a Poisson process. What are its properties or assumptions?
 (b) Suppose that customers enter a certain shop at the rate of 30 persons / hr. Using the Poisson distribution, calculate the probability that in a 3-minute interval, no more than 5 persons enter the shop.

(P.U., B.A., B.Sc.)

- 8.49 (a) Flaws in plywood occur at random with an average of one flaw per 50 square feet. Find the probability that a 4-feet x 8-feet sheet will have no flaws? At most one flaw?
 (b) A doctor receives an average of 3 telephone calls from 9 p.m. until 9 a.m. Assuming intervals of calls are a Poisson process, what is the probability that the doctor can not be disturbed by a call if she goes to bed at midnight and rises at 6 a.m.?

(P.U., B.A., B.Sc.)

- 8.50 A computer system in a company has a breakdown once in 25 days, on the average. Breakdowns are a Poisson process, what is the probability of (i) exactly one breakdown in the next 10 days? (ii) more than one breakdown in the next 10 days?

The number of cars passing over a toll bridge during the time interval 10 to 11 a.m. is 300. The cars pass individually and collectively at random. Find the probability that

- not more than 4 cars will pass during 1-minutes interval 10:45 to 10:46.
- 5 or more cars will pass during the same interval.

Find the m.g.f. about the mean for a Poisson distribution and use it to deduce the moment ratios β_1 and β_2 for the distribution.

- Use m.g.f. to prove that the sum of two independent Poisson variables is a Poisson variable.
- Suppose that the probability of an insect laying x eggs is a Poisson distribution with mean m , and that the probability of an egg developing is p . Assuming technical independence of the eggs, show that the probability of a total of x survivors is given by the Poisson distribution with parameter mp .
(P.U., M.A. Stat. 1969)
- What is a negative binomial experiment and what are its properties? Derive the negative binomial distribution.

The probability that a swimmer will succeed in swimming across a lake is 0.4. What is the probability that the tenth swimmer is the fourth one to cross the lake?

If X has a negative binomial distribution, then show that $E(X) = kp/p$ and $\text{Var}(X) = kp/p^2$. If $E(X) = 10$ and $\sigma^2 = 3$, can X have a negative binomial distribution?

Find the probability that a person flipping a coin gets the third head on the seventh flip.

In each of a succession of independent trials, the probability of a certain event A is p . Trials are continued until the event A has been observed exactly k times. If X be the number of trials, show that the distribution of X is

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \dots$$

Find the expected value and standard deviation of X .

The probability that a person will install a black telephone in a residence is estimated to be 0.3. Find the probability that the 10th phone installed in a new sub-division is the 5th black phone.

Describe the negative binomial distribution and show that its variance is greater than its mean.

Calculate the first four cumulants for the negative binomial distribution.

(P.U., M.A. Stat. 1969)

What is a geometric experiment and what are its properties? Derive the geometric probability distribution.

Show that the mean and variance of the geometric distribution are $\mu = 1/p$ and $\sigma^2 = p(1-p)$.

When flipping an unbiased coin, determine the probability that the first head occurs on the third trial.

CHAPTER 9

CONTINUOUS PROBABILITY DISTRIBUTIONS

CONTINUOUS PROBABILITY DISTRIBUTIONS

INTRODUCTION

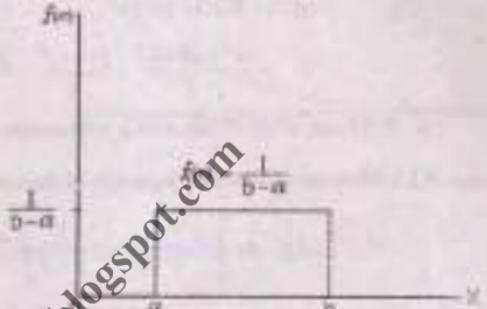
In this chapter, we shall consider some important continuous probability distributions or density functions which are met in practice. Of all the continuous probability distributions, the uniform distribution is perhaps the most important distribution which is used extensively in solving problems both in reliability and in statistical inference.

UNIFORM DISTRIBUTION

The density function of a continuous r.v. X is called a *uniform distribution* when between the two any two subintervals of the same length containing X , have the same probability.

Alternatively, a r.v. X is said to be *uniformly distributed* if its density function is given as

$$\begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}$$



This distribution derives its name from the fact that its density is constant or uniform over the interval $[a, b]$ and is 0 elsewhere.

called the *rectangular distribution* because its total probability is confined to a rectangle whose base is equal to $(b-a)$ and height $1/(b-a)$. The parameters of this distribution are a and b . Since X with this distribution is random variable, therefore we must have that

$$\int_a^b f(x) dx = \int_a^b \frac{1}{b-a} dx = \frac{1}{b-a} [x]_a^b = \frac{b-a}{b-a} = 1$$

which makes sense in the study of rounding off errors, etc. Its distribution function will be

$$F(x) = \begin{cases} 0, & \text{for } x < a, \\ \frac{x-a}{b-a}, & \text{for } a \leq x \leq b, \\ 1, & \text{for } x > b. \end{cases}$$

Properties of the Uniform Distribution.

Let X have the uniform distribution over $[a, b]$. Then its mean is $\frac{a+b}{2}$ and variance is $\frac{(b-a)^2}{12}$.

$$E(X) = \int_a^b f(x) dx = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b$$

$$\Rightarrow \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}, \text{ the midpoint of the interval.}$$

And $\text{Var}(X) = E(X^2) - [E(X)]^2$, where

$$\begin{aligned} \text{Now } E(X^2) &= \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x^2 dx = \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b \\ &= \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3} \\ \therefore \text{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12} \end{aligned}$$

2. The shape of the distribution is rectangular.

9.2.2. Moment Generating Function of the Uniform Distribution. The m.g.f. is obtained as

$$\begin{aligned} M_x(t) &= E[e^{tx}] = \int_a^b e^{tx} \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b e^{tx} dx = \frac{1}{b-a} \left[\frac{1}{t} e^{tx} \right]_a^b \\ &= \frac{e^{tb} - e^{ta}}{(b-a)t} \end{aligned}$$

Putting $a = 0$ and $b = 1$, we obtain the m.g.f. of a uniform distribution in the interval $[0, 1]$. The k th moment about the origin is

$$\mu'_k = \int_a^b x^k \cdot \frac{1}{b-a} dx = \frac{1}{k+1}$$

9.3 EXPONENTIAL DISTRIBUTION

A random variable X is said to have an exponential distribution with parameter λ , if its p.d.f. is defined by

$$\begin{cases} f(x) = \lambda e^{-\lambda x}, & \text{for } x > 0, \\ = 0, & \text{elsewhere,} \end{cases}$$

where $\lambda > 0$. The p.d.f. may also be written as

$$\begin{cases} f(x) = \frac{1}{\beta} e^{-x/\beta}, & \text{for } x > 0, \\ = 0, & \text{elsewhere.} \end{cases}$$

The function $f(x)$ is a proper p.d.f. since

$$\int_0^{\infty} f(x) dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x} \right]_0^{\infty} = 1$$

The distribution function of the exponential r.v. X is given by

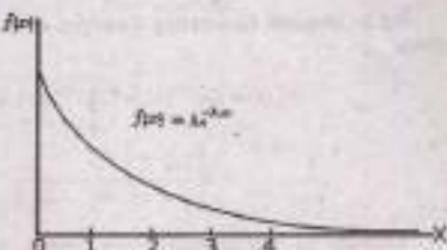
$$F(x) = P(X \leq x)$$

$$= \int_0^x \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t} \right]_0^x$$

$$= 1 - e^{-\lambda x}, \quad \text{for } x > 0,$$

$$= 0, \quad \text{elsewhere,}$$

$$\therefore F(x) = 1 - e^{-\lambda x}.$$



9.3.1 Properties. The following are some of the important properties of the exponential distribution with parameter λ .

- 1. The mean and standard deviation of the exponential distribution are equal.

$$\text{Now } E(X) = \mu = \int_0^{\infty} x f(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx$$

To integrate it by parts, we use the formula $\int u dv = uv - \int v du$ and make the substitution, $du = dv$ and $x = u$, so that $v = -e^{-\lambda x}$ and $du = dx$. Then we have

$$\begin{aligned} \int_0^{\infty} x \lambda e^{-\lambda x} dx &= \left[-x e^{-\lambda x} \right]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx \\ &= 0 + \left[\frac{-e^{-\lambda x}}{\lambda} \right]_0^{\infty} = \frac{1}{\lambda} \end{aligned}$$

$$\therefore E(X) \text{ or } \mu = \frac{1}{\lambda}$$

Again $\text{Var}(X) = E(X^2) - [E(X)]^2$, where

$$\begin{aligned} E(X^2) &= \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx \\ &= \lambda \left[x^2 \left(-\frac{1}{\lambda} e^{-\lambda x} \right) \right]_0^{\infty} + \int_0^{\infty} 2x \lambda e^{-\lambda x} dx \quad (\text{integrating by parts}) \\ &= 0 + \frac{2}{\lambda^2} E(X) = \frac{2}{\lambda^2} \end{aligned}$$

$$\text{Var}(\lambda) = \sigma^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda} \right)^2 = \frac{1}{\lambda^2}$$

and hence $\sigma = \frac{1}{\lambda}$.

2. The distribution is extremely skewed and thus there does not exist any mode.

9.3.2 Moment Generating Function of Exponential Distribution. The *m.g.f.* is obtained as follows:

$$\begin{aligned} M_{\lambda}(t) &= E[e^{tX}] = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{-(\lambda-t)x} dx \\ &= \lambda \left[\frac{-e^{-(\lambda-t)x}}{\lambda-t} \right]_0^{\infty} = \frac{\lambda}{\lambda-t} \quad \text{for } t < \lambda. \end{aligned}$$

Example 9.1 The duration of long-distance telephone calls is found to be exponentially distributed with a mean of 3 minutes. What is the probability that a call will last (i) more than 3 minutes, (ii) more than 5 minutes?

Let X be the exponential r.v. with parameter λ . Then the mean, i.e., $\frac{1}{\lambda} = 3$, so that $\lambda = \frac{1}{3}$.

Now (i) the probability that a call will last more than 3 minutes, is given by

$$P(X > 3) = \int_3^{\infty} \left(\frac{1}{3} \right) e^{-x/3} dx = \left[-e^{-x/3} \right]_3^{\infty} = 0.3679$$

and (ii) the probability that a call will last more than 5 minutes, is given by

$$P(X > 5) = \int_5^{\infty} \left(\frac{1}{3} \right) e^{-x/3} dx = \left[-e^{-x/3} \right]_5^{\infty} = e^{-1.7} = 0.1827.$$

9.4 GAMMA AND BETA DISTRIBUTIONS

The gamma and beta distributions derive their names from the well-known gamma and beta functions which are very important in many areas of probability theory and mathematics. Therefore, before proceeding to these distributions, it is appropriate to review the gamma and beta functions and some of their main properties.

9.4.1 Gamma Function. The *gamma function* for any number $n > 0$, denoted by $\Gamma(n)$, is defined by

$$\Gamma(n) = \int_1^{\infty} z^{n-1} e^{-z} dz$$

We can give an alternative meaning to $\Gamma(n)$ by proving that $\Gamma(n+1) = n!$

Now by definition, $\Gamma(n+1) = \int_0^\infty x^n e^{-x} dx, \quad n > 0$

To integrate the above integral by parts, we use the formula $\int u dv = uv - \int v du$ and make the substitution $u = x^n, dv = e^{-x} dx$ so that $du = nx^{n-1} dx$ and $v = -e^{-x}$. Then we obtain

$$\begin{aligned}\Gamma(n+1) &= \left[-x^{n-1} e^{-x} \right]_0^\infty + \int_0^\infty nx^{n-1} e^{-x} dx \\ &= n \int_0^\infty x^{n-1} e^{-x} dx \\ &= n\Gamma(n) \text{ by the definition of } \Gamma\text{-function.}\end{aligned}$$

$$\begin{aligned}\Gamma(n) &= \int_0^\infty x^{n-1} e^{-x} dx \\ &= \left[-x^{n-1} e^{-x} \right]_0^\infty + \int_0^\infty (n-1)x^{n-2} e^{-x} dx \\ &= (n-1)\Gamma(n-1)\end{aligned}$$

Let n be a positive integer. Then repeating the application of $\Gamma(n) = (n-1)\Gamma(n-1)$, we get

$$\begin{aligned}\Gamma(n+1) &= n \cdot (n-1) \Gamma(n-1) \\ &= n(n-1)(n-2) \dots 3.2.1 = n!\end{aligned}$$

Using $n=0$, we find that

$$\Gamma(0) = \int_0^\infty e^{-x} dx = \left[-e^{-x} \right]_0^\infty = -e^{-\infty} + 1 = 1$$

Hence $\Gamma(n+1) = n(n-1)(n-2) \dots 3.2.1 = n!$

Writing $x = y^2$ in $\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx$, we have

$$\Gamma(n) = 2 \int_0^\infty e^{-y^2} y^{2n-1} dy.$$

Another form of Γ -function.

Using $n = \frac{1}{2}$ in this form, we obtain

$$\Gamma\left(\frac{1}{2}\right) = 2 \int_0^\infty e^{-y^2} dy = \int_0^\infty e^{-y^2} dy = \sqrt{\pi}.$$

9.4.2 Beta Function. The beta function for any two positive numbers m, n , denoted by $B(m, n)$, is defined by

$$B(m, n) = \int_0^1 x^{m-1} (1-x)^{n-1} dx, \quad \text{for } m > 0, n > 0.$$

When $m = n = 1$, $B(1, 1) = \int_0^1 x^0 (1-x)^0 dx = 1$

Writing $x = 1 - z$, we find that

$$\begin{aligned} B(m, n) &= - \int_1^0 (1-z)^{n-1} z^{m-1} dz = \int_0^1 (1-z)^{n-1} z^{m-1} dz \\ &= B(n, m) \end{aligned}$$

Hence the B -function is symmetrical about m and n .

Now let $x = \sin^2 \theta$ so that $dx = 2 \sin \theta \cos \theta d\theta$. Then the substitution gives

$$B(m, n) = 2 \int_0^{\pi/2} \sin^{2m-1} \theta \cos^{2n-1} \theta d\theta$$

Putting $m = n = \frac{1}{2}$, we find that $B\left(\frac{1}{2}, \frac{1}{2}\right) = 2 \int_0^{\pi/2} d\theta = \pi$

Again let $x = \frac{1}{1+z}$, so that $\frac{-1}{(1+z)^2} dz$. Then

$$B(m, n) = \int_0^{\infty} \frac{z^{m-1}}{(1+z)^{m+n}} dz, \text{ which is also } B\text{-function}$$

It is interesting to note that the gamma and beta functions are related according to the following formula:-

$$B(m, n) = \frac{\Gamma(m) \cdot \Gamma(n)}{\Gamma(m+n)}, \quad \text{for } m > 0, n > 0.$$

Putting $m = n = \frac{1}{2}$, we have

$$B\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{\Gamma\left(\frac{1}{2}\right) \cdot \Gamma\left(\frac{1}{2}\right)}{\Gamma(1)} = \left[\Gamma\left(\frac{1}{2}\right)\right]^2$$

$$\text{But } \Gamma\left(\frac{1}{2}, \frac{1}{2}\right) = \pi$$

$$\text{Hence } \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \text{ or } \int_0^\infty x^{n-1} e^{-x^2} dx = \sqrt{\pi}$$

This is a very important result.

9.4.3 Gamma Distribution. A continuous r.v. X is said to have a gamma distribution with parameter $n > 0$, if its p.d.f. is defined by

$$f(x) = \begin{cases} \frac{1}{\Gamma(n)} x^{n-1} e^{-x} & \text{for } 0 \leq x < \infty \\ 0, & \text{elsewhere} \end{cases}$$

gamma variable with parameter n is usually denoted by $\gamma(n)$. A straightforward integration shows that $\int f(x) dx = 1$ and hence it represents a p.d.f.

The distribution function $F(x)$ is

$$F(x) = \int_0^x \frac{1}{\Gamma(n)} x^{n-1} e^{-x} dx,$$

$$= 0, \quad x < 0$$

This is also called the Incomplete Gamma Function and has been tabulated by Karl Pearson.

9.4.4 Properties of Gamma Distribution. The important properties of the gamma distribution are as follows:

1. The mean and variance of the gamma distribution are equal to its parameter n .

$$\text{Now, } \mu = E(X) = \int_0^\infty x f(x) dx = \int_0^\infty x \cdot \frac{1}{\Gamma(n)} x^{n-1} e^{-x} dx$$

$$= \int_0^\infty \frac{1}{\Gamma(n)} x^n e^{-x} dx = \frac{\Gamma(n+1)}{\Gamma(n)} = \frac{n\Gamma(n)}{\Gamma(n)} = n$$

$$\text{and } \text{Var}(X) = E(X^2) - [E(X)]^2, \text{ where}$$

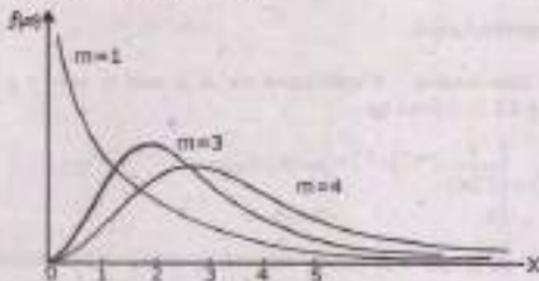
$$E(X^2) = \int_0^\infty x^2 \frac{1}{\Gamma(n)} x^{n-1} e^{-x} dx$$

$$= \frac{\Gamma(n+2)}{\Gamma(n)} = \frac{(n+1)(n)\Gamma(n)}{\Gamma(n)} = n(n+1)$$

$$\sigma^2 = m(m+1) - m^2 = m.$$

Hence mean and variance are each equal to m .

2. The curve of $f(x)$ is asymptotic to the X -axis. If $m > 1$, the curve has a mode at $x=0$. If $m \geq 2$, it touches the X -axis at the origin.



3. The curve becomes asymptotic to both axes when m lies between zero and one.

4. Reproductive Property. The sum of two independent Gamma distributions with parameters m and n is a Gamma distribution with parameter $(m+n)$.

9.4.5 Moment Generating Function of Gamma Distribution. The m.g.f. of X with respect to origin is

$$M(t) = E(e^{tX}) = \int_0^\infty e^{tx} \frac{1}{\Gamma(m)} x^{m-1} e^{-x} dx \\ = \int_0^\infty \frac{1}{\Gamma(m)} e^{-(1-t)x} dx.$$

Let $u = x(1-t)$, so $du/dx = \frac{du}{1-t}$. Then substituting these values, we get

$$M(t) = \frac{1}{\Gamma(m)} \int_0^\infty e^{-u} \left(\frac{u}{1-t} \right)^{m-1} \frac{du}{1-t} \\ = \frac{1}{(1-t)^m} \int_0^\infty \frac{1}{\Gamma(m)} u^{m-1} e^{-u} du \\ = (1-t)^{-m}, \text{ provided that } |t| < 1.$$

On differentiating $M(t)$ r times with respect to t and putting $t=0$, we find,

$$\mu'_r = m(m+1)\dots(m+r-1)$$

Thus $\mu'_1 = m$, $\mu'_2 = m(m+1)$ and hence $\mu'_2 = \mu'_2 - \mu'^2_1 = m$, etc.

The cumulant generating function for this distribution with respect to origin is:

$$\begin{aligned}K(t) &= \log_s M_X(t) = -n \log(1-t) \\&= n \left[t + \frac{t^2}{2} + \frac{t^3}{3} + \dots \right] = n \sum_{r=1}^{\infty} \frac{t^r}{r!} (r-1)!\end{aligned}$$

Thus the r th cumulant is given as

$$K_r = n(r-1)! = n \Gamma(r).$$

9.4.6 Beta Distribution of the First Kind. A continuous r.v. X is said to have a beta distribution with two parameters m and n , if its p.d.f. is defined by

$$f(x) = \begin{cases} \frac{1}{B(m, n)} x^{m-1} (1-x)^{n-1}, & 0 \leq x \leq 1; m, n > 0 \\ 0, & \text{elsewhere} \end{cases}$$

This distribution is known as a *beta distribution of the first kind* and its variable of the first kind is simply referred to as $\beta_1(m, n)$. It is a proper probability density function since the area under the curve is one.

The distribution function $F(x)$ is given by

$$F(x) = \begin{cases} 0, & \text{for } x < 0, \\ \int_0^x \frac{1}{B(m, n)} t^{m-1} (1-t)^{n-1} dt, & \text{for } 0 \leq x \leq 1 \\ 1, & \text{for } x > 1 \end{cases}$$

This is also called the *complete beta function*, and it has been extensively tabulated.

9.4.7 Properties of $\beta_1(m, n)$. The main properties of this distribution are given below:

- The mean and variance of this distribution are $\frac{m}{m+n}$ and $\frac{mn}{(m+n)^2(m+n+1)}$ respectively.

They are computed as below:

$$\begin{aligned}\mu &= E(X) = \int_0^1 \frac{1}{B(m, n)} x^{m-1} (1-x)^{n-1} dx \\&= \int_0^1 \frac{1}{B(m, n)} x^m (1-x)^{n-1} dx = \frac{B(m+1, n)}{B(m, n)} \\&= \frac{m \Gamma(m) \Gamma(n)}{(m+n) \Gamma(m+n)} = \frac{\Gamma(m+n)}{\Gamma(m) \Gamma(n)} \cdot \frac{m}{m+n}\end{aligned}$$

$$\text{Var}(X) = \sigma^2 = E(X^2) - [E(X)]^2, \text{ where}$$

$$E(X^2) = \int x^2 f(x) dx$$

$$= \int_0^1 \frac{1}{B(m, n)} \cdot x^{m+1} (1-x)^{n-1} dx$$

$$= \frac{B(m+2, n)}{B(m, n)} = \frac{m(m+1)}{(m+n)(m+n+1)}$$

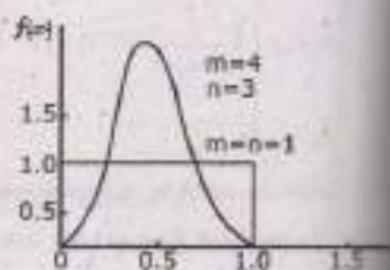
$$\therefore \sigma^2 = \frac{m(m+1)}{(m+n)(m+n+1)} - \left(\frac{m}{m+n} \right)^2 = \frac{mn}{(m+n)^2 (m+n+1)}$$

2. Higher Moments. The r th moment about the origin 0 is given by

$$\begin{aligned} \mu_r &= E(X^r) = \frac{1}{B(m, n)} \int_0^1 x^r \cdot x^{m-1} (1-x)^{n-1} dx \\ &= \frac{B(m+r, n)}{B(m, n)} = \frac{\Gamma(m+r)}{\Gamma(m)} \frac{\Gamma(m+n)}{\Gamma(m+r+n)} \\ &= \frac{m(m+1) \dots (m+r-1)}{(m+n) (m+n-1) \dots (m+n+r-1)} \end{aligned}$$

It should be noted that the m.g.f. of this distribution does not have a simple form.

3. The shape of the beta distribution for $m=4, n=3$ is indicated in the figure.
4. If m and n are both greater than 1, it has a modal value at $x = \frac{n-1}{m+n-2}$. If $m = n = 1$, it reduces to the uniform distribution over the unit interval. The curve of $f(x)$ touches the X -axis at $x=0$, when $m>2$.



9.4.8 Beta Distribution of Second Kind. A continuous r.v. X is said to have a beta distribution of the second kind with parameters m and n , if its p.d.f. is defined by

$$f(x) = \begin{cases} \frac{1}{B(m, n)} \frac{x^{m-1}}{(1+x)^{m+n}}, & \text{for } 0 \leq x < \infty, m, n > 0, \\ 0, & \text{otherwise.} \end{cases}$$

A beta variable of the second kind is generally denoted by $\beta_2(m, n)$. To check that the function represents a proper probability distribution, we observe that

$$\int_0^1 f(x) dx = \int_0^1 \frac{1}{B(m, n)} \cdot \frac{x^{m-1}}{(1+x)^{m+n}} dx$$

Let $1+x = \frac{1}{y}$, so that $x = \frac{1-y}{y}$ and $dx = -\frac{dy}{y^2}$. Then substitution gives

$$\int_0^1 f(x) dx = \int_0^1 \frac{1}{B(m, n)} \cdot y^{m-1} (1-y)^{n-1} dy = 1.$$

4.4.9 Properties of $\beta_2(m, n)$

The important properties of beta distribution of the second kind are

Moments about the origin are easily calculated as follows:

$$\begin{aligned} E(X^r) &= \int_0^1 \frac{1}{B(m, n)} \cdot \frac{x^{m+r-1}}{(1+x)^{m+n}} \cdot x^r dx \\ &= \frac{1}{B(m, n)} \int_0^1 y^{m+r-1} (1-y)^{n+r-1} dy, \text{ on putting } 1+x = \frac{1}{y} \\ &= \frac{1}{B(m, n)} \int_0^1 z^{m+r-1} (1-z)^{n+r-1} dz, \text{ where } z = 1-y \\ &= \frac{B(m+r, n+r)}{B(m, n)} = \frac{m(m+1) \dots (m+r-1)}{(m+1)(m+2) \dots (m+r)}, \quad r < n. \end{aligned}$$

If $n > 1$, the distribution is unimodal with a mode at $x = \frac{m-1}{n+1}$. If $m = 1$, the distribution is J-shaped.

The curve of $f(x)$ is asymptotic to the X-axis and it touches it at the origin if $m > 2$.

The curve touches the Y-axis at the origin if $1 < m < 2$, and the curve becomes asymptotic to both axes when m lies between 0 and 1.

FINAL DISTRIBUTION

The normal probability distribution, which is considered the cornerstone of the modern statistical theory, was discovered by Abraham de Moivre (1667–1754) as the limiting form of the binomial distribution by increasing n , the number of trials, to a very large number for a fixed value of p . But his work remained unnoticed until 1924 when it was found in a library by Karl Pearson (1857–1936). The French mathematician Laplace (1749–1827) is also associated with the derivation of the normal distribution. This distribution is also called the Gaussian distribution in honour of the great German

mathematician Carl F. Gauss (1777–1855), who also derived its equation mathematically. The probability distribution of the errors of measurement. It was Karl Pearson who in 1893 called it *normal distribution* and is best known by this name today.

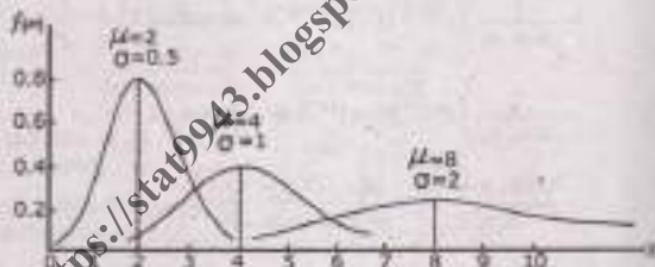
A normal distribution is defined by the p.d.f.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \text{ for } -\infty < x < \infty, \text{ and } \sigma > 0$$

where μ is the mean, σ is the standard deviation and $\pi (=3.1416)$ and $e (=2.7183)$ are constants.

Obviously a normal distribution is characterized by two parameters μ and σ , its mean and standard deviation. Since $\int f(x) dx = 1$ (see properties), the function $f(x)$ is a proper probability density function.

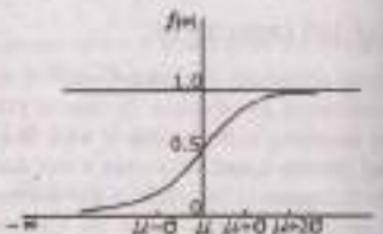
The normal distribution having mean μ and variance σ^2 is usually denoted by $N(\mu, \sigma^2)$. The $N(\mu, \sigma^2)$ means that a r.v. X is normally distributed with mean μ and variance σ^2 . The graph of a normal distribution, which is symmetrical bell-shaped curve, is called the *normal curve*. The basic shape of the normal curve are determined by μ and σ . To put it differently, μ changes the position of the normal curve along horizontal axis while σ determines the horizontal spread. A sketch for three values of μ and σ is given below.



The distribution function of the normal probability distribution is given by

$$F(x) = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-(t-\mu)^2/2\sigma^2} dt,$$

which is sketched below:



This curve is the ogive of the normal curve.

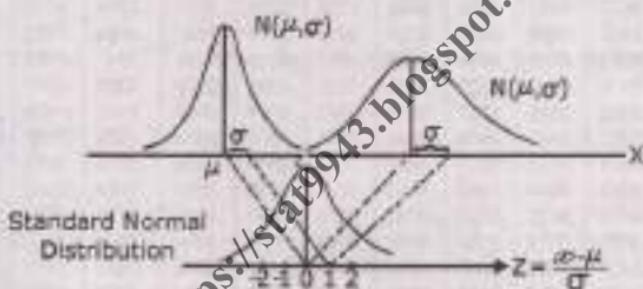
9.5.1 Standardized Normal Distribution. A normal probability distribution depends on the values of the parameters μ and σ^2 and the various possible values for these two parameters will result in an infinite number of different normal distributions. The r.v. $Z = \frac{X - \mu}{\sigma}$, as we have seen, has zero mean and unit variance. Every normally distributed r.v. X with mean $= \mu$ and variance $= \sigma^2$ is therefore easily transformed into a new normal r.v. Z with zero mean and unit variance by using the following expression

$$Z = \frac{X - \mu}{\sigma}$$

Then the p.d.f. of Z , denoted by $\phi(z)$ (ϕ is pronounced *phi*) becomes

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \text{for } -\infty < z < \infty.$$

The following figures illustrate the transformation of the original normal distribution into the standard normal distribution.



The probability distribution of Z which has zero mean and unit variance, is called the *standard normal distribution or unit normal distribution* and is denoted by $N(0, 1)$. The distribution of the standard normal distribution, usually denoted by $\Phi(z)$ (Φ is capital ϕ) is

$$\Phi(z) = P(Z < z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt.$$

Values of $\Phi(z)$ are tabulated for positive values of z . (Table 9.1 contains these values). The values of $\Phi(z)$ for negative values of z are obtained from the identity $\Phi(-z) = 1 - \Phi(z)$. It should be noted that $\Phi(\infty) = 1$ and for any a and b (positive or negative) $P(a < Z < b) = \Phi(b) - \Phi(a)$.

Table 9.1. Cumulative Standard Normal Probabilities



<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	<i>Value of Φ(z)</i>
0	.5000	.5049	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	.5399
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	.5793
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	.6179
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	.6554
.4	.6554	.6593	.6632	.6664	.6700	.6736	.6772	.6808	.6844	.6879	.6915
.5	.6915	.6959	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224	.7257
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549	.7580
.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852	.7881
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133	.8159
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389	.8413
1.0	.8413	.8438	.8461	.8483	.8508	.8531	.8554	.8577	.8599	.8621	.8643
1.1	.8643	.8665	.8686	.8708	.8729	.8751	.8770	.8790	.8810	.8830	.8849
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9013	.9032
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177	.9192
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319	.9332
1.5	.9332	.9343	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441	.9452
1.6	.9452	.9463	.9474	.9484	.9493	.9505	.9515	.9525	.9535	.9545	.9554
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633	.9641
1.8	.9641	.9648	.9654	.9664	.9671	.9678	.9686	.9693	.9699	.9706	.9713
1.9	.9713	.9718	.9724	.9732	.9738	.9744	.9750	.9756	.9763	.9767	.9772
2.0	.9772	.9778	.9784	.9788	.9793	.9798	.9803	.9808	.9812	.9817	.9821
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857	.9861
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890	.9893
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916	.9918
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936	.9938
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952	.9953
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964	.9965
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9996	.9996
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
<i>z</i>		1.292	1.345		1.360		2.326		2.378		
$\Phi(z)$		0.90	0.95		0.975		0.99		0.999		
$\Phi(-z) = 1 - \Phi(z)$		0.10	0.05		0.025		0.01		0.001		

9.5.2 Properties of Normal Distribution. The main properties of the normal distribution are given below:

1. The function $f(x)$ defining the normal distribution is a proper p.d.f., i.e. $f(x) \geq 0$ and the total area under the normal curve is unity.

Proof. Clearly $f(x)$ is always non-negative, and the total probability (area) is

$$\int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$

Let $z = \frac{x-\mu}{\sigma}$. Then $\sigma dz = dx$. Substituting these values, we get

$$\begin{aligned}\text{Area} &= \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2} \sigma dz \\ &= \frac{1}{\sqrt{2\pi}} \left[\int_0^\infty e^{-z^2/2} dz + \int_0^\infty e^{-z^2/2} dz \right]\end{aligned}$$

The function $\int_0^\infty e^{-z^2/2} dz$ being an even function of z , can be by letting $w = -z$, written as

$$\int_0^\infty e^{-z^2/2} dz = \int_0^\infty e^{-w^2/2} dw. \text{ Then}$$

$$\text{Area} = 2 \cdot \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-z^2/2} dz$$

Let $v = \frac{1}{2} z^2$, so that $dv = zdz$. Then

$$\begin{aligned}\text{Area} &= \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^\infty e^{-v} \frac{dv}{\sqrt{2v}} = \frac{1}{\sqrt{\pi}} \int_0^\infty v^{-1/2} e^{-v} dv \\ &= \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1}{2}\right) = \frac{1}{\sqrt{\pi}} \cdot \sqrt{\pi} = 1.\end{aligned}$$

Thus the total area (probability) under the normal curve is unity and hence the function $f(x)$ defines a p.d.f.

2. The mean and variance of the normal distribution are μ and σ^2 respectively.

Proof. By definition, $\mu = E(X) = \int x f(x) dx$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot e^{-(1+\mu)^2/(2\sigma^2)} dx$$

Let $x = \frac{z - \mu}{\sigma}$. Then $x - \mu \neq 0$ and $dx = \sigma dz$.

(Limits:

when $x = \infty$, $z = \infty$;

when $x = -\infty$, $z = -\infty$)

Therefore $E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + z\sigma) e^{-z^2/2} dz$

$$= \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-z^2/2} dz$$

The first integral represents μ times the area under a normal curve with zero mean and variance and hence is equal to μ . The second integral being an odd function, equals zero. Thus

$E(X) = \mu$, i.e. μ is the mean of the normal distribution.

And: $\text{Var}(X) := E(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-(1+\mu)^2/(2\sigma^2)} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (z - \mu)^2 e^{-z^2/2} dz, \quad \text{on putting } z = \frac{x - \mu}{\sigma}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz,$$

To integrate by parts we use the formula: $\int u dv = uv - \int v du$ and make the substitution $dv = ze^{-z^2/2} dz$ and $u = z$ so that $v = -e^{-z^2/2}$ and $du = dz$. Then

$$\text{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}} \left[z \cdot e^{-z^2/2} \right]_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-z^2/2} dz = 0 + \sigma^2 = \sigma^2$$

Hence $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

3. The median and the mode of the normal distribution are each equal to μ , the mean of the distribution.

Now, the median, a , is given by $\int_{-\infty}^a f(x) dx = \frac{1}{2}$. Therefore

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^a e^{-(1+(x-\mu)^2/2\sigma^2)} dx = \frac{1}{2}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\frac{\mu-a}{\sigma}}^0 e^{-t^2/2} dt = \frac{1}{2}, \text{ on putting } t = \frac{x-\mu}{\sigma}.$$

But we know from the symmetry of the standard normal distribution that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-t^2/2} dt = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-t^2/2} dt = \frac{1}{2}$$

$$\frac{\mu-a}{\sigma} = 0 \text{ or } a = \mu, \text{ i.e. } \mu \text{ is the median of the distribution.}$$

Again the mode, if any, is that value of x for which $f'(x) = 0$ and $f''(x) < 0$.

$$f'(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \left(\frac{x-\mu}{\sigma^2} \right) \left(-\frac{2}{\sigma} \right)$$

$$= -\frac{1}{\sigma^3\sqrt{2\pi}} (x-\mu) e^{-(x-\mu)^2/2\sigma^2}$$

On setting $f'(x) = 0$, we see that $x = \mu$.

Differentiating, we obtain

$$f''(x) = \frac{1}{\sigma^3\sqrt{2\pi}} \left[e^{-(x-\mu)^2/2\sigma^2} - e^{-(x-\mu)^2/2\sigma^2} \frac{(x-\mu)^2}{\sigma^2} \right]$$

$$= -\frac{1}{\sigma^5\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \left[1 - \frac{(x-\mu)^2}{\sigma^2} \right]$$

On setting $x = \mu$ in $f''(x)$, we see that $f''(\mu) < 0$. Thus $x = \mu$ is the mode of the normal distribution. Since mean = median = mode, the normal distribution is unimodal.

The mean deviation of the normal distribution is approximately $\frac{4}{5}$ of its standard deviation.

Proof. Now the mean deviation of the normal random variable X from the mean, μ , is given by

$$\begin{aligned}
 M.D. &= E|X - \mu| = \int_{-\infty}^{\infty} |x - \mu| \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \\
 &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{-z^2/2} dz, \quad \text{where } z = \frac{x-\mu}{\sigma} \\
 &= \frac{\sigma}{\sqrt{2\pi}} \left[\int_0^{\infty} z e^{-z^2/2} dz + \int_0^{\infty} z e^{-z^2/2} dz \right] \\
 &= \frac{2\sigma}{\sqrt{2\pi}} \int_0^{\infty} z e^{-z^2/2} dz = \sigma\sqrt{2/\pi} \left[-e^{-z^2/2} \right]_0^{\infty} \\
 &= \sigma\sqrt{2/\pi} = 0.7979\sigma = \frac{4}{5}\sigma, \text{ approximately}
 \end{aligned}$$

5. The normal curve has points of inflection which are equidistant from the mean.

Proof: The point of inflection by which we mean a point at which the concavity changes obtained by solving the equation $f''(x) = 0$.

Differentiating the function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, we get

$$f'(x) = \frac{-1}{\sigma^3\sqrt{2\pi}} (\mu-x)e^{-(x-\mu)^2/2\sigma^2}$$

Equating $f'(x)$ to zero, we see that $x = \mu$. We also observe that $f'(x) > 0$ for $x < \mu$ and $f'(x) < 0$. Thus the maximum of the function $f(x)$ is at $x = \mu$ and its value is $\frac{1}{\sigma\sqrt{2\pi}}$.

To find the points of inflection, we take the second derivative. Thus

$$\begin{aligned}
 f''(x) &= \frac{-1}{\sigma^5\sqrt{2\pi}} \left[\frac{(x-\mu)^3}{\sigma^2} e^{-(x-\mu)^2/2\sigma^2} + 3e^{-(x-\mu)^2/2\sigma^2} \right] \\
 &= \frac{-1}{\sigma^5\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \left[1 - \frac{(x-\mu)^2}{\sigma^2} \right]
 \end{aligned}$$

Equating $f''(x)$ to zero, we are left with the following equation

$$1 - \frac{(x-\mu)^2}{\sigma^2} = 0$$

which gives $(x - \mu)^2 = \sigma^2$

$$\text{or } x - \mu = \pm \sigma \text{ or } x = \mu + \sigma \text{ or } \mu - \sigma.$$

At these two points, the values of the function $f(x)$ is $\frac{1}{\sigma\sqrt{2\pi e}}$.

Hence the two points of inflection of normal curve are $\left[\mu - \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right]$ and $\left[\mu + \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right]$.

In other words, the points of inflection occur on the right and on the left of the mean distance equal to standard deviation and thus the graph of the normal curve is bell-shaped.

6. For the normal distribution, the odd order moments about the mean are all zero and the even order moments are given by

$$\mu_{2n} = (2n-1)!(2n-3) \dots 5.3.1 \cdot \sigma^{2n}.$$

Proof. The odd order moments about the mean are given by

$$\begin{aligned}\mu_{2n+1} &= E[(X-\mu)^{2n+1}] \\ &= \int_{-\infty}^{\infty} (x-\mu)^{2n+1} \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} dx \\ &= \int_{-\infty}^{\infty} (\sigma z)^{2n+1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \sigma dz, \text{ where } z = \frac{x-\mu}{\sigma} \\ &= \frac{\sigma^{2n+1}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n+1} e^{-z^2/2} dz, \\ &= 0, \text{ because the integral is an odd function of } z.\end{aligned}$$

$$\mu_1 = 0 = \mu_3 = \dots$$

Even order moments about the mean are obtained as below:

$$\begin{aligned}\mu_{2n} &= E[(X-\mu)^{2n}] \\ &= \int_{-\infty}^{\infty} (x-\mu)^{2n} \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} dx \\ &= \int_{-\infty}^{\infty} (x\sigma)^{2n} \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2} \sigma dx, \text{ on putting } z = \frac{x-\mu}{\sigma}\end{aligned}$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n} \cdot e^{-z^2/2} dz,$$

$$= \frac{2\sigma^{2n}}{\sqrt{2\pi}} \int_0^{\infty} z^{2n} \cdot e^{-z^2/2} dz,$$

Let $y = \frac{z^2}{2}$. Then $dy = z dz$, and therefore

$$\mu_{2n} = \frac{\sigma^{2n}}{\sqrt{2\pi}} 2^{n+1/2} \int_0^{\infty} y^{(n+1)/2-1} e^{-y} dy$$

$$= \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \int_0^{\infty} y^{(n+1)/2-1} e^{-y} dy$$

$$= \frac{2^n \cdot \sigma^{2n} \Gamma(n+1/2)}{\sqrt{\pi}}$$

$$= \frac{2^n \cdot \sigma^{2n}}{\sqrt{n!}} \left(\frac{2n-1}{2}\right) \left(\frac{2n-3}{2}\right) \cdots \left(\frac{3}{2}\right) \left(\frac{1}{2}\right) \sqrt{\pi}$$

$$= (2e-1)(2n-1)\dots(1) \sigma^{2n}$$

$$\left(\because \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \right)$$

Putting $n=1$ and 2, we get $\mu_1 = \sigma^2$ and $\mu_3 = 3\sigma^4$.

Hence $\beta_1 = 0$ meaning that skewness is zero, and $\beta_2 = \frac{\mu_3}{\mu_1^2} = \frac{3\sigma^4}{\sigma^4} = 3$, i.e. the normal curve has platykurtosis.

This is, in fact, the reason of the choice of the apparently arbitrary value 3 in the definition of platykurtosis, and leptokurtosis.

7. If $X \sim N(\mu, \sigma^2)$ and if $Y = a + bX$, then $Y \sim N(a + b\mu, b^2\sigma^2)$.

Proof. Finding expectation and variance of Y , we get:

$$E(Y) = E(a + bX) = a + b\mu, \quad (\because a \text{ and } b \text{ are constants})$$

$$\text{and} \quad \text{Var}(Y) = E(Y^2) - [E(Y)]^2 = b^2\sigma^2$$

Now the function $X \mapsto a + bX$ may be either a decreasing or increasing function according whether b is negative or positive. If Y is normally distributed, then its p.d.f. is given by

$$h(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x^2-y^2)^2}{2b^2}} \cdot \frac{1}{|b|} \quad \left[\because h(y) = f(x) \left| \frac{dx}{dy} \right| \right]$$

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

which represents the p.d.f. of a r.v. that is $N(a + b\mu, b^2\sigma^2)$

It follows that, if X is $N(\mu, \sigma^2)$ and if $Z = \frac{X-\mu}{\sigma}$, Z is $N(0, 1)$

8. The sum of independent normal variables is a normal variable. Stated differently, if X_1 is $N(\mu_1, \sigma_1^2)$ and X_2 is $N(\mu_2, \sigma_2^2)$, then for independent X_1 and X_2 , $X_1 + X_2$ is $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
9. No matter what the values of μ and σ are, areas under normal curve remain in certain fixed proportions within a specified number of standard deviations on either side of μ . For example, the interval:
 - i) $\mu \pm \sigma$ will always contain 68.26%,
 - ii) $\mu \pm 2\sigma$ will always contain 95.44%,
 - iii) $\mu \pm 3\sigma$ will always contain 99.73%.

Practically all of the area is between $\mu - 3\sigma$ and $\mu + 3\sigma$; the range of the distribution is therefore approximately 6 standard deviations (theoretically curve goes from $-\infty$ to $+\infty$), and we usually cut the graph at these points. This is a very important property of the normal distribution as most of the significance for large samples are based on it.

10. The Quartile Deviation, Q , is found

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu+Q} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{2}$$

$$\text{or } \frac{1}{\sqrt{2\pi}} \int_0^{Q/\sigma} e^{-z^2/2} dz = \frac{1}{4}, \text{ where } z = \frac{x-\mu}{\sigma}$$

we find, from area table, that $\frac{Q}{\sigma} = 0.6745$ or $Q = 0.6745\sigma$, which is also called the Probable Error, a term not used nowadays.

This also gives the values of the quartiles which are:

$$Q_1 = \mu - 0.6745\sigma \text{ and } Q_3 = \mu + 0.6745\sigma$$

11. The normal curve approaches, but never really touches, the horizontal axis on either side of the mean towards plus and minus infinity, that is the curve is asymptotic to the horizontal axis as $x \rightarrow \pm\infty$.

Example 9.2 For a certain normal distribution, the first moment about 10 is 40 and the fourth about 50 is 48. Find its mean and standard deviation.

For a normal distribution, the first moment about an arbitrary mean α is given by

$$\mu_1' = \int (\lambda - \alpha) f(x) dx, \text{ where } f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\begin{aligned} \text{Now } 40 &= \int (x - 10) f(x) dx = \int x f(x) dx - 10 \int f(x) dx \\ &= \mu - 10, \text{ where } \mu \text{ denotes mean.} \\ \mu &= 40 + 10 = 50 \end{aligned}$$

The fourth moment about 50 (which is mean) implies that $\mu_4 = 48$. But for a normal distribution, fourth moment about the mean is $3\sigma^4$. Therefore, we have

$$3\sigma^4 = 48, \text{ which gives } \sigma = 2$$

Hence mean = 50 and standard deviation = 2

9.5.3 Moment Generating and Cumulant Generating Functions of the Normal Distribution

Let the r.v. X be $N(\mu, \sigma^2)$. Then the m.g.f. of X with respect to origin is given by

$$M_0(t) = E(e^{tX}) = \frac{1}{\sigma\sqrt{2\pi}} \int e^{tx} e^{-x^2/2\sigma^2} dx$$

$$\text{Let } z = \frac{x-\mu}{\sigma} \text{ so that } dx = \sigma dz. \text{ Then}$$

$$\begin{aligned} M_0(t) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(\mu+\sigma z)} e^{-z^2/2} \sigma dz \\ &= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\mu t + \frac{1}{2}\sigma^2 z^2} dz \\ &= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2 - 2\mu\sigma z + \mu^2 + \sigma^2)} dz \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-\mu)^2} dz \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \end{aligned}$$

The m.g.f. with respect to mean, μ , is

$$M_\mu(t) = E[e^{t(X-\mu)}] = e^{-\mu t} E[e^{tX}]$$

$$\begin{aligned}
 &= e^{\mu t + \frac{\sigma^2 t^2}{2}} e^{\mu + \frac{1}{2}\sigma^2 t^2} = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \\
 &= 1 + \frac{\left(\frac{1}{2}\sigma^2 t^2\right)}{1!} + \frac{\left(\frac{1}{2}\sigma^2 t^2\right)^2}{2!} + \cdots + \frac{\left(\frac{1}{2}\sigma^2 t^2\right)^n}{n!}
 \end{aligned}$$

Since $\mu_{2n+1} = 0$; and

$$\begin{aligned}
 \mu_{2n} &= \text{coefficient of } \frac{t^{2n}}{(2n)!} \\
 &= \left(\frac{1}{2}\sigma^2\right)^n \frac{(2n)!}{n!} = \frac{\sigma^{2n}}{2^n} \frac{(2n)!}{n!} \\
 &= 1, 3, 5, \dots, (2n-1) \sigma^{2n} \quad \text{for all } n \geq 0
 \end{aligned}$$

The cumulant generating function is given by

$$\kappa(t) = \log_e M_1(t) = \mu t + \frac{1}{2}t^2\sigma^2$$

Thus $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$ and $\kappa_r = 0$ for $r \geq 3$.

all the cumulants after the second are equal to zero.

Example 9.3 If the p.d.f. of the r.v. X is

$$f(x) = \frac{1}{\sqrt{32\pi}} e^{-\frac{(x+7)^2}{32}} \quad -\infty < x < \infty$$

mean, variance and moment generating function:

The function $f(x)$ may be written as

$$f(x) = \frac{1}{4\sqrt{2\pi}} e^{-\frac{(x+7)^2}{16}}$$

Comparing with the general form of the normal distribution, we find that $\mu = -7$ and $\sigma^2 = 16$.
 $X \sim N(-7, 16)$.

Substituting the values of mean and variance in the relation $m.g.f. = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$, we get

$$m.g.f. = e^{-7t + \frac{1}{2}(16t^2)} = e^{-7t + 8t^2}$$

Example 9.4 Let $X \sim N(0, 1)$ mean that X has a normal distribution with zero mean and unit variance. What will be the distribution of $2X - 3$, $\frac{7}{8}X + 5$ and $4X$?

Given that $E(X) = 0$ and $\text{Var}(X) = 1$.

Let $Y = 2X - 3$. Then $E(Y) = E(2X - 3) = 2E(X) - 3 = -3$, and

$$\text{Var}(Y) = \text{Var}(2X - 3) = 2^2 \text{Var}(X) = 4(1) = 4.$$

Hence the distribution of $2X - 3$ is $N(-3, 4)$.

Similarly, we find that $\frac{7}{8}X + 5$ is $N\left(5, \frac{49}{64}\right)$ and $4X$ is $N(0, 16)$.

Example 9.5 If X is $N(0, 1)$, then find the distribution of X^2 .

The distribution of X is

$$dF = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx, \quad -\infty < x < \infty$$

Let $y = x^2$ so that $x = \sqrt{y}$ and $dx = \frac{1}{2\sqrt{y}} dy$. Then the distribution of Y is

$$\begin{aligned} dF' &= \frac{1}{\sqrt{2\pi}} e^{-y/2} \cdot \frac{dy}{\sqrt{y}}, \quad 0 \leq y < \infty \\ &= \frac{1}{\sqrt{2\pi}} e^{-y/2} \cdot y^{1/2} dy \\ &= \frac{1}{\Gamma(1/2)} e^{-y/2} y^{1/2} dm, \text{ where } m = y \text{ and } 0 \leq m < \infty. \end{aligned}$$

This is a Gamma variable with parameter $\frac{1}{2}$.

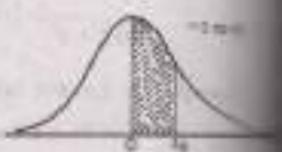
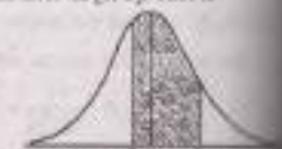
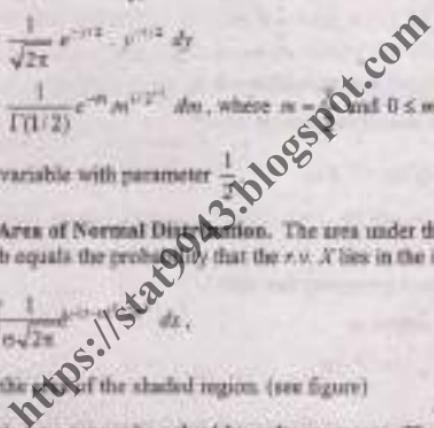
9.5.4 Tabulated Areas of Normal Distribution. The area under the normal curve between two ordinates at $X=a$ and $X=b$ equals the probability that the r.v. X lies in the interval $[a, b]$. That is

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx,$$

which is represented by the area of the shaded region (see figure).

But integrals of this type cannot be solved by ordinary means. They are, however, evaluated by methods of numerical integration, and numerical approximations for some tabulation have been made for quick reference.

Table 9.2 on page (365) gives the areas (probabilities) for the standard normal distribution from the mean, $z=0$ to a specified positive value of z , say z_0 . Since normal curves are symmetrical, therefore $P(0 \leq z) = P(0 \leq -z)$. That is why the areas for negative values of z are not tabulated. It is important to note that this single table for the standard normal distribution suffices for the calculation of probabilities for any normal distribution. Hence, to use the table of areas for the standard normal distribution, the values of the r.v. X in any problem are changed to the values of the standard normal variable Z and the desired probabilities are



obtained from Table 9.2. Thus, to find $P(a < X \leq b)$, we would change X into Z as follows:

$$\begin{aligned} P(a < X \leq b) &= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right), \\ &= P\left(\frac{a-\mu}{\sigma} < Z \leq \frac{b-\mu}{\sigma}\right). \end{aligned}$$

Since $\frac{a-\mu}{\sigma}$ and $\frac{b-\mu}{\sigma}$ are the z -values of the standard normal variable Z . In practice, a normal curve search for the given problem, showing under the X -scale, a scale for the corresponding values of z will help in solving the problem.

Table 9.2 Areas under the Unit Normal Curve

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0159	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0388	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2388	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3021	0.3048	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3263	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3507	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3688	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3905	0.3922	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4081	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4223	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4358	0.4370	0.4382	0.4394	0.4406	0.4418	0.4430	0.4441
1.6	0.4452	0.4463	0.4474	0.4485	0.4495	0.4505	0.4515	0.4525	0.4535	0.4543
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4631
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4758	0.4762	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4865	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4923	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4963	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4980	0.4980	0.4981
2.9	0.4981	0.4982	0.4983	0.4983	0.4984	0.4984	0.4985	0.4985	0.4985	0.4986
3.0	0.49865	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.49903	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993

Example 9.6: Let the r.v. Z have the standard normal distribution. Find

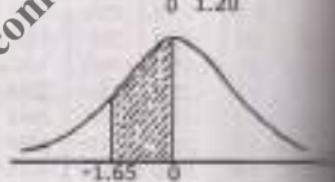
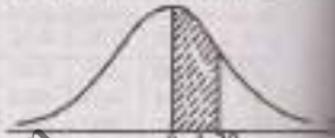
- $P(0 \leq Z \leq 1.20)$,
- $P(-1.65 \leq Z \leq 0)$,
- $P(0.6 \leq Z \leq 1.67)$,
- $P(-1.30 \leq Z \leq 2.18)$,
- $P(-1.96 \leq Z \leq -0.84)$,
- $P(Z \geq 1.96)$, and
- $P(Z \leq -2.15)$.

First we draw the normal curve sketch, shading the desired area (probability) for each part.

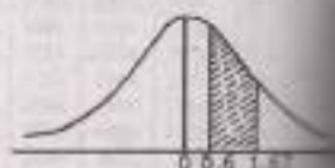
- To find $P(0 \leq Z \leq 1.20)$, in Table 9.2 page 365 we move downward the columns marked Z until 1.2 is reached, and then move across that row to the column headed 0.00 to find entry 0.3849. Therefore $P(0 \leq Z \leq 1.20) = 0.3849$.
- Since the normal curve is symmetrical about the mean, therefore area between $z = 0$ and positive value of z is equal to the area between $z = 0$ and a negative value of z of the same magnitude.

Hence, using Table 9.1 page 365 we have
 $P(-1.65 \leq Z \leq 0) = P(0 \leq Z \leq 1.65) = 0.4500$

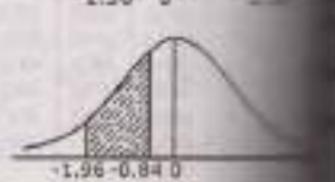
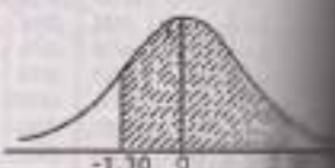
$$\begin{aligned} \text{iii)} \quad & P(0.6 \leq Z \leq 1.67) \\ &= P(0 \leq Z \leq 1.67) - P(0 \leq Z \leq 0.6) \\ &= 0.4525 - 0.3255 \\ &= 0.2268 \text{ (From area tables)} \end{aligned}$$



$$\begin{aligned} \text{iv)} \quad & P(-1.30 \leq Z \leq 2.18) \\ &= P(-1.30 \leq Z \leq 0) + P(0 \leq Z \leq 2.18) \\ &= 0.4032 + 0.4854 \\ &= 0.8886 \text{ (From area tables)} \end{aligned}$$



$$\begin{aligned} \text{v)} \quad & P(-1.96 \leq Z \leq -0.84) \\ &= P(-1.96 \leq Z \leq 0) - P(-0.84 \leq Z \leq 0) \\ &= 0.4759 - 0.2995 \\ &= 0.1755 \text{ (From area tables)} \end{aligned}$$



CONTINUOUS PROBABILITY DISTRIBUTIONS

$$\text{(i)} \quad P(Z \geq 1.96) = 0.5 - P(0 \leq Z \leq 1.96)$$

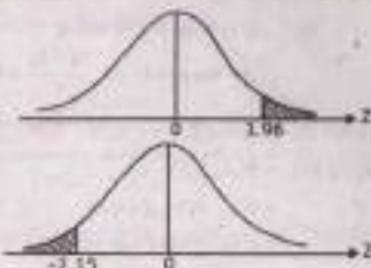
$$= 0.5 - 0.4750$$

$$= 0.0250$$

$$\text{(ii)} \quad P(Z \leq -2.15) = 0.5 - P(-2.15 \leq Z \leq 0)$$

$$= 0.5 - 0.4842$$

$$= 0.0158$$



Example 9.7 A random variable X is normally distributed with $\mu=50$ and $\sigma^2=25$. Find the probability (a) that it will fall between (i) 0 and 40; (ii) 55 and 100; (b) that it will be (i) large than 54, smaller than 57.

We draw the normal curve sketch showing x and z values, and the desired area for each part. With $\mu=50$ and $\sigma=5$, we have

$$z = \frac{x - \mu}{\sigma}$$

$$\text{(i)} \quad \text{At } x=0, \text{ we compute } z = \frac{0-50}{5} = -10, \text{ and}$$

$$\text{at } x=40, \text{ we find } z = \frac{40-50}{5} = -2.0$$

Using Table 9.2 page (365), we have

$$P(0 \leq X \leq 40)$$

$$= P(-10 \leq Z \leq -2)$$

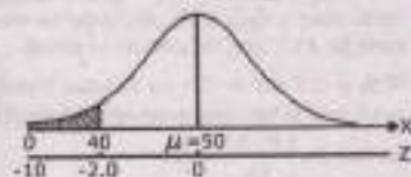
$$= P(-10 \leq Z \leq 0) - P(-2 \leq Z \leq 0)$$

$$= 0.5 - 0.4770 = 0.0228$$

(ii) We have for $x=55$,

$$z = \frac{55-50}{5} = +1.0$$

$$\text{For } x=100, z = \frac{100-50}{5} = 10.0$$



and the corresponding z values are shown in the figure.

Using Table 9.2, we have

$$P(50 \leq X \leq 100) = P(1.0 \leq Z \leq 10.0)$$

$$= P(0 \leq Z \leq 10.0) - P(0 \leq Z \leq 1.0)$$

$$= 0.5 - 0.3413 = 0.1587$$

- b) i) With $\mu = 50$ and $\sigma = 5$, we have
 for $x = 54$, $z = \frac{54 - 50}{5} = 0.80$.

Hence using Table 9.2, we have

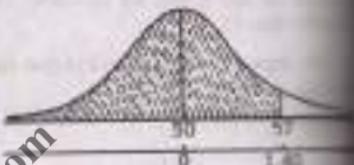
$$\begin{aligned} P(X \geq 54) &= P(Z \geq 0.8) \\ &= 0.5 + P(0 \leq Z \leq 0.8) \\ &= 0.5 + 0.2881 = 0.7881. \end{aligned}$$

ii) At $x = 57$, $z = \frac{57 - 50}{5} = 1.40$.



Therefore using Table 9.2, we have

$$\begin{aligned} P(X \leq 57) &= P(Z \leq 1.40) \\ &= 0.5 + P(0 \leq Z \leq 1.40) \\ &= 0.5 + 0.4192 \\ &= 0.9192. \end{aligned}$$



Example 9.8 The length of life for an automatic dishwasher is approximately normally distributed with a mean of 3.5 years and a standard deviation of 1.0 year. If this type of dishwasher is guaranteed 12 months, what fraction of the sales will require replacement?

The fraction of sales requiring replacement is equal to the area under the normal curve for $X \leq 1$ year, the guaranteed period.

With $\mu = 3.5$ and $\sigma = 1.0$, we then can transform $x = 1.0$ year to the corresponding z values. Thus, we find $z = \frac{1.0 - 3.5}{1.0} = -2.5$.



The x value and the corresponding z values are indicated in the figure. From Table 9.2, we have

$$\begin{aligned} P(X \leq 1.0) &= P(Z \leq -2.5) \\ &= 0.5 - P(-2.5 \leq Z \leq 0) \\ &= 0.5 - 0.4938 = 0.0062. \end{aligned}$$

Hence 0.62% of sales need replacement before 12 months.

Example 9.9 The mean height of soldiers is 68.22 inches with a variance of 10.8 (in.)². Assume the distribution of heights to be normal, how many soldiers in a regiment of 1000 would you expect over 6 feet tall?

With $\mu = 68.22$ and $\sigma^2 = 10.8$ (in.)², we first compute the z value.

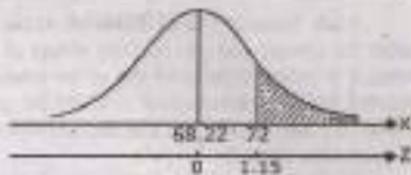
At $x = 72$ (6 ft.), we have

$$z = \frac{72 - 68.22}{\sqrt{10.8}} = \frac{3.78}{3.29} = 1.15$$

The x values and the corresponding z values are shown in normal curve sketch, and we need the right tail area which is shaded.

From using Table 9.2 page (365), we find

$$\begin{aligned} P(Y \geq 72) &= P(Z \geq 1.15) = 0.5 - P(0 \leq Z \leq 1.15) \\ &= 0.5 - 0.3749 = 0.1251 \end{aligned}$$



If there are 1,000 soldiers in the regiment, then number expected to be over 6 feet (or 72 inches) is $0.1251 \times 1000 = 125$.

Example 9.16 If the moment generating function of $X \sim M(t) = e^{166+20t^2}$, find (i) $P(170 < X < 200)$, (ii) $P(148 \leq X \leq 172)$.

Comparing $M(t) = e^{166+20t^2}$ with the M.g.f. of $M(\mu, \sigma^2)$, we find $\mu = 166$ and $\sigma^2 = 400$.

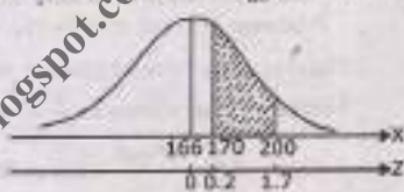
To find the desired probabilities, we transform x values to z values using $z = \frac{x - 166}{20}$. Therefore

At $x = 170$, we get

$$z = \frac{170 - 166}{20} = 0.2, \text{ and}$$

at $x = 200$ we find

$$z = \frac{200 - 166}{20} = 1.7.$$



Using Table 9.2 page (365), we find

$$\begin{aligned} P(170 < X < 200) &= P(0.2 < Z < 1.7) = P(0 < Z < 1.7) - P(-0.2 < Z < 0) \\ &= 0.4550 + 0.0793 = 0.3761 \end{aligned}$$

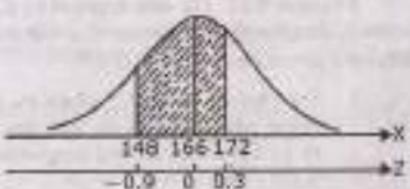
The figure illustrates the problem.

At $x = 148$, we compute

$$\frac{148 - 166}{20} = -0.9, \text{ and}$$

at $x = 172$, we find

$$\frac{172 - 166}{20} = +0.3.$$



Using Table 9.2 page (365), we get

$$\begin{aligned} P(148 \leq X \leq 172) &= P(-0.9 < Z < 0.3) \\ &= P(-0.9 < Z < 0) + P(0 < Z < 0.3) \\ &= 0.3159 + 0.1179 = 0.4338 \end{aligned}$$

9.5.5 Inverse Use of Table of Areas under Normal Curve. When we use the Table of Areas under the normal curve to find the values of z corresponding to a given probability in the body of the table, it is called the *inverse use of the area Table*. The value of z given that $P(0 \leq Z \leq z_0) = k$ is denoted by the symbol $(z | P = k)$. If the given probability does not appear in the table, the closest probability may be taken to find the corresponding value of z . The z value thus obtained is then converted to an x value, using the formula

$$z = \frac{x - \mu}{\sigma}$$

which may be written as $x = \mu + z\sigma$.

The value of z is positive when x lies to the right of μ , and it is negative when x lies to the left of μ . A sketch of the standard normal curve showing the given probability and the location of z on the correct side of μ helps in solving the given problem.

Example 9.11. In a normal distribution, $\mu = 40$ and $P(25 \leq X \leq 55) = 0.8662$. Find $P(20 \leq X \leq 60)$.

Given $\mu = 40$ and $P(25 \leq X \leq 55) = 0.8662$, where X is a normal random variable.

We find that 25 and 55 lie on either side of the mean $\mu = 40$ at equal distance, so that

$$P(\text{between } X=25 \text{ and } \mu=40) = P(\mu=40 \text{ and } X=55) = 0.8662/2$$

Therefore, using Table 9.2 inversely, we find $(z | P = 0.331) = 1.50$

Substitution in the relation $z = \frac{x - \mu}{\sigma}$ gives

$$1.50 = \frac{25 - 40}{\sigma} \quad \text{which yields } \sigma = 10$$

$$\begin{aligned} \text{Now } P(20 \leq X \leq 60) &= P\left(\frac{20 - 40}{10} \leq \frac{X - 40}{10} \leq \frac{60 - 40}{10}\right) = P(-2 \leq Z \leq 2) \\ &= P(-Z \leq 0) + P(0 \leq Z \leq 2) \\ &= 0.4772 + 0.4772 = 0.9544 \end{aligned}$$

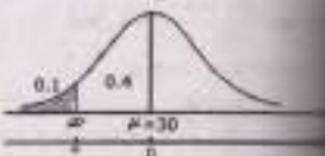
Example 9.12. The time required by a nurse to inject a shot of penicillin has been observed to be normally distributed, with a mean of $\mu = 30$ seconds and a standard deviation of $\sigma = 10$ seconds. Find the following percentiles: (i) 10^{th} ; (ii) 90^{th} .

- Let x be the point P_{10} , where P_{10} is a point at or below which 10% of the area lies. Then the area to the left of x is 0.1 and area between μ and x is $0.5 - 0.1 = 0.4$.

Looking at Table 9.2 page (365) we find that a probability of 0.4 does not appear in the table, so we take the closest probability to 0.4 , which is 0.3997 .

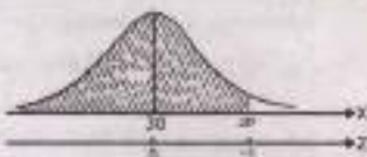
Thus, using Table 9.2 inversely we find $(z | P = 0.3997) = 1.28$.

Since x lies to the left of μ , therefore z is negative at this point.



Hence $x = \mu + z\sigma = 30 + (-1.28)(10) = 17.2$ seconds.

- (ii) Let x be the point P_{90} , where P_{90} is a point at or below which 90% of the area lies. Then the area to the left of x is 0.9 and the area between μ and x is $0.9 - 0.5 = 0.4$.



From Table 9.2 page (365), we find $(zP = 0.4) = 1.28$. Since x lies to the right of μ , therefore z is positive at this point.

$$x = \mu + z\sigma = 30 + (1.28)(10) = 42.8 \text{ seconds.}$$

Example 9.13 In a normal distribution with $\mu=13.5$ and $\sigma=3.6$, find two points such that a single observation has 95% chance for falling between them.

Let x_1 and x_2 be the two points between which the probability of an observation falling is 0.95. As the curve is symmetrical, so half of 0.95, i.e. 0.475 is the area lying on each side of μ .



Using Table 9.2 inversely, we find

$$(zP = 0.475) = 1.96$$

The point x_1 is to the left of μ , so $z = -1.96$ and at x_2 we have to the right of μ , $z = 1.96$. Therefore

$$x_1 = \mu + z\sigma = 13.5 + (-1.96)(3.6) = 6.4 \text{ and}$$

$$x_2 = \mu + z\sigma = 13.5 + (1.96)(3.6) = 20.6.$$

Example 9.14 An athlete finds that in a high jump he can clear a height of 1.68m once in five attempts and a height of 1.52m nine times out of ten attempts. Assuming the heights he can clear in jumps form a normal distribution, estimate the mean and standard deviation of the distribution.

Let X denote the height the athlete can clear in various jumps. Then X is $N(\mu, \sigma^2)$ where μ and σ are unknown.

Now there is a probability of $\frac{1}{5} = 0.2$, if he can clear a height of 1.68m, i.e. $P(X > 1.68) = 0.2$ and

a probability of $\frac{9}{10} = 0.9$, if he can clear a height of 1.52m, that is $P(X \geq 1.52) = 0.9$, implying that $P(X < 1.52) = 0.1$.

The probability (area) between μ and $x = 1.68$ is 0.3 , and between μ and $x = 1.52$ is $0.5 - 0.1 = 0.4$.

From Table 9.2 page (365) inversely we find that $(zP = 0.4) = 0.84$.

Since x lies to the right of μ , therefore z is positive at this point.

and $(zP = 0.1) = 1.28$, since x lies to the left of μ , so z is negative at this point.



Substitution in the relation $x = \mu + z\sigma$, gives

$$\mu + 0.84\sigma = 1.68$$

$$\mu - 1.28\sigma = 1.52$$

Subtracting, we get $2.12\sigma = 0.16$ or $\sigma \approx 0.075$.

Putting $\sigma \approx 0.075$ in $\mu + 0.84\sigma = 1.68$, we obtain

$$\mu + 0.84(0.075) = 1.68$$

$$\text{or } \mu = 1.68 - 0.063 = 1.617$$

Hence the estimated value of μ is 1.617 and of σ is 0.075.

Example 9.15 A collection of human skulls is divided into three classes A, B and C according to the value of a "length-breadth index" X . Skulls with $X < 75$ are classified as A (long-headed), those with $75 < X < 80$ as B (medium) and those with $X > 80$ as C (short-headed). The percentages in the three classes in this collection are 58, 38 and 4. Find approximately the mean and the standard deviation of X , with the assumption that X is normally distributed.

Let μ and σ be the mean and the standard deviation of the normal distribution respectively. Then the area of skull A, whose length-breadth index is under 75 is 0.58, the area of skull B whose index lies between 75 and 80 is 0.38, and the area under skull C, whose index is over 80 is 0.04. The accompanying sketch of the normal curve shows the given information.



The area between μ and $X=75$ is $0.58 - 0.5 = 0.08$, while the area between μ and $X=80$ is $0.68 - 0.58 = 0.10$. In Table 9.2, we find that these probabilities do not appear there, so we take the closest probabilities to 0.08 and 0.10 which are 0.0793 and 0.4599. Therefore

$$(z_1)^P = 0.0793, 0.20 \text{ and } (z_2)^P = 0.4599 = 1.75$$

These z values are positive as the z values lie to the right of μ .

Since $x = \mu + z\sigma$, therefore we get

$$\mu + 0.20\sigma = 75, \text{ and}$$

$$\mu + 1.75\sigma = 80$$

Solving these two equations simultaneously, we obtain

$$\mu = 74.4 \text{ and } \sigma = 3.23$$

Example 9.16 A lawyer commutes daily from his suburban home to his midtown office. On average, the trip one way takes 24 minutes, with a standard deviation of 3.8 minutes. Assume the distribution of trip-times to be normally distributed.

- a) What is the probability that a trip will take at least $\frac{1}{2}$ hour?

- If the office opens at 9:00 AM and he leaves his house at 8:42 AM daily, what percentage of the time is he late for work?
- If he leaves the house at 8:35 AM and coffee is served at the office from 8:50 AM until 9:00 AM, what is the probability that he misses coffee?
- Find the length of time above which we find the slowest 15% of the trips.
- Find the probability that 2 of the next 3 trips will take at least $\frac{1}{2}$ hour.

(P.U., M.Sc. 1989; I.I.T., M.Sc. 1986, 93)

Let the r.v. X denote the trip time in minutes. Then $X \sim N[24, (3.8)^2]$.We need to calculate the probability that a trip will take at least $\frac{1}{2}$ hour, i.e. $P(X \geq 30)$.

$$P(X \geq 30) = P\left(\frac{X - 24}{3.8} \geq \frac{30 - 24}{3.8}\right) = P(Z \geq 1.58) = 0.0571$$

He leaves home at 8:45 AM and the office opens at 9:00 AM implies that he has 15 minutes to reach the office. He will be late for work if he takes more than 15 minutes. Thus we need $P(X > 15)$.

$$\begin{aligned} P(X \geq 15) &= P\left(\frac{X - 24}{3.8} \geq \frac{15 - 24}{3.8}\right) = P(Z \geq -2.37) \\ &= 0.5 + P(0 \leq Z \leq 2.37) = 0.5 + 0.9911 = 0.9911 \end{aligned}$$

The percentage of the time he will be late for work is 99.11%.

He leaves home at 8:35 AM and coffee is served from 8:50 AM until 9:00 AM. He will miss coffee if he reaches office after 9:00 AM, i.e. if he takes 25 minutes or more time. Thus we need $P(X \geq 25)$.

$$\begin{aligned} P(X \geq 25) &= P\left(\frac{X - 24}{3.8} \geq \frac{25 - 24}{3.8}\right) = P(Z \geq 0.26) \\ &= 0.5 - P(0 \leq Z \leq 0.26) = 0.5 - 0.1026 = 0.3974 \end{aligned}$$

To calculate the length of time (x) above which we find the slowest 15% of the trips, we need to calculate the value of x such that $P(X \geq x) = 0.15$.

$$\text{Then } P(X \geq x) = 0.15 \text{ or } P\left(\frac{X - 24}{3.8} \geq \frac{x - 24}{3.8}\right) = 0.15$$

$$\therefore P\left(Z \geq \frac{x - 24}{3.8}\right) = 0.15$$

From the Tables of Standard Normal distribution that the value of z corresponding to a tail area of 0.14 is -1.04.

$$x = \mu + z\sigma = 24 + (-1.04)(3.8) = 27.952 \text{ minutes}$$

- c) Using the binomial distribution with $n=3$ and $p=0.0571$ (where p is the probability the bus will take at least $\frac{1}{2}$ hour), we find that the probability that 2 out of the next 3 trips will take at least $\frac{1}{2}$ hour is $\binom{3}{2} (0.0571)^2 (0.9429) = 0.0092$.

9.5.6 Normal Approximation to the Binomial Distribution. The binomial distribution can be closely approximated by the normal distribution when n is sufficiently large and p not close to zero. As a rule of thumb, the normal distribution provides a reasonable approximation to the binomial distribution if both np and nq are equal to or greater than 5.

The probability for a binomial random variable X to take the value x is:

$$f(x) = \binom{n}{x} p^x q^{n-x}, \text{ for } 0 \leq x \leq n \text{ and } q = 1-p$$

The variable X has a mean of $\mu = E(X) = np$ and $\text{Var}(X) = npq$.

To derive the limiting form, we define a new random variable Z by the relation

$$Z = \frac{X - np}{\sqrt{npq}}$$

Now, obviously $E(Z) = 0$ and $\text{Var}(Z) = 1$.

Thus we have to show that Z is normally distributed in the limit as $n \rightarrow \infty$ and p is fixed (this is actually the de Moivre-Laplace theorem). We shall use the moment generating function (mgf) of the theorem.

The mgf of Z is

$$\begin{aligned} M(t) &= E(e^{tZ}) = E\left[e^{t(X-np)/\sqrt{npq}}\right] \\ &= e^{-np/\sqrt{npq}} \left[E\left[e^{t(X-np)/\sqrt{npq}}\right]\right] \\ &= e^{-np/\sqrt{npq}} \sum_{x=0}^{\infty} \binom{n}{x} p^x q^{n-x} e^{tx/\sqrt{npq}} \\ &= e^{-np/\sqrt{npq}} \sum_{x=0}^{\infty} \binom{n}{x} \left(pe^{t/\sqrt{npq}} \right)^x q^{n-x} \\ &= e^{-np/\sqrt{npq}} \left(q + pe^{t/\sqrt{npq}} \right)^n \\ &= \left(qe^{-p/\sqrt{npq}} + pe^{t/\sqrt{npq}} \right)^n \\ &= \left[q \sum_{r=0}^{\infty} \left(\frac{-pt}{\sqrt{npq}} \right)^r / r! + p \sum_{r=0}^{\infty} \left(\frac{pt}{\sqrt{npq}} \right)^r / r! \right]^n \end{aligned}$$

$$\begin{aligned}
 &= \left[q \left(1 - \frac{pt}{\sqrt{npq}} + \frac{1}{2!} \left(\frac{pt}{\sqrt{npq}} \right)^2 - \frac{1}{3!} \left(\frac{pt}{\sqrt{npq}} \right)^3 + \frac{1}{4!} \left(\frac{pt}{\sqrt{npq}} \right)^4 - \dots \right) \right. \\
 &\quad \left. + p \left(1 + \frac{qt}{\sqrt{npq}} + \frac{1}{2!} \left(\frac{qt}{\sqrt{npq}} \right)^2 + \frac{1}{3!} \left(\frac{qt}{\sqrt{npq}} \right)^3 + \frac{1}{4!} \left(\frac{qt}{\sqrt{npq}} \right)^4 + \dots \right) \right] \\
 &= \left[1 + \frac{t^2}{2!} \cdot \frac{1}{n} + \frac{t^2}{2!} \cdot \frac{q-p}{n\sqrt{npq}} + \frac{t^4}{4!} \cdot \frac{q^2 - qp + p^2}{n^2 qp} + \dots \right].
 \end{aligned}$$

Using logs to the base e , we have

$$\begin{aligned}
 \log_e M(t) &= n \log_e \left[1 + \frac{t^2}{2!} \cdot \frac{1}{n} + \frac{t^2}{2!} \cdot \frac{q-p}{n\sqrt{npq}} + \frac{t^4}{4!} \cdot \frac{1-3pq}{n^2 qp} + \dots \right] \\
 &= \frac{t^2}{2!} + \frac{t^2}{2!} \cdot \frac{q-p}{\sqrt{npq}} + \frac{t^4}{4!} \cdot \frac{1-5pq}{npq} + \dots
 \end{aligned}$$

$$\lim_{n \rightarrow \infty} \log_e M(t) = \frac{t^2}{2!} \text{ or } \lim_{n \rightarrow \infty} M(t) = e^{t^2/2}$$

We know that $e^{t^2/2}$ is the m.g.f. of a standard normal variable.

Hence we find that in the limit Z has a standard normal distribution.

It is important to note that a binomial variable is discrete whereas the normal curve probability is a density for an interval. Therefore, using normal curve areas to approximate binomial probabilities, a discrete value of the binomial variable is to be replaced by an interval before the z values are computed. Usually, a discrete value x becomes the interval from $x-0.5$ to $x+0.5$; and this sort of adjustment is called continuity correction. Thus, the discrete value 5, adjusted means 4.5 to 5.5.

Example 9.17 A fair coin is tossed 20 times. Find the probability that the number of heads is between 10 and 14 inclusive by using (a) the binomial distribution, (b) the normal distribution to the binomial distribution.

Let X denote the number of heads occurring. Then the p.d. of X is

$$P(X = x) = f(x) = \binom{20}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{20-x}$$

and the desired probability is $P(10 \leq X \leq 14)$.

$$\text{Now } P(10 \leq X \leq 14) = \sum_{x=10}^{14} \binom{20}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{20-x}$$

$$\begin{aligned}
 &= \binom{20}{10} \left(\frac{1}{2}\right)^{10} + \binom{20}{11} \left(\frac{1}{2}\right)^{11} + \binom{20}{12} \left(\frac{1}{2}\right)^{12} + \binom{20}{13} \left(\frac{1}{2}\right)^{13} + \binom{20}{14} \left(\frac{1}{2}\right)^{14} \\
 &= 0.1762 + 0.1602 + 0.1201 + 0.0739 + 0.0370 \\
 &= 0.5674
 \end{aligned}$$

Alternatively

$$\begin{aligned}
 P(10 \leq X \leq 14) &= \sum_{x=10}^{14} b(x; 20, 0.5) \\
 &= \sum_{x=10}^{14} b(x; 20, 0.5) - \sum_{x=0}^9 b(x; 20, 0.5) \\
 &= 0.9793 - 0.4119 \quad (\text{from binomial tables}) \\
 &= 0.5674
 \end{aligned}$$

- b) Since $np = 20(0.5) = 10 > 5$ and $nq = 20(0.5) = 10 \geq 5$, so we will use the normal distribution to approximate the binomial distribution.

Now $\mu = np = 20(0.5) = 10$, and

$$\sigma = \sqrt{npq} = \sqrt{20(0.5)(0.5)} = 2.24.$$

For the normal approximation, the interval of discrete value $10 \leq X \leq 14$ is replaced by the interval $9.5 \leq X \leq 14.5$. We compute the z -scores as below:

At $x = 9.5$, we find $z = \frac{9.5 - 10}{2.24} = -0.22$, and

at $x = 14.5$, we get $z = \frac{14.5 - 10}{2.24} = +2.01$

Hence, using Table 9.2, we find

$$\begin{aligned}
 P(10 \leq X \leq 14) &= P(-0.22 \leq Z \leq 2.01) \\
 &= P(-0.22 \leq Z \leq 0) + P(0 \leq Z \leq 2.01) \\
 &= 0.0871 + 0.4778 = 0.5649
 \end{aligned}$$

Hence the probability of obtaining heads between 10 and 14 is 0.5649.

Example 9.18 A pair of dice is rolled 180 times. Use the normal approximation method to find the probability that a total of 7 occurs. (i) at least 25 times, (ii) between 33 and 41 times inclusive, (iii) exactly 30 times.

Let X denote the number of times a total of 7 occurs when a pair of dice is rolled. Then X is a binomial variable with $n=180$ and $p = \frac{1}{6}$ (probability of getting a total of 7 with 2 dice).

- Since n is large and p is not too small, so we use the normal approximation method with $\mu = 180 \times \frac{1}{6} = 30$, and

$$\sigma = \sqrt{npq} = \sqrt{180 \times \frac{1}{6} \times \frac{5}{6}} = 5,$$

and the desired probabilities.

- The interval at least 25 includes 25; therefore it starts at 24.5 and extends to ∞ , i.e. at least 25 becomes the interval 24.5 to ∞ .

The corresponding z value is $z = \frac{24.5 - 30}{5} = -1.1$.

Hence, using Table 9.2, we find

$$\begin{aligned} P(\text{at least } 25) &= \sum_{x=25}^{\infty} p\left(x; 180, \frac{1}{6}\right) = P(-1.1 \leq Z \leq \infty) \\ &= P(-1.1 \leq Z \leq 0) + P(0 \leq Z \leq \infty) \\ &= 0.3643 + 0.5 = 0.8643 \end{aligned}$$

The probability of obtaining at least 25 times seven is 0.8643.

- The interval of discrete values $33 \leq X \leq 41$ is replaced by the interval $32.5 \leq X \leq 41.5$, and the corresponding z values are:

$$\text{at } x = 32.5, z = \frac{32.5 - 30}{5} = 0.5 \text{ and}$$

$$\text{at } x = 41.5, z = \frac{41.5 - 30}{5} = 2.3.$$

Hence using Table 9.2, we find

$$\begin{aligned} P(33 \leq X \leq 41) &= P(0.5 \leq Z \leq 2.3) \\ &= P(0 \leq Z \leq 2.3) - P(0 \leq Z \leq 0.5) \\ &= 0.4893 - 0.1915 = 0.2978. \end{aligned}$$

- The discrete value 30, adjusted for continuity, becomes the interval 29.5 to 30.5, and the corresponding z values are:

$$\text{at } x = 29.5, z = \frac{29.5 - 30}{5} = -0.1 \text{ and}$$

$$\text{at } x = 30.5, z = \frac{30.5 - 30}{5} = 0.1.$$

Hence using Table 9.2, we obtain

$$\begin{aligned} P(X=30) &= P(-0.1 \leq Z \leq 0.1) \\ &= 2(0.0398) = 0.0796. \end{aligned}$$

9.5.7 Normal Approximation to the Poisson Distribution. The Poisson distribution $p(x; \mu)$ can also be approximated by the normal distribution when $\mu \rightarrow \infty$. The probability for a Poisson r.v. X to take the value x is

$$P(X = x) = \frac{e^{-\mu} \mu^x}{x!}, \quad \text{for } x = 0, 1, 2, \dots, \infty$$

and $E(X) = \text{Var}(X) = \mu$.

Let $Z = \frac{X - \mu}{\sqrt{\mu}}$ so that $E(Z) = 0$ and $\text{Var}(Z) = 1$.

In order to show that as $\mu \rightarrow \infty$, the limiting distribution of Z is standard normal distribution we use the moment generating function.

Now, the m.g.f. of Z is

$$\begin{aligned} M(t) &= E(e^{tZ}) = E\left[e^{t(X-\mu)/\sqrt{\mu}}\right] \\ &= e^{-t\sqrt{\mu}} E\left[e^{t(X/\sqrt{\mu})}\right] \\ &= e^{-t\sqrt{\mu}} \sum_{x=0}^{\infty} e^{t(x/\sqrt{\mu})} \frac{e^{-\mu} \mu^x}{x!} \\ &= e^{-t\sqrt{\mu}} e^{-\mu} \frac{\left(\frac{te^{t/\sqrt{\mu}}}{\sqrt{\mu}}\right)^x}{x!} \\ &= e^{-t\sqrt{\mu}} e^{-\mu} \left[1 + \mu e^{t/\sqrt{\mu}} + \frac{(\mu e^{t/\sqrt{\mu}})^2}{2} + \dots\right] \\ &= e^{-t\sqrt{\mu}} e^{-\mu} e^{\mu e^{t/\sqrt{\mu}}} \\ &= e^{-t\sqrt{\mu}} e^{\mu(e^{t/\sqrt{\mu}} - 1)} \end{aligned}$$

Taking logs to the base e , we get

$$\begin{aligned} \log_e M(t) &= -t\sqrt{\mu} + \mu(e^{t/\sqrt{\mu}} - 1) \\ &\approx -t\sqrt{\mu} + \mu \left[1 + \frac{t}{\sqrt{\mu}} + \frac{t^2}{2\mu} + \frac{t^3}{3\mu\sqrt{\mu}} + \dots - 1\right] \end{aligned}$$

$$\begin{aligned} &= -t\sqrt{\mu} + t\sqrt{\mu} + \frac{t^2}{2!} + \frac{t^3}{3!\sqrt{\mu}} + \frac{t^4}{4!\mu} + \dots \\ &= \frac{t^2}{2!} + \frac{t^3}{3!\sqrt{\mu}} + \frac{t^4}{4!\mu} + \dots \end{aligned}$$

Therefore $\lim_{\mu \rightarrow \infty} \log M(t) = \frac{t^2}{2!}$ or $\lim_{\mu \rightarrow \infty} M(t) = e^{t^2/2}$

which is the m.g.f. of the standard normal distribution.

Hence we see the Poisson distribution approaches the normal distribution as $\mu \rightarrow \infty$.

Example 9.19 The number of calls received by an office switch board per hour follows a Poisson distribution with parameter 25. Find the probabilities that in one hour (a) there are between 23 and 26 (inclusive), (b) more than 30 calls, using the normal approximation to the Poisson distribution.

Let X be the r.r. the number of calls received in one hour. Then X is $p(x; 25)$. We require (a) $P(23 \leq X \leq 26)$ and (b) $P(X > 30)$.

Using the normal approximation, X is $N(25, 25)$.

- (a) $P(23 \leq X \leq 26)$ becomes on continuous scale $P(22.5 \leq X \leq 26.5)$. The z-values are:

$$\text{at } x = 22.5, z = \frac{22.5 - 25}{5} = -0.5, \text{ and}$$

$$\text{at } x = 26.5, z = \frac{26.5 - 25}{5} = 0.3.$$

$$P(22.5 \leq X \leq 26.5) = P(-0.5 \leq Z \leq 0.3)$$

$$= P(-0.5 < Z < 0.5) + P(0 < Z < 0.3)$$

$$= 0.1915 + 0.1179 = 0.3094$$

- (b) $P(X > 30)$ becomes on continuous scale $P(X > 30.5)$

$$P(X > 30.5) = P\left(\frac{X - 25}{5} > \frac{30.5 - 25}{5}\right) = P(Z > 1.1)$$

$$= 0.5 - P(0 < Z \leq 1.1)$$

$$= 0.5 - 0.3643 = 0.1357$$

9.5.8 Fitting a Normal Distribution

There are two possible situations to deal with.

- Given an observed frequency distribution with k classes, we need to calculate frequencies which we would expect for a normal distribution with the same mean and standard deviation as the given data.

- b) Given only values of mean μ and variance σ^2 , we need to determine the number of classes (and hence their width) for the presentation of the distribution.

Case (a). To fit a normal distribution to an observed frequency distribution when neither the mean nor the variance σ^2 is known, we proceed as below:

- We estimate the two parameters μ and σ^2 by calculating \bar{x} and s^2 from the observed frequency distribution.
- We calculate the standard normal z-values corresponding to the upper class-boundaries subtracting the estimated mean from each upper class-boundary and dividing by s . Since normal distribution is defined from $-\infty$ to $+\infty$, we therefore extend the lowest class-boundary to $-\infty$ and the last (highest) class forward to $+\infty$.
- We find the cumulative probability $P(Z \leq z) = \Phi(z)$ associated with each z-value, using Tables of standard normal distribution.
- We then obtain the probability, p (or area) for each class by successive subtraction of $\Phi(z)$.
- We finally get the expected (theoretical) frequencies by multiplying each of the probabilities by the total frequency of the given distribution.

This procedure is also known as fitting a normal distribution to a given frequency distribution *area method*.

Case (b). To fit a normal distribution when we are given only the values of the two parameters μ and σ^2 , we proceed as follows:

- We determine the practical range of the distribution as $\mu - 3\sigma$ to $\mu + 3\sigma$, since most normal distributions are practically covered by a range of 6 standard deviations.
- We choose all the classes (between 6 and 15 classes) of a convenient width within the range.
- We proceed as in Case (a) above to calculate the probabilities of the theoretical probabilities of the distribution. The expected frequencies can be obtained when the total number of observations (total frequency) is available.

The following examples illustrate the procedure for fitting a normal distribution:

Example 9.20 Fit a normal distribution to the following frequency distribution of weights:

Weight (kg)	28–31, 32–35, 36–39, 40–43, 44–47, 48–51, 52–55, 56–59, 60–63, ...
f_i	1 14 56 172 245 263 134 67 23

We first calculate the mean and standard deviation of this distribution to estimate the two parameters of the normal distribution.

Computation of mean and standard deviation.

Weight (kg)	x_i	f_i	$u_i = \frac{(x_i - 49.5)}{4}$	$f_i u_i$	$f_i u_i^2$
28-31	29.5	1	-5	-5	25
32-35	33.5	14	-4	-56	224
36-39	37.5	56	-3	-168	504
40-43	41.5	172	-2	-344	688
44-47	45.5	245	-1	-245	245
48-51	49.5	262	0	0	0
52-55	53.5	156	1	156	156
56-59	57.5	67	2	134	268
60-63	61.5	23	3	69	207
64-67	65.5	3	4	12	48
Σ	—	1000	—	-477	2365

$$\text{Now, } \bar{x} = \mu + \frac{\sum f_i u_i}{n} \times h \\ = 49.5 + \frac{(-477)}{1000} \times 4 = 49.5 - 1.79 = 47.71 \text{ kg.}$$

$$\sigma = \sqrt{\frac{\sum f_i u_i^2}{n} - \left(\frac{\sum f_i u_i}{n} \right)^2} \\ = 4 \sqrt{\frac{2365}{1000} - \left(\frac{-477}{1000} \right)^2} = 4 \sqrt{2.365 - 0.1998} \\ = 4 \sqrt{2.1652} = 4 \times 1.47 = 5.88 \text{ kg.}$$

The necessary calculations for the expected frequencies of the fitted distribution are shown below:

Weight (kg)	upper class boundary	$\frac{s}{(\mu + x_i h - \bar{x})}$	$P(Z < z) \Phi(z)$	Probability (\hat{p})	Expected frequency ($\hat{p} \times \sum f$)
-∞-27	27.5	-3.44	0.0003	0.0003	—
28-31	31.5	-2.76	0.0029	0.0029	3
32-35	35.5	-2.08	0.0188	0.0188	16
36-39	39.5	-1.40	0.0808	0.0808	62
40-43	43.5	-0.71	0.2389	0.1581	158
44-47	47.5	-0.03	0.4880	0.2491	249
48-51	51.5	+0.64	0.7389	0.2509	251
52-55	55.5	1.32	0.9066	0.1677	168
56-59	59.5	2.01	0.9778	0.0713	71
60-63	63.5	2.68	0.9963	0.0185	18
64-67	67.5	3.37	0.9996	0.0033	4
68-∞	∞	=	1.0000	0.0004	—
Σ	—	—	—	—	1,000

Example 9.21. Fit a normal distribution given mean $\mu = 60$ and standard deviation $\sigma = 2.5$.

Since the bulk of the normal distribution lies between $\mu - 3\sigma$ and $\mu + 3\sigma$, therefore the range classes would be $60 \pm 3(2.5)$ i.e. 52.5 to 67.5.

As range = 15, we use 8 classes with a common width $k=2$. We then construct the classes 52.5–54.5, 54.5–56.5, ..., 66.5–68.5. Following are the necessary computations:

Classes	upper class boundary	$\frac{z}{k} = \frac{x.c.b. - \mu}{\sigma}$	$P(Z < z) = \Phi(z)$	Probability (f_i)
Up to 52.5	52.5	-3.0	0.0013	0.0013
52.5–54.5	54.5	-2.2	0.0139	0.0126
54.5–56.5	56.5	-1.4	0.0808	0.0669
56.5–58.5	58.5	-0.6	0.2743	0.1935
58.5–60.5	60.5	0.2	0.5793	0.3050
60.5–62.5	62.5	1.0	0.8413	0.2620
62.5–64.5	64.5	1.8	0.9663	0.1228
64.5–66.5	66.5	2.6	0.9933	0.0312
Over 66.5	∞	∞	0.9999	0.0047

If the total frequency were available, we could have obtained the expected frequencies by multiplication of the probabilities by the total frequency.

9.5.9 Ordinates of Normal Distribution. The ordinate (height) is the value of $\phi(x)$ corresponding to a specified value of x . For convenience, the ordinates of the standard normal curve at various distances from the mean have been tabulated. Table 9.3 on page (383) gives the ordinates obtained from ϕ -function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \text{where } z = \frac{x - \mu}{\sigma}$$

for different positive values of z . Because of symmetry, ordinates at positive values of z equal to ordinates at negative values of z .

We calculate the heights of the ordinates for an observed frequency distribution having k classes with common class-width k by:

- calculating \bar{x} and s , as estimates of μ and σ , from the given distribution;
- converting the class-marks, x_i , into standard normal z -values by the relation $z = \frac{x_i - \bar{x}}{s}$;
- finding the ordinates $\phi(z)$ corresponding to each z -value from the *Table of Ordinates*;
- multiplying $\phi(z)$ by the factor $\frac{n}{k}$, where n is the total frequency.

We need the heights of a number of ordinates when we wish to draw the graph of the fitted curve.

9.3 Ordinates of the Normal Curve $\phi(z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.3989	.3989	.3989	.3986	.3984	.3982	.3980	.3977	.3973	
0.1	.3970	.3965	.3961	.3956	.3951	.3947	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3887	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3796	.3778	.3763	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3604	.3589	.3572	.3555	.3538
0.5	.3521	.3603	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3293	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3013	.2999	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2563	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2395	.2371	.2347	.2322	.2299	.2275	.2251	.2227	.2201
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1893	.1872	.1849	.1825	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1668	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1479	.1456	.1435	.1413	.1394	.1374		.1354	.1335
1.5	.1293	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1143	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.9889	.9773	.9557
1.7	.0940	.0925	.0910	.0893	.0878	.0863		.0833	.0818	.0804
1.8	.0796	.0775	.0751	.0738	.0714	.0701	.0687	.0664	.0641	.0619
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0489	.0478	.0468	.0455	.0448
2.1	.0440	.0431	.0422	.0413	.0404	.0395	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0324	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0228
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167		.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0123	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0098	.0096	.0093	.0091	.0089	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0023
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

Reproduced from Table II of Fisher and Yates, "Statistical Tables for Biological, Agricultural and Medical Research", published by Oliver and Boyd, Ltd., Edinburgh, by permission of the author and publishers.

Example 9.22 Find the ordinates of the standard normal curve at (i) $z = 0.64$, (ii) $z = -0.08$.

- To find the ordinate at $z = 0.64$, in Table 9.3 on page 383, we move downward until the column marked Z to reach the entry 0.6, and then move across that row to the column headed 0.04 to find entry 0.3251, which is the desired ordinate.
- Similarly, ordinate at $z = 2.18$ is 0.0371.
- By symmetry, (ordinate at $z = -0.08$) = (ordinate at $z = 0.08$)
 $= 0.3977$

Example 9.23 Find the ordinates of the frequency distribution of weights given in Example 9.1.

We have calculated $\bar{x} = 47.71$ kg and $s = 5.88$ kg for the distribution of weights in Example 9.1.

The procedure for calculating the heights of the ordinates for the given frequency distribution is shown in the table below:

x_i	$x_i - \bar{x}$	$\phi(z_i)$	Ordinates $\frac{nh}{s} \phi(z_i)$
29.5	-3.10	0.046	2.24
33.5	-2.42	0.113	14.49
37.5	-1.74	0.0678	59.73
41.5	-1.06	0.2275	154.76
45.5	-0.38	0.3712	252.52
49.5	0.30	0.3814	258.45
53.5	0.98	0.2468	167.89
57.5	1.66	0.1006	68.44
61.5	2.35	0.0252	17.14
65.5	3.03	0.0041	2.79

The entries in the last column of the table have been obtained by multiplying each value of $\phi(z_i)$ by $\frac{nh}{s}$, i.e. $\frac{1000 \times 4}{5.88} = 680.27$.

EXERCISES

OBJECTIVE

- Answer 'True' or 'False'. If the statement is not true then replace the underlined word or words that make the statement true:
 - The mean and standard deviation of the exponential distribution are equal.
 - All continuous random variables follow a Normal Probability Distribution.

Not all normal distributions can be transformed to a Standard Normal Distribution.

For any continuous random variable X ; $P(X > 1.0) = P(X \geq 1.0)$.

The area under the normal curve left to its mean is -0.5.

The normal distribution is symmetrical about zero.

The mean, median and mode of normal probability distribution are not equal.

The mean of the standard Z score is zero and its standard deviation equal to one.

The total area under the curve of any normal distribution is two.

If the computed value of Z is zero, then the value of normal random variable is greater than the mean of this variable.

The total area under any normal curve is always 0.5.

Z scores for a standard normal r.v. is set of all whole numbers.

The second quartile of the standard normal variable is larger than its median.

Transforming a normal distribution to a standard normal distribution will not change the mean of the distribution.

The standard score unit is the same as the data unit.

MULTIPLE CHOICE QUESTIONS

The area under the standard normal curve between -3.0 and -2.0 is:

- a) 0.0228
- b) 0.4472
- c) .02165
- d) 0.3413

Which is the characteristic of the normal distribution?

- a) It is bell shaped and symmetric curve.
- b) For any normal r.v. X , $P(X \leq \mu) = P(X \geq \mu)$
- c) The total area under the curve is unity.
- d) All of above.

All normal distributions are:

- a) Symmetrical.
- b) Having two parameters μ and σ^2 .
- c) Bell shaped.
- d) All of above.

- iv) For a standard normal probability distribution the mean and standard deviation are
- $\mu = 1$ and $\sigma = 1$
 - $\mu = 0$ and $\sigma = 1$
 - $\mu = 30$ and $\sigma = 10$
 - All of above.
- v) The middle area under the normal curve with $\mu \pm 2\sigma$ is
- 0.6827
 - 1.0000
 - 0.9545
 - 0.9973
- vi) For a normal distribution with $\mu = 50$ and $\sigma = 10$, how much area will be between $X = 50$?
- 0.35
 - 0.95
 - 0
 - 0.5
- vii) In a normal distribution, mean deviation is equal to
- 1σ
 - 0.8σ
 - 0.6745σ
 - 2.0σ
- viii) The normal distribution will be less spread out when
- The mean is small
 - The median is small
 - The mode is small
 - The standard deviation is small
- ix) The lifetime of general tires is normally distributed with an average of 40,000 hours and a standard deviation of 5000 kilometers. The probability that a randomly selected tire will last more than 50,000 kilometers is
- 0.6739
 - 0.9772
 - 0.0228
 - 0.1460

Which of the following statements is correct for standard normal distribution?

- $P(Z > -2.0) = P(Z > 2.0)$
- $P(Z > -2.0) = P(Z < 2.0)$
- $P(Z > -2.0) = P(Z < -2.0)$
- All of above

EFFECTIVE

- i) Find the mean, variance and m.g.f. of the uniform distribution.
- ✓ Find the moment generating function and the first four moments of the rectangular distribution on $(-1/2, 1/2)$.
- ✓ Describe an exponential distribution and derive its mean and standard deviation.
- Suppose the average length of life of a colour television tube is 12 months. What is the probability that the length of life is equal to or greater than 18 months? Assume an exponential distribution.
- If X has an exponential distribution given by

$$f(x) = \frac{1}{2} e^{-x/2}, \quad 0 \leq x < \infty.$$

what are the mean, variance and m.g.f. of X ? also calculate $P(X>3)$ and $P(X>5 | X>2)$.

- The distance, x , in kilometers travelled by customers to the "Cheap Supermarket" are distributed with the density function

$$f(x) = \begin{cases} \frac{1}{2} e^{-x/2}, & 0 \leq x < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the proportion of customers travelling less than 1 kilometer and the proportion travelling more than 75 kilometers to the Super market.

- Show that the standard deviation of the density function $f(x) = \alpha e^{-\alpha x}$, where x takes all values from 0 to ∞ , is $\frac{1}{\alpha}$.

Write down the m.g.f. for the distribution given by

$$f(x) = \frac{1}{\alpha} e^{-x/\alpha}, \quad 0 \leq x < \infty.$$

Derive the first four moments about the mean.

- $f(x) = xe^{-x}, \quad 0 \leq x \leq \infty,$
 $= 0, \quad \text{otherwise.}$

Find the first four moment using moment generating function. (P.U., B.A. (Hons.) Part-III, 1963)

- 9.8 Show that for the exponential distribution

$$dy = y_0 e^{-y/\sigma} dy, \quad 0 \leq y \leq \infty, \sigma > 0,$$

y_0 being constant, the mean and the standard deviation are each equal to σ and the interquartile range is $\sigma \log_e 3$. Also find μ' , and show that $\beta_1 = 4$ and $\beta_2 = 9$.

(P.U., B.A., Hons.)

- 9.9 a) Let X have probability density

$$f(x) = \frac{1}{2} e^{-|x|}, \quad -\infty < x < \infty.$$

Find the expectation and variance of X' .

- b) The density function of a random variable X is given by

$$f(x) = \frac{\theta}{e^{-\theta} + e^{\theta}}, \quad -\infty < x < \infty.$$

Find (i) the constant θ , (ii) the probability that in two independent observations X one value lies below 1.

(P.U., B.A., Hons.) (Hons.) Part I, 2000

- 9.10 a) The Gamma distribution is given by

$$f(x) = \frac{1}{\Gamma(n)} e^{-x} x^{n-1}, \quad 0 < x < \infty,$$

$$\text{where } \Gamma(n) = \int_0^\infty t^{n-1} e^{-t} dt.$$

Show that the mean and the variance are each equal to n and the third moment about the mean is $2n$.

- b) Find the moments of the Gamma distribution.

- c) Find the r th moment of the Gamma distribution. (i) without using the mgf, (ii) using the mgf.

- b) Show that the sum of two independent Gamma variables with parameters n and m is a Gamma variable with parameter $n+m$.

- 9.11 a) The Beta distribution of the first kind is given by

$$f(x) = \frac{1}{B(l, m)} x^{l-1} (1-x)^{m-1}, \quad 0 < x < 1,$$

$$\text{where } B(l, m) = \int_0^1 t^{l-1} (1-t)^{m-1} dt.$$

Show that the mean is $\frac{l}{l+m}$ and the variance is $\frac{lm}{(l+m)^2(l+m+1)}$.

- b) Find the harmonic mean of the Beta distribution.
 (P.U., B.A. (Hons.) Part-III, 1967)
- x) Defining a $\gamma(m)$ variate as one with a probability density function of the form

$$f(x) = \frac{x^{m-1} e^{-x}}{\Gamma(m)}, \text{ for } x \geq 0, m \text{ being a +ve constant.}$$

Obtain the distribution of $\frac{X}{Y}$ where X and Y are independent $\gamma(m)$ and $\gamma(n)$ variates respectively.

- b) Show that the mean values of the positive square root of a $\gamma(l)$ variate is $\frac{l+1/2}{\Gamma(l)}$. Hence prove that the mean deviation of normal variate from its mean is $\sigma\sqrt{2/\pi}$.
 (P.U., B.A. (Hons.) Part-III, 1966)
- a) Define the Normal Distribution and obtain its mean and variance.
 b) Show that for the normal distribution, the mean, mode and median are the same.
 (B.Z.U., B.A./B.Sc. 1976)

State the mathematical equation of the Normal distribution and prove that

- i) the area under the normal curve is unity;
 ii) the normal curve has points of inflection which are equidistant from the mean;
 iii) all odd order moments about mean vanish.
 (P.U., B.A./B.Sc. 1986, 88, 93)

The continuous random variable X has p.d.f. $f(x)$, where

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

- a) Derive the mean and the variance of the distribution.
 b) Show that the maximum value of $f(x)$ occurs at $x=\mu$.
 c) Show that there are points of inflection at $x=\mu+\sigma$ and $x=\mu-\sigma$.
 If $f(x) = ke^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is the equation of a normal curve, find the value of k , the mean and standard deviation.
 d) Show that for the normal distribution, the mean deviation from the mean is approximately $\frac{4}{5}$ of its standard deviation.
 (P.U., B.A./B.Sc. 1983, 87)

Show that, for the normal distribution, the quartile deviation, the mean deviation and the standard deviation are approximately in the ratio 10:12:15.
 (B.Z.U., B.A./B.Sc. 1990)

- 9.17 a) Obtain the moment generating function of the standardized normal distribution.
 b) Show that for the normal distribution, moments of odd order about mean are all zero and moments of even order are given by

$$\mu_m = \left(\frac{\sigma^2}{2} \right)^{m/2} \frac{(2\pi)^{m/2}}{(m)!}$$

- 9.18 a) Show that the m.g.f. of the normal distribution with mean μ and variance σ^2 , is
 $M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$
 b) What percentage of the normal distribution with mean μ and variance σ^2 is located between the points (i) μ and $(\mu + 1.54\sigma)$, (ii) $(\mu - 1.73\sigma)$ and $(\mu + 0.56\sigma)$? (P.U., B.A., B.S., 2008)

- 9.19 a) If X is normally distributed with zero mean and $\sigma = 0.6$, find $P(X > 0)$ and $P(0.2 < X < 1.8)$.
 b) For a normal distribution with mean 1 and standard deviation 3, find the probabilities
 i) $P(3.43 \leq X \leq 6.19)$, ii) $P(-1.43 \leq X \leq 1.19)$.

- 9.20 A normal distribution has mean = 12 and $\sigma = 2$, find the area under the curve
 a) from $X=10$ to $X=13.5$, b) from $X=11.4$ to $X=14.2$, c) from $X=6$ to $X=18$.
 9.21 Let X be $N(100, 225)$. Find the following probabilities:
 a) $P(X \leq 92.5)$, b) $P(X \geq 105)$, c) $P(X \geq 124)$,
 d) $P(112 \leq X \leq 128.5)$, e) $P(125 \leq X \leq 127.5)$, f) $P(X \geq 70)$.

- 9.22 a) If X is $N(\mu, \sigma^2)$ and if $Y = \mu + b$, then show that Y is $N(\mu + b, \sigma^2 \pi^2)$.
 b) Let $Y = 5X + 10$ and X is normally distributed with a mean 10 and variance 25. Find the following:
 (i) $P(Y \leq 54)$, (ii) $P(Y \geq 68)$, (iii) $P(52 \leq Y \leq 67)$. (B.Z.U., B.A., B.S., 2008)

- 9.23 a) Scores on a certain nation-wide college entrance examination follow a normal distribution with a mean of 500 and a standard deviation of 100. Find the probability that a student's score (i) over 630, (ii) less than 250, and (iii) between 325 and 675.

- b) Given that the height of college boys is normally distributed with mean $5' - 2''$ and standard deviation $4''$, and that the minimum height required for joining the NCC is $5' - 7''$. Find the percentage of boys who would be rejected on account of their height.

- 9.24 a) If the heights (X) of college students are normally distributed with mean 65 and standard deviation 4, find the probability that (i) $X > 65$ and (ii) $65 \leq X \leq 70$.

- b) If the m.g.f. of X is $M(t) = e^{4t + 16t^2}$, find
 $P(-4 \leq X \leq 16)$ and $P(-10 \leq X \leq 0)$.

- 125 Suppose that weights of 2000 male students are normally distributed with mean 155 pounds and standard deviation 20 pounds. Find the number of students with weights (i) less than or equal to 180 pounds, (ii) between 120 and 130 pounds, (iii) between 150 and 175 pounds, (iv) greater than or equal to 200 pounds.
- 126 a) The mean life of stockings used by an army was 40 days, with a standard deviation of 8 days. Assume the life of the stockings follows a normal distribution. If 100,000 pairs are issued, how many would need replacement before 35 days? After 46 days?
(B.Z.U., B.A./B.Sc. 1990)
- b) The time taken by a milkman to deliver milk to the GOR Estate is normally distributed with mean 12 minutes and standard deviation 2 minutes. He delivers milk everyday. Estimate the number of days during the year when he takes (i) longer than 17 minutes, (ii) less than 10 minutes, (iii) between 9 and 13 minutes.
- 127 The scores made by candidates in a certain test are normally distributed with mean 500 and standard deviation 100.
- a) What percent of the candidates received scores (i) greater than 700, (ii) less than 400, (iii) between 400 and 600, (iv) which differ from mean by more than 150?
(~~✓~~) MHTCET 47% less from mean by no more than 150
b) If a candidate gets a score of 680, what percent of the candidates have higher scores than he?
(P.U., B.A./B.Sc. 1980)

A man goes by car to his office, and the route through city centre takes him, on the average, 27 minutes with a standard deviation of 5 minutes. With the opening of a new ring road, the man can bypass the congestion of the city centre, but the journey now takes, on the average, 29 minutes with a standard deviation of 2 minutes. Assuming that both journey times are normally distributed, determine which route is the better one if a man has (i) 28 minutes, and (ii) 32 minutes to reach his office for an appointment.
(I.U., M.Sc. 1990)

Hint. The better route is one that gives the smaller probability of the man's being late for appointment.

- a) A random variable X has $E(X) = 100$, $\sigma = 16$. If $P(X > a) = 0.5636$, find the value of a .
- b) A possible measure of kurtosis (i.e. flatness) is given by $k = \frac{Q.D.}{P_{90} - P_{10}}$, where Q.D. is the Semi-Interquartile range, and P_i 's are the percentiles. Use the standard normal table to estimate the value of k for a normal distribution.

In a normal distribution with $\mu = 47.6$ and $\sigma = 16.2$, find (i) the probability that a single observation will be larger than 50, (ii) two points such that a single observation has a 97% probability of falling between them, (iii) P_{10} , P_{50} and P_{90} .
(P.U., B.A./B.Sc. 1976)

- A random sample of 1,000 rods are tested for their length and it is found that the mean and the standard deviation are 14.40 metres and 2.30 metres respectively. If the lengths of the rods are normally distributed, then find (i) how many rods will be between 12 and 16 metres? (ii) what are the chances that any rod selected at random will be 15 metres length or above?

- b) The mean score of 1000 students appearing for an examination is 34.4 and the standard deviation is 16.6. How many candidates may be expected to obtain marks between 30 and 60 assuming the normality of the distribution? Under the same assumptions, determine also the limits of marks of the central 70% of the candidates.
- 9.32 A soft drink machine is regulated so that it discharges an average of 200 milliliters per cup. The amount of drink is normally distributed with a standard deviation equal to 15 milliliters.
- what fraction of the cups will contain more than 240 milliliters?
 - what is the probability that a cup contains between 191 and 209 milliliters?
 - how many cups will likely overflow if 230 milliliters cups are used for the next 1000 cups?
 - below what value do we get the smallest 25% of the drinks?
- (P.U., B.A./B.Sc. 2016)
- 9.33 a) The heights of applicants to the police force are normally distributed with mean 170 cm and standard deviation 3.8cm. If 30% of applicants are rejected because they are too small, what is the minimum acceptable height for the police force?
- b) The average life of a certain type of small motor is 10 years with standard deviation 1 year. The manufacturer replaces free all motors that fail while under guarantee. If he is willing to replace only 3% of the motors that fail, how long a guarantee should he offer? Assume that the lives of the motors follow a normal distribution.
- 9.34 An architect is designing the interior doors of a men's gymnasium. He wants to make them tall enough so that 95 percent of the men using the doors will have at least a one-foot clearance. Assuming that the heights will be normally distributed, with a mean of 70 inches and a standard deviation of 3 inches, how high must the architect make the doors? (P.U., B.A./B.Sc. 2016)
- 9.35 a) In a normal distribution the lower and upper quartiles are respectively 8 and 17. Find the mean and standard deviation of the normal distribution. (B.I.S.E., Lahore 2016)
- b) In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.
- c) Let X be normally distributed with mean μ and variance σ^2 , so that $P(X < 89) = 0.05$ and $P(X > 94) = 0.05$. Find μ and σ^2 . (P.U., B.A./B.Sc. 2016)
- 9.36 Assuming that the number of marks scored by a candidate is normally distributed, find the mean and the standard deviation, if the number of first class students (60% or more marks) is 90, the number of failed students (less than 30% marks) is 90 and the total number of candidates appearing for the examination is 450.
- 9.37 A boy is trying to climb a slippery pole and finds that he can climb to a height of at least once in five attempts, and to a height of at least 1.70m nine times out of ten attempts. Assuming that the heights he can reach in various attempts form a normal distribution, calculate the mean and standard deviation of the distribution. Calculate also the heights that the boy can expect to reach once in one thousand attempts. (I.I.T., M.Sc. 2016)

- Q.38 a) Prove that the Binomial distribution $(q+p)^n$ tends to become a normal distribution for large values of n .
 b) Supposing that the death rate from Malaria is 20%, find the probability that the number of deaths in a particular village is between 70 and 80 out of 300.
 (P.U., B.Sc. Hon. Part-II, 1972)
- Q.39 a) Explain the reason why the correction for continuity is usually made when we apply the normal approximation to the binomial distribution.
 b) Find the probability that 200 tosses of a fair coin will result in (i) between 80 and 120 heads inclusive, (ii) less than 90 heads, and (iii) exactly 100 heads.
 (P.U., B.A./B.Sc. 1977)
- Q.40 a) A coin is tossed 200 times. Find the probability of getting (i) between 105 and 110 heads inclusive, (ii) less than 95 heads.
 (P.U., B.A./B.Sc. 1983)
 b) Find the probability of obtaining between six and nine heads inclusive in 15 tosses of an ideal coin by applying (i) the binomial distribution, (ii) the normal approximation to the binomial, with correction for continuity.
 (P.U., B.A./B.Sc. 1978)
- Q.41 a) A telephone exchange receives, on average, 5 calls per minute. Find the probability that in a 20-minute period no more than 102 calls are received.
 b) If X is $B(x; 20, 0.4)$ find $P(6 \leq X \leq 10)$. Then find the approximations to this probability using (i) the Poisson distribution, (ii) the normal distribution.
 c) Find the ordinates of the normal curve (i) $z=0.064$, (ii) $z=1.27$, (iii) $z=0.84$ and (iv) $z=-2.08$.
 d) Fit a normal distribution given mean $\mu=27.0$, standard deviation $\sigma=2.2$ and total frequency=209.

The following table gives the distribution of statures among the first year students of a university

Stature (in.)	61	62	64	65	66	67	68	69	70	71	72	73	74
Frequency	2	10	11	38	57	93	106	126	109	87	75	23	9

- a) Test the normality of the distribution by comparing the proportion of the cases lying between $\bar{X} \pm s$, $\bar{X} \pm 2s$, $\bar{X} \pm 3s$ for the distribution and for the normal distribution.
 b) Fit a normal distribution to the data, using area Tables.

Fit a normal distribution by area method to the following data. Also calculate ordinates.

Stature	40 ₋	45 ₋	50 ₋	55 ₋	60 ₋	65 ₋	70 ₋	75 ₋	80 ₋	85 ₋
Frequency	7	8	25	38	51	60	45	32	10	4

Compare actual and expected frequencies on a graph.

- 9.42 Calculate the frequencies of the normal distribution which has the same total frequency, mean and standard deviation as the following distribution.

Group	10 _—	12 _—	14 _—	16 _—	18 _—	20 _—	22 _—	24 _—	26 _—
<i>f</i>	4	30	106	206	272	219	120	37	6

CHAPTER 10

**SIMPLE
REGRESSION AND
CORRELATION**

SIMPLE REGRESSION AND CORRELATION

1.1 INTRODUCTION

The term regression was introduced by the English biometrist, Sir Francis Galton (1822–1911) to describe a phenomenon which he observed in analyzing the heights of children and their parents. He found that, though tall parents have tall children and short parents have short children, the average height of children tends to step back or to regress toward the average height of all men. This tendency toward the average height of all men was called a regression by Galton.

Today, the word regression is used in a quite different sense. It investigates the dependence of one variable, conventionally called the dependent variable, on one or more other variables, called independent variables, and provides an equation to be used for estimating or predicting the average value of the dependent variable from the known values of the independent variable. The dependent variable is assumed to be a random variable whereas the independent variables are assumed to have fixed values, i.e. they are chosen non-randomly. The relation between the expected value of the dependent variable and the independent variable is called a regression relation. When we study the dependence of a variable on a single independent variable, it is called a simple or two-variable regression. When the dependence of a variable on two or more than two independent variables is studied, it is called multiple regression. Furthermore, when the dependence is represented by a straight line equation, the regression is said to be linear, otherwise it is said to be curvilinear.

It is relevant to note that in regression study, a variable whose variation we try to explain is a dependent variable while an independent variable is a variable that is used to explain the variation in the dependent variable.

Some more terminology: The dependent variable is also called the regressand, the predicted, the outcome or the explained variable whereas the independent or the non-random variable is also referred to as regressor, the predictor, the regressor variable or the explanatory variable.

1.2 DETERMINISTIC AND PROBABILISTIC RELATIONS OR MODELS

The relationship among variables may or may not be governed by an exact physical law. For instance, let us consider a set of n pairs of observation (X_i, Y_i) . If the relation between the variables is linear, then the mathematical equation describing the linear relation is generally written as

$$Y_i = a + bX_i$$

where a is the value of Y when X equals zero and is called the Y -intercept, and b indicates the change in Y for a one-unit change in X and is called the slope of the line. Substituting a value for X in the equation, we can completely determine a unique value of Y . The linear relation in such a case is said to be a deterministic model. An important example of the deterministic model is the relationship between Celsius and Fahrenheit scales in the form of $F = 32 + \frac{9}{5}C$. Another example is the area of a circle expressed by the formula, $\text{area} = \pi r^2$. Such relations cannot be studied by regression.

In contrast to the above, the linear relationship in some situations is not exact. For example, we cannot precisely determine a person's weight from his height as the relationship between them is not exactly linear and so follow an exact linear form. The weights for given values of age are reasonably assumed to follow a normal distribution with measurement of random errors. The deterministic relation in such cases is then modified to allow for the presence of random errors.

for the inexact relationship between the variables and we get what is called a *non-deterministic probabilistic model* as

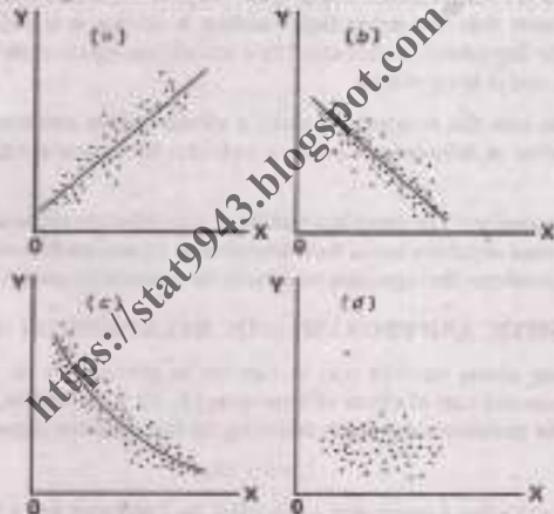
$$Y_i = \alpha + \beta X_i + e_i, \quad (i = 1, 2, \dots, n)$$

where e_i 's are the unknown random errors.

10.3 SCATTER DIAGRAM

A first step in finding whether or not a relationship between two variables exists, is to plot a pair of independent-dependent observations $(X_i, Y_i), i = 1, 2, \dots, n$ as a point on graph paper, using X -axis for the regression variable and the Y -axis for the dependent variable. Such a diagram is called *scatter diagram* or a *scatter plot*. If a relationship between the variables exists, then the points in scatter diagram will show a tendency to cluster around a straight line or some curve. Such a line or curve around which the points cluster, is called the *regression line* or *regression curve* which can be used to estimate the expected value of the random variable Y from the values of the nonrandom variable X .

The scatter diagrams shown below reveal that the relationship between two variables is (a) positive and linear, in (b) is negative and linear, in (c) is curvilinear and in (d) there is no relationship.



10.4 SIMPLE LINEAR REGRESSION MODEL

We assume that the linear relationship between the dependent variable Y_i and the value x_i of regressor X is

$$Y_i = \alpha + \beta X_i + e_i,$$

where the X_i 's are fixed or predetermined values,

the Y_i 's are observations randomly drawn from a population,

the e_i 's are error components or random deviations,

α and β are population parameters, α is the intercept and the slope β is called regression coefficient, which may be positive or negative depending upon the direction of the relationship between X and Y .

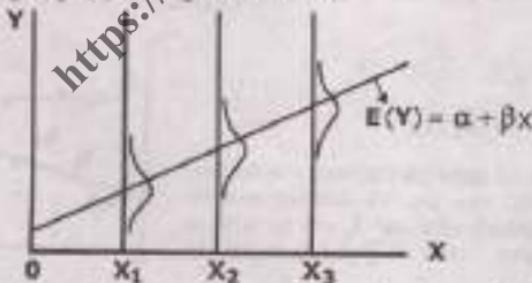
Furthermore, we assume that

- i) $E(\epsilon_i) = 0$, i.e. the expected value of error term is zero, it implies that the expected value of Y is related to X in the population by a straight line;
- ii) $\text{Var}(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$ for all i , i.e. the variance of error term is constant. It means that the distribution of error has the same variance for all values of X . (Homoscedasticity assumption);
- iii) $E(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$, i.e. error terms are independent of each other (assumption of no serial or auto correlation between ϵ 's);
- iv) $E(X, \epsilon_i) = 0$, i.e. X and ϵ are also independent of each other;
- v) ϵ_i 's are normally distributed with a mean of zero and a constant variance σ^2 . This implies that Y values are also normally distributed. The distributions ϵ and π are identical except that they have different means. This assumption is required for estimation and testing of hypothesis on linear regression.

According to this population regression model, each Y_i is an observation from a normal distribution $= \alpha + \beta X$ and variance $= \sigma^2$. Thus the relation may be expressed alternatively as

$$E(Y) = \alpha + \beta X$$

implies that the expected value of Y is linearly related to X and the observed value of Y deviates from the line $E(Y) = \alpha + \beta X$ by a random component π_i , i.e. $Y_i = Y_i - (\alpha + \beta X_i)$. The following graph shows the assumed line, giving $E(Y)$ for the given values of X .



In practice, we have a sample from some population, therefore we desire to estimate the regression line from the sample data. Then the basic relation in terms of sample data may be

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i + \hat{\epsilon}_i$$

$\hat{\alpha}$ and $\hat{\beta}$ are the estimates of α , β and ϵ_i . The estimated regression is generally written as

Many possible regression lines could be fitted to the sample data, but we choose that particular line which best fits that data. The best regression line is obtained by estimating the regression parameters by the most commonly used method of least squares which we describe in the following subsection.

10.4.1 An Aside—The Principle of Least Squares. The principle of least squares (LS) is the method of determining the values of the unknown parameters that will minimize the sum of squares of residuals (residues) where errors are defined as the differences between observed values and the corresponding values predicted or estimated by the fitted model equation.

The parameter values thus determined, will give the least sum of the squares of residuals known as least squares estimates. The method of least squares that gets its name from the minimization of a sum of squared deviations is attributed to Karl F. Gauss (1777–1855). Some people believe that this method was discovered at the same time by Adrien M. Legendre (1752–1833), Pierre-Simon Laplace (1749–1827) and others. Markov's name is also mentioned in connection with its further development. In recent years, efforts have been made to find better methods of fitting but the least squares method remains dominant and is used as one of the important methods of estimating the population parameters.

10.4.2 Least-Squares Estimates in Simple Linear Regression. Let there be a set of observations (X_i, Y_i) , $i = 1, 2, \dots, n$.

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

or in terms of sample data as

$$Y_i = \alpha + b X_i + e_i$$

where α and b are the least-squares estimates of α and β , e_i , commonly called residual, is the error of the observed Y_i from its estimate provided by $\hat{Y}_i = \alpha + b X_i$.

According to the principle of least-squares, we determine those values of α and b which minimize the sum of squares of the residuals. In other words, the best regression line is the one that minimizes the sum of the squares of the vertical deviations between the observed values Y_i and the corresponding values predicted by the regression model, i.e., $\hat{Y}_i = \alpha + b X_i$. That is the least squares line minimizes

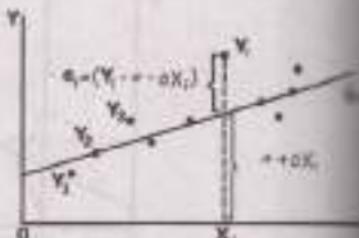
$$\begin{aligned} S(a, b) &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - \alpha - b X_i)^2 \end{aligned}$$

As α and b , the two quantities that determine the line, vary, $S(a, b)$ will vary too. We therefore consider $S(a, b)$ as a function of a and b , and we wish to determine at what values of a and b , it will be minimum.

Minimizing $S(a, b)$, we need to set its partial derivatives w.r.t a and b equal to zero. Thus,

$$\frac{\partial S(a, b)}{\partial a} = 2 \sum (Y_i - \alpha - b X_i) (-1) = 0, \text{ and}$$

$$\frac{\partial S(a, b)}{\partial b} = 2 \sum (Y_i - \alpha - b X_i) (-X_i) = 0$$



Simplifying, we obtain the following two equations, called the *normal equations* (the word *normal* here in the sense of regular or standard).

$$\Sigma Y = n\alpha + b \sum X, \text{ and } \sum X_i Y_i = a \sum X_i + b \sum X_i^2.$$

These two normal equations are solved simultaneously for values of a and b either by direct solution or by using determinants.

- **Direct Elimination:** Multiplying the first equation by $\sum X$, and the second equation by n , we get $\sum X \sum Y = n\alpha \sum X + b(\sum X)^2$ and $n \sum X Y = n\alpha \sum X + nb \sum X^2$.

Subtracting, we get

$$n \sum XY - \sum X \sum Y = b(n \sum X^2 - (\sum X)^2)$$

$$b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

is the least-squares estimate of the regression co-efficient β .

Similarly, we get

$$a = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n \sum X^2 - (\sum X)^2}$$

least-squares estimate of α .

Alternatively, we divide the first normal equation by n , and get the least-squares estimate of α as

$$-\bar{b}\bar{X}$$

This also shows that the estimated regression line passes through (\bar{X}, \bar{Y}) , the means of the data.

By means of determinants, the solution is

$$b = \begin{vmatrix} \sum XY & \sum Y \\ \sum X & n \end{vmatrix} = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}, \text{ and}$$

$$a = \begin{vmatrix} \sum X^2 & \sum XY \\ \sum X & \sum Y \end{vmatrix} = \frac{(\sum X^2)(\sum Y) - (\sum X)(\sum XY)}{n \sum X^2 - (\sum X)^2}.$$

These estimates give us the regression equation

$$\begin{aligned} Y_i &= a + bX_i \\ &= \bar{Y} + b(X_i - \bar{X}). \end{aligned}$$

$$\text{where } b = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}.$$

Since Y is a random variable, therefore the deviations in the Y direction are taken into account in determining the best-fitting line.

It is very important to note that, when both X and Y are observed at random, i.e. the samples are from a bivariate population, there are two regression equations, each obtained by choosing one variable as dependent whose average value is to be estimated and treating the other as independent. In case of a single random variable, the single regression equation is used to estimate values of either the dependent or the independent variable. In case of two regression lines, it is common to denote the regression coefficients of Y on X and of X on Y by b_{xy} and b_{yx} respectively.

Example 10.1. Compute the least squares regression equation of Y on X for the following data. What is the regression coefficient and what does it mean?

X	5	6	8	10	12	13	15	16	17
Y	16	19	23	28	36	41	44	45	50

The estimated regression line of Y on X is

$$\hat{Y} = a + bX,$$

and the two normal equations are

$$\begin{aligned}\sum Y &= a n + b \sum X \\ \sum XY &= a \sum X + b \sum X^2\end{aligned}$$

To compute the necessary summations, we arrange the computations in the table below.

X	Y	XY	X^2
5	16	80	25
6	19	114	36
8	23	184	64
10	28	280	100
12	36	432	144
13	41	533	169
15	44	660	225
16	45	720	256
17	50	850	289
Total	302	3853	1308

$$\text{Now } \bar{X} = \frac{\sum X}{n} = \frac{102}{9} = 11.33, \bar{Y} = \frac{\sum Y}{n} = \frac{302}{9} = 33.56.$$

$$b = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2} = \frac{9(3853) - (102)(302)}{9(1308) - (102)^2}$$

$$= \frac{34677 - 30804}{11772 - 10404} = \frac{3873}{1368} = 2.831, \text{ and}$$

$$x = \bar{Y} - b \bar{X} = 33.56 - (2.831)(11.33) = 1.47.$$

the desired estimated regression line of Y on X is

$$\hat{Y} = 1.47 + 2.831X.$$

The estimated regression co-efficient, $b = 2.831$, which indicates that the values of Y increase by units for a unit increase in X .

Example 10.2 In an experiment to measure the stiffness of a spring, the length of the spring under loads was measured as follows:

$X = \text{Load (lb)}$	3	5	6	9	10	12	15	20	22	28
$Y = \text{length (in)}$	10	12	15	18	20	22	27	30	32	34

Find the regression equations appropriate for predicting

- i. the length, given the weight on the spring;
- ii. the weight, given the length of the spring.

(W.P.C.S., 1964)

The data come from a bivariate population, i.e. both X and Y are random, therefore there are two lines. To find the regression equation for predicting length (Y), we take Y as dependent variable and treat X as independent variable (i.e. non-random). For the second regression, the choice of variables is reversed.

The computations needed for the regression lines are given in the following table.

X	Y	X^2	Y^2	XY
3	10	9	100	30
5	12	25	144	60
6	15	36	225	90
9	18	81	324	162
10	20	100	400	200
12	22	144	484	264
15	27	225	729	405
20	30	400	900	600
22	32	484	1024	704
28	34	784	1156	932
Total	130	220	2288	5486
				3467

The estimated regression equation appropriate for predicting the length, Y , given the weight X , is

$$\hat{Y} = a_0 + b_{12} X,$$

$$\text{where } b_{12} = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2} = \frac{(10)(3467) - (130)(220)}{(10)(2288) - (130)^2}$$

$$= \frac{6670}{5980} = 1.02, \text{ and}$$

$$a_0 = \bar{Y} - b_{12} \bar{X} = 22 - (1.02)(13) = 8.74$$

Hence the desired estimated regression equation is

$$\hat{Y} = 3.74 + 1.02X$$

- ii) The estimated regression equation appropriate for predicting the weight, X , given the length, Y ,

$$X = a_1 + b_{XY} Y,$$

$$\text{where } b_{XY} = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum Y^2 - (\sum Y)^2} = \frac{(10)(2467) - (130)(220)}{(10)(5456) - (220)^2}$$

$$= \frac{6070}{6460} = 0.94, \text{ and}$$

$$a_1 = \bar{X} - b_{XY} \bar{Y} = 13 - (0.94)(22) = -7.68$$

Hence $\hat{X} = 0.94Y - 7.68$ is the estimated regression equation appropriate for predicting the weight given the length (Y).

10.4.3 Properties of the Least-Squares Regression Line. The least-squares linear regression has the following properties:

$$b_{YX} \neq b_{XY}$$

- ii) The least squares regression line always goes through the point (\bar{X}, \bar{Y}) , the means of the variables.
- iii) The sum of the deviations of the observed values of Y_i from the least squares regression line is always equal to zero, i.e. $\sum(Y_i - \hat{Y}_i) = 0$.
- iv) The sum of the squares of the deviations of the observed values from the least squares regression line is a minimum, i.e. $\sum(Y_i - \hat{Y}_i)^2 = \text{minimum}$.
- v) The least-squares regression line obtained from a random sample is the line of best fit if a and b are the unbiased estimates of the parameters α and β .

10.4.4 Standard Deviation of Regression or Standard Error of Estimate. The observed values of (X_i, Y_i) do not all fall on the regression line but they scatter away from it. The degree of dispersion (or dispersion) of the observed values about the regression line is measured by what is called the standard deviation of regression or the standard error of estimate of Y on X . For the population data, the standard deviation that measures the variation of observations about the true regression line $E(Y) = \alpha + \beta X$ is denoted by $\sigma_{Y/X}$ and is defined by

$$\sigma_{Y/X} = \sqrt{\frac{\sum(Y_i - (\alpha + \beta X_i))^2}{N}}$$

where N is the population size.

For sample data, we estimate $\sigma_{Y/X}$ by $s_{Y/X}$ which is defined as

$$s_{Y/X} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}},$$

$\hat{Y} = a + bX$, the estimated regression line. This is actually an unbiased estimate of $\sigma_{F,N}$, the standard deviation about the regression line. The standard error of estimate, s_{est} , will be zero if all the observed values fall on the regression line. It is interesting to note that the ranges $\hat{Y} \pm 2s_{\text{est}}$ and $\hat{Y} \pm 3s_{\text{est}}$ contain about 68%, 95.4% and 99.7% observations respectively.

To find $\sum(Y_i - \hat{Y}_i)^2$, we have to calculate \hat{Y}_i from the estimated regression line for the observed value of X_i , which is not an easy task. We therefore use an alternative form obtained as below:

$$\begin{aligned} \sum(Y_i - \hat{Y}_i)^2 &= \sum(Y_i - a - bX_i)^2 \\ &= \sum Y_i(Y_i - a - bX_i) - a \sum(Y_i - a - bX_i) - b \sum X_i(Y_i - a - bX_i) \\ &= \sum Y_i^2 - a \sum Y_i - b \sum X_i Y_i - a(\sum Y_i - nb) - b(\sum X_i) - b(\sum X_i Y_i - a \sum X_i - b \sum X_i^2) \\ \text{and } \sum Y_i - nb - b \sum X_i &= 0 \text{ and } \sum X_i Y_i - a \sum X_i - b \sum X_i^2 = 0 \text{ as they are the normal equations.} \end{aligned}$$

$$\sum(Y_i - \hat{Y}_i)^2 = \sum Y_i^2 - a \sum Y_i - b \sum X_i Y_i$$

$$\text{Since } s_{\text{est}} = \sqrt{\frac{\sum Y_i^2 - a \sum Y_i - b \sum X_i Y_i}{n-2}},$$

n is the number of pairs.

Example 10.3 Using the data in Example 10.1,

find the values of \hat{Y} and show that $\sum(Y_i - \hat{Y}) = 0$, and

compute the standard error of estimate s_{est} . $= 3$

The calculations needed to find the values of \hat{Y} and the standard error of estimate s_{est} are given in the following:

X	Y	\hat{Y} $(=1.47+2.831X)$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$	Y^2
5	16	15.625	0.375	0.140625	256
6	19	18.456	0.544	0.295936	361
8	23	24.118	-1.118	1.240924	529
10	28	29.280	-1.780	3.168400	784
12	36	35.442	0.558	0.311364	1296
13	41	38.273	2.727	7.436529	1681
15	44	43.915	0.065	0.004225	1936
16	45	46.766	-1.766	3.118756	2025
17	50	49.597	0.403	0.162409	2500
102	302	301.992	0.008	15.888168	910368

- i) The estimated values \hat{Y} appear in the third column of the table on page 483, and $\Sigma(\hat{Y} - \bar{Y})^2$ turns out to be 0.008. This small difference is due to rounding off.
- ii) The standard error of estimate of \hat{Y} in X is

$$s_{\hat{Y}|x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{15.888168}{7}} = \sqrt{2.269738} = 1.51$$

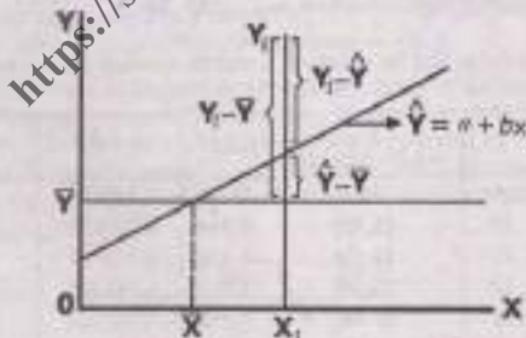
Using the alternative form for the calculation of $s_{\hat{Y}|x}$, we get

$$\begin{aligned}s_{\hat{Y}|x} &= \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n-2}} \\&= \sqrt{\frac{11368 - (1.47)(302) - (2.83)(3853)}{9-2}} \\&= \sqrt{\frac{16.217}{7}} = \sqrt{2.316714} = 1.52.\end{aligned}$$

10.4.5 Co-efficient of Determination. The variability among the values of the dependent variable Y , called the *total variation*, is given by $\Sigma(Y - \bar{Y})^2$. This is composed of two parts (i) that which is explained by (associated with) the regression line, i.e., $\Sigma(\hat{Y} - \bar{Y})^2$, (ii) that which the regression line fails to explain, i.e., $\Sigma(Y - \hat{Y})^2$ (see figure). In symbols,

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y - \hat{Y})^2 + \Sigma(\hat{Y} - \bar{Y})^2$$

$$\text{Total variation} = \text{Unexplained variation} + \text{Explained variation}$$



The co-efficient of determination which measures the proportion of variability of the values of the dependent variable (Y) explained by its linear relation with the independent variable (X), is defined as the ratio of the explained variation to the total variation. We use the symbol ρ^2 for the population pa-

and symbol r^2 for the estimate obtained from sample. Thus the sample co-efficient of determination is given by

$$\begin{aligned} r^2 &= \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} \\ &= 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}. \end{aligned}$$

An alternative form for calculating the coefficient of determination is

$$r^2 = \frac{n \sum Y + b \sum XY - (\sum Y)^2 / n}{\sum Y^2 - (\sum Y)^2 / n}.$$

When all the observed values fall on the regression line, then $Y = \hat{Y}$ and $\sum(Y - \hat{Y})^2 = \sum(\hat{Y} - \bar{Y})^2$, and hence $r^2 = 1$. When the observed values are such that $\hat{Y} = \bar{Y}$, then $\sum(\hat{Y} - \bar{Y})^2 = 0$, and hence $r^2 = 0$. This shows that $0 \leq r^2 \leq 1$. A value of $r^2 = 1$, signifies that 100% of the variability in the dependent variable is associated with the regression equation. When $r^2 = 0$, it means that none of the variability in the dependent variable is explained by X -variable. A value of $r^2 = 0.93$, indicates that 93% of the variability in Y is explained by its linear relationship with the independent variable X and 7% of the variation is due to chance or other factors.

Example 10.4 Taking length (Y) as dependent variable for the data in Example 10.2, calculate (i) total variation, (ii) unexplained variation, (iii) the explained variation, and (iv) the co-efficient of determination and interpret the coefficient.

In Example 10.2, we found that

$$\sum Y = 220, \sum Y^2 = 5486, \sum XY = 3467, b = 1.02, a = 8.74 \text{ and } n = 10.$$

We now find

$$\begin{aligned} \text{Total variation} &= \sum(Y - \bar{Y})^2 = \sum Y^2 - (\sum Y)^2 / n \\ &= 5486 - (220)^2 / 10 = 646. \end{aligned}$$

$$\begin{aligned} \text{Unexplained variation} &= \sum(Y - \hat{Y})^2 = \sum Y^2 - a \sum Y - b \sum XY \\ &= 5486 - (8.74)(220) - (1.02)(3467) \\ &= 5486 - 54,99.14 = 26.86. \end{aligned}$$

$$\begin{aligned} \text{Explained variation} &= \text{Total variation} - \text{unexplained variation} \\ &= 646 - 26.86 = 619.14. \end{aligned}$$

The coefficient of determination, r^2 , is given by

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

$$= \frac{619.14}{646} = 0.958$$

A value of $r^2 = 0.958$ indicates that 95.8% of the variability in Y , the length of the spring, is demonstrated by its linear relationship with X , the weight on the spring.

10.5 CORRELATION

Correlation, like covariance, is a measure of the degree to which any two variables vary together. In other words, two variables are said to be correlated if they tend to simultaneously vary in one direction. If both the variables tend to increase (or decrease) together, the correlation is said to be positive or positive, e.g. the length of an iron bar will increase as the temperature increases. If one variable tends to increase as the other variable decreases, the correlation is said to be negative or inverse, e.g. the volume of gas will decrease as the pressure increases. It is worth remarking that in correlation, we assess the strength of the relationship (or interdependence) between two variables; both the variables are random variables, and they are treated symmetrically, i.e. there is no distinction between dependent and independent variable. In regression, by contrast, we are interested in determining the dependence of one variable that is random, upon the other variable that is non-random or fixed, and in predicting the expected value of the dependent variable by using the known values of the other variable.

10.5.1 Pearson Product Moment Correlation Co-efficient. A numerical measure of strength of the linear relationship between any two variables is called the Pearson's product moment correlation co-efficient or sometimes, the coefficient of simple correlation or total correlation. The sample correlation coefficient for n pairs of observations (X_i, Y_i) , usually denoted by the letter r , is defined as

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

The population correlation co-efficient for a bivariate distribution, denoted by ρ , has already been defined as

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

For computational purposes, we have an alternative form of r as

$$r = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

$$= \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}}$$

This is a more convenient and useful form, especially when \bar{X} and \bar{Y} are not integers. The coefficient of correlation r is a pure number (i.e. independent of the units in which the variables are measured) and it assumes values that can range from +1 for perfect positive linear relationship, to -1 for perfect negative linear relationship with the intermediate value of zero indicating no linear relationship between X and Y . The sign of r indicates the direction of the relationship or correlation.

It is important to note that $r = 0$ does not mean that there is no relationship at all. For example, if all observed values lie exactly on a circle, there is a perfect non-linear relationship between the variables but r will have a value of zero as r only measures the linear correlation.

The linear correlation co-efficient, is also the square root of the linear co-efficient of determination.

$$\text{we have } \hat{Y} = \bar{Y} + b(X - \bar{X})$$

$$\text{or } \hat{Y} - \bar{Y} = b(X - \bar{X})$$

Squaring both sides, we get

$$(\hat{Y} - \bar{Y})^2 = b^2(X - \bar{X})^2$$

Substituting in the ratio, we find

$$\begin{aligned} \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} &= \frac{b^2 \sum(X - \bar{X})^2}{\sum(Y - \bar{Y})^2} \\ &= \frac{\sum(X - \bar{X})^2}{\sum(Y - \bar{Y})^2} \left[\frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2} \right] \\ &= \left[\frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(Y - \bar{Y})^2 \sum(X - \bar{X})^2}} \right] = r^2 \end{aligned}$$

Example 10.5 Calculate the product moment co-efficient of correlation between X and Y from the following data:

	1	2	3	4	5
\bar{Y}	2	5	3	8	7

(P.U., B.A./B.Sc. 1973)

The calculations needed to compute r are given below:

X	Y	$(X - \bar{X})$	$(X - \bar{X})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
1	2	-2	4	-3	9	6
2	5	-1	1	0	0	0
3	3	0	0	-2	4	0
4	8	1	1	3	9	3
5	7	2	4	2	4	4
15	25	0	10	0	16	13

Here $\bar{X} = \frac{\sum X}{n} = \frac{15}{5} = 3$, and $\bar{Y} = \frac{\sum Y}{n} = \frac{25}{5} = 5$

$$\therefore r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}} = \frac{13}{\sqrt{10 \times 26}} = \frac{13}{16.1} = 0.8$$

Alternatively, the following table is set up for calculation of r .

X	Y	X^2	Y^2	XY
1	2	1	4	2
2	5	4	25	10
3	3	9	9	9
4	8	16	64	32
5	7	25	49	35
15	25	55	151	88

$$r = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

$$= \frac{88 - (15)25/5}{\sqrt{[55 - (15)^2/5][151 - (25)^2/5]}} = \frac{88 - 75}{\sqrt{10 \times 26}} = 0.8$$

10.5.2 Correlation and Causation. The fact that correlation exists between two variables does not imply any cause-and-effect relationship. Two unrelated variables such as the sale of bananas and death rate from cancer in a city, may produce a high positive correlation which may be due to some unknown variable (namely, the city's population). The larger the city, the more consumption of bananas and the higher will be the death rate from cancer. Clearly, this is a false or a merely incidental correlation which is the result of a third variable, the city size. Such a false correlation between two uncorrelated variables is called noncausal or spurious correlation. We therefore should be very careful in interpreting the correlation coefficient as a measure of relationship or interdependence between two variables.

10.5.3 Properties of r . The sample correlation co-efficient r has the following properties:

- The correlation co-efficient r is symmetrical with respect to the variables X and Y , i.e., $r_{XY} = r_{YX}$.
- The correlation co-efficient lies between -1 and $+1$, i.e., $-1 \leq r \leq +1$.
- The correlation co-efficient is independent of the origin and scale.

Proof: Let u and v be the two new variables defined by $u = \frac{X-a}{h}$ and $v = \frac{Y-b}{k}$ so that $X = u + ah$ and $Y = b + kv$, where a and b are the new origins and h and k are the units of measurement.

Let r_{uv} denote the correlation co-efficient between X and Y and r_{uv} the correlation co-efficient between u and v .

Substituting these values in r_{XY} , we get

$r_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$, we get

$$r_{xy} = \frac{\sum[(a + bu) - (a + bv)][(b + ku) - (b + kv)]}{\sqrt{\sum[(a + bu) - (a + bv)]^2 \sum[(b + ku) - (b + kv)]^2}}.$$

where $\bar{X} = a + bu$ and $\bar{Y} = b + kv$. Therefore

$$r_{xy} = \frac{ab \sum(u - \bar{u})(v - \bar{v})}{ab \sqrt{\sum(u - \bar{u})^2 \sum(v - \bar{v})^2}} = r_{uv}$$

This property is very useful in numerical evaluation of r , since due to this property, we can choose convenient origin and scale.

- In case of a bivariate population where both X and Y are random variables, r is the geometric mean between the two regression co-efficients.

That is, if b_{yx} is the regression coefficient of the regression line of Y on X and b_{xy} is the regression coefficient of the regression line of X on Y , and r is the coefficient of correlation, then $r^2 = b_{yx}b_{xy}$ implies $r = \pm \sqrt{b_{yx}b_{xy}}$.

Since the signs of the regression coefficients depend on the same expression $\sum(Y - \bar{Y})(Y - \bar{Y})$ so b_{yx} and b_{xy} are both positive or b_{yx} and b_{xy} are both negative. Therefore

$$r = +\sqrt{b_{yx}b_{xy}}, \text{ if } b_{yx} \text{ and } b_{xy} \text{ are positive,}$$

$$r = -\sqrt{b_{yx}b_{xy}}, \text{ if } b_{yx} \text{ and } b_{xy} \text{ are negative.}$$

The value of r always takes the sign sign as the regression coefficients.

The regression co-efficients of the regression lines for a bivariate population, by using the value of the correlation co-efficient, may be expressed as

$$b_{yx} = r \frac{S_y}{S_x}; b_{xy} = r \frac{S_x}{S_y}$$

$$\bar{Y} - \bar{Y} = r \frac{S_y}{S_x} (\bar{X} - \bar{X}); \text{ and } \bar{X} - \bar{X} = r \frac{S_x}{S_y} (\bar{Y} - \bar{Y}),$$

The letters have their usual meaning.

Example 10.6 Calculate the co-efficient of correlation between the values of X and Y given below:

X	78	89	97	69	59	79	68	61
Y	125	137	156	112	107	136	123	108

Let $a = X - 69$ and $v = Y - 112$. Then $r_{xy} = r_{uv}$. The calculations needed to find r are given in the next page.

X	Y	u	v	u^2	v^2	uv
78	125	9	13	81	169	117
89	137	20	25	400	625	500
97	156	28	44	784	1936	1232
69	112	0	0	0	0	0
59	107	-10	-5	100	25	50
79	136	10	24	100	576	540
68	123	-1	11	1	121	-11
62	108	-8	-4	64	16	32
600	1004	48	108	1530	3468	2160

Now $r = \frac{\sum uv - (\sum u)(\sum v)/n}{\sqrt{\left[\frac{\sum u^2 - (\sum u)^2}{n} \right] \left[\sum v^2 - (\sum v)^2 \right]}}$

$$= \frac{2160 - \frac{48 \times 108}{8}}{\sqrt{\left[\frac{1530 - (48)^2}{8} \right] \left[\frac{3468 - (108)^2}{8} \right]}}$$

$$= \frac{2160 - 648}{\sqrt{(1530 - 288) \times (3468 - 1458)}} = \frac{1512}{\sqrt{841 \times 1910}} = \frac{1512}{910} = 0.96.$$

Hence the correlation coefficient between X and Y is 0.96.

Example 10.7 If b_{xy} is the regression coefficient of X_1 on X_2 , then calculate the product moment coefficient of correlation in each case given

- i) $b_{12} = -0.1, b_{21} = -0.4$
- ii) $b_{12} = 0.27, b_{21} = 0.6$
- iii) $b_{12} = 0.67, b_{21} = 0.35$.

The product moment coefficient of correlation between X_i and X_j is given by

$$r_{ij} = \sqrt{b_{ij} \times b_{ji}}$$

- i) Here $b_{12} = -0.1$, and $b_{21} = -0.4$

$$r_{12} = -\sqrt{(-0.1)(-0.4)} = -0.20.$$

r is negative since both regression coefficients are negative.

- ii) Here both regression coefficients are positive, so r is positive. Thus

$$r_{12} = +\sqrt{b_{12} \times b_{21}} = +\sqrt{(0.27)(0.6)} = +0.40.$$

(ii) Here we have

$$r_{12} = \sqrt{(0.67)(0.38)} = 0.50 \quad (\because b_{11} \text{ and } b_{22} \text{ are positive})$$

10.5.4 Correlation Co-efficient for Grouped Data. In a simple frequency table, the data are arranged with respect to one variable only. If the arrangement is made according to two variables simultaneously in say, m columns and k rows, the frequency table thus obtained is called a correlation table or a bivariate frequency array. The number of observations falling in the (l, j) th cell, is called the (l, j) th cell frequency and is denoted by f_{ij} . The correlation co-efficient, if it exists, can be calculated from such a two-way frequency table by using the class midpoints as the value of the observations. The formula for r then becomes

$$r = \frac{\sum f_{ij} X_i Y_j - \frac{1}{n} (\sum f_{ij} X_i)(\sum f_{ij} Y_j)}{\sqrt{\left[\sum f_{ij} X_i^2 - \frac{1}{n} (\sum f_{ij} X_i)^2 \right] \left[\sum f_{ij} Y_j^2 - \frac{1}{n} (\sum f_{ij} Y_j)^2 \right]}}$$

where $f_{ij} = \sum_{j=1}^m f_{ij}$, the frequency of Y values, $f_{i.} = \sum_{j=1}^k f_{ij}$, the frequency of X values and n is the total frequency.

Example 10.8 Calculate the co-efficient of correlation from the table given below:

Grades in Statistics (Y)	Grades in Mathematics (X)						Total
	40-49	50-59	60-69	70-79	80-89	90-99	
90-99	--	--	--	2	4	4	10
80-89	--	--	1	4	6	5	16
70-79	--	--	3	10	8	1	24
60-69	1	4	9	5	2	--	21
50-59	5	6	6	2	--	--	17
40-49	3	5	4	--	--	--	12
Total	7	15	25	23	20	10	100

(P.U., B.A./B.Sc. 1968)

Let us introduce two new variables u and v given by the relations $u = \frac{X - 64.5}{10}$ and $v = \frac{Y - 74.5}{10}$. The calculations needed for finding r are arranged in the table on page (412).

T_i	λ_j	44.5	54.5	64.5	74.5	84.5	94.5	f_{ij}	$f_{i\cdot}v_j$	$f_{i\cdot}v_j^2$	$f_{i\cdot}f_{j\cdot}v_i$	
	v_i	-2	-1	0	1	2	3					
94.5	2	-1	-	-	-	[8] 2	[16] 4	[24] 4	10	20	40	44
84.5	1	-	-	-	[0] 1	[8] 4	[12] 6	[18] 5	16	16	16	31
74.5	0	-	-	-	[0] 2	[0] 10	[0] 8	[0] 1	26	0	0	0
64.5	-1	[2] 1	[4] 4	[0] 9	[−2] 5	[−4] 2	---	21	−21	21	−3	
54.5	-2	[12] 3	[12] 6	[0] 6	[−4] 2	---	---	17	−34	68	20	
44.5	-3	[18] 3	[18] 5	[0] 4	---	---	---	12	−36	108	33	
	$f_{i\cdot}$	7	15	25	25	20	9	100	−55	253	125	
	$f_{i\cdot}v_i$	−14	−15	0	25	30	64					↑
	$f_{i\cdot}v_i^2$	28	15	0	25	80	90	236				
	$f_{i\cdot}f_{j\cdot}v_i$	32	31	0	24	39	125					← Check

The number in the corner of each cell represents the product $f_{ij}v_iv_j$ where f_{ij} is the cell frequency. Thus $f_{1,1}v_{1,1} = 2(1)(2) = 4$ and $f_{1,5}v_{1,5} = 2(1)(2) = 16$ and so on. The totals in the last column and the last row are equal and represent $\sum f_{i\cdot}f_{j\cdot}v_i$.

$$\text{Now } r_{xx} = r_{yy} = \frac{\sqrt{n(\sum v_i^2 - (\sum v_i)^2)}(\sum f_{ij})(\sum f_{ij}v_iv_j)}{\sqrt{[(\sum f_{ij})^2 - (\sum f_{ij}v_i)^2][(\sum f_{ij})^2 - (\sum f_{ij}v_j)^2]}} \quad (\text{subscripts dropped for convenience in practice})$$

$$= \frac{(100)(125) - (64)(-55)}{\sqrt{[(100)(236) - (64)^2][(100)(253) - (-55)^2]}}$$

$$= \frac{16020}{\sqrt{19304 \times 22275}} \approx 0.77$$

Example 18.9 (a) Correlation between X and Y is r , show that correlation between $aX + b$ and $cY + d$ is r or $-r$ according as a and b have the same or different signs.

- b) Find correlation between X and Y connected by

$$aX + bY + c = 0.$$

i) Let $u = aX$, so that $\bar{u} = a\bar{X}$,

and $v = bY$, so that $\bar{v} = b\bar{Y}$.

Then $(u - \bar{u}) = a(X - \bar{X})$ and $(v - \bar{v}) = b(Y - \bar{Y})$.

By definition, we have

$$\begin{aligned} r_{uv} &= \frac{\sum(u - \bar{u})(v - \bar{v})}{\sqrt{\sum(u - \bar{u})^2 \sum(v - \bar{v})^2}} \\ &= \frac{ab \sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{a^2 \sum(X - \bar{X})^2 b^2 \sum(Y - \bar{Y})^2}} \\ &= \frac{ab}{\sqrt{a^2 b^2}} r_{xy} \\ &= +r, \text{ if } a \text{ and } b \text{ are of the same signs.} \\ &= -r, \text{ if } a \text{ and } b \text{ are of the different signs.} \end{aligned}$$

ii) We are given $aX + bY = c = 0$.

Thus $a \sum X + b \sum Y + nc = 0$, where n is the number of pairs of values (X_i, Y_i) .

Dividing by n , we get

$a\bar{X} + b\bar{Y} - c = 0$, \bar{X} and \bar{Y} being the means of X and Y sets of observations. Subtracting, we have

$$a(\bar{X} - \bar{Y}) + b(\bar{Y} - \bar{Y}) = 0$$

$$\text{or } (\bar{Y} - \bar{Y}) = -\frac{a}{b}(\bar{X} - \bar{Y})$$

$$\text{Now } r_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

$$= \frac{-\frac{a}{b} \sum(X - \bar{X})^2}{\sqrt{\left[\sum(X - \bar{X})^2\right] \left[\frac{a^2}{b^2} \sum(X - \bar{X})^2\right]}} = \frac{-a/b}{\sqrt{\frac{a^2}{b^2}}} = -\frac{a}{b}$$

$= -1$, if a and b are of the same sign.

$= +1$, if a and b are of the opposite signs.

RANK CORRELATION

Sometimes, the actual measurements or counts of individuals or objects are either not available or assessment is not possible. They are then arranged in order according to some characteristic of the object. Such an ordered arrangement is called a ranking and the order given to an individual or object is its rank. The correlation between two such sets of rankings is known as Rank Correlation.

10.6.1 Derivation of Rank Correlation. Let a set of n objects be ranked with respect to character A as $x_1, x_2, \dots, x_n, \dots, x_m$ and according to character B as $y_1, y_2, \dots, y_n, \dots, y_m$. We assume that no two more objects are given the same ranks (i.e. are tied). Then obviously x_i and y_i are some two numbers in $1, 2, \dots, n$.

Since both x_i and y_i are the first n natural numbers, therefore, we have

$$\begin{aligned}\sum_{i=1}^n x_i &= \sum_{i=1}^n y_i = \sum_{i=1}^n i = 1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}, \\ \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i^2 = \sum_{i=1}^n (i)^2 = 1^2 + 2^2 + 3^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}, \\ \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum y_i)^2}{n} \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12}.\end{aligned}$$

Let d_i denote the difference in ranks assigned to the i th individual or object, i.e. $d_i = x_i - y_i$.

$$\begin{aligned}\text{Then } \sum_{i=1}^n d_i^2 &= \sum_{i=1}^n (x_i - y_i)^2 \\ &= \sum (x_i^2 + y_i^2 - 2x_i y_i) = \sum x_i^2 + \sum y_i^2 - 2\sum x_i y_i\end{aligned}$$

Substituting for $\sum x_i^2$ and $\sum y_i^2$, we get

$$\sum_{i=1}^n d_i^2 = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)(2n+1)}{6} - 2\sum x_i y_i$$

$$\text{or } \sum x_i y_i = \frac{n(n+1)(2n+1)}{6} - \frac{1}{2} \sum d_i^2$$

The product moment coefficient of correlation between the two sets of rankings is

$$r_p = \frac{\sum x_i y_i - \frac{1}{n} (\sum x_i)(\sum y_i)}{\sqrt{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] \left[\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right]}}$$

Substitution gives

$$\begin{aligned}r_p &= \frac{\left[\frac{n(n+1)(2n+1)}{6} - \frac{1}{2} \sum d_i^2 \right] - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}} \\ &= \frac{\left[\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} \right] - \frac{1}{2} \sum d_i^2}{\frac{n(n^2-1)}{12}}\end{aligned}$$

$$= \frac{\frac{n(n^2 - 1)}{12} - \frac{1}{2} \sum d_i^2}{\frac{n(n^2 - 1)}{12}} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

This formula is usually denoted by r_s in order to have a distinction. It is often called Spearman's coefficient of rank correlation, in honour of the psychometrician Charles Edward Spearman (1845), who first developed the procedure in 1904.

It is to be noted that $\sum d_i^2$ has the least value and is zero when the numbers are in complete agreement. When they are in complete disagreement, $\sum d_i^2$ attains the maximum value and is equal to

Substituting these values in the formula, we see that

$$r_s = 1 \text{ for } \sum d_i^2 = 0, \text{ and}$$

$$r_s = -1 \text{ for } \sum d_i^2 = \frac{n(n^2 - 1)}{3}$$

Thus r_s also lies between -1 and +1.

Example 10.10 Find the co-efficient of rank correlation from the following rankings of 10 students in Statistics and Mathematics.

Statistics (x):	1	2	3	5	6	7	8	9	10
Mathematics (y):	2	4	3	7	5	8	10	6	9

(P.U., B.A. (Hons.) Part-I, 1964)

We calculate the co-efficient of rank correlation as follows:

x_i	y_i	$d_i = x_i - y_i$	d_i^2
1	2	-1	1
2	4	-2	4
3	3	0	0
4	1	3	9
5	7	-2	4
6	5	1	1
7	8	-1	1
8	10	-2	4
9	6	3	9
10	9	1	1
—	—	0	24

Hence, using Spearman's co-efficient of rank correlation, we get

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 34}{10 \times 99} = 1 - 0.2 = +0.8.$$

This indicates a high correlation between Statistics and Mathematics.

10.6.2 Rank Correlation for Tied Ranks. The Spearman's co-efficient of rank correlation applies only when no ties are present. In case there are ties in ranks, the ranks are adjusted by assigning the mean of the ranks which the tied objects or observations would have if they were ordered. For example, if two objects or observations are tied for fourth and fifth, they are both given the mean rank of 4.5 and 5, i.e. 4.5. The sum of adjusted ranks remains $\frac{n(n+1)}{2}$ but $\sum(x_i - \bar{x})^2 = \sum(y_i - \bar{y})^2 = \frac{n(n^2 - 1)}{12}$. It has been shown that each set of ties involving t observations reduces the values of d^2 by a quantity equal to $\frac{1}{12}(t^2 - t)$. In such a situation, one of the following two methods is to be used.

First, for each tie, add a quantity $\frac{1}{12}(t^2 - t) \approx \sum d^2$ before substituting the values in the formula for the rank correlation coefficient.

Spearman's co-efficient of rank correlation in order to adjust the formula for the tied observations.

Second, use the product moment co-efficient of correlation to find the correlation between the sets of adjusted ranks.

Example 10.11 Two members of a selection committee rank eight persons according to their suitability for promotion as follows:

Persons	A	B	C	D	E	F	G	H
Member 1	1	2.5	2.5	4	5	6	7	8
Member 2	2	4	1	3	6	6	6	8

Calculate the co-efficient of rank correlation.

We observe that both the sets of rankings contain ties. The coefficient of rank correlation is therefore calculated as below:

Rank	Member 1	Member 2	d_i	d_i^2
A	1	2	-1	1
B	2.5	4	-1.5	2.25
C	2.5	1	1.5	2.25
D	4	3	1	1
E	5	6	-1	1
F	6	6	0	0
G	7	6	1	1
H	8	8	0	0
Σ	36	36	0	8.5

For tie between B and C, (first rankings) $t=2$ and for E, F and G (second rankings) $t=3$. The quantity to be added to $\sum d^2$ is

$$\frac{1}{12}(2^2 - 2) + \frac{1}{12}(3^2 - 3) = 2.5.$$

$$\text{Hence } r_s = 1 - \frac{6(8.5 + 2.5)}{8(64 - 1)} = 1 - \frac{66}{504} = 1 - 0.131 = 0.869.$$

Alternative Method:

We see that the first member has tied *B* and *C*, while the second member has tied *E*, *F* and *G*. Let us estimate the ranks given by the first member by x_i and those of second member by y_i . Then we proceed as follows:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	1	4	2
2.5	4	6.25	16	10
2.5	1	6.25	1	2.5
4	3	16	9	12
5	6	25	36	30
6	6	36	36	36
7	6	49	36	42
8	8	64	64	64
36	36	203.5	246	198.5

Hence the co-efficient of rank correlation is

$$\begin{aligned}
 r &= \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sqrt{[\sum x_i^2 - (\sum x_i)^2/n][\sum y_i^2 - (\sum y_i)^2/n]}} \\
 &= \frac{198.5 - (36)(36)/8}{\sqrt{[203.5 - (36)^2/8][202 - (36)^2/8]}} \\
 &= \frac{198.5 - 162}{\sqrt{(203.5 - 162)(202 - 162)}} = \frac{36.5}{\sqrt{(41.5)(40)}} \\
 &= \frac{36.5}{40.47} = 0.896,
 \end{aligned}$$

which indicates a high degree of agreement between the two members.

16.3 Co-efficient of Concordance. The Spearman's co-efficient of rank correlation measures the agreement between two sets of rankings only, but in practice, the individuals or objects are sometimes ranked by more than two people. We then need a co-efficient to measure agreement among more than two sets of rankings. Such a co-efficient is obtained as below:

Let there be n rankings of n individuals or objects instead of two. Obviously in case of complete agreement, the rank totals will form the series $n, 2n, 3n, \dots, nn$.

The mean of these totals is

$$\bar{X} = (m + 2m + 3m + \dots + nm) / n = u.$$

$$= \frac{m(1+2+3+\dots+n)}{n} = \frac{m(n+1)}{2},$$

and the variance of these sums, which is the maximum possible, is

$$\begin{aligned}\text{Var(Total)} &= \frac{1}{n} [m^2 + (2m)^2 + (3m)^2 + \dots + (nm)^2] - \left[\frac{m(n+1)}{2} \right]^2 \\ &= \frac{m^2[1^2 + 2^2 + 3^2 + \dots + n^2]}{n} - \left[\frac{m(n+1)}{2} \right]^2 \\ &= \frac{m^2(n+1)(2n+1)}{6} - \frac{m^2(n+1)^2}{4} = \frac{m^2(n^2-1)}{12}.\end{aligned}$$

But the totals of observed ranks will not necessarily be the same. Let Σ denote the sum of squares of deviations of the totals of the observed ranks from their common mean, i.e. $\frac{m(n+1)}{2}$.

(Co-efficient of Concordance) W , is defined as the ratio of the variance of the totals of the observed ranks to the variance in case of complete agreement. Thus, we have

$$W = \frac{\Sigma}{n} = \frac{m^2(n^2-1)}{12} = \frac{12S}{m^2(n^2-n)}.$$

This co-efficient is due to Maurice G. Kendall (1907-1983) and varies from 0 to 1. When $W=1$ it represents complete agreement.

 **Example 18.12** The following data give rankings of six persons for their ability by three judges P , Q and R . Calculate the co-efficient of concordance.

Person	A	B	C	D	E	F
Judge P	3	1	6	2	5	4
Judge Q	4	3	2	5	1	6
Judge R	2	1	6	5	4	3

(P.U., B.A. (Hons.), Part-II, 2008)

Here the totals of the observed ranks are 9, 5, 14, 12, 10 and 13; $m=3$ and $n=6$ so the mean $= \frac{m(n+1)}{2} = \frac{3(6+1)}{2} = 10.5$.

$$\begin{aligned}\text{Thus } \Sigma &= (9-10.5)^2 + (5-10.5)^2 + (14-10.5)^2 + (12-10.5)^2 + (10-10.5)^2 + (13-10.5)^2 \\ &= (-1.5)^2 + (-5.5)^2 + (3.5)^2 + (1.5)^2 + (-0.5)^2 + (2.5)^2 = 53.50\end{aligned}$$

$$\text{Hence } W = \frac{12S}{m^2(n^2-n)} = \frac{12 \times 53.5}{9(216-6)} = \frac{642}{1890} = +0.34.$$

EXERCISES

SUBJETIVE

Answer 'True' or 'False'. If the statement is not true then replace the underlined words with words that make the statement true:

- i) A high value of correlation between Y and X indicates a high likelihood of a cause and effect relationship between Y and X.
- ii) Correlation analysis finds the equation of the line for two variables.
- iii) The co-efficient of correlation lies between 0 and +1.
- iv) The correlation co-efficient is not independent of the origin and scale.
- v) Regression analysis measures the strength of the linear relationship between two variables.
- vi) In regression analysis X and Y must both be normally distributed.
- vii) The method of least squares gives the line of best fit.
- viii) If the co-efficient of determination r^2 is equal to 5, then it indicates that 50% of the variation is due to chance or other factors.
- ix) If the slope of the regression line has a negative sign, then the coefficient of determination also is negative.
- x) If all the points in a scatter diagram fall on the regression line, then the standard error of estimate equals positive value.

MULTIPLE CHOICE QUESTIONS

When the slope of regression line is negative, the following statistic is also negative

- a) r
- b) r^2
- c) Standard error of estimate
- d) Standard error of slope co-efficient

If there is no linear relationship between the two variables then which one of the following does not hold?

- a) $a = 0$
- b) $b = 0$
- c) $r^2 = 0$
- d) The regression line is either vertical or horizontal.

- iii) If the correlation co-efficient $r = 0.7$, then the proportion of variation for Y explained by X is
a) 0.49
b) 0.50
c) 0.70
d) $\sqrt{0.70}$
- iv) The dependent variable is also known as
a) Explained variable
b) Response variable
c) Predicted variable
d) All of above
- v) In the regression equation $\hat{Y} = \alpha + \beta X + e$, both X and Y variables are
a) Random
b) Fixed
c) X is fixed and Y is random
d) Y is fixed and X is random
- vi) The variation of the Y values around the regression line is measured by
a) $\sum(Y - \bar{Y})^2$
b) $\sum(Y - \hat{Y})^2$
c) $\sum(\hat{Y} - \bar{Y})^2$
d) None of above
- vii) If both the dependent and independent variables increase simultaneously, the coefficient will be in the range of
a) 0 to +1
b) 0 to -1
c) 1 to 2
d) -1 to +1
- viii) Which of the following statements is incorrect about correlation coefficient?
a) It passes through the means of the data
b) It is symmetrical with respect to X and Y
c) It is independent of origin and scale
d) It is the geometric mean between the two regression coefficients

- (i) If the unexplained variation between variables X and Y is 0.40 then r^2 is
a) 0.75
b) 0.60
c) 0.40
d) None of the above
- (ii) The strength of a linear relationship between two variables Y and X is measured by
a) r^2
b) b_{yx}
c) r
d) None of above

OBJECTIVE

- (i) Explain what is meant by (i) regression, (ii) regressand, (iii) regressor, and (iv) regression co-efficient.
- (ii) Differentiate between a deterministic and a probabilistic relationship, giving examples.
- (iii) What is a scatter diagram? Describe its role in the theory of regression.
- (iv) What is a linear regression model? Explain the assumptions underlying the linear regression model.
- (v) Explain the principle of least-squares.
- (vi) Explain briefly how the principle of least-squares is used to find a regression line based on a sample of size n . Illustrate on a rough sketch the distances whose squares are minimized, taking care to distinguish the dependent and independent variables.
- (vii) Find least-squares estimates of parameters in a simple linear regression model $Y = \alpha + \beta X_i + e_i$, where e_i s are distributed independently with mean zero and constant variance.
- (viii) What are the properties of the least-squares regression line?
(P.U., B.A./B.Sc. 1992)
- (ix) Show that the regression line passes through the means of observations.
(P.U., D.Sc. 1962)
- (x) Describe briefly how you would obtain the line of regression of one variable (Y) on another variable (X), using the method of least-squares.
(P.U., B.A./B.Sc. 1975)
- (xi) What is meant by the standard error of estimate? If the regression line of Y on X is given by $\hat{Y} = a + bX$, prove that the standard error of estimate $s_{e,a}$ is given by

$$s_{e,a} = \sqrt{\frac{\sum Y^2 - n \sum Y - b \sum XY}{n-2}}$$

10.6 Given the following set of values:

X	20	11	15	10	17	19
Y	5	13	14	17	8	9

- a) Determine the equation of the least squares regression line.
- b) Find the predicted values of Y for $X = 10, 11, 15, 17, 19, 20$.
- c) Use the predicted values found in (b) to find the standard error of estimate.

10.7 Given these ten pairs of (X, Y) values:

X	1	1	2	3	4	4	5	6	6	7
Y	2.1	2.5	3.1	3.0	3.8	3.2	4.3	3.9	4.4	4.8

- a) Plot a scatter diagram for the above data.
- b) Carry out the necessary computations to obtain the least-squares estimates of the parameters in the simple linear regression $\hat{Y} = a + bX + e$.
- c) Compute the residuals and verify that they add to zero.
- d) Use the regression equation to predict the values of Y when $X=10$.

10.8 For each of the following data, determine the estimated regression equation $\hat{Y} = a + bX$.

- a) $\bar{X} = 10$; $\bar{Y} = 20$; $\sum XY = 1,000$; $\sum X^2 = 200$; $n = 10$.
- b) $\sum X = 528$; $\sum Y = 11,720$; $\sum XY = 194,400$; $\sum X^2 = 11,440$; $n = 32$.
- c) $\sum X = 1,239$; $\sum Y = 79$; $\sum XY = 1,232$; $\sum X^2 = 17,322$; $\sum Y^2 = 293$; $n = 100$.
- d) $n = 10$; $\sum X = 1710$; $\sum Y = 1500$; $\sum X^2 = 293,162$; $\sum Y^2 = 59,390$; $\sum XY = 130,621$.
- e) $\bar{X} = 52$; $\bar{Y} = 237$; $\sum (X - \bar{X})^2 = 2800$; $\sum (X - \bar{X})(Y - \bar{Y}) = 9871$.

10.9 The owner of a retail organization is interested in the relationship between price at which a commodity is offered for sale and the quantity sold. The following sample data have been collected:

Price	25	45	30	50	35	40	65	75	70	60
Quantity sold	118	105	112	100	111	108	95	88	91	96

- a) Plot a scatter diagram for the above data.
- b) Using the method of least squares, determine the equation for the estimated regression line. Plot this line on the scatter diagram.
- c) Calculate the standard deviation of regression, s_{xy} . (B.Z.U., M.A. Econ., 2008)

10.10 Given the following sets of values:

Y	6.5	5.3	8.6	1.2	4.2	2.9	1.1	3.0
X	3.2	2.7	4.5	1.6	2.0	1.7	0.6	1.9

- a) Compute the least-squares regression equation for \bar{Y} values on X values, that is the equation $\bar{Y} = a + bX$.
- b) Compute the standard error of estimate, s_e .
- c) Compute the least-squares regression equation for \bar{X} values on \bar{Y} values, that is the equation $\bar{X} = a_1 + b_1\bar{Y}$.
- d) Compute the standard error of estimate, s_{e_1} .
- (11) a) Explain what is meant by the co-efficient of determination.
- b) Compute the co-efficient of determination for the following data and interpret the co-efficient.

Income (X) (000)	10	20	30	40	50	60
Expenditure (Y) (000)	7	21	23	34	6	53

- (12) a) What is the total variation, the explained variation and the unexplained variation?
- b) Compute (i) the total variation, (ii) the explained variation and (iii) the unexplained variation for the data in (10.1)(b). How much of the variability in Y is explained by the linear regression model?
- (13) a) Differentiate between regression and correlation, giving examples. (P.U., B.A./B.Sc. 1979)
- b) Describe the properties of the correlation coefficient.
- c) What values may r assume? Interpret the meaning when $r = -1, 0, +1$. (P.U., B.A./B.Sc. 1980)

- (14) Define the terms correlation and product moment co-efficient of correlation. Prove that the correlation co-efficient is independent of the origin and scale. (P.U., B.A./B.Sc. 1981)
- (15) Compute the correlation co-efficient between the variables X and Y represented in the following table:

X	2	4	5	6	8	11
Y	18	12	0	8	7	5

- (16) Multiply each X value by 2 and add 6. Multiply each Y value by 3 and subtract 15. Find the correlation co-efficient between the two new sets of values, explaining why you do or do not obtain the same result as in part (b).
- (17) Show that, if r_{xy} is the correlation co-efficient calculated from a set of paired data $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, then r_{xy} , the correlation co-efficient for $x_i = aX_i + b$ and $y_i = cY_i + d$ (with $a \neq 0$ and $c \neq 0$), is given by $r_{xy} = r_{x'y'}$.
- (18) Calculate the correlation co-efficient by first multiplying each X and Y by 10 and then subtracting 70 from each X and 60 from each Y .

X	8.2	9.6	7.0	9.4	10.9	7.1	9.0	8.6	8.4	10.5
Y	8.7	9.6	6.9	8.5	11.3	7.6	9.2	8.3	8.4	12.4

- 10.16 a) Explain the term correlation. It is known that $r_{xy} = 0.7$. Find (i) r_{xG} (ii) r_{yv} , where $x = G$ and $v = 3Y$.
- b) Calculate the coefficient of correlation for a sample of 20 pairs of observations, given that

$$\bar{X} = 2, \bar{Y} = 8, \sum X^2 = 180, \sum Y^2 = 1424 \text{ and } \sum XY = 464. \quad (\text{P.U., B.Sc. Hon. 2010})$$

- 10.17 The following data were computed from personnel records of a manufacturing firm:

X = number of years of service, Y = weekly wage rate

$n = 25$; $\sum X = 2,433$; $\sum Y = 4,245$; $\sum X^2 = 281,019$; $\sum Y^2 = 841,786$ and $\sum XY = 482,788$.

- i) Compute the correlation co-efficient.
- ii) If the correlation co-efficient indicates that there does exist a relationship between X and Y , compute the least-squares line of regression. What do the values of a and b signify?

(P.U., B.A./B.Sc. 2010)

- 10.18 Find the product moment co-efficients of correlation between traffic density and accident rate, given the following information available. Find also the coefficient of determination and interpret it.

Traffic Density	30	35	40	45	50	60	70	80	90
Accident Rate	2	4	5	5	8	15	24	30	32

- ✓ 10.19 Given marks as

Student	1	2	4	5	6	7	8	9	10	11	12
Economics paper	36	42	41	46	59	46	65	31	68	41	70
Physics Paper	62	42	60	53	36	50	42	66	44	58	65

Find the co-efficient of correlation and interpret it.

- 10.20 Calculate the co-efficient of correlation and obtain the lines of regression of the following data:

Price (X)	3	4	5	6	7	8	9	10	11	12
Demand (Y)	25	24	20	20	19	17	16	13	10	6

(P.U., M.A. Econ. 2010)

- 10.21 a) Find the correlation co-efficient between X and Y , given

X	5	12	4	16	18	21	22	23	25
Y	11	16	15	20	17	19	25	24	21

(P.U., M.A. Econ. 2010)

- b) Find the co-efficient of correlation between persons employed and cloth manufactured in a textile mill. Interpret the result.

Persons employed	137	209	113	189	176	200	219
Cloth manufactured ('000 yds)	23	47	22	40	39	51	49

(P.U., B.A./B.Sc. 1980)

- 22 The following table gives the distribution of the total population and those who are wholly or partially blind among them. Find out if there is any relation between age and blindness.

Age	No. of Persons in thousand	Blind
0 - 9	100	55
10 - 19	60	40
20 - 29	40	40
30 - 39	36	40
40 - 49	24	16
50 - 59	11	12
60 - 69	6	18
70 - 79	3	15

(P.U., B.A./B.Sc. 1983)

Hint. First calculate the numbers of blind per 1000 and then correlate with the midpoints of age groups.

- 23 A computer while calculating the correlation co-efficient between two variables X and Y from 23 pairs of observations obtained the following sum:

$$\sum X = 125, \sum Y = 650, \sum Y^2 = 100, \sum Y^2 = 460, \sum XY = 508$$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X

9

8

Y

14

6

X

8

6

Y

12

8

correlation.

(P.C.S. 1972; P.U., B.A./B.Sc. 1974)

- 24 If the equations of the least squares regression lines are:

i) $\bar{Y} = 20.8 - 0.219X$ (\bar{Y} on X), and $\bar{X} = 16.2 - 0.785\bar{Y}$ (\bar{X} or \bar{Y});

ii) $\bar{Y} = 2.64 + 0.643\bar{X}$ (\bar{Y} on X), and $\bar{X} = -1.91 + 0.917\bar{Y}$ (\bar{X} or \bar{Y});

iii) $\bar{Y} = 1.94\bar{X} + 10.83$ (\bar{Y} on X), and $\bar{X} = 0.15\bar{Y} + 6.18$ (\bar{X} or \bar{Y});

iv) $\bar{Y} = 15 - 1.96\bar{X}$ (\bar{Y} on X), and $\bar{Y} = 15.91 - 2.22\bar{X}$ (\bar{X} or \bar{Y}).

Find the product moment coefficient of correlation in each case.

- 10.25 Find the co-efficient of correlation for the frequency distribution of two variables given below following table.

$X \backslash Y$	5 - 14	15 - 24	25 - 34	35 - 44	45 - 54
0 - 9	3	1	—	—	—
10 - 19	12	8	14	1	—
20 - 29	2	13	40	12	3
30 - 39	—	3	40	27	7
40 - 49	—	—	6	6	4

Practical

Also find the regression equation of Y and X .

- 10.26 Compute correlation co-efficient from the following correlation table for weights and heights of women students.

Height in inches	Weight in pounds					
	90	110	130	150	170	190
57	—	—	—	—	1	—
60	8	21	8	—	—	—
63	3	50	57	—	—	2
66	1	24	54	19	3	—
69	—	1	5	3	—	—
72	—	—	2	—	—	1

(P.U., B.A./B.Sc. 1960)

- 10.27 a) Describe in simple words the following concepts:

(i) co-efficient of correlation, (ii) scatter diagram, (iii) least squares principle
 (iv) estimate of regression co-efficient.

- b) The following results were obtained for a bivariate frequency distribution after making a transformation $u = \frac{x - 1250}{500}$ and $v = \frac{y - 500}{300}$; $n = 66$, $\sum fuv = -4$, $\sum f u^2 = 109$, $\sum f v^2 = 115$, $\sum f uv = 91$. Calculate the coefficient of correlation and obtain the equations of the lines of regression in the simplest form.

(P.U., B.A./B.Sc. 1960)

- 10.28 If X_1 , X_2 and X_3 are uncorrelated variables, each having the same standard deviation, obtain the co-efficient of correlation between $(X_1 + X_2)$ and $(X_2 + X_3)$. (P.U., B.A. Hon. Part-II, 1960)

- 10.29 a) What is rank correlation? Derive Spearman's co-efficient of rank correlation.

(P.U., B.A./B.Sc. 1960, 71, 82, 84, 85)

- b) The ranks of the same 16 students in Mathematics and Physics were as follows:

(1, 1); (2, 10); (3, 3); (4, 4); (5, 5); (6, 7); (7, 2); (8, 6); (9, 8); (10, 11); (11, 15); (12, 9); (13, 14); (14, 12); (15, 16); (16, 13); the two numbers within brackets denoting the ranks of the same student in Maths. and Physics respectively. Calculate the rank correlation co-efficient for proficiencies of this group in two subjects. (P.U., B.A./B.Sc. 1960)

- 30 a) If n pairs of values of two variables a and b are given, where each variable is ranked in order (1 to n), show that the co-efficient of correlation between ranks is given by

$$\rho_r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where d = the difference between the ranks of a and b .

(P.U., B.A./B.Sc., 1989)

- b) Obtain the product moment coefficient of correlation between the following values:

a	7.4	9.0	11.0	2.5	4.6	6.5
b	8.5	6.1	2.4	6.7	12.6	3.3

Rank the values and hence find a rank correlation coefficient between the two sets.

- 31 a) Describe circumstances in which you would use: (i) rank correlation co-efficient; (ii) product moment correlation coefficient.

- b) The following table shows how 10 students, arranged in alphabetical order, were ranked according to their achievements in both laboratory and lecture portions of a statistics course. Find the co-efficient of rank correlation.

Laboratory	8	3	9	2	7	10	6	1	5
Lecture	9	5	10	1	8	7	3	4	2

(P.U., B.A./B.Sc., 1989)

- 32 Ten competitors in a beauty contest are ranked by three judges in the following order:

First Judge	1	6	5	3	2	4	9	7	8
Second Judge	3	5	8	4	7	10	2	1	6
Third Judge	6	4	8	1	2	3	10	5	7

Use the rank correlation co-efficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

(P.U., B.A./B.Sc., 1960; B.Sc. (Hons.) Part-I, 1971)

- 33 In a pointing competition various entries are ranked by three judges. Use Spearman's rank correlation co-efficient to discuss which pair of judges has the nearest approach to common tastes.

Entry	A	B	C	D	E	F	G	H	K	I
Judge X	5	2	6	8	1	7	4	9	3	10
Judge Y	1	7	6	10	4	5	3	8	2	9
Judge Z	6	4	9	8	1	2	3	10	5	7

(P.U., D.St., 1964)

- 34 a) What are tied ranks? Explain how you would find the co-efficient of rank correlation for tied ranks.

- b) Compute the co-efficient of rank correlation for the following ranks:

X	8	3	6.5	3	6.5	9	3	1	3
Y	8	9	6.5	2.5	4	3	6.5	1	2.5

10.35 Establish the formula for the "co-efficient of concordance". Find the same for the following data:

X	1	2	3	4	5	6	7	8	9	10
Y	7	10	4	1	6	8	9	5	2	3
Z	9	6	10	3	5	4	7	8	2	1

+

CHAPTER 11

**MULTIPLE
REGRESSION AND
CORRELATION**

MULTIPLE REGRESSION AND CORRELATION

INTRODUCTION

The technique of simple regression which involves one dependent variable and one independent variable is often inadequate in most real-world situations where a variable depends upon two or more independent variables or regressors. For example, the yield of a crop depends upon the fertility of the fertilizer applied, rainfall, quality of seed, etc. Likewise, the systolic blood pressure of a person depends upon one's weight, age, etc. In such cases, the technique of simple regression may be expanded to include several independent variables. A regression which involves two or more independent variables is called a multiple regression. This, in case of multiple linear

regression where k independent variables influence the dependent variable Y , the general format of model is

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \epsilon_i, \quad (i = 1, 2, \dots, n)$$

where ϵ_i 's are the random errors,

α and β_j 's are the unknown population parameters, α is the intercept and $\beta_1, \beta_2, \dots, \beta_k$ are the regression coefficients for variables X_1, X_2, \dots, X_k respectively,

$X_{1i}, X_{2i}, \dots, X_{ki}$ are the fixed values of k independent variables. The first of the two subscripts attached to each regressor denotes the variable and the second refers to the observation number.

We assume that

(i) $E(\epsilon_i) = 0$ for all i . This implies that for given values of X_i 's,

$$E(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$$

(ii) $\text{Var}(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$ for all i , i.e. the variance of error terms is constant.

(iii) $E(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$, i.e. error terms are independent of each other.

(iv) $E(X_i \epsilon_i) = 0$ for all regressors X_i and each X variable are independent.

(v) ϵ_i 's are normally distributed with a mean of zero and a constant variance, σ^2 .

We assume further in a multiple regression model that there exists no exact linear relationship between any two of the regressors.

The corresponding regression equation estimated from sample data then takes the following form

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$$

and $\hat{\alpha}$'s and $\hat{\beta}_j$'s are the least-squares estimates of the population parameters α and β_j 's. The parameters β_j 's are also called the partial regression coefficients as $\hat{\beta}_j$ or its estimate $\hat{\beta}_j$ measures the change in the mean value of \hat{Y} for a unit change in X_j while all other variables remain unchanged.

MULTIPLE LINEAR REGRESSION WITH TWO REGRESSORS

For two independent variables X_1 and X_2 , the predicting equation for an individual \hat{Y} value is

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \epsilon_i$$

and the estimated multiple linear regression based on sample data is

$$\hat{Y} = \alpha + b_1 X_{1i} + b_2 X_{2i}$$

for a set of n observations, each of which is a number triple (Y_i, X_{1i}, X_{2i}) . The error or residual is given as

$$e_i = Y_i - \hat{Y}_i = Y_i - (\alpha + b_1 X_{1i} + b_2 X_{2i})$$

Using the least-squares criterion, we determine those values of α , b_1 and b_2 which will minimize the sum of squared residuals, $\sum e_i^2$. To minimize $\sum e_i^2$, we find $\frac{\partial \sum e_i^2}{\partial \alpha}$, $\frac{\partial \sum e_i^2}{\partial b_1}$ and $\frac{\partial \sum e_i^2}{\partial b_2}$ and set equal to zero.

Thus

$$\frac{\partial \sum e_i^2}{\partial \alpha} = -2 \sum [Y_i - (\alpha + b_1 X_{1i} + b_2 X_{2i})] = 0,$$

$$\frac{\partial \sum e_i^2}{\partial b_1} = -2 \sum X_{1i} [Y_i - (\alpha + b_1 X_{1i} + b_2 X_{2i})] = 0,$$

$$\frac{\partial \sum e_i^2}{\partial b_2} = -2 \sum X_{2i} [Y_i - (\alpha + b_1 X_{1i} + b_2 X_{2i})] = 0.$$

Simplifying, we get the following three normal equations:

$$\sum Y = n\alpha + b_1 \sum X_1 + b_2 \sum X_2,$$

$$\sum X_1 Y = \alpha \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2,$$

$$\sum X_2 Y = \alpha \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2.$$

The values of α , b_1 and b_2 are determined by solving these three normal equations simultaneously and are substituted into

$$\hat{Y} = \alpha + b_1 X_1 + b_2 X_2$$

to obtain the estimated multiple linear regression equation.

Example 11.1 A statistician wants to predict the incomes of restaurants, using two independent variables: the number of restaurant employees and restaurant floor area. He collected the following data:

Income (000) Y	Floor area (000 sq. ft) X_1	Number of employees X_2
30	10	15
22	5	8
16	10	12
7	3	7
14	2	10

Calculate the estimated multiple linear regression equation (i.e. $\hat{Y} = a + b_1 X_1 + b_2 X_2$) for the above data.

The estimated multiple linear regression equation is

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

where a , b_1 and b_2 are the least squares estimates of α , β_1 and β_2 . The three normal equations are:

$$\sum Y = a n + b_1 \sum X_1 + b_2 \sum X_2$$

$$\sum X_1 Y = a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2,$$

$$\sum X_2 Y = a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2.$$

The calculations needed to find a , b_1 and b_2 are showing in the following table:

Y	X_1	X_2	X_1^2	X_2^2	$X_1 X_2$	$X_1 Y$	$X_2 Y$
30	10	15	100	225	150	300	450
22	5	8	25	64	40	110	176
16	10	12	100	144	120	160	192
7	3	7	9	49	21	21	49
14	2	10	4	100	20	28	140
89	30	52	230	582	351	619	1007

Substituting the sums in the normal equations we get

$$5a + 30b_1 + 52b_2 = 89$$

$$30a + 230b_1 + 351b_2 = 619$$

$$52a + 351b_1 + 582b_2 = 1007$$

Solving them simultaneously we obtain

$$a = -1.33, b_1 = 0.38 \text{ and } b_2 = 1.62.$$

Hence the desired estimated multiple linear regression is

$$\hat{Y} = -1.33 + 0.38X_1 + 1.62X_2.$$

11.2.1 Expression of Multiple Linear Regression in Deviation Form. The computational procedure is considerably simplified by working with the deviations from the respective means of the variables. With two independent variables, the estimated multiple regression equation is

$$\hat{Y}_j = a + b_1 \bar{X}_1 + b_2 \bar{X}_2, \quad (j = 1, 2, \dots, n)$$

As the regression equation goes through the point of means, we have

$$\hat{Y} = a + b_1 \bar{X}_1 + b_2 \bar{X}_2$$

Subtracting, we get

$$y_i = b_1 x_{1i} + b_2 x_{2i},$$

where $\hat{y}_i = \bar{Y} - \bar{x}_1 x_{1i} = \bar{X}_2 - \bar{X}_1$ and $x_{2i} = X_{2i} - \bar{X}_2$.

Then $e_i = y_i - \hat{y}_i = y_i - b_1 x_{1i} - b_2 x_{2i}$, and

$$\sum e_i^2 = \sum (y_i - b_1 x_{1i} - b_2 x_{2i})^2.$$

Differentiating $\sum e_i^2$ partially w.r.t b_1 and b_2 , and equating to zero, we get

$$\frac{\partial \sum e_i^2}{\partial b_1} = -2 \sum x_{1i} (y_i - b_1 x_{1i} - b_2 x_{2i}) = 0$$

$$\frac{\partial \sum e_i^2}{\partial b_2} = -2 \sum x_{2i} (y_i - b_1 x_{1i} - b_2 x_{2i}) = 0$$

which yield, on simplification, the following two normal equations:

$$\sum x_{1i} y = b_1 \sum x_{1i}^2 + b_2 \sum x_{1i} x_{2i},$$

$$\sum x_{2i} y = b_1 \sum x_{1i} x_{2i} + b_2 \sum x_{2i}^2,$$

where the subscript i is dropped for convenience in printing.

Solving these two equations simultaneously, we get

$$b_1 = \frac{(\sum x_{1i} y)(\sum x_{2i}^2) - (\sum x_{1i} y)(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}, \text{ and}$$

$$b_2 = \frac{(\sum x_{1i} y)(\sum x_{1i}^2) - (\sum x_{1i} y)(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}.$$

Then a , the constant, is determined by

$$a = \bar{Y} - b_1 \bar{x}_1 + b_2 \bar{x}_2.$$

This is an alternative approach to solving the normal equations directly.

Example 11.2 Compute the estimated multiple linear regression, $\hat{Y} = a + b_1 X_1 + b_2 X_2$, from the data in Example 11.1, using the multiple regression in the deviation form.

In Example 11.1, we found that

$$\sum Y = 89, \sum X_1 = 30, \sum X_2 = 52, \sum X_1^2 = 238, \sum X_2^2 = 582,$$

$$\sum X_1 X_2 = 351, \sum X_1 Y = 619, \sum X_2 Y = 1007 \text{ and } n = 5$$

Now we first calculate

$$\sum x_1^2 = \sum X_1^2 - \frac{(\sum X_1)^2}{n} = 238 - \frac{(30)^2}{5} = 58,$$

$$\sum x_2^2 = \sum X_2^2 - \frac{(\sum X_2)^2}{n} = 582 - \frac{(52)^2}{5} = 41.2,$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{n} = 251 - \frac{(30)(52)}{5} = 39,$$

$$\sum x_1 Y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{n} = 619 - \frac{(30)(89)}{5} = 85,$$

$$\sum x_2 Y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{n} = 1007 - \frac{(52)(89)}{5} = 81.4.$$

Next, we compute the regression co-efficients and constant as follows:

$$\begin{aligned} b_0 &= \frac{(\sum x_1 Y)(\sum x_1^2) - (\sum x_1 Y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \\ &= \frac{(85)(41.2) - (81.4)(39)}{(58)(41.2) - (39)^2} = \frac{327.4}{868.6} = 0.38, \end{aligned}$$

$$\begin{aligned} b_1 &= \frac{(\sum x_2 Y)(\sum x_1^2) - (\sum x_1 Y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \\ &= \frac{(81.4)(58) - (85)(39)}{(58)(41.2) - (39)^2} = \frac{148.4}{868.6} = 1.62, \end{aligned}$$

$$\text{and } \bar{y} = \bar{Y} - b_0 \bar{X}_1 + b_1 \bar{X}_2 \\ = 17.8 - (0.38)(10.4) + (1.62)(10.4) = -1.33$$

Hence the desired multiple linear regression equation is

$$\hat{Y} = -1.33 + 0.38 X_1 + 1.62 X_2.$$

It is to be noted that we have exactly the same results as previously.

11.2.2 Standard Error of Estimate. The standard error of estimate is the standard deviation of regression. It measures the dispersion of Y values about the population multiple regression line. For a multiple regression with two independent variables X_1 and X_2 , it is denoted symbolically as $\sigma_{Y|X_1, X_2}$, where the subscripts indicate that Y is regressed against two independent variables X_1 and X_2 ; usually, the value of $\sigma_{Y|X_1, X_2}$ is not known, it is therefore estimated from sample data.

The standard error of estimate (unbiased estimate), denoted by $s_{Y|X_1, X_2}$ is given by

$$s_{Y|X_1, X_2} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n-3}}.$$

but it can be computed more readily by using the following relation:

$$\begin{aligned}\sum(Y - \hat{Y})^2 &= \sum(Y - (a + b_1 X_1 + b_2 X_2))^2 \\&= \sum Y^2 - a \sum Y - b_1 \sum X_1 Y - b_2 \sum X_2 Y.\end{aligned}$$

A larger value of $s_{Y|X_1, X_2}$ means that the multiple regression equation is of little use in estimation or prediction.

11.2.3 Co-efficient of Multiple Determination and Multiple Correlation. The co-efficient of multiple determination, which measures as in the case of simple regression, the proportion of variability in the values of the dependent variable Y explained by its linear relation with the independent variables, is defined by the ratio of the variation in Y explained by the regression equation to the total variation. In multiple regression with two regressors X_1 and X_2 , the co-efficient of multiple determination is denoted symbolically by $R_{Y|X_1, X_2}^2$ and is computed by

$$R_{Y|X_1, X_2}^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2},$$

where $\hat{Y} = a + b_1 X_1 + b_2 X_2$, but it can be readily computed by using the relation

$$\sum(\hat{Y} - \bar{Y})^2 = a \sum Y + b_1 \sum X_1 Y + b_2 \sum X_2 Y - (\bar{Y})^2/n.$$

The co-efficient of multiple determination lies between 0 and 1, and has same meaning as in simple linear regression.

The positive square root of the co-efficient of multiple determination, i.e., $\sqrt{R_{Y|X_1, X_2}^2}$ is called the co-efficient of multiple correlation. R_Y measures the degree of association between Y and both the regressors X_1 and X_2 combined, and is always taken to be positive.

Example 11.3 Compute the standard error of estimate, co-efficient of multiple determination and co-efficient of multiple correlation for the data in Example 11.1.

For the data in Example 11.1, we found from the regression calculation, that

$$\sum Y = 89, \sum Y^2 = 1885, n = 5, a = -1.33,$$

$$\sum X_1 Y = 519, \sum X_2 Y = 1007, b_1 = 0.38, b_2 = 1.62$$

Therefore $s_{Y|X_1, X_2} = \sqrt{\frac{\sum Y^2 - a \sum Y - b_1 \sum X_1 Y - b_2 \sum X_2 Y}{n-3}}$

$$= \sqrt{\frac{1885 - (-1.33)(89) - (0.38)(519) - (1.62)(1007)}{5-3}}$$

$$= \sqrt{\frac{136.81}{2}} = \sqrt{68.405} = 8.27$$

which is the standard deviation of the multiple regression.

The coefficient of multiple determination is

$$\begin{aligned}
 R_{r,12}^2 &= \frac{\sum(\bar{Y} - \hat{Y})^2}{\sum(Y - \bar{Y})^2} \\
 &= \frac{n \sum Y + b_1 \sum X_1 Y + b_2 \sum X_2 Y - (\sum Y)^2 / n}{\sum Y^2 - (\sum Y)^2 / n} \\
 &= \frac{(-1.33)(89) + (0.38)(619) + (1.02)(1007) - (89)^2 / 5}{1885 - (89)^2 / 5} \\
 &= \frac{163.99}{300.80} = 0.55
 \end{aligned}$$

This means that 55% of the variability in income is explained by its linear relationship with floor area and the number of employees.

The co-efficient of multiple correlation, $R_{r,12}$ is

$$R_{r,12} = \sqrt{0.55} = 0.74.$$

11.2.4 Subscript Notation. For the purposes of generalization and change of variables, it is convenient to adopt a notation due to G. Udny Yule (1873-1951). This notation involves subscripts. For example, the individual Y value in case of the multiple linear regression with two independent variables, is denoted as

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Using Yule's notation, this can be written as

$$X_{i1} = \beta_{1,23} + \beta_{2,12} X_2 + \beta_{3,12} X_3 + \epsilon_i$$

the variables are numbered 1, 2 and 3 by the use of subscripts. The subscripted number 1 denotes dependent variable, 2 and 3 denote the independent variables X_1 and X_2 respectively, and $\beta_{1,23}$ is the value of X_1 when X_2 and X_3 are both equal to zero.

There are three subscripts attached to each parameter. The subscripts preceding the point are called primary subscripts and those following the point are known as secondary subscripts. The dependent variable is always indicated by the first primary subscripts, while the second primary subscript indicates to which the β co-efficient is attached. The secondary subscript(s) indicates which other variable(s) has been included in the regression equation. The secondary subscript, if more than one, may appear in any order.

The advantage of this notation is that it indicates the number of variables involved in the regression equation and also shows which is the dependent variable and which are the independent variables.

The estimated multiple regression equation of \hat{Y}_1 on X_1 and X_2 is

$$\hat{Y}_1 = b_{1,23} + \beta_{1,12} X_1 + \beta_{2,12} X_2$$

It should be noted that in general $b_{1,2,3}$ is different from $b_{1,2,3}$.

Allowing a change of variables, the estimated regression equation of X_2 on X_1 and X_3 is given by

$$\hat{X}_2 = b_{2,13}X_3 + b_{2,1,3}X_1.$$

Similarly, the estimated regression equation of X_3 on X_1 and X_2 is

$$X_3 = b_{3,12} + b_{3,1,2}X_1 + b_{3,2}X_2.$$

If the two variables, measured from their means, be x_1 and x_2 then the two simple regression equations of x_1 on x_2 and of x_2 on x_1 are

$$x_1 = b_{1,2}x_2 \text{ and } x_2 = b_{2,1}x_1.$$

The residuals may be expressed as

$$r_{1,2} = x_1 - b_{1,2}x_2 \text{ and } r_{2,1} = x_2 - b_{2,1}x_1.$$

If x_1 , x_2 and x_3 are three variables, measured from their respective means, then the regression equation of x_1 on x_2 and x_3 is

$$x_1 = b_{1,2,3}x_3 + b_{1,2,3}x_2,$$

and its residual is expressed by

$$r_{1,2,3} = x_1 - b_{1,2,3}x_3 - b_{1,2,3}x_2.$$

The two normal equations may be written as

$$\sum x_j x_{1,23} = 0 \text{ and } \sum x_j r_{1,23} = 0.$$

11.2.5 Properties of Residuals. The residuals or errors have the following properties

1. "The sum of the products of corresponding values of a variable and a residual is zero, provided the subscript of the variable is included among the secondary subscripts of the residual."

Let the regression equation (in deviation form) of x_1 on x_2 and x_3 be

$$x_1 = b_{1,2,3}x_3 + b_{1,2,3}x_2.$$

Then the two normal equations for determining the b 's are:

$$\sum x_j x_{1,23} = \sum x_j r_{1,23},$$

where $x_{1,23} = x_1 - b_{1,2,3}x_3 - b_{1,2,3}x_2$.

Similarly, the normal equations for the regression of x_2 on x_1 and x_3 and of x_3 on x_1 and x_2 are

$$\sum x_j x_{2,13} = 0 = \sum x_j r_{2,13},$$

$$\sum x_j x_{3,12} = 0 = \sum x_j r_{3,12}.$$

2. The sum of the products (or covariance) of two residuals remains unchanged by omitting by one residual any or all of secondary subscripts which are common to both".

Let the residual defined as $x_{1,2} = x_1 - b_{1,2}x_2$ be considered.

- * Then $\sum x_{1,2} x_{1,23} = \sum x_{1,23} (x_1 - b_{1,2}x_2)$
 $= \sum x_1 x_{1,23} - b_{1,2} \sum x_2 x_{1,23}$

The second term vanishes as $\sum x_1 x_{1,23} = 0$

$$\text{Then } \sum x_{1,2} x_{1,23} = \sum x_1 x_{1,23}$$

$$\begin{aligned}\text{Again } \sum x_{1,2} x_{1,23} &= \sum x_{1,23} (x_1 - b_{123} x_2 - b_{132} x_3) \\ &= \sum x_1 x_{1,23} - b_{123} \sum x_2 x_{1,23} - b_{132} \sum x_3 x_{1,23}\end{aligned}$$

Here again the second and third terms vanish due to their being normal equations.

$$\text{Hence } \sum x_{1,2} x_{1,23} = \sum x_1 x_{1,23} -$$

1. "The sum of the products (or covariance) of two residuals is zero provided all the subscripts of one residual are included among the secondary subscripts of the second."

Let us consider the residuals defined by $x_{1,2}$ and $x_{1,23}$.

$$\text{Then } \sum x_{1,2} x_{1,23} = \sum x_{1,2} (x_1 - b_{112} x_2 - b_{122} x_3)$$

But this vanishes because of normal equation and property 1.

$$\text{Similarly, } \sum x_{1,23} x_{1,23} = 0.$$

11.2.4. Multiple Regression in terms of Linear Correlation Coefficients. The multiple regression equation of a variable, say X_1 , on other variables, say X_2 and X_3 , can be sometimes expressed in terms of r_{12} , r_{13} and r_{23} , the linear correlation coefficients. The simple regression equation (in deviation form) of x_1 on x_2 and x_3 is given by

$$x_1 = b_{123} x_2 + b_{132} x_3$$

The two normal equations are obtained as

$$\sum x_1 x_2 - b_{123} \sum x_2^2 + b_{132} \sum x_3^2$$

$$\sum x_1 x_3 - b_{123} \sum x_2 x_3 + b_{132} \sum x_2^2$$

Let S_i^2 be the variance of x_i , let r_{ij} be the linear correlation co-efficient between x_i and x_j . Then using the normal equations in terms of variances and linear correlation co-efficient, we get

$$nr_{12} S_1 S_2 = n b_{123} S_2^2 + n b_{132} r_{23} S_1 S_3$$

$$nr_{13} S_1 S_3 = n b_{123} r_{12} S_2 S_3 + n b_{132} S_2^2$$

Simplification gives

$$r_{12} S_1 = b_{123} S_2 + b_{132} r_{12} S_3, \text{ and}$$

$$r_{13} S_1 = b_{123} r_{12} S_2 + b_{132} S_3$$

Solving these equations simultaneously for b 's, we get

$$b_{123} = \frac{S_1}{S_2} \left(\frac{r_{12} - r_{13} r_{23}}{1 - r_{12}^2} \right) \text{ and}$$

$$b_{1,2,3} = \frac{S_1}{S_3} \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right)$$

Substituting these values in the regression equation, we obtain

$$x_1 = \left(\frac{S_1}{S_2} \right) \left(\frac{r_{12} - r_{13}r_{21}}{1 - r_{21}^2} \right) x_2 + \left(\frac{S_1}{S_3} \right) \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) x_3$$

Or dividing both sides of the equation by S_1 , we get

$$\frac{x_1}{S_1} = \left(\frac{r_{12} - r_{13}r_{21}}{1 - r_{21}^2} \right) \left(\frac{x_2}{S_2} \right) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) \left(\frac{x_3}{S_3} \right)$$

as the multiple regression of x_1 on x_2 and x_3 in terms of standard deviations and the linear combined co-efficients of the variables involved. Similarly, the other two multiple regression equations of x_2 on x_1 and x_3 and of x_3 on x_1 and x_2 are obtained as

$$\frac{x_2}{S_2} = \left(\frac{r_{12} - r_{13}r_{21}}{1 - r_{21}^2} \right) \left(\frac{x_1}{S_1} \right) + \left(\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2} \right) \left(\frac{x_3}{S_3} \right), \text{ and}$$

$$\frac{x_3}{S_3} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) \left(\frac{x_1}{S_1} \right) + \left(\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2} \right) \left(\frac{x_2}{S_2} \right)$$

To obtain the regression equations in terms of original values, we replace x_1 by $X_1 - \bar{X}_1$, x_2 by $X_2 - \bar{X}_2$, and x_3 by $X_3 - \bar{X}_3$, respectively.

11.3 MULTIPLE CORRELATION CO-EFFICIENT

(The co-efficient of multiple correlation measures the degree of relationship between a variable and its estimate from the regression equation) In other words, it is a product moment correlation between a variable, say x_1 , and its value expressed by the regression equation $x_1 = b_{12,3}x_2 + b_{13,2}x_3$. The co-efficient of multiple correlation between x_1 and the variables x_2 and x_3 combined, is denoted symbolically by $R_{1,2,3}$.

Let us denote the estimated value of x_1 by \hat{x}_1 . Then by definition,

$$R_{1,2,3} = \frac{\text{Cov}(x_1, \hat{x}_1)}{\sqrt{\text{Var}(x_1) \text{Var}(\hat{x}_1)}} = \frac{\sum x_1 \hat{x}_1}{\sqrt{\sum x_1^2 \sum (\hat{x}_1)^2}}$$

$$\begin{aligned} \text{Now } \sum x_1 \hat{x}_1 &= \sum x_1 (x_1 - x_{1,23}) && (\because x_1 = x_{1,23} - x_{1,23}) \\ &= \sum x_1^2 - \sum x_1 x_{1,23} \\ &= \sum x_1^2 - \sum x_{1,23} \bar{x}_{1,23} && (\because \sum x_1 x_{1,23} = \sum x_{1,23} \bar{x}_{1,23}) \\ &= n(S_1^2 - S_{1,23}^2). \end{aligned}$$

where $S_{e,23}^2$ is the sample variance of residuals.

$$\text{Also } \sum(x_i)^2 = \sum(x_i - x_{1,23})^2 \\ = \sum x_1^2 + \sum x_{1,23}^2 - 2\sum x_1 x_{1,23} \\ = \sum x_1^2 + \sum x_{1,23}^2 - 2\sum x_{1,23} \\ = \sum x_1^2 + \sum x_{1,23}^2 - n(S_1^2 - S_{1,23}^2) \quad (\because \sum x_1 x_{1,23} = \sum x_{1,23} x_{1,23}) \\ = \sum x_1^2 + \sum x_{1,23}^2 - n(S_1^2 - S_{1,23}^2).$$

and $\sum x_1^2 = nS_1^2$

Substituting these values in the formula, we get

$$R_{1,23} = \frac{S_1^2 - S_{1,23}^2}{S_1 \sqrt{S_1^2 - S_{1,23}^2}} = \left(1 - \frac{S_{1,23}^2}{S_1^2} \right)^{1/2}$$

Similarly, we get $R_{2,31}^2 = 1 - \frac{S_{2,31}^2}{S_2^2}$.

The quantity $S_{1,23}^2$ can be expressed in terms of the simple correlation co-efficients between the pairs of variables as below:

$$S_{1,23}^2 = \frac{1}{n} \sum x_{1,23}^2 = \frac{1}{n} \sum (x_1 - b_{1,23}x_2 - b_{1,12}x_3)^2 \\ = \frac{1}{n} \sum x_1(x_1 - b_{1,23}x_2 - b_{1,12}x_3) \\ = \frac{1}{n} \sum x_1^2 - b_{1,23}^2 \sum x_1 x_2 - b_{1,12}^2 \sum x_1 x_3 \\ = S_1^2 - b_{1,23}^2 S_2 r_{12} - b_{1,12}^2 S_3 r_{13}$$

(second property of residuals)

$$= S_1^2 - b_{1,23}^2 S_2 r_{12} - b_{1,12}^2 S_3 r_{13}$$

Substituting the values of $b_{1,23}$ and $b_{1,12}$ in terms of simple correlation co-efficient and simplifying, we get

$$S_{1,23}^2 = S_1^2 \left[\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{23}r_{13}}{1 - r_{23}^2} \right]$$

$$\text{Hence: } R_{1,23}^2 = 1 - \frac{S_1^2 (1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{23}r_{13})}{S_1^2 (1 - r_{23}^2)}$$

$$= \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}, \text{ so that}$$

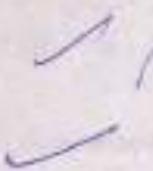
$$R_{1,23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

It should be noted that $R_{1,2,3}$ is necessarily positive or zero as the term $\sum z_i t_i$, being equal to Σz_i^2 , cannot be negative. If $R_{1,2,3} = 1$, the $S_{1,2,3}^2 = 0$, i.e. all the residuals $r_{1,2,3}$ are zero; the observed and estimated values of x_1 coincide. The multiple correlation in this case, is called perfect, indicating a linear relationship between the variables.

Similarly, by the change of variables, we get

$$R_{1,2,3} = \sqrt{\frac{r_{11}^2 + r_{22}^2 - 2r_{12}r_{11}r_{22}}{1 - r_{11}^2}}, \text{ and}$$

$$R_{1,3,2} = \sqrt{\frac{r_{11}^2 + r_{33}^2 - 2r_{13}r_{11}r_{33}}{1 - r_{11}^2}}.$$



Example 11.4 An instructor of mathematics wished to determine the relationship of grades on final examination to grades on two quizzes given during the semester. Calling X_1 , X_2 and X_3 the grades of a student on the first quiz, second quiz and final examination respectively, he made the following computations for a total of 120 students.

$$\bar{X}_1 = 6.8 \quad S_1 = 1.0 \quad r_{12} = 0.60$$

$$\bar{X}_2 = 7.0 \quad S_2 = 0.8 \quad r_{13} = 0.75$$

$$\bar{X}_3 = 74 \quad S_3 = 9.0 \quad r_{23} = 0.05$$

- Find the least-squares regression equation of X_3 on X_1 and X_2 .
- Estimate the final grades of two students who scored respectively (1) 9 and 7, and (2) 4 and 6 on the two quizzes.
- Compute $R_{1,2,3}$. (B.Sc. Engg.)
- Since the standard deviations and linear correlation co-efficients are given, therefore the estimated regression equation of X_3 on X_1 and X_2 is

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

Now $\frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2} = \frac{0.75 - (0.60)(0.05)}{1 - (0.60)^2} = \frac{0.31}{0.64}, \text{ and}$

$$\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2} = \frac{0.05 - (0.60)(0.75)}{1 - (0.60)^2} = \frac{0.23}{0.64}.$$

Substituting these values, we get

$$\frac{X_3 - 74}{9.0} = \left(\frac{0.31}{0.64} \right) \left(\frac{X_1 - 6.8}{1.0} \right) + \left(\frac{0.23}{0.64} \right) \left(\frac{X_2 - 7.0}{0.8} \right)$$

or $X_3 - 74 = 4.36(X_1 - 6.8) + 4.04(X_2 - 7.0)$
 $= 4.36 X_1 - 29.648 + 4.04 X_2 - 28.28$

$$\text{e} \quad \hat{X}_1 = 16.07 + 4.36 X_1 + 4.04 X_2$$

at the desired least squares regression equation of \hat{X}_1 on X_1 and X_2 .

b) Student 1: When $X_1 = 9$ and $X_2 = 7$, we get

$$\hat{X}_1 = 16.07 + 4.36(9) + 4.04(7) = 83.59 \approx 84$$

Student 2: When $X_1 = 4$ and $X_2 = 8$, we get

$$\hat{X}_1 = 16.07 + 4.36(4) + 4.04(8) = 65.83 \approx 66.$$

c) The co-efficient of multiple correlation $R_{1,12}$ is

$$\begin{aligned} R_{1,12} &= \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}}{1 - r_{12}^2}}, \\ &= \sqrt{\frac{(0.70)^2 + (0.65)^2 - 2(0.60)(0.65)(0.70)}{1 - (0.60)^2}} \\ &= \sqrt{\frac{0.3665}{0.64}} = \sqrt{0.5727} = 0.757. \end{aligned}$$

Relationship b/w any two variables by neglecting the effect of 3rd variable.

4 PARTIAL CORRELATION

A partial correlation measures the degree of linear relationship between any two variables in a variable problem under the condition that no common relationship or influence with all other variables (or some of them) has been removed. And differently, if there are three variables X_1 , X_2 and X_3 , then the correlation between X_1 and X_2 after removing the linear effect of X_3 from X_1 and from X_2 is partial correlation. The sample co-efficient of partial correlation measuring the strength of the relationship (correlation) between X_1 and X_2 , when the influence of X_3 has been removed, is denoted statistically by $r_{12,3}$. By removing its influence, we mean subtracting the fixed regression \hat{X}_1 from the observed values X_1 , obtaining the residual – a part of X_1 not explained by X_3 .

To derive the co-efficient of partial correlation $r_{12,3}$, we use the variables x_1 , x_2 and x_3 which are scores from their means. The linear regressions of x_1 on x_3 and of x_2 on x_3 are $x_1 = b_{13}x_3$ and $x_2 = b_{23}x_3$. Neglecting the linear effect of x_3 from x_1 and from x_2 and denoting the residuals by x_{13} and x_{23} , we get

$$x_{13} = x_1 - b_{13}x_3, \text{ and } x_{23} = x_2 - b_{23}x_3.$$

These residuals may be written as

$$x_{13} = x_1 - r_{13} \frac{S_1}{S_3} x_3, \text{ and } x_{23} = x_2 - r_{23} \frac{S_2}{S_3} x_3.$$

Now the co-efficient of partial correlation is the product moment correlation co-efficient between x_{13} and x_{23} . Thus by definition

$$r_{12,3} = \frac{\sum x_{13}x_{23}}{\sqrt{\sum x_{13}^2 \sum x_{23}^2}}.$$

$$\begin{aligned}
 \text{Now } \sum x_{13}x_{13} &= \sum \left[x_1 - r_{12} \frac{S_1}{S_2} x_2 \right] \left[x_1 - r_{12} \frac{S_1}{S_2} x_2 \right] \\
 &= \sum \left[x_1 x_2 - r_{12} \frac{S_1}{S_2} x_1 x_2 - r_{12} \frac{S_1}{S_2} x_1 x_2 + r_{12}^2 \frac{S_1^2}{S_2^2} x_2^2 \right] \\
 &= \sum x_1 x_2 - r_{12} \frac{S_1}{S_2} \sum x_1 x_2 - r_{12} \frac{S_1}{S_2} \sum x_1 x_2 + r_{12} r_{23} \frac{S_1 S_2}{S_2^2} \sum x_2^2 \\
 &= n [r_{12} S_1 S_2 - r_{12} r_{13} S_1 S_2 - r_{12} r_{23} S_1 S_2 + r_{12} r_{23} S_1 S_2] \\
 &= n S_1 S_2 (r_{12} - r_{13} r_{23})
 \end{aligned}$$

$$\begin{aligned}
 \text{And } \sum x_{13}^2 &= \sum \left[x_1 - r_{12} \frac{S_1}{S_2} x_2 \right]^2 \\
 &= \sum x_1^2 - r_{12}^2 \frac{S_1^2}{S_2^2} \sum x_2^2 - 2r_{12} \frac{S_1}{S_2} \sum x_1 x_2 \\
 &= n(S_1^2 + r_{12}^2 S_2^2 - 2r_{12}^2 S_1^2) \\
 &= nS_1^2(1 - r_{12}^2).
 \end{aligned}$$

$$\text{Similarly, } \sum x_{23}^2 = nS_2^2(1 - r_{23}^2).$$

Substituting these values in the formula, we obtain

$$r_{12,3} = \frac{S_1 S_2 (r_{12} - r_{13} r_{23})}{S_1 S_2 \sqrt{r_{12}^2 (1 - r_{23}^2)}} \quad \checkmark$$

$$\rho_{12,3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} \quad \checkmark$$

Alternatively, The partial correlation co-efficient between x_1 and x_3 when the influence of x_2 has been eliminated, is also defined as the geometric mean of the regression co-efficients $b_{12,3}$ and $b_{21,3}$, i.e., two partial regression lines of x_1 on x_2 and of x_2 on x_1 , respectively, i.e.

$$\begin{aligned}
 r_{12,3} &= \sqrt{b_{12,3} \times b_{21,3}} \\
 &= \sqrt{\frac{S_1}{S_2} \left(\frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \right)} \sqrt{\frac{S_2}{S_1} \left(\frac{r_{12} - r_{13} r_{23}}{1 - r_{12}^2} \right)} \\
 &= \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}
 \end{aligned}$$

$(r_{12,3}$ has the same sign as $b_{12,3}$ and $b_{21,3}$)

In a similar way, we can prove that

$$r_{12} = \frac{r_{13} - r_{11}r_{33}}{\sqrt{1 - r_{11}^2}\sqrt{1 - r_{33}^2}}, \text{ and } r_{23} = \frac{r_{13} - r_{11}r_{33}}{\sqrt{1 - r_{11}^2}\sqrt{1 - r_{33}^2}}.$$

Example 11.5 From the following data, determine the linear regression equations of X_1 on X_2 and of X_2 on X_3 .

X_1	7	12	14	17	20
X_2	4	7	8	9	12
X_3	1	2	4	5	8

Find the deviations of observed values of X_1 from the regression, viz., \bar{X}_{13} . Repeat the same for X_2 , obtain \bar{X}_{23} . Determine the simple correlation co-efficient between the two sets of deviations X_1 and X_2 . (P.U., B.A. B.Sc. 1977)

The estimated simple regression equation of X_1 on X_3 is

$$\bar{X}_1 = b_{13} + b_{23} \bar{X}_3,$$

$$b_{13} = \frac{n \sum X_1 X_3 - (\sum X_1)(\sum X_3)}{n \sum X_3^2 - (\sum X_3)^2} \quad \text{and} \quad b_{23} = \bar{X}_2 - b_{13} \bar{X}_3.$$

The estimated simple regression equation of X_2 on X_3 is

$$\bar{X}_2 = b_{23} + b_{13} \bar{X}_3,$$

$$b_{23} = \frac{n \sum X_2 X_3 - (\sum X_2)(\sum X_3)}{n \sum X_3^2 - (\sum X_3)^2} \quad \text{and} \quad b_{13} = \bar{X}_1 - b_{23} \bar{X}_3.$$

Computations needed to find b 's are given in the table below:

X_1	X_2	X_3	$X_1 X_3$	$X_2 X_3$	X_1^2
7	4	1	7	4	1
12	7	2	24	14	4
14	8	4	56	32	16
17	9	5	85	45	25
20	12	8	160	96	64
70	40	20	332	191	110

$$\bar{X}_1 = \frac{\sum X_1}{n} = \frac{70}{5} = 14, \bar{X}_2 = \frac{\sum X_2}{n} = \frac{40}{5} = 8 \text{ and } \bar{X}_3 = 4.$$

And the regression co-efficients are obtained as

$$b_{11} = \frac{(5)(332) - (70)(20)}{(5)(110) - (20)^2} = \frac{260}{150} = 1.73,$$

$$b_{12} = 14 - (1.73)(4) = 7.08,$$

$$b_{21} = \frac{(5)(191) - (40)(20)}{(5)(110) - (20)^2} = \frac{155}{150} = 1.03, \text{ and}$$

$$b_{22} = 8 - (1.03)(4) = 3.88.$$

Hence the desired regression equations are

$$\hat{X}_1 = 7.08 + 1.73X_1 \text{ and } \hat{X}_2 = 3.88 + 1.03X_2.$$

Next, we compute the residuals $X_{13} = X_1 - 7.08 - 1.73X_2$ and $X_{23} = X_2 - 3.88 - 1.03X_1$, and the correlation between them. The necessary computations are given in the following table:

X_1	\hat{X}_1	X_2	X_{12}	X_{21}	$X_{13}X_{23}$	X_{13}^2	X_{23}^2
7	4	1	-1.81	-0.91	1	3.2761	0.8281
12	7	2	1.46	1.06	-0.476	2.1316	1.1236
14	8	4	0	0	0	0	0
17	9	5	1.27	0.03	0.0381	1.6129	0.0009
20	12	8	-0.92	-0.12	0.1104	0.8464	0.0144
70	40	20		0	3.2670	7.8670	1.9670

Hence the co-efficient of correlation between X_{13} and X_{23} , which is the co-efficient of partial correlation between X_1 and X_2 when the influence of X_3 has been removed, is obtained as

$$r_{123} = \frac{r_{12} - r_{11}r_{22}}{\sqrt{(\sum X_{13}^2)(\sum X_{23}^2)}} \quad (\because \sum X_{13} = \sum X_{23} = 0)$$

$$= \frac{3.2670}{\sqrt{(7.8670)(1.9670)}} = \frac{3.2670}{3.9340} = 0.83$$

Example 11.6 Given $r_{12} = 0.492$, $r_{13} = 0.927$ and $r_{23} = 0.758$, find all the partial co-efficients.

$$\begin{aligned} \text{We have } r_{123} &= \frac{r_{12} - r_{11}r_{22}}{\sqrt{1 - r_{11}^2}\sqrt{1 - r_{22}^2}} = \frac{0.492 - (0.927)(0.758)}{\sqrt{1 - (0.927)^2}\sqrt{1 - (0.758)^2}} \\ &= \frac{-0.2107}{\sqrt{0.1407 \times 0.4254}} = -0.86 \end{aligned}$$

$$r_{311} = \frac{r_{31} - r_{32}r_{13}}{\sqrt{1-r_{31}^2}\sqrt{1-r_{13}^2}} = \frac{0.758 - (0.492)(0.927)}{\sqrt{1-(0.492)^2}\sqrt{1-(0.927)^2}}$$

$$= \frac{0.302}{\sqrt{0.7579 \times 0.1407}} = 0.92; \text{ and}$$

$$r_{312} = \frac{r_{31} - r_{32}r_{13}}{\sqrt{1-r_{32}^2}\sqrt{1-r_{13}^2}} = \frac{0.927 - (0.758)(0.492)}{\sqrt{1-(0.758)^2}\sqrt{1-(0.492)^2}}$$

$$= \frac{0.5541}{\sqrt{0.4254 \times 0.7579}} = 0.98.$$

Example 11.7. Show that if $x_1 = ax_2 + bx_3$, the three partial correlations are numerically equal to having the sign of a , r_{112} the sign of b and r_{113} the opposite sign of $a.b$.

In the multiple regression equation $x_1 = ax_2 + bx_3$, we treat x_1 as dependent and x_2 and x_3 as independent variables. Let the three variables be measured from their respective means.

Ignoring and summing over all values, we get

$$\begin{aligned} \sum x_1^2 &= a^2 \sum x_2^2 + b^2 \sum x_3^2 && (\text{the product of values as } x_1 \text{ and } x_2 \text{ are independent}) \\ &= n(a^2 S_1^2 + b^2 S_2^2) \end{aligned}$$

Multiplying the given equation by x_1 and summing, we have

$$\begin{aligned} \sum x_1 x_1 &= a \sum x_1^2 && (\sum x_1 x_2 = 0, \text{ as } x_1 \text{ and } x_2 \text{ are independent}) \\ &= n a S_1^2 \end{aligned}$$

$$\begin{aligned} r_{112} &= \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \sum x_2^2}} = \frac{a S_1}{\sqrt{S_1^2 (a^2 S_1^2 + b^2 S_2^2)}} \\ &= \frac{a S_1}{\sqrt{a^2 S_1^2 + b^2 S_2^2}} = \frac{a S_1}{w}, \text{ where } w^2 = a^2 S_1^2 + b^2 S_2^2. \end{aligned}$$

Similarly, $r_{113} = \frac{b S_1}{w}$ and $r_{123} = 0$.

$$\text{Since } r_{113} = \frac{r_{11} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}} = \frac{\frac{a S_1}{w} - 0}{\sqrt{(1-0)\left(1-\frac{b^2 S_2^2}{w^2}\right)}}$$

$$= \frac{a}{\sqrt{w^2}} = \pm 1, \text{ according as } a \text{ is } +ve \text{ or } -ve.$$

In other words, $r_{12,3}$ has the sign of a .

$$\text{Again, } r_{12,3} = \frac{r_{12} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}} = \frac{0 - \frac{ab}{w} \frac{bS_2}{w}}{\sqrt{1 - \frac{a^2S_1^2}{w^2}} \sqrt{1 - \frac{b^2S_2^2}{w^2}}} \\ = \frac{-abS_2}{\sqrt{a^2b^2S_1^2S_2^2}} = \frac{-ab}{\sqrt{a^2b^2}}.$$

$\therefore a^2b^2$ is always positive, therefore, $\sqrt{a^2b^2}$ is always positive.

Now ab may be positive or negative.

Thus $r_{12,3}$ has the sign opposite to ab or $\frac{a}{b}$.

$$\text{Similarly, } r_{31,2} = \frac{r_{31} - r_{31}r_{21}}{\sqrt{(1-r_{31}^2)(1-r_{21}^2)}} \\ r_{31,2} = \frac{\frac{bS_1}{w} - 0}{\sqrt{1 - \frac{a^2S_1^2}{w^2}}} \\ = \frac{b}{\sqrt{1 - \frac{a^2S_1^2}{w^2}}}, \text{ according as } b \text{ is +ve or -ve.}$$

Hence the result.

11.4.1 Relationship between Multiple and Partial Correlation Co-efficients: correlation co-efficients can be connected with the various partial correlation co-efficients. For we have shown earlier that

$$1 - R_{12,3}^2 = \frac{S_{121}^2}{S_1^2},$$

$$\begin{aligned} \text{where } nS_{121}^2 &= \sum r_{121} - \sum r_{11,3}r_{21,3} && \text{(second property of residuals)} \\ &= \sum r_{121}(x_3 - b_{11,3}x_1 - b_{21,3}x_2) \\ &= \sum r_{121}^2 - b_{11,3}^2 \sum r_{11,3}^2 - b_{21,3}^2 \sum r_{21,3}^2 && \text{because of the properties of residuals.} \\ &= \sum r_{121}^2 \left[1 - b_{11,3}^2 \frac{\sum r_{11,3}^2}{\sum r_{121}^2} \right] \\ &= nS_{12}^2 [1 - b_{11,3}b_{21,3}] = nS_{12}^2 (1 - r_{12,3}^2) \end{aligned}$$

$$\begin{aligned} S_{123}^2 &= S_{12}^2(1 - r_{123}^2) \\ &= S_1^2(1 - r_{12}^2)(1 - r_{123}^2) \\ \therefore 1 - R_{123}^2 &= (1 - r_{12}^2)(1 - r_{123}^2). \end{aligned}$$

Using in the same way, we can find

$$1 - R_{1234}^2 = (1 - r_{12}^2)(1 - r_{123}^2)(1 - r_{1234}^2).$$

CURVILINEAR REGRESSION

Sometimes a scatter diagram indicates that the relationship between the two variables will be more easily described by a non-linear regression line. When this occurs, either we may transform one or both of the variables so that the transformed data appear approximately linear or we may use a general equation. In the former case, the estimating equation may be an exponential or a logarithmic curve. In the latter case, the estimating equation may be

$$\hat{Y} = a + bX + cX^2.$$

Here a and c are the least-squares estimates of the population parameters in

$$E(Y) = a + bX + cX^2.$$

They are determined from the following set of normal equations:

$$\sum Y = a n + b \sum X + c \sum X^2$$

$$\sum XY = a \sum X + b \sum X^2 + c \sum X^3$$

$$\sum X^2 Y = a \sum X^2 + b \sum X^3 + c \sum X^4$$

The quadratic equation may also be changed into a multiple linear form

$$\hat{X}_1 = a_{123} + b_{123} X_1 + b_{123} X_{12},$$

—

$\hat{X}_1 = \hat{Y}$, $a_{123} = a$, $b_{123} = b_1$, $b_{123} = c$, $X_1 = X$ and $X_{12} = X^2$. A number of other curvilinear forms are available. The co-efficient of determination and standard error of estimate can be obtained in the same way as in the case of linear regressions.

EXERCISES

TRUE/FALSE

State 'True' or 'False'. If the statement is not true then replace the underlined words with words which make the statement true:

A partial correlation coefficient measures the degree of relationship between a variable and its deviation from the regression line.

- ii) A multiple correlation coefficient measures the degree of linear relationship between all variables in a multivariable problem when the influence with all other variables has been removed.
- iii) The multiple correlation coefficient is the square of the coefficient of multiple determination.
- iv) The multiple correlation coefficient R^2 will be negative in sign when all of the two variable correlation coefficients are negative in sign.
- v) The regression sum of squares in case of multiple regression is the explained variation.
- vi) For a multiple regression analysis, if $\sum(Y - \bar{Y})^2 = 50$ and $\sum(Y - \hat{Y})^2 = 20$, then the multiple coefficient of determination R^2 is equal to 0.20.
- vii) The standard error of estimate in multiple regression has $n-k$ degrees of freedom.
- viii) The standard error of estimate is a measure of scatter of the observations about the regression line.
- ix) The regression coefficients are the other name for multiple regression coefficients.
- x) In a multiple regression the addition of new variables will always reduce the standard error estimate.

b) MULTIPLE CHOICE QUESTIONS

- i) The range of multiple correlation coefficient is
 - a) -1 to +1
 - b) 0 to + ∞
 - c) 0 to 1
 - d) none of above
- ii) The range of partial correlation coefficient is
 - a) 0 to 1
 - b) -1 to +1
 - c) 0 to ∞
 - d) -1 to 0
- iii) If the multiple correlation coefficient $R_{k,n} = 1$, then it implies a
 - a) perfect relationship
 - b) high relationship
 - c) weak linear relationship
 - d) perfect linear relationship

- i) In the regression analysis, the explained variation of the dependent variable Y is given by
- $\sum(Y - \bar{Y})^2$
 - $\sum(Y - \hat{Y})^2$
 - $\sum(\hat{Y} - \bar{Y})^2$
 - $\sum(Y - \hat{Y})$
- ii) Which of the following is not a standard deviation?
- Standard error of the slope coefficient
 - Mean square errors
 - Standard error of estimator
 - Standard deviation of the Y variable
- iii) The coefficient of determination in multiple regression is given by
- $R_{Y,12}^2 = 1 - (SST / SSE)$
 - $R_{Y,12}^2 = 1 - (SSR / SST)$
 - $R_{Y,12}^2 = 1 - (SSE / SSR)$
 - $R_{Y,12}^2 = 1 - (SSE / SST)$
- iv) The slope b_1 in the multiple regression equation $\hat{Y} = b_0 + b_1X_1 + b_2X_2$ measures
- the amount of variation in \hat{Y} explained by X_1
 - the change in \hat{Y} per unit change in X_1
 - the change in \hat{Y} per unit change in X_1 , holding X_2 constant
 - the change in \hat{Y} per unit change in X_2 , holding X_1 constant
- v) The predicted value of \hat{Y} for $X_1 = 1$, $X_2 = 5$, and $X_3 = 10$ by using the regression line $\hat{Y} = 30 - 10X_1 + 18X_2 - 7.5X_3$ is
- 45
 - 15
 - 35
 - 50
- vi) Which of the following statements remains always true?
- The coefficient of multiple determination will increase when new variables are added
 - The coefficient of multiple determination will decrease when new variables are added
 - The adjusted coefficient of multiple determination will not decrease when new variables are added
 - Both a and c above

x) Which of the following relationship holds?

- a) $r_{12,3} = \sqrt{b_{12,2} \times b_{21,2}}$
- b) $r_{12,3} = \sqrt{b_{12,2} \times b_{31,2}}$
- c) $r_{12,3} = \sqrt{b_{21,2} \times b_{31,2}}$
- d) All of above

SUBJECTIVE

- 11.1 a) What is a multiple regression? Explain the basic differences between simple regression and multiple regression.
 b) What is meant by the co-efficient of multiple determination and multiple correlation?
 c) Explain the assumptions underlying a multiple linear regression model.

- 11.2 Carryout the necessary computations to obtain the least-squares estimates of the parameters in the multiple regression model $\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, given

\bar{Y}	12	10	9	13	7
X_1	2	2	3	4	4
X_2	1	1	0	1	1

(B.Z.U., M.A. Econ. 1990)

- 11.3 Given the data

\bar{Y}	2	7	8	5
X_1	6	8	6	5
X_2	1	1	3	4

(B.Z.U., M.A. Econ. 1990)

- 11.4 Given the following data

X_1	1	4	1	3	2	4
X_2	1	8	3	5	6	10
X_3	2	8	1	7	4	6

- a) Find the least-squares regression line where X_1 is the dependent variable and X_2 and X_3 are independent variables.
- b) Calculate the standard error of estimate, $s_{e,12}$.
- c) Calculate the co-efficient of multiple determination and multiple correlation and interpret the result.

The following table shows the corresponding values of three variables X_1 , X_2 and X_3 .

X_1	3	5	6	8	12	14
X_2	16	10	7	4	2	2
X_3	90	72	54	42	30	12

a) Find the regression equation of X_3 on X_1 and X_2 .

b) Estimate X_3 when $X_1 = 10$ and $X_2 = 6$.

c) Compute $R_{3,12}$ and $r_{3,12}$.

(I.I.T., M.Sc. 1991)

The following data were collected to determine a suitable regression equation relating the length of an infant, L (cm), to age, X_1 (days), and weight at birth, X_2 (kg):

L	57.5	52.8	61.3	67.8	53.5	62.7	56.2	58.5	69.2
X_1	78	69	77	88	67	80	74	94	102
X_2	2.75	2.15	4.41	5.52	3.21	4.32	2.31	4.30	3.71

d) Fit a least-squares regression equation of the form

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

Predict the average length of infants who are 75 days old and weighed 3.15 kg at birth.

Calculate the standard error of estimate s_e .

Define the multiple correlation co-efficient and prove that

$$R_{(2)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}}{1 - r_{12}^2}$$

Calculate the multiple correlation co-efficient $R_{1,23}$ of X_1 on X_2 and X_3 from the following data:

X_1	4	3	2	4	6	7	8
X_2	2	5	3	2	1	4	5
X_3	8	10	13	5	17	16	20

(P.U., B.A. (Hons.) Part-I, 1969)

The following data represent concomitant values of three variables.

X_1	32	18	52	16	42	48
X_2	3	7	5	1	4	6
X_3	2	4	2	5	3	9

Calculate all the multiple correlation coefficients, working out the usual simple correlation co-efficients.

(B.Z.U., M.A. Econ. 1991)

- b) Given $\bar{X}_1 = 20$, $S_1 = 1.0$, $r_{12} = -0.20$.

$$\bar{X}_2 = 36, \quad S_2 = 2.0, \quad r_{13} = 0.40.$$

$$\bar{X}_3 = 12, \quad S_3 = 1.5, \quad r_{23} = 0.50.$$

Find the regression equation of X_3 on X_1 and X_2 .

(P.U., B.A./B.Sc. 1987)

- 11.9 a) Distinguish between the simple and the multiple correlation co-efficients.

- b) If b_{ij} is the regression co-efficient of X_i on X_j , then calculate the multiple correlation co-efficient of X_3 with X_1 and X_2 , where

$$b_{12} = 0.75, \quad b_{13} = 0.58, \quad b_{23} = 0.88,$$

$$b_{21} = 0.53, \quad b_{31} = 1.63, \quad \text{and } b_{32} = 1.30$$

(P.U., B.A./B.Sc. 1987)

- c) Three variables have in pairs simple correlation coefficients: $r_{12} = 0.60$, $r_{13} = 0.70$, $r_{23} = 0.65$. Find the multiple correlation coefficient $R_{1,23}$ of X_3 on X_1 and X_2 .

(P.U., B.A./B.Sc. 1987)

- 11.10 a) Three variable have in pairs simple correlation coefficients given by

$$r_{12} = 0.8, \quad r_{13} = -0.7, \quad r_{23} = -0.9.$$

Find the multiple correlation co-efficient $R_{1,23}$ of X_3 on X_1 and X_2 .

(P.U., B.A./B.Sc. 1987)

- b) Calculate the multiple correlation co-efficient of X_3 on X_1 and X_2 and the partial correlation co-efficient $r_{12|3}$ from the values given below:

$$b_{13} = -0.1, \quad b_{23} = -0.4, \quad b_{33} = 0.3,$$

$$b_{31} = 0.6, \quad b_{32} = 0.67, \quad b_{12} = 0.38$$

(P.U., B.A./B.Sc. 1987)

- 11.11 a) Explain what is meant by partial correlation. Establish a formula for the co-efficient of partial correlation.

- b) From the following data, determine the linear regression equations of X_3 on X_1 and X_2 :

X_1	5	9	7	10	12	8	6	10
X_2	10	12	8	9	11	7	5	8
X_3	2	6	4	5	7	6	4	6

Find the deviations of observed values of X_1 from the regression equation, viz. \hat{X}_{11} , from the same for X_2 , i.e. obtain \hat{X}_{21} . Determine the simple correlation co-efficient between two sets of deviations \hat{X}_{11} and \hat{X}_{21} .

- 11.12 The following means, standard deviations and correlations are found for

X_1 = Seed-hay crops in cwt.s per acre,

X_2 = Spring rainfall in inches,

X_3 = Accumulated temperature above 42°F in spring in a certain district in England for years.

$$\bar{X}_1 = 28.02, \quad S_1 = 4.42, \quad r_{12} = 0.80,$$

$$\bar{X}_2 = 4.91, \quad S_2 = 1.10, \quad r_{13} = -0.40,$$

$$\bar{X}_3 = 594, \quad S_3 = 85, \quad r_{23} = -0.56.$$

Find the partial correlation and the regression equation for hay-crop on spring rainfall and accumulated temperature. (P.U., B.A./B.Sc. 1974)

- 11.13 The following values represent sample values of 450 college students in which the three variables represent marks obtained (X_1), general intelligence scores (X_2) and hours of study (X_3). Find the regression equation for estimating marks obtained. Find all three partial correlations and interpret them in the light of the corresponding simple correlations.

$$\bar{X}_1 = 18.5, \quad S_1 = 11.2, \quad r_{12} = 0.60,$$

$$\bar{X}_2 = 100.6, \quad S_2 = 15.8, \quad r_{13} = 0.32,$$

$$\bar{X}_3 = 24, \quad S_3 = 6.0, \quad r_{23} = 0.35.$$

(P.U., M.A. Stat., 1960)

- 11.14 a) Prove that a variable and a residual are uncorrelated if the subscript of the variable is included among the secondary subscripts of the residual.
 b) Given the equations of the three regression planes as

$$z_1 = 0.41 x_2 - 0.25 x_3,$$

$$z_2 = 0.56 z_1 - 0.025 x_3,$$

$$z_3 = 1.04 z_1 - 0.05 x_2.$$

Calculate the partial correlation co-efficients. Do we have sufficient data to determine the correlation co-efficients r_{23}, r_{13} and r_{12} ? (P.U., B.A. (Hons.), Part-I, 1970)

- 11.15 a) If $X_1 = a + b_{123} X_2 + b_{13} X_3$ and $X_2 = d + b_{23} X_3 + b_{21} X_1$ are the regression equations of X_1 on X_2 and X_3 , and of X_2 on X_3 and X_1 , respectively, prove that $r_{123}^2 = b_{13} \times b_{21}$.
 b) Is it possible to obtain the following from a set of data?
 (i) $r_{12} = 0.6, r_{23} = 0.8, r_{13} = -0.5$.
 (ii) $r_{23} = 0.7, r_{13} = -0.4, r_{12} = 0.6$.
 (iii) $r_{21} = 0.01, r_{13} = 0.66, r_{23} = -0.76$.

- 11.16 If X_1, X_2 and X_3 are three correlated variables, where $S_1=1, S_2=1.3, S_3=1.9$ and $r_{12}=0.370, r_{13}=-0.641$, and $r_{23}=-0.736$, find r_{123} . If $X_0=X_1+X_3$, obtain r_{10}, r_{20} and r_{30} . Verify that the two partial correlation co-efficients are equal and explain this result. (M.Sc. Stat., P.U., 1972; I.U., 1990, 92, 94)

- 11.17 a) Differentiate between multiple correlation and partial correlation.
 b) If $R_{123} = 1$, prove that (i) $R_{210} = 1$ and (ii) $R_{132} = 1$.
 c) If $R_{123} = 0$, does it necessarily follow that $R_{123} = 0$?

- d) If $r_{12} = r_{23} = r_{31} = r \neq 1$, then show that $R_{1,23} = R_{2,31} = R_{3,12} = \frac{r\sqrt{2}}{\sqrt{1+r}}$. Discuss the case when $r=1$.
 (B.Sc. Eng. 1976)

- 11.18 a) Show that if r_{12} is zero, $r_{1,23}$ will not be zero unless one at least of r_{13} and r_{23} is zero.
 b) If the relation $aX_1 + bX_2 + cX_3 = 0$ holds true for all sets of values of X_1, X_2 and X_3 , find out the three partial correlation co-efficients.

- T1.19 Show that the correlation co-efficient between the residuals $\pi_{1,23}$ and $\pi_{2,13}$ is equal and opposite to that between $x_{1,3}$ and $x_{2,3}$.
 (P.U., M.A. 1963)

Solution. The co-efficient of correlation between $x_{1,23}$ and $x_{2,13}$ is given by

$$\begin{aligned} \frac{\text{Cov}(\pi_{1,23}, \pi_{2,13})}{\sqrt{\text{Var}(\pi_{1,23}) \text{Var}(\pi_{2,13})}} &= \frac{1}{n} \frac{\sum \pi_{1,23} \pi_{2,13}}{S_{1,23} S_{2,13}} \\ &= \frac{1}{n} \frac{\sum x_{1,23}(x_1 - b_{123}x_2 - b_{312}x_3)}{S_{1,23} S_{2,13}} \\ &= \frac{1}{n} \frac{0 - b_{123} \sum x_{1,23}^2 - 0}{S_{1,23} S_{2,13}} = \frac{b_{123} S_{1,23}}{S_{2,13}} \end{aligned}$$

Substituting the values of $S_{1,23}$ and $S_{2,13}$ and simplifying, we get

$$\text{Corr.} = -b_{123} \left[\frac{S_{1,2} \sqrt{1 - r_{12}^2}}{S_1 \sqrt{1 - r_{13}^2}} \right] = -b_{123} \frac{S_{1,2}}{S_{1,3}}$$

Again the co-efficient of correlation between $x_{1,3}$ and $x_{2,3}$ is

$$\frac{\text{Cov}(x_{1,3}, x_{2,3})}{\sqrt{\text{Var}(x_{1,3}) \text{Var}(x_{2,3})}} = \frac{1}{n} \frac{\sum x_{1,3} x_{2,3}}{S_{1,3} S_{2,3}} = b_{123} \frac{S_{1,2}}{S_{1,3}}$$

Hence the result:

- 11.20 Using the method of least-squares, fit a quadratic model $Y = \alpha + \beta_1 X + \beta_2 X^2 + \epsilon$ to the following data.

X	-2	-1	0	1	2
f	0.4	1.3	2.2	2.5	3.0

Also calculate the standard error of estimate.

CHAPTER 12

**CURVE FITTING BY
LEAST SQUARES**

CURVE FITTING BY LEAST SQUARES

1.1 INTRODUCTION

Let us suppose that we wish to approximate (describe) a certain type of function that best expresses the association that exists between variables. A scatter plot of the set of values of the variables makes it possible to visualize a smooth curve that effectively approximates the given data set. A more useful way to represent this sort of approximating curve is by means of an equation or a formula. A term applied to the process of determining the equation and/or estimating the parameters appearing in the equation of the approximating curve, is commonly called *curve fitting*.

It is relevant to point out that the relationship between the variables may be *functional* or *non-functional*. In functional relationship, a variable Y has a *true value* corresponding to each possible value of another variable X , i.e. there is no question of random variation in the values of Y , and we make no probabilistic assumptions in this respect. In this chapter, we shall limit our discussion to some functional relationships, i.e. problems of approximation and not of regression (already discussed earlier). Such relationships which are common in the natural sciences may be *linear* or *non-linear*.

1.2 APPROXIMATING CURVES AND THE PRINCIPLE OF LEAST SQUARES

The data sets encountered in practice greatly vary in nature. It is therefore necessary to decide what type of approximating curve and equation should be used. For this purpose, some of many non types of approximating curves and their equations are given below:

Straight line or linear curve,

Parabola of second degree or quadratic curve,

Parabola of third degree or cubic curve,

Exponential curve,

Logarithmic or power curve,

Hyperbola,

and so on.

In these equations, F is the dependent variable and X , the independent variable. In some situations, however, the variables X and F can be reversed.

We may approximate a given set of data by drawing a *free hand curve*, covering most of the points. But it is clear that different individuals would draw different curves according to their personal taste. Therefore this procedure of fixing a curve is not satisfactory.

The principle of least squares is applicable to curve fitting where the purpose is simply one of fitting (or approximation) of a set of observations. Accordingly, we choose to determine the values of the parameters in the equations of approximating curves so as to make the sum of squares of residuals a minimum. A residual has been defined as the difference between the observed value and the corresponding value of the approximating curve.

1.2.1 Fitting a Straight Line. A straight line is the simplest type of approximating curve and its equation is written as

$$Y = a + bX$$

where the values of a and b are to be determined.

Given a pairs of observations (X_i, Y_i) , $i = 1, 2, \dots, n$ to which we wish to fit a straight line. We determine the values of a and b by the principle of least squares, which calls for the minimization of Σ

the sum of squares of the differences between the actual Y_i values and the corresponding values given by $\hat{Y} = a + bX_i$. That is we minimize

$$S = \sum_{i=1}^n (Y_i - a - bX_i)^2$$

To do so, we need to solve the two equations $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$.

That is $\frac{\partial S}{\partial a} = 2 \sum(Y_i - a - bX_i)(-1) = 0$, and

$$\frac{\partial S}{\partial b} = 2 \sum(Y_i - a - bX_i)(-X_i) = 0.$$

which on simplification become

$$\sum Y = n a + b \sum X$$

$$\sum XY = a \sum X + b \sum X^2$$

Solving these two normal equations simultaneously, we get

$$b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} \quad \text{and} \quad a = \bar{Y} - b \bar{X}$$

The value of a indicates that the least squares line passes through the means of observation (\bar{X}, \bar{Y}) .

It should be noted that, when the origin is believed to lie on the curve, the straight line which is simply $\hat{Y} = bX$ and the sum of squares of deviations to be minimized is

$$S = \sum (Y - bX)^2.$$

For a minimum value of S , $\frac{\partial S}{\partial b}$ must be zero, that is

$$\frac{\partial S}{\partial b} = 2 \sum (Y - bX)(-X) = 0, \text{ which gives } \sum XY = b \sum X^2$$

as the normal equation and whence $b = \frac{\sum XY}{\sum X^2}$.

The sum of squares of residuals for a straight line is

$$\begin{aligned} S &= \sum (Y - a - bX)^2 \\ &= \sum [(Y - a) - b(X - \bar{X})]^2 = \sum Y^2 - a \sum Y - b \sum XY. \end{aligned}$$

Example 12.1 Fit a straight line by the method of least squares to the following data:

X	1	2	3	4	5
Y	3	4	6	9	10

Also find the sum of squares of residuals.

Let the equation of the straight line to be fitted to the data, be $Y = a + bX$, where a and b are to be evaluated.

The normal equations for determining a and b are

$$\sum Y = na + b \sum X,$$

$$\sum XY = a \sum X + b \sum X^2.$$

We now calculate $\sum X$, $\sum X^2$, $\sum Y$ and $\sum XY$ as below:

	X	Y	XY	X^2	Y^2
	1	3	3	1	9
	2	4	8	4	16
	3	6	18	9	36
	4	9	36	16	81
	5	10	50	25	100
Σ	15	32	115	55	242

Thus the normal equations become

$$5a + 15b = 32$$

$$15a + 55b = 115$$

Solving these two equations simultaneously, we obtain

$$a = 0.7 \text{ and } b = 1.9$$

Hence the equation of the required straight line is

$$Y = 0.7 + 1.9X$$

The sum of squares of residuals is given by

$$\begin{aligned} S &= \sum (Y_i - a - bX_i)^2 \\ &= \sum [Y_i(Y_i - a - bX_i)] = \sum Y^2 - a \sum Y - b \sum XY \end{aligned}$$

$$S = 242 - 0.7(32) - 1.9(115)$$

$$= 242 - 240.9 = 1.1.$$

12.2.2 Fitting a Second Degree Parabola. The simplest type of a non-linear approximating curve is a second degree parabola that has the equation

$$Y = a + bX + cX^2$$

the values of a , b and c are to be determined.

Let us suppose that we wish to fit this parabolic curve to n pairs of observations (X_i, Y_i) , $i = 1, 2, \dots, n$. Then we need to find those values of a , b and c which will minimize the sum of squares of errors between actual Y values and corresponding values obtained by $a + bX + cX^2$ (the principle of least squares). That is we minimize

$$S = \sum (Y_i - a - bX_i - cX_i^2)^2$$

Minimizing S , we need to set its partial derivatives w.r.t a , b and c equal to zero. Thus

$$\frac{\partial S}{\partial a} = 2 \sum (Y_i - a - bX_i - cX_i^2)(-1) = 0,$$

$$\frac{\partial S}{\partial b} = 2 \sum (Y_i - a - bX_i - cX_i^2)(-X_i) = 0, \text{ and}$$

$$\frac{\partial S}{\partial c} = 2 \sum (Y_i - a - bX_i - cX_i^2)(-X_i^2) = 0.$$

Simplifying, we get the following three normal equations

$$\sum Y = a + b \sum X + c \sum X^2$$

$$\sum XY = a \sum X + b \sum X^2 + c \sum X^3$$

$$\sum X^2 Y = a \sum X^2 + b \sum X^3 + c \sum X^4$$

These equations are solved simultaneously to determine the values of a , b and c .

The sum of squares of residuals in case of second degree parabola is given by

$$S = \sum (Y_i - a - bX_i - cX_i^2)^2 = \sum [Y_i(Y_i - a - bX_i - cX_i^2)] \\ = \sum Y_i^2 - a \sum Y_i - b \sum XY_i - c \sum X^2 Y_i$$

Example 12.2 Fit a second-degree parabola to the following data, taking X as independent variable.

X	0	2	3	4
Y	1.0	1.3	2.5	6.3

(P.U., B.A./B.Sc. 1961)

Let the equation of the second-degree parabola be

$$Y = a + bX + cX^2$$

The normal equations

$$\sum Y = a + b \sum X + c \sum X^2$$

$$\sum XY = a \sum X + b \sum X^2 + c \sum X^3$$

$$\sum X^2 Y = a \sum X^2 + b \sum X^3 + c \sum X^4$$

The computations involved are shown in the following table.

	X	Y	XY	X^2	$X^3 Y$	X^4	X^6
	0	1.0	0	0	0	0	0
	1	1.3	1.3	1	1.3	1	1
	2	1.3	2.6	4	5.2	8	16
	3	2.5	7.5	9	22.5	27	81
	4	6.3	25.2	16	100.8	64	256
Total	10	12.9	37.1	30	130.3	100	354

Putting these values in the normal equations, we get

$$12.9 = 5a + 10b + 30c$$

$$37.1 = 10a + 30b + 100c$$

$$130.3 = 30a + 100b + 354c$$

Solving them as simultaneous equations in a , b and c , we obtain

$$a = 1.42, b = -1.07, \text{ and } c = 0.55$$

Hence the equation of the required second degree parabola is

$$Y = 1.42 - 1.07X + 0.55X^2$$

Example 12.3 Fit an equation of the form $Y = aX^2 + bX^3$ to the following data:

X	0	1	2	3	4	5
Y	1	5	12	20	25	36

Also find the sum of squares of residuals.

The curve to be fitted is $Y = aX^2 + bX^3$.

The normal equations are

$$\sum X^2 Y = a \sum X^4 + b \sum X^5 \quad \text{and} \quad \sum XY = a \sum X^3 + b \sum X^2$$

The arithmetic is arranged in the table below:

X	Y	X^2	X^3	XY	$X^2 Y$	$X^3 Y$	X^4
0	1	0	0	0	0	0	1
1	5	1	1	5	5	5	25
2	12	4	8	16	24	48	144
3	20	9	27	60	180	400	400
4	25	16	64	100	400	640	625
5	36	25	125	180	900	1250	1250
15	99	55	225	975	369	1533	2491

Substitution gives

$$975a + 225b = 1533$$

$$225a + 55b = 369$$

Solving them simultaneously, we get

$$a = 0.4006 \text{ and } b = 5.0703$$

Hence the desired equation is $Y = 0.4006X^2 + 5.0703X$.

The sum of squares of residuals is given by

$$S = \sum (Y - aX^2 - bX)^2 = \sum [Y(Y - aX^2 - bX)]$$

(P.U., B.A./B.Sc. 1993)

$$\begin{aligned}
 &= \sum Y^2 - n \sum X^2 Y - b \sum XY \\
 &= 2491 - (0.40)(1533) - (5.07)(369) \\
 &= 2491 - 613.2 - 1870.83 = 6.97
 \end{aligned}$$

12.2.3 Fitting of Higher Degree Parabolic Curves. A parabolic curve of degree p , approximating a set of observations $\{(X_i, Y_i), i=1, 2, \dots, n\}$ has the equation

$$Y_i = a + bX_i + cX_i^2 + \dots + kX_i^p$$

where a, b, c, \dots, k are the unknown quantities and where $k = 0$, and $n \geq p+1$.

The problem is to determine the $(p+1)$ unknown quantities a, b, c, \dots, k in such a way that the resulting values of Y should be as close as possible to the observed values. We, therefore, take the sum of squares of the residuals, i.e.

$$S = \sum_{i=1}^n (Y_i - a - bX_i - cX_i^2 - \dots - kX_i^p)^2$$

which is a function of a, b, c, \dots, k as $\{X_i, Y_i\}$ are certain numbers. The principle of least-squares suggests the selection of that parabolic curve that minimizes S , the sum of squares of differences between the observed values of Y and the corresponding values calculated from the curve. To minimize S , we take $\frac{\partial S}{\partial a}, \frac{\partial S}{\partial b}, \frac{\partial S}{\partial c}, \dots, \frac{\partial S}{\partial k}$ and set them equal to zero. Simplification leads to the following $(p+1)$ equations:

$$\begin{aligned}
 \sum Y &= na + b \sum X + c \sum X^2 + \dots + k \sum X^p \\
 \sum XY &= a \sum X + b \sum X^2 + c \sum X^3 + \dots + k \sum X^{p+1} \\
 \sum X^2 Y &= a \sum X^2 + b \sum X^3 + c \sum X^4 + \dots + k \sum X^{p+2}
 \end{aligned}$$

$$\sum X^p Y = a \sum X^p + b \sum X^{p+1} + c \sum X^{p+2} + \dots + k \sum X^{2p}$$

These are the normal equations for fitting the parabolic curve of degree p . Solving these simultaneously, we determine a, b, c, \dots, k .

For the particular case $p = 3$, the normal equations for fitting the cubic curve $Y_i = a + bX_i + cX_i^2 + dX_i^3$ become

$$\begin{aligned}
 \sum Y &= na + b \sum X + c \sum X^2 + d \sum X^3 \\
 \sum XY &= a \sum X + b \sum X^2 + c \sum X^3 + d \sum X^4 \\
 \sum X^2 Y &= a \sum X^2 + b \sum X^3 + c \sum X^4 + d \sum X^5 \\
 \sum X^3 Y &= a \sum X^3 + b \sum X^4 + c \sum X^5 + d \sum X^6
 \end{aligned}$$

Similarly, parabolic curves of higher degree may be fitted.

The sum of squares of residuals in case of cubic parabola is given by

$$\begin{aligned} S &= \sum(Y - a - bX - cX^2 - dX^3)^2 \\ &= \sum Y^2 - a \sum Y - b \sum XY - c \sum X^2Y - d \sum X^3Y \end{aligned}$$

A better fit. It is important to note that the sum of squares of residuals enables us to make some comparison. A simple way of judging whether a straight line, a quadratic parabola or a cubic parabola is likely to give the better fit, is to calculate the sum of squares of residuals in each case. The smaller the sum of squares, the better is the fit.

12.2.4 Change of Origin and Unit. The computational labour may be reduced by a suitable choice of origin and unit. If the given values of X_i ($i=1, 2, \dots, n$) are equally spaced with a common interval k and n is an odd number of values, say, $n=2k+1$, the normal equations are simplified by taking the mid value of X as the origin and the common interval k as unit of measurement. That is, if X_i be the i th value, then $u_i = (X_i - X_0)/k$ takes the values $-k, -(k-1), \dots, -2, -1, 0, 1, 2, \dots, (k-1), k$. Hence we get $\sum u_i = 0 = \sum u_i^3 = \dots$. If instead, n is an even number, say $n=2k$, we take the origin at the mean of the two extreme values of X and $k/2$ as the new unit. The values of u_i then become $-(2k-1), -(2k-3), \dots, -3, -1, 1, 3, 2k-3, 2k-1$, so that $\sum u_i = 0 = \sum u_i^3 = \dots$ (Also see chapter 13).

Example 12.4 The profits, $\text{£}P$, of a certain company in the X th year of its life are given by

X	1	2	3	4	5
P	2500	2800	3300	3900	4600

Taking $u = X-3$ and $v = (P-3300)/100$, find the cubic parabolic curve of v on u in the form $v=a+bu+cu^2+du^3$ based on the curve of V on X .

(P.U., B.A./B.Sc. (Hons) 1964)

Since $n=2k+1$ (given), so we find that the sum of odd powers of u are zero, i.e. $\sum u = 0 = \sum u^3$.

The normal equations are thus reduced to

$$\sum v = nu + c \sum u^2,$$

$$\sum uv = b \sum u^2,$$

$$\sum u^2 v = a \sum u^3 + c \sum u^4.$$

These are computed in the following table:

X	u	P	v	u^2	u^4	uv	u^2v
1	-2	2500	-8	4	16	16	-32
2	-1	2800	-5	1	1	5	-5
3	0	3300	0	0	0	0	0
4	1	3900	6	1	1	6	6
5	2	4600	13	4	16	26	52
Σ	0	—	6	10	34	53	21

Substituting these values in the normal equations, we get

$$\begin{aligned} 5a + 10c &= 9, \\ 10b &= 53, \\ 10g + 34c &= 21. \end{aligned}$$

Solving them, we find $a = -0.086$, $b = 5.3$ and $c = 0.643$.

The equation of the required parabolic curve is therefore

$$v = -0.086 + 5.3u + 0.643u^2.$$

In order to deduce the parabolic curve of Y and X we replace u by $X-3$ and v by $\frac{Y-3300}{100}$ in the above relation. Thus we obtain

$$\frac{Y-3300}{100} = -0.086 + 5.3(X-3) + 0.643(X-3)^2$$

Simplifying, we get

$$Y = 2280 + 144.2X + 64.3X^2$$

as the required parabolic curve of Y on X .

12.3 EXPONENTIAL CURVES

Equations in which one of the variable quantities occurs as an exponent such as $Y = ae^{bx}$, are called exponential equations and graphs showing these equations as exponential curves. Exponential curves are used to describe a relation in which one variable grows approximately a geometric progression, while the other forms an arithmetic progression. Data of this hybrid type frequently occurs in the fields of business, banking and economics. In the equation $Y = ae^{bx}$, the letter c is a fixed constant, usually either 10 or e . a and b are determined from the data, a and b are estimated by method of least-squares, we have to minimize S , where

$$S = \sum (Y_i - ae^{bx_i})^2$$

Finding the partial derivatives with respect to a and b , and equating them to zero, we get

$$\frac{\partial S}{\partial a} = 2 \sum (Y_i - ae^{bx_i}) [-e^{bx_i}] = 0, \text{ and}$$

$$\frac{\partial S}{\partial b} = 2 \sum (Y_i - ae^{bx_i}) [-ae^{bx_i} \cdot X_i] = 0.$$

Simplifying, we get

$$\sum Y_i e^{bx_i} = a \sum e^{bx_i}, \text{ and}$$

$$\sum X_i Y_i e^{bx_i} = a \sum X_i e^{bx_i}.$$

It is difficult to solve these equations as the solution requires tedious numerical methods. The solution simplifies if the non-linear curve may be reduced to the linear form by some transformation of one or both the variables. The equation $Y = ae^{bx}$ can be linearized by taking logarithms to the base 10, of both sides. Thus the exponential curve becomes

$$\log Y = \log a + (b \log e) X$$

which may be written as

$$Y' = A + BX'$$

where $Y' = \log Y$, $A = \log a$ and $B = b \log e$. But this is the equation of a straight line in $\log Y$ and X' . Hence the method of fitting an exponential curve to the observed set of data is to fit a straight line to the logarithms of the y_i 's. It should be noted that it is the deviations of $\log Y$, and not of Y , which are being minimized. It is relevant to point out that log form is better for calculating the values from the fitted curve.

We give some of the more common non-linear curves with suitable transformations to convert them into linear form $Y' = A + BX'$.

Non-linear Form	Transformation	Linearized Form
$Y = ax^b$	$Y' = \log Y, A = \log a,$ $X' = \log X$	$Y' = A + BX'$
$Y = ab^x$	$Y' = \log Y, A = \log a,$ $B = \log b$	$Y' = A + BX'$
$Y = \frac{1}{a+bX}$ or $\frac{1}{Y} = a+bX$	$Y' = \frac{1}{Y}$	$Y' = a+BX'$
$\frac{1}{Y} = a + \frac{b}{1+X}$	$Y' = \frac{1}{Y}, X' = \frac{1}{1+X}$	$Y' = a+BX'$
$Y = a + b\sqrt{X}$	$Y' = \sqrt{Y}$	$Y' = a+BX'$
$Y = aX^b + bX$	$Y' = \frac{Y}{X}$	$Y' = aX' + b$

It is worth remarking that, if the variable Y incorporates an element of random variation, we introduce a random error term ϵ . The equations become

$$Y = a + bX + \epsilon$$

$$Y = a + bX + cX^2 + \epsilon \text{ etc.}$$

which will be very similar to the regression models discussed in an earlier chapter.

Example 12.5 Fit an exponential curve $Y = ae^{bx}$ to the following data:

X	1	2	3	4	5	6
Y	1.6	4.5	13.8	40.2	125.0	363.0

(P.U., B.A./B.Sc. (Hons.), 1962; B.Z.U., 1976)

We can write the given equation as

$$\log Y = \log a + (b \log e) X$$

or

$$Y' = A + BX' \quad (\text{From eqn of a st. line})$$

where $Y' = \log_{10} Y, A = \log_{10} a$ and $B = b \log_{10} e$.

As the equation is linear in Y' , $\log Y$ and X , therefore the two normal equations are

$$\sum Y' = aA + b \sum X'$$

$$\sum XY' = A \sum X + B \sum X^2.$$

The necessary calculations are shown in the following table:

	X	Y	X^2	$Y' (= \log Y)$	XY'
	1	1.6	1	0.4041	0.4041
	2	4.5	4	0.6532	1.3064
	3	13.8	9	1.1399	3.4197
	4	40.2	16	1.6042	6.4168
	5	125.0	25	2.0969	10.4845
	6	363.0	36	2.5599	15.3594
Total	21	----	91	8.2582	37.1909

Substituting these values, the normal equations become

$$6A + 21B = 8.2582$$

$$21A + 91B = 37.1909$$

Solving these equations simultaneously, we get

$$A = -0.2805, \text{ and } B = 0.4734$$

$$a = \text{anti-log } A = \text{anti-log} (-0.2805)$$

$$= \text{anti-log } (-7.195) = 0.32$$

$$\text{and } b = 0.4734; \quad b = 0.4734 \text{ or } b = 1.09 \quad (\because \log_{10} e = 0.4343)$$

Hence the equation of the curve fitted to the data is

$$Y = 0.32 (x)^{1.09}$$

Example 12.6 Fit an equation of the form $Y = aX^b$ to the following data:

X	1	2	3	4	5	6
Y	2.98	4.26	5.21	6.10	6.80	7.56

We may reduce the given equation to a linear form by taking log₁₀ to the base 10. Thus

$$\log Y = \log a + b \log X$$

$$\text{or } Y' = A + BX'$$

where $Y' = \log Y$, $A = \log a$ and $X' = \log X$.

As the equation is linear in $T' = \log Y$ and $X' = \log X$, therefore the two normal equations are

$$\sum Y' = n\alpha + b \sum X'$$

$$\sum X'Y' = A \sum X'^2 + b \sum X'^3$$

The following table contains the necessary calculations:

X'	$X' (=\log X)$	Y'	$Y' (= \log Y)$	$X'Y'$	X'^2
1	0	2.98	0.4742	0	0
2	0.3010	4.26	0.6294	0.189449	0.0906
3	0.4771	5.21	0.7368	0.341986	0.2276
4	0.6021	6.10	0.7853	0.472829	0.3625
5	0.6990	6.80	0.8325	0.581918	0.4886
6	0.7732	7.50	0.8751	0.681003	0.6056
Σ	2.8574	—	4.3133	2.267186	1.7749

Substituting these summations, we get

$$64 = 2.8574b - 4.$$

$$2.8574A + 1.7749b = 2.2672.$$

Solving them simultaneously and taking antilog of A , we get

$$a = 2.98 \text{ and } b = 0.5144$$

Hence the required equation is

$$Y = 2.978(X)^{0.5144} \\ \sim 3X^{0.52} \text{ approximately.}$$

OTHER TYPES OF CURVES

Some other types of curves frequently encountered in applied statistics are the following:

(1.4.1) **Modified Exponential Curve.** A modified exponential curve, which is obtained by adding k to an exponential curve, is defined by the relation

$$Y = k + ab^x.$$

If it describes a set of data, the absolute growth of which decreases by a constant proportion when x is negative and " N " is less than one.

The first method to fit this curve is to transform it into a linear form by taking logarithms of both members to use the least-squares method. But this method is difficult for practical use. In the second method we need three equations, because there are three constants k , a and b which are to be determined. Observed data are therefore divided into three equal parts, leaving one or two values at the beginning,

if necessary, to obtain the three equations, the criterion of fit being that the three partial totals of the trend values must equal those of the original data.

Let n denote the number of values in each third of the data.

Then the first equation is

$$\begin{aligned}\sum_1 Y &= nk + a + ab + ab^2 + ab^3 + \dots + ab^{n-1} \\&= nk + a[1 + b + b^2 + b^3 + \dots + b^{n-1}] \\&= nk + a \left[\frac{b^n - 1}{b - 1} \right] \quad (\because \frac{b^n - 1}{b - 1} = 1 + b + b^2 + \dots + b^{n-1})\end{aligned}$$

In a similar way, the other two equations are obtained as

$$\sum_2 Y = nk + ab^* \left(\frac{b^* - 1}{b - 1} \right) \text{ and}$$

$$\sum_3 Y = nk + ab^{**} \left(\frac{b^{**} - 1}{b - 1} \right)$$

Now we find the constant k , a and b .

Subtracting the first equation from the second, we get

$$\sum_1 Y - \sum_2 Y = a \left(\frac{b^* - 1}{b - 1} \right) - a \left(\frac{(b^*)^2 - 1}{b - 1} \right)$$

Again, subtracting the second equation from the third one, we get

$$\sum_2 Y - \sum_3 Y = \frac{(b^{**} - 1)^2}{b - 1}$$

Dividing, we have

$$\frac{\sum_1 Y - \sum_2 Y}{\sum_2 Y - \sum_3 Y} = \left[ab^* \cdot \frac{(b^*)^2 - 1}{b - 1} \right] + \left[a \frac{(b^{**} - 1)^2}{b - 1} \right] = b^*$$

which gives

$$b = \sqrt{\frac{\sum_1 Y - \sum_2 Y}{\sum_2 Y - \sum_3 Y}}$$

Finally,

$$a = (\sum_1 Y - \sum_2 Y) \frac{b - 1}{(b^*)^2}, \text{ and}$$

$$k = \frac{1}{n} \left[\sum_1 Y - \left(\frac{b^* - 1}{b - 1} \right) a \right]$$

12.4.2 The Compertz Curve, named after Benjamin Compertz, is given by the equation

$$Y = ka^{b^X},$$

where k , a and b are constants. The equation is changed to modified exponential equation by taking logarithms of both sides. Thus

$$\log Y = \log k + b^X \log a$$

$$\text{or } Y' = k' + a' b^X$$

where $Y' = \log Y$, $k' = \log k$ and $a' = \log a$.

The Compertz curve, which increases first at an increasing rate, then increases at a decreasing rate until it reaches a maximum level, is frequently used in business and actuarial work.

12.4.3 The Logistic Curve, which is widely used to represent growth, is defined by the relation

$$Y = \frac{k}{1 + bc^X},$$

$$\text{on inverting, we get } \frac{1}{Y} = \frac{1}{k} + \frac{b}{k} c^X \\ = k' + a c^X,$$

where $k' = \frac{1}{k}$ and $a = \frac{b}{k}$. This is similar in form to the modified exponential curve if $\frac{1}{Y}$ is expressed as a function of X and the same method of fitting may therefore be applied with the reciprocals $\frac{1}{Y}$ instead of Y .

The use of this curve to analyse population and biological growth was advocated by Raymond Pearl and L.J. Reed. It should be noted that the logistic curve has four different stages, viz., (i) a period of relatively slow growth, (ii) then a period of accelerated growth (iii) then a period of decelerated growth and (iv) finally a period of stability, when the curve does not go up at all. The growth of human population and that of economic variables are appropriately described by the curve as they conform to these stages.

12.4.4 The Makeham Curve is defined as

$$Y = kx^A b^X$$

in the logarithmic form;

$$\log Y = \log k + A \log x + X \log b \\ = A + CX + Bc^X$$

where $A = \log k$, $C = \log x$ and $B = \log b$.

This type of curve, which is actually a combination of a straight line with a Compertz curve, is used in actuarial and insurance work.

12.5 CRITERIA FOR A SUITABLE CURVE

Frequently, we are required to choose a suitable form of curve to obtain a reasonable fit to the observed sets of data in two variables. The suitability of several curves may be determined by examining the differences in the values of the dependent variable Y . The first difference, denoted by ΔY read as delta Y is defined by $\Delta Y = Y_{n+1} - Y_n$, the second difference is defined by $\Delta^2 Y = \Delta Y_{n+1} - \Delta Y_n$, and so on. A straight line has the property that its first difference is equal to b (a constant), a second degree parabola has the property that its second difference is equal to $2c$ (a constant) and, in general, a parabolic curve of the n th degree has the property that its n th differences are constant. Thus we fit:

- i) a straight line, if the first differences between successive values are approximately constant;
- ii) a second degree parabola, if the second differences are approximately constant;
- iii) a third degree parabola, if the third differences prove to be constant;
- iv) an exponential curve, if the first differences of the logarithms are approximately constant;
- v) a log parabola $Y = ab^x c^{x^2}$, if the second differences of the logarithms of the Y -values tend to be constant;
- vi) a modified exponential curve, if each first difference is a constant percentage of the preceding first difference;
- vii) a Gompertz curve, if the first differences of logarithms are changing by a constant percentage;
- viii) a logistic curve, if the first differences of the reciprocals are changing by a constant percentage; and
- ix) a reciprocal line $\frac{1}{Y} = a + bX$, if the reciprocals of the data show a straight line when plotted on a graph.

12.6 FINDING PLAUSIBLE VALUES BY THE PRINCIPLE OF LEAST-SQUARES

The principle of least-squares can also be applied to find the most satisfactory values of the unknown quantities from a number of independent linear equations in the unknowns when the number of equations is greater than the number of unknowns.

Suppose there are k unknown quantities X_1, X_2, \dots, X_k and let the n observed relations where $n > k$ be

$$a_1 X_1 + b_1 X_2 + \dots + f_1 X_k = l_1$$

$$a_2 X_1 + b_2 X_2 + \dots + f_2 X_k = l_2$$

$$a_n X_1 + b_n X_2 + \dots + f_n X_k = l_n$$

where a 's, b 's, ..., f 's are constants.

When $n > k$, i.e. the number of equations is greater than the number of unknowns, there may not exist a unique solution. In such cases, we therefore try to find those values of X_1, X_2, \dots, X_k which will simultaneously satisfy the given set of independent linear equations as nearly as possible. Such values obtained by the least-squares method and are called the best or most plausible values.

The least-squares criterion calls for the selection of those values of X_1, X_2, \dots, X_k which make the sum of squares of the discrepancies D_i^2 , also called errors or residuals, a minimum, where

$$D_i = a_1 X_1 + b_1 X_2 + \dots + f_i X_k - l_i \quad i = 1, 2, \dots, n$$

In other words, we have to select those values of X_1, X_2, \dots, X_k which minimize

$$S = \sum_{i=1}^n D_i^2 = \sum_{i=1}^n (a_1 X_1 + b_1 X_2 + \dots + f_i X_k - l_i)^2$$

It is obvious that $S = f(X_1, X_2, \dots, X_k)$, that is, the sum of squares of residuals is some function of X_1, X_2, \dots, X_k . If S is to have a minimum value, it is necessary that its partial derivatives with respect to X_1, X_2, \dots, X_k , if they exist, vanish there; hence X_1, X_2, \dots, X_k must satisfy the equations

$$\frac{\partial S}{\partial X_1} = 2 \sum a_1 (a_1 X_1 + b_1 X_2 + \dots + f_i X_k - l_i) = 0$$

$$\frac{\partial S}{\partial X_2} = 2 \sum b_1 (a_1 X_1 + b_1 X_2 + \dots + f_i X_k - l_i) = 0$$

$$\frac{\partial S}{\partial X_k} = 2 \sum f_i (a_1 X_1 + b_1 X_2 + \dots + f_i X_k - l_i) = 0$$

The equations given above may be written in the standard form as

$$X_1 \sum a_i^2 + X_2 \sum a_i b_i + \dots + X_k \sum a_i f_i = \sum a_i l_i$$

$$X_1 \sum a_i b_i + X_2 \sum b_i^2 + \dots + X_k \sum b_i f_i = \sum b_i l_i$$

$$X_1 \sum a_i f_i + X_2 \sum b_i f_i + \dots + X_k \sum f_i^2 = \sum f_i l_i$$

These simultaneous equations obtained by minimizing process, are the normal equations which are simultaneously solved to obtain the best or the most plausible values of X_1, X_2, \dots, X_k .

It should be noted that the normal equations for a set of variables are obtained by multiplying each equation by the co-efficient of the respective variable in the equations and adding them together. This is a convenient way for remembering the normal equations.

Example 12.7 Apply the principle of least-squares to solve

$$2X + Y = 0, \quad 3X - 2Y = 0, \quad -X + Y = -2.$$

(P.U., B.A., B.Sc.: 1971, 75)

There are 3 linear equations and 2 unknown variables X and Y , therefore we apply the least-squares method to get the most plausible values of X and Y .

$$\text{Now } S = (2X + Y - 0)^2 + (3X - 2Y - 0)^2 + (-X + Y + 2)^2$$

The normal equations are $\frac{\partial S}{\partial X} = 0$ and $\frac{\partial S}{\partial Y} = 0$,

$$\text{i.e. } 2(2X + Y) + 3(3X - 2Y) - (-X + Y + 2) = 0$$

$$\text{and } (2X + Y) - 2(3X - 2Y) + (-X + Y + 2) = 0$$

$$\text{or } 14X - 5Y = 2 \text{ and } -5X + 6Y = -2.$$

Solving these equations simultaneously, we get:

$$X \approx 0.034, \text{ and } Y = -0.305.$$

Example 12.8 Find the most plausible values of X and Y from the following equations:

$$X + Y - 3 = 0$$

$$3X + 2Y - 4 = 0$$

$$2X - 3Y + 1 = 0$$

(P.U., B.A. (Hons.) Part-I, 1963; B.A./B.Sc. 1977)

We first find the normal equation for X . Multiplying each equation by the co-efficient of X in it, we have

$$X + Y = 3$$

$$9X + 6Y = 12$$

$$4X - 6Y = -2$$

Adding, we get $14X - Y = 13$, which is the normal equation for X .

We then find the normal equation for Y . Again multiplying each equation by the co-efficient of Y in it, we get

$$X + Y = 3$$

$$6X + 4Y = 8$$

$$-5X + 9Y = 3$$

Adding them together, we have $-X + 14Y = 8$ as the normal equation for Y .

Thus the two normal equations are

$$14X - Y = 13$$

$$-X + 14Y = 8$$

Solving them simultaneously, we obtain

$$X \approx 0.97 \text{ and } Y \approx 0.64$$

which is the required solution.

EXERCISES

- 12.1. a) What is meant by Curve Fitting? (P.U., B.A./B.Sc. 1982)
- b) Explain the principle of Least Squares with particular reference to a straight line fit. In what sense, does it give the "best" solution? (P.U., B.A./B.Sc. 1982)

- c) Fit a straight line to the following data:

X	1	2	3	4	5	6
Y	2	6	7	8	10	11

Calculate the values of Y for each value of X , obtain the values of residuals e_i 's and check that $\sum e_i = 0$.

- 12.2 a) By means of Least Squares, show how a straight line can be fitted to a set of given observations, and obtain the normal equations.

- b) Prove that a least squares line always passes through the point (\bar{X}, \bar{Y}) .

(P.U., B.A./B.Sc. 1978)

- c) Fit a straight line to the following data and plot on the graph paper the actual and calculated values.

X	0	1	2	3	4	5	6	7	8
Y	5	11	8	14	10	16	2	20	15

- 13 a) Write down the equation of a straight line through the origin and derive an expression for finding its slope by the principle of least squares. (P.U., B.A./B.Sc. 1991)

- b) Fit a least-squares line to the following data:

Year (X)	1	2	3	4	5	6	7	8	9
Output (Y)	1	3	2	4	3	5	4	6	5

Measure the deviations from the fitted line and find the sum of squared deviations.

- 14 a) Find the normal equations which determine the values of a and b in least squares line $Y = a + bX$, and show that the sum of squares of residuals from the least squares line is given by

$$S = \sum Y^2 - a \sum Y - b \sum XY$$

- b) Fit a straight line to the following data:

X	0	5	10	15	20	25
Y	12	15	17	22	24	36

(P.U., B.A./B.Sc. 1962; 80)

- a) Fit the least squares line for 20 pairs of observations having $\bar{X} = 2$, $\bar{Y} = 1$, $\sum X^2 = 180$ and $\sum XY = 404$. (P.U., B.A./B.Sc. 1986)

- b) Given:

X	1	2	3	4	5
Y	8	9	13	18	27

Fit $Y = a + bX$ by least-squares and estimate Y for $X=6$. Also fit $Y = c + dX$ and use this equation to estimate Y for $X=6$. Account for the difference in two estimates.

(P.U., B.A. (Hons.) Part-II, 1963-S)

- 12.6 a) Find the normal equations for a , b and c that will minimize

$$S = \sum(Y - (a + bX + cX^2))^2$$

- b) Show that the sum of squares of residuals for a second degree parabola is

$$S = \sum Y^2 - a \sum Y - b \sum XY - c \sum X^2 Y$$

- c) Fit a parabola of the form $Y = a + bX + cX^2$ to the data:

X	-2	-1	0	1	2
Y	-5	-2	1	2	1

- 12.7 a) By means of the principle of least squares, show how a parabola of second order can be fitted to a set of n observations (X_i, Y_i) and obtain the normal equations.

- b) For 5 pairs of observations, it is given that A.M. of X series is 2 and A.M. of Y series is 10. It is also known that:

$$\sum X^2 = 30, \sum X^3 = 100, \sum X^4 = 354, \sum XY = 242, \sum Y^2 = 850$$

Fit a second degree parabola, taking X as the independent variable.

(P.U., B.A./B.Sc., 2008)

- c) Fit a second degree parabola to the following data:

X	0	2	3	4
Y	1	5	10	22

(P.U., B.A. (Part I), 2008)

- 12.8 Fit a second degree parabola to the following seven pairs of values:

X	1.5	2.0	2.5	3.0	3.5	4.0
Y	1.1	1.3	1.6	2.0	2.7	3.4

(P.U., B.A./B.Sc., 2008)

- 12.9 Fit a second degree parabola to the following data:

X	8	12	16	20	24	28	32	36	40
Y	2.4	4.8	8.3	9.5	11.2	24.3	22.2	21.2	25.4

(P.U., B.A. (Hons.), 2008)

- 12.10 The profits, EY , of a certain company in the X th year of its life are given by:

X	1	2	3	4	5
Y	1250	1400	1650	1950	2300

(P.U., B.A. (Hons.), 2008)

Taking $u = X - 3$, $v = (Y - 1650)/50$, show that the parabolic curve of v on u is

$$v = 0.086 + 5.36 u + 0.643 u^2$$

and deduce that the parabolic curve of Y on X is

$$Y = 1340 + 72.14X + 32.14X^2$$

(P.U., B.A. (Hons.), 2008)

CURVE FITTING BY LEAST SQUARES

Q.11 Fit, by the method of least-squares,

- the straight line of best fit,
- the 2nd degree parabola of best fit, to the following data:

X	29	25	30	35	40	45	50	55
Y	240	315	403	450	488	520	525	532

Also calculate the sum of squares of residuals in the two cases.

Q.12 Fit a straight line and parabolas of the second and third degrees to the following data, taking X to be the independent variable;

X	0	1	2	3	4
Y	1	1.8	1.3	2.5	6.3

and calculate the sum of squares of residuals in the three cases.

Q.13 a) You are given data in two variables X and Y and you have to take a decision about fitting a suitable trend. How will you proceed?

(P.U., B.A./B.Sc. 1987)

b) Given the following pairs of values of X and Y .

X	0	1	2	3	4
Y	10	17	28	45	62

Fit a suitable curve.

(P.U., B.A./B.Sc. 1976)

Q.14 Explain the principle of least squares and use it to obtain the normal equations when a cubic parabola is fitted to n pairs of observations.

Q.15 Fit a curve of the form $Y = ab^x$ to the following data in which Y represents the number of bacteria per unit volume existing in a culture at the end of X hours:

	0	1	2	3	4
Y	73	91	112	131	162

Estimate the value of Y when $X=5$ and 6.

(P.C.S. 1972; P.U., B.A./B.Sc. 1978)

The number (Y) of bacteria per unit volume present in a culture after X hours is given in the following table:

No. of hours (X)	0	1	2	3	4	5	6
No. of bacteria per unit volume (Y)	32	47	65	92	132	190	275

Fit a least-squares curve having the form $Y = ab^x$ to the data. Estimate the value of Y when $X=7$.

(P.U., B.A./B.Sc. 1969, 79, 80)

- 12.16 Fit a simple exponential to the following data for a growing plant by taking the logarithms of the exponential equation.

Day	0	1	2	3	4	5	6	7	8
Height	0.75	1.20	1.75	2.50	3.45	4.70	6.20	8.25	11.50

- 12.17 Fit a curve of the type $Y = ab^x$ to the following data.

X	1	2	3	4	5	6	7
Y	10	12.2	14.5	17.3	21.0	25.0	29.0

- 12.18 The following data represent the enrolments at a small liberal arts college during the past seven years:

X (years)	1	2	3	4	5	6	7
Y (enrolments)	364	341	393	457	548	670	882

Use the method of least-squares to estimate a curve of the form $Y = ab^x$ and predict the enrolments seven years from now. (P.U., B.A./B.Sc. 1979)

- 12.19 a) Given $n=8$, $\sum X = 16$, $\sum X^2 = 204$, $\sum X^3 = 582$, $\log Y = 23$, $\sum X \log Y = 104$. Fit a suitable curve. (P.U., B.A./B.Sc. Hon. 1980)

- b) Fit a curve of form $Y = a + b\sqrt{X}$ to the following data:

X	1.20	3.20	3.40	4.70	5.30
Y	6.60	8.03	8.95	10.09	10.56

- 12.20 Fit a curve $Y = aX^b$ to the following data.

X	1	2	3	4	5	6
Y	1200	900	600	200	110	50

- 12.21 Fit a curve of the form $Y = aX^b$ to the following data on the unit cost in dollars of producing electronic components and the number of units produced.

Lot size (X)	50	100	250	500	1000
Unit cost (Y)	108	53	24	9	5

Use the result to estimate the unit cost for a lot of 400 components.

(P.U., B.A./B.Sc. 1979)

- 12.22 It is thought that two physical quantities X and Y should be connected by a relation of the form $Y = aX^b$. The experimental values are:

X	0.5	1.5	2.5	5.0	10.0
Y	3.4	7.0	12.8	29.8	68.2

Find the best values of a and b .

¹

(P.U., B.A./B.Sc. 1979)

- 22) The discharge of a capacitor through a resistance gave the following results:

t (seconds)	0.5	0.8	1.4	2.0	2.5
v (volts)	9.1	8.5	7.5	6.7	6.1

Fit a curve of the type $v = ae^{-rt}$ to these data.

- 23) a) Fit a curve of the type $Y = ae^{bX}$ to the following data:

X	1	2	3	4	5
T	27	73	200	545	1484

where $e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$.

- b) Obtain the values of V from the approximating line for various values of X . Do the deviations of the observed values of V from the corresponding calculated values add to zero? Explain your result.

(P.U., B.A./B.Sc. 1977)

- 24) Estimate the constant of Parclo Curve, $n = AX^n$, which fits the data below:

Income (₹)	Number (n)
150	14,000
500	25,000
1,000	173,000
2,000	35,560

- 25) The pressure (p) of a gas and its volume (v) are known to be related by an equation of the form $pv^y = \text{constant}$. From the following data, find the value of y by fitting a straight line to the logarithms of p and v , taking v to be the independent variable.

p (kg per sq. cm)	0.5	1.0	1.5	2.0	2.5	3.0
v (litres)	1.62	1.00	0.75	0.62	0.52	0.46

- 26) a) Derive the least-squares equations for fitting a curve of the type $\frac{1}{Y} = a + bX$ to a set of n observations. Also find the values of a and b .

- b) Fit a reciprocal curve $\frac{1}{Y} = a + bX$ to the following data:

X	0	1	4	6	12	16
Y	10	8	5	4	2.5	2

- 27) Find the normal equations for determining a , b and c from the linear equation $Y = a + bX_1 + cX_2$.

- b) Find the least-squares fit $Y = a + bX_1 + cX_2$, given

Y	2	5	7	8	5
X_1	8	8	6	5	3
X_2	0	1	1	3	4

- 12.29 a) What is the modified exponential curve? Describe the method of fitting it.

(P.U., B.A. (Hons.) Part-II, 1962)

- b) Derive the least-squares equations for fitting a modified exponential, $Y = c + ze^{kY}$ to a set of n observations, and indicate why these equations would be difficult to solve.

- 12.30 Write a critical note on the law of growth as portrayed by the logistic curve and the Gompertz curve. (P.U., B.A./B.Sc. 1962)

- 12.31 Use the "principle of least-squares" to find the normal equations when the number of equations is greater than the number of unknown quantities. * (P.U., B.A./B.Sc. 1981, 84, 85, 86)

- 12.32 a) Explain the method of least-squares. Apply it to solve the equations

$$X + 7Y = 17, \quad 2X - Y = 6, \quad 3X - 2Y = -1$$

(P.U., B.A./B.Sc. 1970)

- b) Find the most plausible values of X and Y from the following equations. Also compute the sum of squares of residuals.

$$2X + Y = 4.8, \quad -X + 3Y = 3$$

$$3X - 2Y = -2.1, \quad 3X + Y = 8.0,$$

(P.U., B.A./B.Sc. 1981, 86)

- 12.33 a) Find the most plausible values of X and Y from the following equations:

$$X + Y = 3.01, \quad 2Y - Y' = 0.03$$

$$X + 3Y = 7.02, \quad 3X + Y = 4.97$$

- b) Obtain the best possible values of X and Y from:

$$2X + Y = 1.02, \quad 3X - Y = 10.02,$$

$$X + 2Y = 5.02, \quad 3X + 2Y = 0.97.$$

(P.U., B.A./B.Sc. 1970)

- 12.34 Form normal equations and solve

$$X + 2Y + Z = 1, \quad 2X + Y + Z = 4,$$

$$-X + Y + 2Z = 3, \quad 4X + 2Y - 5Z = -7.$$

(P.U., B.A./B.Sc. 1962, 60, 70)

- 12.35 Find the most plausible values of X , Y and Z from the following equations:

$$X - Y + 2Z = 3, \quad 3X + 2Y - 5Z = 5,$$

$$4Y + Z + 4Z = 21, \quad -X + 3Y + 3Z = 14.$$

(P.U., B.A./B.Sc. 1970)



Contents

CHAPTER 13

**TIME SERIES
ANALYSIS**

13.1 INTRODUCTION

A time series consists of numerical data collected, observed or recorded at more or less regular intervals of time each hour, day, month, quarter or year. More specifically, it is any set of data in which observations are arranged in a chronological order. Examples of time series are the hourly temperature recorded at a locality for a period of years, the weekly prices of wheat in Lahore, the monthly consumption of electricity in a certain town, the monthly total of passengers carried by rail, the quarterly sales of a certain fertilizer, the annual rainfall at Karachi for a number of years, the enrolment of students in a college or university over a number of years and so forth.

The analysis of a time series is a process by which a set of observations in a time series is analysed. Time series analysis is rather a difficult topic but we shall limit our discussion to the basics of time series analysis.

The observations in a time series, denoted by $Y_1, Y_2, \dots, Y_n, \dots$, are usually made at equally spaced units of time or they are associated with equal intervals of time (t). Given an observed time series, the first step in analysing a time series is to plot the given series on a graph taking time intervals (t) along the X -axis, as the independent variable, and the observed values (Y_t) on the Y -axis, as the dependent variable. Such a graph will show various types of fluctuations and other points of interest.

It is worthwhile to note that the middle of the period is taken to represent the data for that period. For example, the yearly data corresponds to June 30 or July 1, the middle of a calendar year and monthly data to the middle of the month, i.e. the 15th day of the month.

Example 13.1. The following table shows the number of bags (hundreds) of fertilizer sold by a certain dealer. Plot these data as a time series and comment on the graph.

Year	Sales			
	I	II	III	IV
2001		98	79	106
2002	79	122	101	143
2003	94	141	128	160
2004	125	143	135	187

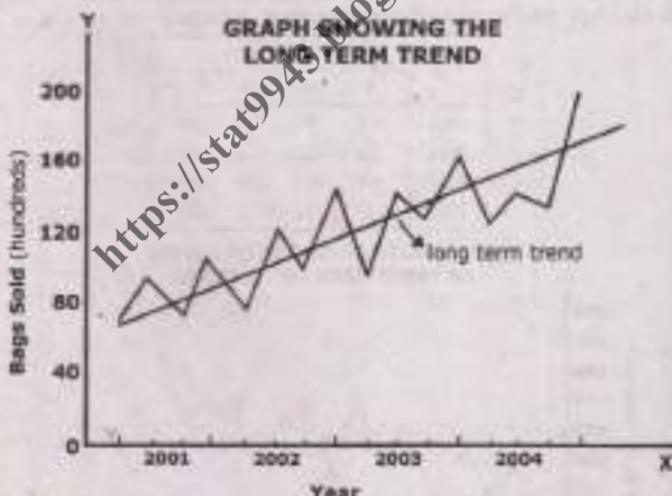


The histogram obtained by plotting the given time series, shows that the sales have risen for the second quarter, have fallen for the third quarter and then have risen to a higher point for the fourth quarter every year. The graph also reveals that the sales, on the whole, have risen over 4 years. The graph further suggests that by smoothing out the irregularities, the annual rate at which the sales have increased, can be ascertained.

13.2 COMPONENTS OF A TIME SERIES

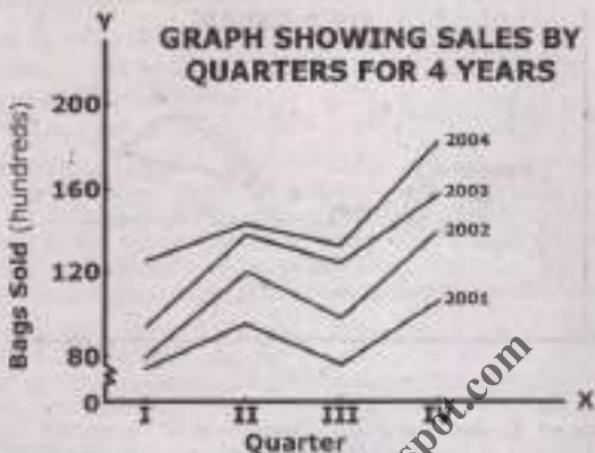
A typical time series may be regarded as composed of four basic types of movements, usually called components of a time series. The four components are: *secular trend (T)*, *seasonal variations (S)*, *cyclical fluctuations (C)* and *irregular or random variations (I)*. These components are assumed to be the outcomes of distinct causes of variation. All four of these components are not necessarily present in a time series occurring in practice. Let us discuss each of these components in turn.

13.2.1 Secular Trend. A *secular trend (T)* is a long-term movement that persists for many years and indicates the general direction of the change of observed values; in other words, it refers to a smooth broad movement of a time series in the same direction, showing a gradual rise or fall within the data. The secular trend generally dominates other variations in the long run (it covers a fairly long period of time, not less than 10 years). The long-term trend is a peculiar characteristic of most of the economic variables such as sales (see figure).



prices, industrial production, capital formation, etc. Analysing the trend component helps in ascertaining the rate of change to be used for further estimates. It also helps in business planning and in analysing other variations.

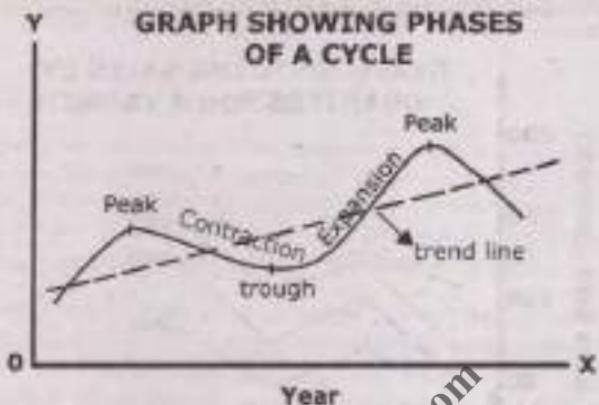
13.2.2 Seasonal Variations. The seasonal variations (S), which are mainly caused by the change in seasons, are short-term movements occurring in a periodic manner. These fluctuations are repeated with more or less the same intensity within a specific period of one year or shorter (see figure).



The main causes for seasonal variations are the weather condition, the religious festivals and the social customs. Examples of seasonal variations are the sales of wheat which fall after the harvesting season and rise before the sowing time, the sales of soft drinks which are high in the summer and low in the winter, investments in Savings Certificates which are high in the months of May and June and low in other months, and so forth. The concept of seasonal variation is customarily broadened to include the more or less regular fluctuations of shorter duration occurring within a day, a week, a month, a quarter and so forth. Examples of such variations are the daily variations in temperature or the monthly variations in bank deposits.

13.2.3 Cyclical Fluctuations. The cyclical fluctuations (C) are the long-period oscillations about the long-term trend, which tend to occur in a more or less regular pattern over a period of certain number of years. The so-called business cycles which represent alternating periods of prosperity and depression, provide an important example of cyclical movements. A cycle, as it is known, is said to be completed when it begins with a peak (a peak is a value which is greater than the two-neighbouring values), the curve reaches a low point, called a trough (a trough is a value which is lower than the two-neighbouring values) and then rising again reaches the next peak. The period either from peak to peak or from trough to trough, is usually referred to as the duration of a cycle. Cycles have a duration of three to four years from two to ten years or even a longer period. In general, a complete business cycle has the following four phases: (i) the period of prosperity, (ii) the period of contraction, (iii) the period of

recession or depression and (iv) the period of recovery or expansion, which finally develops into a period of prosperity.



13.2.4 Irregular or Random Variations. These variations are irregular and unsystematic in nature. They occur in a completely unpredictable manner and they are caused by some unusual events such as floods, droughts, strikes, fire, earthquakes, war, disease, political events and the like. These variations are also known as the偶然的, 隨機的, 離散的 variations. It is difficult to make a study of such non-recurring variations, though they can be easily identified.

13.3 TIME SERIES DECOMPOSITION

A time series analysis is mainly concerned with the decomposition of the observed series into its components so as to estimate their separate effects. To do this, we must make assumptions about the relationship existing among the various components. Accordingly, these components are assumed to have either the multiplicative relationship (also called multiplicative model) or the additive relationship (also called additive model).

In multiplicative (decomposition) model, we assume that each observed value T_t at any time t is determined by the product of the measures of all the four components T, S, C and I . Symbolically,

$$T_t = T \times S \times C \times I_t = TSCI_t$$

where the observed values and T -values are stated in original units but the other components, i.e., S, C and I are expressed in percentages. On the other hand, in the additive (decomposition) model, each value that we observe, is thought of as being a sum of all the four components, i.e.

$$Y_t = T + S + C + I_t$$

where the components T, S, C and I are assumed to be mutually independent. Before analyzing a time series, it is often desirable to adjust the data for calendar variations, for holidays, for price changes and so forth. Such adjustments help in removing the effects of certain false differences.

13.4 ANALYSING THE SECULAR TREND

The analysis of a trend component involves its measurement and/or elimination from an observed time series data. To measure a trend which can be represented as a straight line or some type of smooth curve, the following methods are used.

- i) The method of freehand curve.
- ii) The method of semi-averages.
- iii) The method of moving averages.
- iv) The method of least-squares.

13.4.1 The Method of Freehand Curve. Plot the given data on a graph paper and join the plotted points by segments of straight line. Observe the up and down movements on the graph and draw a smooth curve or a straight line freehand passing through the plotted points in a way such that the general direction of change in values is indicated. This line may also be drawn by a transparent ruler or by stretching a piece of thread through the central region of the plotted points. The line smoothes out short-run fluctuations. The trend values for the given periods can be read from the graph.

This method is simple and quick. It will be a close approximation to a mathematically based trend drawn with care. It has certain disadvantages. It is a rough and subjective method. As the drawing of trends depends on individual judgment and experience, different persons will draw the graphical trend at different positions with different slopes. Moreover, considerable practice is needed to make a good fit.

13.4.2 The Method of Semi-Averages. Divide the values in the series into two equal parts, including the middle value in each half or omitting it altogether when the number of values is odd. Find the average of the values in each part and place the average values against the respective midpoints of the two parts. Plot these two average values on the graph of the original values, draw a straight line connecting the two points and extend the line to cover the whole series. This is the semi-average trend line which is to be used to read or to compute trend values.

This method is simple and quick. It gives an entirely objective result when the trend is a straight line. It has two disadvantages. The arithmetic mean which is used to average the observed values, is easily affected by abnormally small or large values. The method is only suitable when the trend is linearly linear.

Example 13.2. Compute and insert on graphs the semi-average trends for the following series.

- a) Annual profits in thousands of rupees in a certain business

Year	1973	1974	1975	1976	1977	1978	1979	1980
Profit	85	97	100	90	83	105	112	120

- b) Indoor patients ('000) treated in hospitals in the Punjab

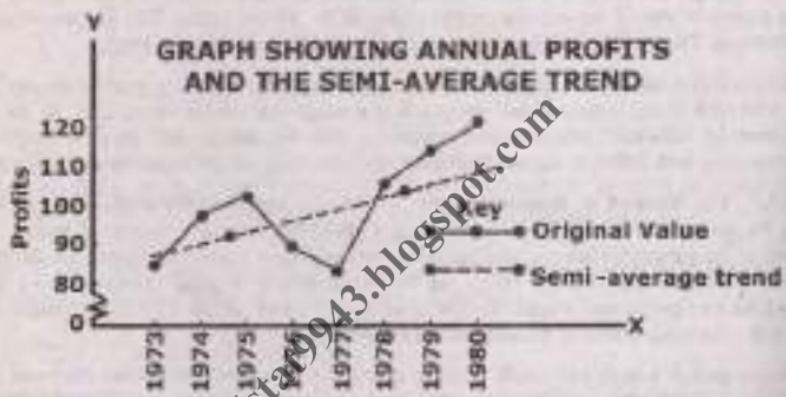
Year	1967	1968	1969	1970	1971	1972	1973	1974	1975
Patients	276	270	260	286	302	321	351	348	346

(Source: Bureau of Statistics, Govt. of the Punjab, 1977)

- c) The data are divided into two equal parts, each consisting of 4 years. Having found the simple averages of the two parts, the first average value is placed opposite the middle of 1974 and 1975, and the second average value is placed opposite the middle of 1978 and 1979. These

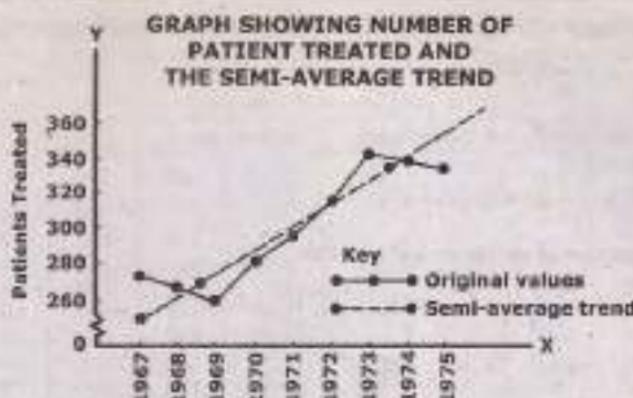
values are then plotted on the graph of the observed series. The line joining these points give the semi-average trend.

Year	Profit	Total	Averages
1973	85		
1974	97		
1975	100		
1976	90		
1977	83		
1978	105		
1979	112		
1980	120		
		372	$372 \div 4 = 93$
		420	$420 \div 4 = 105$



- b) Here the number of years is odd, therefore the middle value is omitted in order to divide values in the series in two equal parts. Having computed the simple averages of the two parts, they are placed opposite the respective midpoints of the two parts. They are plotted on the graph of the original values. The line connecting these points gives the semi-average trend.

Year	Profit	Total	Averages
1967	276		
1968	270		
1969	260		
1970	286		
1971	—		
1972	321		
1973	351		
1974	348		
1975	346		
		1092	$1092 \div 4 = 273$
		1366	$1366 \div 4 = 341.5$



13.4.3. The Method of Moving Averages. The k -period moving averages are defined as the averages calculated by using the k consecutive values of the observed series, then repeating the operation dropping one value at the beginning and including the first value after the preceding total, and so on, carrying on one value to calculate each successive average. This process is continued till the last k consecutive values have been averaged.

Symbolically, the values of the k -period moving averages, denoted by a'_k 's, will be as given below:

$$a'_1 = \frac{1}{k} \sum_{i=1}^k Y_i, \quad a'_2 = \frac{1}{k} \sum_{i=2}^{k+1} Y_i, \quad a'_3 = \frac{1}{k} \sum_{i=3}^{k+2} Y_i, \quad \text{and so on.}$$

In practice, the moving averages may also be obtained by the relations:

$$a'_2 = a'_1 + \frac{Y_{k+1} - Y_1}{k}, \quad a'_3 = a'_2 + \frac{Y_{k+2} - Y_2}{k}, \quad \text{and so on.}$$

An i -period moving average is placed against the middle of its time period. It is relevant to note that the average will correspond directly to the observed value in the series when k is odd and when k is even, the moving average will be placed at points which will be located between two periods. It is then necessary to shift these averages so that they should coincide in time with the observed values in the series. To shift, generally speaking to centre, each average, a 2-period moving average of the already computed k -period moving average is calculated. Then it is called a k -period centred moving average.

For the purposes of illustration, let us first choose k to be odd, say, $k=3$ years. Then we compute the moving averages as

$$a'_1 = \frac{1}{3} (Y_1 + Y_2 + Y_3),$$

$$a'_2 = \frac{1}{3} (Y_2 + Y_3 + Y_4),$$

$$a'_3 = \frac{1}{3} (Y_3 + Y_4 + Y_5),$$

The averages so obtained are placed opposite the middle year of each group i.e. opposite the 2nd year, 3rd year, 4th year, and so on. This process is continued till the last 3 consecutive values have averaged.

Next, we choose k to be even, say, $k=4$ years. Then the 4-year moving averages are computed as

$$a_1 = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4),$$

$$a_2 = \frac{1}{4}(Y_2 + Y_3 + Y_4 + Y_5),$$

$$a_3 = \frac{1}{4}(Y_3 + Y_4 + Y_5 + Y_6),$$

and so on. A 4-year centred moving average, a' , is then

$$\begin{aligned} a' &= \frac{1}{2} \left[\frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4) + \frac{1}{4}(Y_2 + Y_3 + Y_4 + Y_5) \right] \\ &= \frac{1}{8} [Y_1 + 2Y_2 + 2Y_3 + 2Y_4 + Y_5]. \end{aligned}$$

That is, a 4-year centred moving average is closely equivalent to a 5-year weighted moving average with weights 1, 2, 2, 2, 1 respectively. Similarly, the 12-month centred moving average can be obtained by adding the observations for the 13 consecutive months with the central months being counted twice and then dividing the weighted sum by the weights, i.e. 24. In general, a k -period (k is even) centred moving average is equivalent to a $(k+1)$ -period weighted moving average where the $(k-1)$ centre periods are given double weight.

These average values are plotted on the graph. The original values and the line connecting these points is the moving average trend, which smooths out periodic fluctuations of the seasonal and cyclical types present in the series. The line may be continued in the general direction indicated by the original plotting for the purposes of future estimates.

The period of the moving average is chosen in a way that the period over which fluctuations occur, is covered. This is usually equal to the period of at least one cycle. For example, when the yearly data are made up of four quarters or 12 months, 4-quarter or 12-months moving averages should be computed. When the trend is of an exponential type, the moving average are to be computed by using the geometric mean instead of the arithmetic mean.

The method of moving average is easy and simple. The moving averages of appropriate period estimate the combined effects of trend and cyclical components and give a smooth version of the series by removing the seasonal and other effects. It has a number of disadvantages. The moving average does not provide values at the end or the beginning of the original series by half the period. The successive moving averages are unduly affected by large Y -values. This disadvantage may be reduced by using the geometric mean. In the absence of an appropriate period, moving averages may make the resulting series more cyclical than the observed series. It also has the drawback that the method does not provide mathematical expression for the trend, therefore the estimation by extrapolation is subjective.

Example 13.3 Compute (i) 3-year moving averages, (ii) 5-year moving averages and (iii) 7-moving averages for the following data and show the moving-average trends on the graph.

Year: 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989

Values: 18.0 20.5 19.6 24.2 27.8 25.1 25.9 30.2 34.0 36.0

Year: 1990 1991 1992 1993 1994 1995 1996 1997

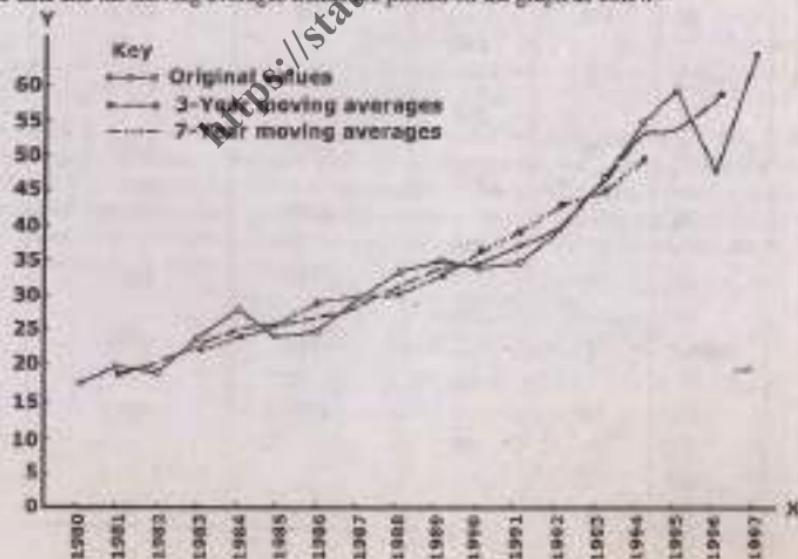
Values: 35.0 35.8 40.9 43.4 55.6 60.4 48.6 68.7

The computation of the 3, 5 or 7-year simple moving averages consists of two steps: (i) computation of a 3, 5, or 7-year moving totals and (ii) division of these moving totals by 3, 5, or 7 to obtain moving averages.

Computation of Moving Averages

Year (t)	Values (Y)	3-year moving		5-year moving		7-year moving	
		Total	Average (trend)	Total	Average (trend)	Total	Average (trend)
1980	18.0	---	---	---	---	---	---
1981	20.5	58.1	19.4	—	—	—	—
1982	19.6	64.3	21.4	110.1	22.0	—	—
1983	24.2	71.6	23.9	117.2	23.4	161.1	23.0
1984	27.8	77.1	25.7	122.6	24.5	173.3	24.8
1985	25.1	78.8	26.3	133.2	26.6	186.8	26.7
1986	25.9	81.2	27.1	143.0	28.6	203.2	29.0
1987	30.2	90.1	30.0	151.2	30.2	214.0	30.6
1988	34.0	100.2	33.4	161.1	32.2	222.2	31.7
1989	36.0	105.0	35.0	171.0	34.2	237.8	34.0
1990	35.0	106.8	35.6	181.7	36.3	260.3	37.2
1991	35.8	111.7	37.2	196.1	39.0	385.7	40.8
1992	40.9	125.1	41.7	215.7	41.1	312.1	44.6
1993	48.4	144.9	48.3	241.1	48.2	324.7	46.4
1994	55.6	164.4	54.8	253.0	50.8	358.4	51.2
1995	60.4	164.6	54.9	248.0	56.3	—	—
1996	48.6	177.7	59.2	—	—	—	—
1997	68.7	—	—	—	—	—	—

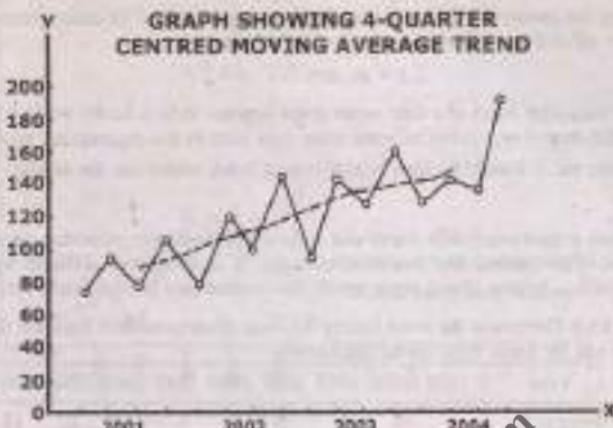
The data and the moving averages trend are plotted on the graph as below:



Example 13.4 Compute the 4-quarter centred moving averages for the time series given in Example 13.1, and show them and the data on a graph.

The 4-quarter centred moving averages appear in the last column of the following table. These are the trend values. They are plotted on the graph of original observations and the trend is shown dashed (see page 487).

Year and quarters (1)	y-values (2)	4-quarter moving totals (3)	4-quarter centred moving totals (4)	4-quarter centred moving averages (5) = col(4) + 8
2001-I	72		--	--
	98	355	--	--
	79	362	717	89.5
	106	386	748	93.5
2002-I	79	408	795	99.2
	122	445	853	106.6
	101	479	905	113.1
	143	523	939	117.4
2003-I	94	506	985	123.1
	41	523	1029	128.6
	128	554	1077	134.6
	160	556	1110	138.8
2004-I	125	563	1119	139.9
	143	590	1153	144.1
	135	--	--	--
	187	--	--	--



13.4.4 The Method of Least-Squares. A trend line, which can be described by a mathematical equation of the form of a straight line, a parabola or an exponential, can be measured using the method of least-squares, by letting time to be the independent variable. Let us suppose that the long-term trend appears to be linear. Then the equation of the least squares line would be

$$T_t = a + b t$$

The values of a and b , the two constants, are determined by solving the following two normal equations simultaneously.

$$\sum T_t = a \sum 1 + b \sum t$$

$$\sum T_t^2 = a \sum 1^2 + b \sum t^2$$

The trend values are computed from this equation by substituting the values of t and are plotted on a graph of the original values.

To simplify computations, the time variable (t) is coded by taking the time deviations from the middle point of the periods and the coded time period is denoted by X . For example, when the number of years is odd, the middle year is taken as 0 or the origin, and when the number of years is even, $X=0$ is taken at the middle of the two middle years. In the latter case, for convenience, the fractional values of X may be multiplied by 2 to express the time units in half-year units. For an illustration, the coded year values X are shown in the table:

Odd number of years		Even number of years		
Year (t)	Coded year (X)	Year (t)	Coded year (X)	X in half-year units
1978	-2	1978	-2.5	-5
1979	-1	1979	-1.5	-3
1980	0	1980	-0.5	-1
1981	1	1981	+0.5	+1
1982	2	1982	1.5	3
		1983	2.5	5

Sometimes, the numbers 1, 2, 3, ..., may also be assigned to X . The coded form of the linear trend would be $\bar{Y}_t = a + bX$ and the two normal equations would reduce to

$$\sum Y = na \quad \text{and} \quad \sum XY = b \sum X^2$$

When the long-term trend of a time series graph appears to be a curve, we fit a parabolic curve or the second or third degree or a curve of some other type such as the exponential, modified exponential, Gompertz, logistic, etc. It should be kept in mind that a trend, which has the smaller $\Sigma(Y - \bar{Y}_t)^2$, is a better fit.

This method is mathematically sound and is therefore useful for prediction purposes. It is also an objective method. The method has two disadvantages. It is sometimes difficult to choose a proper mathematical equation. In case of non-linear trends, the method may involve heavy calculations.

Example 13.5 Determine the trend line by the least-squares method from the following data: (a) the actual values and the linear trend on the same graph.

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003
Price in Rs.	3	6	2	10	7	9	14	12	18

Let the equation of the linear trend be $\bar{Y}_t = a + bX$. Since the number of years in the data is odd, we can assign $X=0$ to the middle year 1999, $X=1, 2, 3, 4$ to the successive years and $X=-1, -2, -3, -4$ to the preceding years. The normal equations then reduce to

$$\sum Y = na \quad \text{and} \quad \sum XY = b \sum X^2 \quad (\sum X = 0)$$

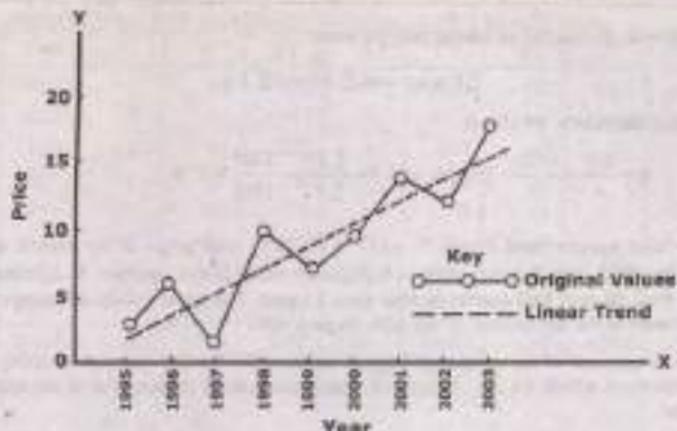
The arithmetic can be arranged as in the table given below:

Year (t)	Coded year (X)	\bar{Y}_t	XY	X^2	Trend $\bar{Y}_t = 9 + 1.7X$
1995	-4	3	-12	16	2.2
1996	-3	6	-18	9	3.9
1997	-2	2	-4	4	5.6
1998	-1	10	-10	1	7.3
1999	0	7	0	0	9.0
2000	1	9	9	1	10.7
2001	2	14	28	4	12.4
2002	3	12	36	9	14.1
2003	4	18	72	16	15.8
Σ	0	81	101	60	81.0

Substituting, we get $a = \frac{\sum Y}{n} = \frac{81}{9} = 9$ and $b = \frac{\sum XY}{\sum X^2} = \frac{101}{60} = 1.7$.

Hence the required equation of the linear trend is $\bar{Y}_t = 9 + 1.7X$,

where the middle year 1999 is taken as $X=0$ and units of X are 1 year. The trend values are obtained by substituting the values of X corresponding to various years into the equation, and are shown in the last column of the table given above. The total of the trend values agree with the total of the original values. The original values and the trend line are graphed on next page.



Example 13.6 Use the method of least-squares to determine the linear trend line for the data in Example 13.1. Compute the trend values and make an estimate of the number of bags of fertilizers sold in 1st and 2nd quarters of 2005.

Let the equation of the linear trend line be $\hat{Y}_t = a + bX_t$. Since the number of quarters in the observed series is even, therefore the middle point of the two middle quarters, i.e. the middle of quarter IV of 2002 and quarter I of 2003, is taken as $X_t = 0$. We then assign $X_t = -1, -3, -5, \dots$, to the preceding quarters and $X_t = 1, 3, 5, \dots$, to the following quarters. For computing the values of a and b in the trend line, the necessary calculations are shown in the table below:

"Computation for linear trend with even number of values".

Year by quarter	Coded quarter (X_t)	t	$X_t Y_t$	X_t^2	Trend $\hat{Y}_t = 119.56 + 2.76X_t$
2001-I	-15	72	-1080	225	78.2
II	-13	98	-1274	169	83.7
III	-11	79	-889	121	89.2
IV	-9	106	-954	81	94.7
2002-I	-7	79	-593	49	100.2
II	-5	122	-610	25	105.8
III	-3	101	-303	9	111.3
IV	-1	143	-143	1	116.8
2003-I	+1	94	+94	1	122.3
II	3	141	423	9	127.8
III	5	128	640	25	133.4
IV	7	160	1120	49	138.9
2004-I	9	125	1125	81	144.4
II	11	143	1573	121	149.4
III	13	135	1755	169	155.4
IV	15	187	2805	225	161.0
Σ	0	1913	3749	1360	1913.0

Since $\sum Y = 0$, the normal equations take the form

$$\sum Y = a + b \sum X^2$$

Substituting the values, we obtain

$$a = \frac{\sum Y}{n} = \frac{1913}{10} = 119.56, \text{ and } b = \frac{\sum XY}{\sum X^2} = \frac{3749}{3360} = 2.76.$$

Thus the least squares trend line is $Y_t = 119.56 + 2.76X$, with origin at the middle of IV quarter 2002 and I quarter 2003, and X is measured in half quarter units. In the equation, $b=2.76$ means the trend line rises by 2.76 each half quarter as time goes forward. The trend values are computed from trend line and shown in the last column of the table on page 489.

To make an estimate of the number of bags of fertiliser for I and II quarters of 2005, we use coded quarter numbers which are $X = 17$ and 19. Substituting these values of X in the equation of trend line, we get

$$\begin{aligned}\text{I quarter 2005: } Y_t &= 119.56 + 2.76(17) \\ &= 166.5 \text{ (hundred bags) and} \\ \text{II quarter 2005: } Y_t &= 119.56 + 2.76(19) \\ &= 172.0 \text{ (hundred bags).}\end{aligned}$$

Example 13.7 Fit a second degree trend curve (parabola) to the following data and compute trend values.

Year	1931	1933	1935	1937	1939	1941	1943	1945
Index of Wholesale Price	96	87	91	102	108	129	307	289

(P.U., B.A.I.E.S., 2005)

Let X and Y denote respectively the coded year number and the index of wholesale prices. The given number of years is given and units of X are 2, therefore, we can assign $X = 0$ to the year 1935, $X = 1, 3, 5, 7$ to the following years and $X = -1, -3, -5, -7$ to the preceding years. Let the equation of the trend curve of second degree fitting the data be

$$Y_t = a + bX + cX^2,$$

where a, b and c are to be computed from the data by the least-squares method.

Since $\sum Y = 0 = \sum X^3$, the normal equation obtained by the method of least-squares, will be

$$\sum Y = a + b \sum X,$$

$$\sum XY = b \sum X^2$$

$$\sum X^2 Y = a \sum X^3 + c \sum X^4$$

The arithmetic involved in computation is arranged in the following table:

Year	X	T	XY	X^2	$X^3 T$	X^4	X^5	Trend
1931	-7	96	-672	49	4704	-343	2401	100.3
1933	-5	87	-435	25	2175	-125	625	83.0
1935	-3	91	-273	9	819	-27	81	81.8
1937	-1	102	-102	1	102	-1	1	96.7
1939	+1	108	-108	1	108	+1	1	127.7
1941	+3	139	-417	9	1251	27	81	174.7
1943	+5	107	-535	25	7675	125	625	237.8
1945	+7	289	-2023	49	14161	343	2401	317.0
Total	0	1219	2601	168	30995	0	6216	1219.0

Substituting these values in the normal equations, we get

$$1219 = 8a + 168c$$

$$2601 = 168a$$

$$30995 = 168a + 6216c$$

Solving these equations, we obtain

$$a = 110.16, b = 15.48, \text{ and } c = -0.75$$

Hence the equation of the second degree trend or the cubic parabola is

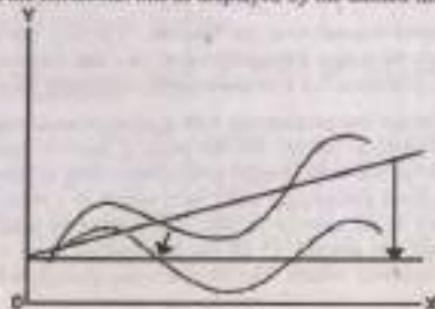
$$T_c = 110.16 + 15.48X + 2.01X^3$$

at the origin $X = 0$ is the year 1938, and X is measured in 1 year units.

The computed trend values are shown in the last column of the table shown above. The total of the values agree with the total of the original values.

DETRENDING

After the determination of a trend by any method, the next step in the analysis of a time series, is to remove its effect from the observed time series. The process of removing the trend component is called **detrending**. A detrended time series is also known as a **stationary time series**. The data in a detrended series would fluctuate around the horizontal line as displayed by the dashed line in the following figure:



The trend component is removed by computing either the deviations from the trend or the ratios to trend depending on whether the components follow the additive or the multiplicative model. When the relationship is additive, we subtract the trend values from the corresponding original observations. If, on the other hand, the model is multiplicative, we divide each of the original observations by the corresponding trend value. The ratios so obtained, are usually expressed as percentages.

13.6 ANALYSING THE SEASONAL VARIATIONS

Having removed the trend component from a time series, we are left with deviations or ratios which are averaged for each season or month to measure the seasonal variation. The deviations or ratios, usually called the seasonal differences with the ratios expressed as percentages, may be called seasonal relatives. That is

$$\text{Seasonal relative} = \frac{\text{original Y - observations}}{\text{corresponding trend value}} \times 100.$$

A measure of variation which is usually computed in index form, is called a seasonal index. To compute seasonal indices, the components of the time series are assumed to follow the multiplicative (decomposition) model.

It is relevant to note that simple averages of the monthly or quarterly values over a period of years are known as the seasonal averages. When these seasonal averages are expressed as percentages of the average of all the seasonal averages, i.e. grand average, they are called the seasonal indices. In symbols,

$$\text{Seasonal Index} = \frac{\text{seasonal average}}{\text{grand average}} \times 100.$$

A seasonal index may also be computed from weekly or daily data. Assuming that the time series components follow the multiplicative (decomposition) model $T = TSCF$, discussed earlier, we give here the various methods available for computing a seasonal index.

- i) The Percentage-of-Annual-Average Method.
- ii) The Ratio-to-Moving-Average method.
- iii) The Ratio-to-Trend Method.
- iv) The Link-Relative Method.

Let us now discuss each of these methods in turn.

13.6.1 The Percentage-of-Annual-Average Method. The first step is to eliminate the effect of the trend. To this end, compute the simple averages for each year and divide each of the given monthly or quarterly observations by the corresponding annual averages, expressing the result as a percentage.

The next step is to average the percentages with a view to removing the cyclical and irregular variations and computing the seasonal indices. For this purpose, sort out these percentages by months or quarters and find the monthly or quarterly average percentages using either the mean or the median or the case of mean, discard the extreme percentages under each month or quarter. If the 12 monthly or quarterly average percentages do not average to 100, adjust them by multiplying each of them by a suitable factor that will make the average of all the percentages equal 100. The resulting percentages are the required seasonal indices. The computational procedure is illustrated by the example 13.8 on the next page.

Example 13.8 Compute the seasonal indices for the data in Example 13.1, using the percentage-of-annual-average method.

We find compute the annual-average as below:

Year	Quarters				Annual or yearly	
	I	II	III	IV	Total	Average
2001	72	98	79	106	355	88.75
2002	79	122	101	143	445	111.25
2003	94	141	128	160	523	130.75
2004	125	143	135	167	590	147.50

We then divide each of the given quarterly observations by the corresponding annual-average and express the result as a percentage. The percentages so obtained appear below:

Year	Quarters				Total
	I	II	III	IV	
2001	81.13	116.42	89.01	119.44	
2002	71.01	109.66	90.79	128.22	
2003	71.89	107.84	97.93	121.47	
2004	84.75	96.65	91.53	116.78	
Total	308.78	424.87	398.31	457.17	
Mean	77.20	106.22	99.58	114.29	400.01

As the total of the average percentages is not equal to the desired total of 400, no adjustment is made. Hence the 4 mean percentages are the required seasonal indices.

13.6.2 The Ratio-to-Moving-Average Method. This is the most frequently used method for theputation of seasonal index numbers. The first step is to eliminate the trend component. To this end, divide the original observations for each month or quarter by the corresponding 12-month or 4-quarter moving average and express the result as a percentage, i.e. compute the seasonal relative for each month or quarter. It is worthwhile to note that each monthly or quarterly value is assumed to consist of product of the effects of T, C, S and I components, and each moving average is a measure of the combined effect of trend and cyclical components, i.e. $T \times C$. Thus dividing the original data by the corresponding moving averages and then multiplying by 100, an estimate of the effect of seasonal and irregular components combined is obtained; that is

$$\frac{\text{original data}}{\text{moving average}} \times 100 = \frac{TCSI}{TC} \times 100 = SI \quad (\text{seasonal relative})$$

The next step is to remove the effects of the irregular component in order to obtain seasonal indices. To achieve this end, arrange the seasonal relatives by months or quarters and find the monthly or quarterly averages, using either the mean or the median. If mean is to be used, compute a modified mean discarding the unusually large or small seasonal relative under each month or quarter so that the range is not disturbed. If these monthly or quarterly averages do not average to 100, then adjust them by multiplying each median or modified mean estimate of seasonal index by the correction factor that will

make the average of all the indices equal 100. The resulting averages give the desired indices of seasonal variations.

Example 13.9 Compute the seasonal indices by using the ratio-to-moving-average method for the data in Example 13.4.

The original data and the 4-quarter centred moving averages which measure the combined effects of trend and cyclical components, appear in columns (2) and (3) respectively in the table below. We then divide each of the original quarterly values by the corresponding centred moving average and express the result as a percentage, i.e. we compute the seasonal relatives, shown in column (4).

Year and quarter (1)	T-values TCSI (2)	4-quarter centred moving average T _C (3)	Seasonal relatives (TCSI/T _C) × 100 = S _t (4)
2001-I	72	—	—
	94	—	—
	79	89.6	88.2
	106	93.5	113.4
2002-I	79	99.2	79.6
	122	106.6	114.4
	101	113.1	89.3
	143	117.4	121.8
2003-I	94	123.1	76.4
	141	126	109.6
	128	128	95.1
	160	138.8	115.3
2004-I	125	139.9	89.3
	143	144.1	99.2
	135	—	—
	181	—	—

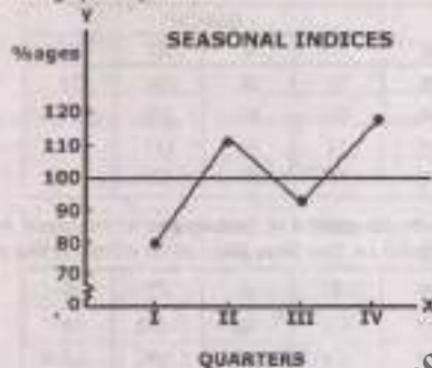
In order to remove the irregular effect and to compute the seasonal index for each quarter, the seasonal relatives are arranged in the following table.

Computation of Seasonal Indices

Year	Quarters				Total
	I	II	III	IV	
2001	—	—	88.2	113.4	
2002	79.6	114.4	89.3	121.8	
2003	76.4	109.6	95.1	115.3	
2004	*89.3	*99.2	—	—	
Total	150.0	224.0	272.8	350.5	
Mean	75.0	112.0	90.87	116.83	397.70
Seasonal Index	78.45	112.65	91.40	117.51	400.01

*Discard these extreme relatives before computing total.

Since the total of the 4 mean percentages is 397.70, we therefore adjust them by multiplying each of the 4 quarterly mean percentages by $400/397.70$ as the sum of the 4-quarterly measures should be 400. Thus the desired seasonal indices are obtained as 78.45, 112.45, 91.40 and 117.51. These indices of seasonal variation are shown graphically below.



13.6.3 The Ratio-to-Trend Method. In this method, the trend values are obtained for each time period by fitting a least-squares trend line either to the observed time series data or to the annual averages. The rest of the computational procedure is the same as that of the ratio-to-moving average method. But this method is inferior to the ratio-to-moving average method as the seasonal index computed by it includes cyclical and irregular variations.

Example 13.10. Compute the indices of seasonal variation by the ratio-to-trend method by fitting a squares straight line trend to the observed time series data in Example 13.1.

We obtain the trend values by fitting a linear trend to the observed time series data (Example 13.6), then divide each observation in the original data by the corresponding trend value and multiply it by 100. The percentages so obtained are arranged by quarters as shown in the following table in order to compute the indices of seasonal variation.

Computation of Seasonal Indices

Year	Quarters				Total
	I	II	III	IV	
2001	*92.1	117.1	88.6	111.9	
2002	78.8	115.3	90.7	*122.4	
2003	76.9	110.3	*96.0	115.2	
2004	86.6	*95.4	86.9	116.1	
Total	242.3	342.7	266.2	343.2	
Mean	60.58	85.68	66.55	85.80	398.13
Seasonal Index	81.15	114.77	89.15	114.94	400.01

*Discard these extreme ratios before computing totals.

The sum of the 4 mean percentages is 398.13. The mean percentages are therefore adjusted by multiplying each of them by the correction factor $400/398.13$ to get the desired sum of 400.

Example 13.11: Obtain the seasonal indices by the ratio-to-trend method and by fitting the least-squares trend line to the annual averages of the following data, showing the amount of the money (£ 000,000) spent upon passenger travel in the United Kingdom at levels of fares, and charges current during the periods given.

Year	Quarters			
	I	II	III	IV
2003	71	89	106	78
2004	71	90	106	79
2005	73	91	111	81
2006	76	97	122	89

We first fit a straight line by the method of least-squares to the annual averages (\bar{Y}), which are assumed to correspond to the midpoint i.e. June 30 or July 1 of the corresponding year.

Year	X	\bar{Y}	\bar{XY}	\bar{X}^2
2003	-3	86	-258	9
2004	-1	87	-87	1
2005	+1	89	89	1
2006	+3	96	288	9
Σ	--	358	-32	20

The equation of the straight line (Linear trend) is $\hat{Y} = a + bX$. Since there is an even number of years, the origin is taken at December 31, 2004/January 1, 2005. The normal equations then reduce to

$$\sum Y = a + b \sum X \quad \text{and} \quad \sum XY = b \sum X^2$$

Substituting the values in the normal equations and solving them, we get $a = 89.5$ and $b = 1.6$.

Thus the required trend is $\hat{Y} = 89.5 + 1.6X$, where X is measured in half years.

This line shows that the values of \hat{Y} increase by 1.6 after every half year or $\frac{1.6}{2} = 0.8$ after one quarter. Assuming that the given quarterly data correspond to the middle of the quarter, we calculate the trend values as below:

When $X = 0$ which corresponds to January 1, 2005; $\hat{Y} = 89.5$

But we need the values of \hat{Y} a half quarter later. Thus

$$\hat{Y} = 89.5 + \frac{1}{2}(0.8) = 89.9$$

This is the trend value corresponding to first quarter of 2005. Now, by successive addition of 0.8, the trend values for the 2nd, 3rd and 4th quarters of 2005 and the quarters of 2006 are worked out, while by successively subtracting 0.8 from 89.9, the trend values for the preceding quarters are found. The quarterly trend values thus found are given on the next page.

Year	Quarters			
	I	II	III	IV
2003	83.5	84.3	85.1	85.9
2004	86.7	87.5	88.3	89.1
2005	89.9	90.7	91.5	92.3
2006	93.1	93.9	94.7	95.5

Dividing each of the actual values by the corresponding trend value and expressing the result as a percentage, we obtain:

Year	Quarters				Total
	I	II	III	IV	
2003	85.0	105.6	124.6	90.8	
2004	81.9	102.3	122.3	88.6	
2005	81.2	100.3	121.3	87.8	
2006	81.6	103.3	128.8	93.2	
Total	329.7	412.0	497.6	364.4	1599.7
Mean	82.4	103.0	124.2	90.1	399.7

The sum of these mean percentages is 199.7, which is so close to the desired 400 that no adjustment is necessary. Hence the desired seasonal indices are 82.4, 103.0, 124.2 and 90.1. Median can also be used to get the seasonal indices.

13.6.4 The Link-Relative Method. This method was at one time the most widely used method as the data for each month or quarter were utilized more completely. But nowadays it is seldom used as its disadvantages outweigh its advantages.

To eliminate the trend component, the computational steps are as follows:

- Compute the link-relatives by expressing each monthly or quarterly value as a percentage of the preceding month or quarterly value.
- Arrange the link-relatives by months or quarters and find an appropriate average of these relatives for each month or quarter. Usually median is used.
- Convert the average (median or mean) relatives into a series of chain relatives by setting the value of January or the first quarter as 100, and carrying the process to include the first unit of the next period.
- A discrepancy due to trend movement (positive or negative) exists between the chain relative for the first January or quarter and that for the next period. Adjust the chain relatives for the trend component by subtracting one-twelfth of the discrepancy from the value of February, two-twelfth from the value of March and so on or by subtracting one-fourth of the discrepancy from the relative of second quarter, two-fourth from the third quarter relative and three-fourth from the fourth quarter relative.

To obtain seasonal indices, reduce the adjusted chain relatives to the same level at January or the first quarter by multiplying each of the adjusted chain relatives by the correction factor that will make the average of all the indices equal 100. These final figures are the desired indices of seasonal variation.

Example 13.12 Obtain the seasonal indices for the data in Example 13.11, using the link-relative method.

Expressing the data for each quarter as a percentage of the data for the preceding quarter, we get the link-relatives as below:

Year	Quarters			
	I	II	III	IV
2003	—	125.4	119.1	73.6
2004	91.0	126.8	120.0	73.1
2005	92.4	124.7	122.0	73.0
2006	93.8	127.6	125.8	73.0
Median	92.4	126.1	121.0	73.1

* Next, we calculate the chain relatives for the four quarterly averages, setting the value of the first quarterly average equal to 100%. The chain relatives are:

Quarter	I	II	III	IV	I
Chain Relative	100	126.1	152.6	111.6	103.1

Continuing the process, the chain relative for the first quarter works out to be 103.1 which as a matter of fact, ought to have been 100. This increase of 3.1% due to the trend component present in the data. An adjustment for the trend therefore becomes necessary. Since the difference is positive so we subtract one-fourth of this from the second quarter figures, two-fourths from the third quarter figures and three-fourths from the fourth quarter. The sum of the adjusted chain relatives is 485.65. The quarterly figures are further adjusted by multiplying each figure by $\frac{400}{485.65}$ so as to get a total of 400. The adjusted figures are given below:

Quarter	I	II	III	IV	Total
Adjusted Chain Relatives	100	125.32	151.05	109.28	485.65
Seasonal Index	82.4	103.2	124.4	90.0	400.0

13.7 DESEASONALIZATION OF DATA

We remove the seasonal effect from an observed time series data to see how things might have been, if there had been no seasonal component. The process of removing the seasonal component from a time series is known as deseasonalization or seasonal adjustment of data and the time series thus obtained is called the deseasonalized or seasonally adjusted time series. To get deseasonalized data, we divide (considering multiplicative model) each value in the original data by the corresponding value of seasonal index and multiply the result by 100. Thus

$$\begin{aligned}\text{Deseasonalized data} &= \frac{\text{period's original value}}{\text{period's seasonal index}} \times 100 \\ &= \frac{\text{TCSI}}{S} \times 100 = TCI \times 100\end{aligned}$$

Thus the deseasonalized data contain the effects of trend, cyclical and irregular components. For example, the deseasonalized data for 2002 of Example 13.9 are shown below:

Quarter	Number of bags of fertilizer (00)	Seasonal Index	Deseasonalized Data
I	79	78.45	101
II	122	112.65	108
III	101	91.40	111
IV	143	117.51	122

We find that the increase from the first quarter to the second quarter of 2002, expected on the basis of seasonal pattern is $\left(79 \times \frac{112.65}{78.45} - 79\right)$, i.e. 34 hundred bags which is less than actual increase of (122-79), i.e. 43 hundred bags. From the second quarter to the third quarter, there is a decrease of (122-101), i.e. 21 hundred bags which is less than the decrease expected on the basis of seasonal pattern amounting to $\left(122 \times \frac{91.40}{112.65} - 122\right)$, i.e. 23 hundred bags of fertilizer. If there had been no seasonal effect, sales for the first quarter of 2002, would have been 101 (hundred) bags of fertilizer.

When the time series components follow the additive model, i.e. $Y=T+C+S+I$, the data are deseasonalized by subtracting the seasonal effects from the corresponding original values.

13.8 ANALYSING THE CYCLICAL VARIATIONS

The cyclical variations can be measured after removing the trend and seasonal components by division and then averaging out irregular variations. The simplest method of obtaining cyclical movement is called the *residual method*. This method consists of removing the effects of trend, seasonal and irregular components from the observed time series data in any order. Any one of the following three procedures can be used to estimate cyclical and irregular movements:

First Procedure:

- Deseasonalize the data, i.e. divide each value of the original data by the corresponding seasonal index to remove the seasonal component: $TCSI + S = TC$.
- Divide the results just obtained, i.e. the deseasonalized data by the corresponding trend value to eliminate the trend component: $TCI + T = CI$.

Second Procedure:

- Remove the trend component by dividing each value of the original data by the corresponding trend value: $TCSI + Y = CSI$.
- Eliminate seasonal variation by dividing the results, i.e. detrended data by the corresponding seasonal index: $CSI + S = CI$.

Third Procedure:

- Multiply each trend value by the corresponding value of the seasonal index to get $T \times S$ values.
- Eliminate trend and seasonal components by dividing each value of original data by the corresponding $T \times S$ value obtained in (i): $TCSI + TS = CI$.

All these three procedures give the same results. In order to remove the irregular variations, if any, take an appropriate moving average of a few months or quarters duration. The resulting quantities in percentage form are called *cyclical relatives or percentages*. Do not divide by t .

Another method for isolating cyclical movements is the *Harmonic analysis*, which is beyond the scope of this text and hence is not discussed. A study of cyclical relatives is useful for economic forecasting.

Example 13.13 Compute the cyclical relatives for the data in Example 13.1.

The process of computing the cyclical-irregular movements and cyclical relatives is shown in the table below. To remove the irregular variations, a three-quarter moving average has been thought appropriate.

Year and quarter	Y-values TSCT	Trend Values T	Seasonal Index (S%)	Trend x Seasonal TSx100 (S)	Cyclical-irregular percentages C% (C)	3-quarter moving Total (T)	Cyclical Relatives C% (C)	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
2001-I	72	78.2	81.15	63.46	113.46	—	—	
	II	98	83.7	114.77	96.06	192.03	104.94	
	III	79	89.2	89.15	79.52	89.4	99.58	
	IV	106	94.7	114.94	108.85	173.38	97.96	
2002-I	79	100.2	81.15	81.15	97.16	295.01	98.34	
	II	122	105.8	114.77	100.47	299.42	99.81	
	III	101	111.3	89.15	99.22	101.79	102.95	
	IV	143	116.8	114.77	134.25	106.32	101.01	
2003-I	94	122.3	81.15	99.25	94.71	297.36	99.12	
	II	141	122.3	114.94	146.68	96.13	298.47	99.49
	III	128	128.4	89.15	118.93	107.63	303.98	101.33
	IV	166	138.9	114.77	159.65	100.23	314.52	104.84
2004-I	125	144.4	81.15	117.18	106.67	290.01	96.67	
	II	143	149.9	114.77	172.68	83.12	287.23	95.74
	III	135	155.4	89.15	138.54	97.44	281.61	93.87
	IV	187	161.0	114.94	185.05	161.85	—	—

13.9 ANALYSING THE IRREGULAR VARIATIONS

The irregular movements of a time series are estimated by dividing the combined cyclical-irregular variations by the corresponding values of the cyclical relatives; that is

$$I = \frac{C \times I}{C}$$

The irregular movements can be shown graphically.

13.10 FORECASTING

In time series analysis, forecasting is a process of assessing the magnitude of a time-series variable which it will assume at some future point of time. Forecasting is based on the assumption that the past pattern and behaviour of a variable will continue in the future. The simple and elementary technique of short-term forecasting involves the components of trend and seasonal index.

Forecasts for a period of 1 year or less can be made by projecting either the least-squares equation to obtain forecasts of trend values (T) or the centred moving averages to obtain forecasts of moving average values ($T \times C$), then multiplying these projected values by the seasonal index for that period and dividing by 100 to obtain a $T \times S$ or $T \times C \times S$ forecast. Thus

$$\text{period's forecast} = \frac{(\text{period's projected trend value}) \times (\text{seasonal index})}{100}$$

It is relevant to note that it is very difficult to forecast cyclical and irregular movements.

13.10.1 Forecasting by Exponential Smoothing. The exponential smoothing is a method of forecasting that assigns positive weights to past and current values only. This technique often provides good short-term forecasts. The exponentially smoothed series \hat{Y}_t from the original time series Y_t is calculated as follows:

$$\begin{aligned}\hat{Y}_1 &= Y_1 \\ \hat{Y}_2 &= w Y_1 + (1-w) \hat{Y}_{1,1} \\ \hat{Y}_3 &= w Y_2 + (1-w) \hat{Y}_{2,1} \\ &\vdots \\ \hat{Y}_t &= w Y_{t-1} + (1-w) \hat{Y}_{t-1,1}\end{aligned}$$

where the weight w , called the *exponential smoothing constant*, is selected so that w is between 0 and 1. The most commonly used value of w is between 0.01 and 0.3. The exponential smoothing has an advantage that no values are lost at either end of the smoothed series.

11 SERIAL CORRELATION

While analyzing a time series data, there is a possibility of dependence (or association) between the successive observations. In case, the successive observations are dependent, one measure of this effect is simple correlation between successive observations. Such a correlation is called a *serial correlation*.

Generally, a serial correlation is defined as correlation between observations ordered in time. Given n observations Y_1, Y_2, \dots, Y_n made over successive time periods, we change the observations into $(n-1)$ pairs such as $(Y_1, Y_2), (Y_2, Y_3), \dots, (Y_{n-1}, Y_n)$. If we regard the first observation in

each pair as (Y_i, Y_{i+k}) , then the other observation is Y_{i+1} , where k can be 1, 2, 3, etc. The correlation between Y_i and Y_{i+1} , i.e. the correlation between successive overlapping pairs, is called the *serial correlation of first order*. The co-efficient of first serial correlation, denoted by r_1 , is generally calculated by the following slightly modified formula

$$r_1 = \frac{\sum_{i=1}^{n-1} (Y_i - \bar{Y})(Y_{i+1} - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$\text{where } \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

This is also called the co-efficient of auto-correlation at lag 1. The terms serial correlation and autocorrelation are used interchangeably.

Likewise, we can find the correlation coefficient between observations, separated by a lag of k time periods, which is given by

$$r_k = \frac{\sum_{i=1}^{n-k} (Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

This is called the coefficient of autocorrelation at lag k or serial correlation of order k .

We can also draw a scatter diagram by plotting the pairs (Y_i, Y_{i+1}) on a graph paper to see whether the successive observations appear to be correlated.

Example 13.14 Sixteen successive observations on a stationary time series are as follows:

1.6, 0.8, 1.2, 0.3, 0.9, 1.1, 1.1, 0.6, 1.5, 0.8, 0.9, 1.2, 0.5, 1.3, 0.6, 1.2

Calculate r_1 , the first serial correlation co-efficient.

The co-efficient of serial correlation of order 1 is given by

$$r_1 = \frac{\sum_{i=1}^{n-1} (Y_i - \bar{Y})(Y_{i+1} - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$\text{where } \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i / n = \frac{16.0}{16} = 1.0$$

The calculations needed to compute r_1 are shown below:

y_i	y_{i+1}	$y_i - \bar{Y}$	$y_{i+1} - \bar{Y}$	$(y_i - \bar{Y})(y_{i+1} - \bar{Y})$	$(y_i - \bar{Y})^2$
1.6	0.8	0.6	-0.2	-0.12	0.36
0.8	1.2	-0.2	0.2	-0.04	0.04
1.2	0.5	0.2	-0.5	-0.10	0.04
0.5	0.9	-0.5	-0.1	+0.05	0.25
0.9	1.1	-0.1	0.1	-0.01	0.01
1.1	1.1	0.1	0.1	-0.01	0.01
1.1	0.6	0.1	-0.4	-0.04	0.01
0.6	1.3	-0.4	0.5	-0.20	0.16
1.5	0.8	0.5	-0.2	-0.10	0.25
0.8	0.9	-0.2	-0.1	+0.02	0.04
0.9	1.2	-0.1	0.2	-0.02	0.01
1.2	0.5	0.2	-0.5	-0.10	0.04
0.5	1.3	-0.5	0.3	-0.06	0.25
1.3	0.8	0.3	-0.2	-0.06	0.09
0.8	1.2	-0.2	0.2	0.04	0.04
1.2	—	0.2	—	-0.08+0.08	0.04
16.0	—	—	—	-0.90	1.64

Substituting the values in the formula, we get

$$r_1 = \frac{-0.59}{1.64} = -0.36$$

Hence the serial correlation of ^{6th} order between the successive observations is found to be -0.55.

EXERCISES

OBJECTIVE

Answer 'True' or 'False'. If the statement is not true then replace the underlined words with words that make the statement true:

- i) Secular trend measures the short-term variation of a time series.
- ii) A typical time series may be regarded as composed of five components.
- iii) A secular trend is mainly caused by the change in seasons.
- iv) Irregular variations can be predicted in time.
- v) A histogram is a graph of time series.
- vi) The addition model of a time series is $Y = TSCI$.

- vii) A main objective of fitting trend lines is to forecast critical turning points.
- viii) Seasonal indexes can be calculated from weekly, monthly and quarterly data only.
- ix) If the yearly sales for year 2007 is Rs.5,00,000/- and the sales index for year 2007 is 60, then the seasonally adjusted sales figure is Rs.4,00,000/-.
- x) Yearly time series contain the following four components: trend, cyclical, seasonal and irregular.

b) MULTIPLE CHOICE QUESTIONS

- i) Decomposing a time series means that past data is distributed into components of:
 - a) Trend, cycles, seasonal and random
 - b) Long term, medium term and short term variations
 - c) Constants and variations
 - d) All of above
- ii) The seasonal variation in the time series is computed by:
 - a) Ratio to moving average method
 - b) Ratio to trend method
 - c) Link relative method
 - d) All of above
- iii) The seasonal variation in the time series is computed by:
 - a) Trend
 - b) Cyclical
 - c) Seasonal
 - d) Irregular
- iv) After detrending the data, the time series (multiplicative model) consists of:
 - a) $Y = TSCI$
 - b) $Y = TSI$
 - c) $Y = CSI$
 - d) None of above
- v) If a time series changes at exact constant percentage then:
 - a) A good fitted trend line cannot be obtained
 - b) A linear line fitted to the data gives a perfect fit
 - c) A linear line fitted to the logarithmic data gives a perfect fit
 - d) A nonlinear is required to be fitted

- vi) Dividing the original time series by moving average, the time series (multiplicative model) consists of
- $Y = S I$
 - $Y = C S$
 - $Y = T S$
 - $Y = T C$
- vii) A company's trend figure for sales for December 2007 is Rs. 200,000. Actual sales during that period were Rs. 1,60,000/- and $C \times I = 0.80$. The value for seasonal index is
- 100
 - 125
 - 60
 - 64
- viii) A second degree trend line is: $\hat{Y} = 15 - 0.11 + 0.05t^2$ where t is sales (in thousands) and t is time (in years) and $t = 1$ for 1995. What is the predicted figure for year 2007?
- 20,000
 - 22,150
 - 18,000
 - 15,000
- ix) If a 4-quarter moving average is projected to obtain short-term forecasts, it contains the following components
- TC
 - TS
 - CSI
 - C only
- x) Exponential smoothing is a forecasting method which
- uses the actual data, not the forecast data
 - requires to fit a mathematical model to the data
 - gives equal weight to all the periods
 - all of above

OBJECTIVE

- a) Define a time series. What are its various components? Describe each carefully.

- b) Associate the following phenomena of business with the components of time series they belong to:
- The prosperity in a Business.
 - The production of sugar recorded for 1956, 1957, ..., 1962.
 - The weekly statement of the sale of pens.
 - The festival sale.
 - The fire in a factory.

(P.U., B.A./B.Sc. 1976)

13.2 a) Define the following terms:

- Time Series Analysis, (ii) Secular trend, (iii) Seasonal variations, (iv) Cyclical fluctuations (v) Irregular movements.
- With which characteristic movement of a time series would you mainly associate each of the following?
 - a fire in a factory delaying production for 3 weeks,
 - an era of prosperity,
 - an after Eid sale in a departmental store,
 - a need for increased wheat production due to constant increase in population,
 - the monthly number of inches of rainfall in a city over a 5-year period,
 - a recession,
 - an increase in employment during summer months,
 - the decline in the death rate due to advance in science,
 - a steel strike,
 - a continually increasing demand for smaller automobiles.

(P.U., B.A. (Part II), 1956, 1962)

13.3 Describe the different components of a time series. Describe various methods of measuring Secular Trend in a time series, giving the advantages and disadvantages for each.

(P.U., B.A./B.Sc. 1961, 63)

13.4 Describe the different components of time series. Discuss the measurement techniques of any one of them.

(P.U., M.A. (Econ.), 1960)

13.5 What do you understand by Time Series Analysis? Discuss how you would analyse a time series to determine the trend and the seasonal variation.

13.6 What is a time series? What are its main components and how will you isolate them?

(P.U., B.A./B.Sc. 1961, 63)

13.7 What is meant by seasonal variation? Explain how seasonal variations are measured and removed from the time series data?

- a) Distinguish between the Additive and Multiplicative models in time series analysis.
 b) When do you compute the deviations from trend and what ratios to trend? Explain how you eliminate the average seasonal variations from the observed values of the time series.

- 13.9 Critically comment on various methods of eliminating seasonal variations from a time series.
(P.U., M.A. (Stat.), 1968)

- 13.10 Plot the following data showing average wages in rupees of some workers during 1990-2001 and find the trend by the method of (i) free-hand curve and (ii) semi-averages.

Year	1990,	1991,	1992,	1993,	1994,	1995,	1996,	1997,	1998,	1999,	2000,	2001
Wages	140,	146	150	155,	200,	235,	260,	280,	290,	330,	325,	340

- 13.11 Name the methods used to measure trends. Determine a trend line by a simple moving average of 3 years from the following data:

Year	1921,	1922,	1923,	1924,	1925,	1926,	1927,	1928,	1929,	1930
Value	102	108	130	140	158	180	196	210	220	230

(P.U., B.A./B.Sc. 1963)

- 13.12 i) Explain briefly the meaning and purpose of moving averages.
ii) The number of items of certain product imported into the United Kingdom is given below in thousands of units:

Year	Number	Year	Number	Year	Number
1951	170	1957	205	1963	135
1952	210	1958	180	1964	80
1953	188	1959	90	1965	60
1954	96	1960	92	1966	107
1955	85	1961	141	1967	140
1956	131	1962	183	1968	124

- a) Calculate 5-year moving averages for the above data. Using these moving averages, determine the trend line.
b) Estimate the number of items imported in 1969. (P.C.S., 1971)

- 13.13 a) What is a Moving Average? Calculate a seven-day moving average for the following record of attendances:

Week	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
1	24	55	29	48	52	55	61
2	27	52	32	43	53	56	65

Plot the given attendances and the moving averages on the same graph.
(B.Z.U., B.A./B.Sc. 1988)

- b) Fit a linear trend to the data given in (a) by least squares. Find trend values as well.
(P.U., B.A./B.Sc. 1998)

- 13.14 Explain the use of moving averages in determining the trend line in a time series. Determine such a line in the following series of values by the use of a simple average of seven consecutive terms:

Year:	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Index:	187	161	149	142	125	129	133	127	130	129	129
Year:	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	
Index:	130	136	152	171	169	218	258	279	295	314	

Would you say that this was a satisfactory line of a trend? If not, why not?

- 13.15 a) Find out and plot the nine-year moving average for the following series:

8, 7, 5, 2, 4, 9, 10, 9, 8, 6, 4, 7, 11, 13, 11, 9, 8, 5, 10, 13, 15, 12, 10, 8, 6, 11, 12, 16.

- b) Compute 4-month centred moving average from the following:

23, 26, 28, 30, 31, 35, 37, 32, 34, 38.

(P.U., B.A. (Hons.), 1987)

- 13.16 The following are the quarterly index numbers of wholesale prices in the U.K. for the years 1951–55:

86, 80, 83, 84, 85, 80, 80, 78, 77, 80, 81, 80, 82, 81, 83, 82, 84, 85, 86.

By a centred moving average of 4, calculate the trend.

(P.M.A. Exam. 1969; B.A./B.Sc. 1973)

- 13.17 Plot the following data as a time series. Compute 4-quarter centred moving average trend and show it on the graph.

Year	Quarters			
	1	2	3	4
2000		71	47	98
2001	25	106	73	231
2002	281	229	200	488
2003	484	447	457	966

- 13.18 The estimated number of visitors ('000) at a holiday resort were as follows:

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Visitors	31	49	74	62	65	73	70	84	86	79

- a) Show by direct numerical calculation that the 2-year centred moving average is equivalent to a 3 year weighted moving average with weights 1, 2, 1 respectively.
 b) Determine a 3 year weighted average if the weights 1, 4, 1 are used.

- 13.19 For the following time series, determine the trend by using the method of (i) semi-averages, (ii) 3-year moving averages, and (iii) least-squares for fitting a straight line:

Year	1968	1969	1970	1971	1972	1973	1974	1975	1976
Values of series	2	4	6	8	7	6	5	10	12

Which of the trend do you prefer, and why?

(P.U., B.A./B.Sc. 1977)

- 13.20-a) Define the following terms:

Secular Trend; Time series Decomposition; Centred Moving Average; Irregular Movements.

- b) For the following time series, determine the trend by using the method of (i) three year moving averages, and (ii) least-squares for fitting a straight line.

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978
Values	12	23	37	48	41	33	48	61	70

Which trend do you prefer, and why?

(P.U., B.A./B.Sc. 1978)

- 13.21-a) The following are the quarterly index numbers of wholesale prices:

Year	1995	1996	1997	1998	1999
Quarterly Index	125	114	99	80	80

Fit a linear trend to these data and add the trend to the original data.

- b) Fit a straight line $T = a + bU$ from the following results, for the years 1948-58 (both inclusive).

$$\sum X = 0, \sum T = 438.9, \sum X^2 = 110, \sum XT = -50.$$

Find out the trend values of T as well.

(P.U., M.A. (Econ.), 1968)

- 12) The following are the annual profits in thousand rupees in a certain business:

Year	1997	1998	1999	2000	2001	2002	2003
Profit	68	101	108	91	113	120	132

- i) Use the method of least-squares to fit a straight line trend and make an estimate of the profits in 2005.

- ii) Fit a parabolic trend.

- iii) Determine which is the better fitting trend.

- 13) The production of vegetables ghee ('000s tons) in Pakistan is given below:

Year	Production	Year	Production
1970-71	136	1974-75	272
1971-72	162	1975-76	277
1972-73	187	1976-77	322
1973-74	225		

Fit a second degree parabola to the data and estimate the production for 1978-79.

(P.U., B.A./B.Sc. 1979)

13.24 Fit a parabola of second order to the following data and find out the trend values.

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Production (in 000 units)	17	20	19	26	24	40	35	35	51	74	65

13.25 Fit a second degree curve $y = a_0 + a_1t + a_2t^2$ to the following data:

Year	2001	2002	2003	2004	2005	2006	2007	2008
Profit	273.7	293.5	315.0	336.8	364.4	394.8	424.2	458.7

Compute the trend values.

13.26 Fit a quadratic parabola to the following series of observations, taking the year as the independent variable.

Year	1924	1927	1930	1933	1936	1939	1942
Index of coal price	187	142	133	120	136	169	279

Use your results to estimate the value of the index for 1935.

(P.U., B.A. (Hons.) Part-I, 1997)

13.27 The population of a country for the years 1911 to 1971 in ten yearly intervals in millions, is 5.38, 7.22, 9.11, 12.70, 17.80, 24.02, and 31.34.

Fit a curve of the type $y = ab^t$ to these data and forecast the population for the year 1991.

(P.U., B.A./B.Sc., 1997)

13.28 a) Define the following terms:

(i) Detrending, (ii) Seasonal variation, (iii) Seasonal Index, (iv) Decaannualization.

b) Compute the seasonal indices for the four quarters by the method of ratio to moving averages.

Year	Quarters			
	1	2	3	4
1949	105	77	68	95
1950	107	83	74	106
1951	117	99	86	112

(P.U., B.A./B.Sc., 1997)

13.29 Compute the seasonal indices for the four quarters by the ratio-to-moving average method from the following data of wholesale prices:

Year	Quarters			
	I	II	III	IV
1961	122	125	118	117
1962	119	114	114	109
1963	105	99	93	89
1964	86	80	83	84

(P.U., B.A./B.Sc. 1973, 80-S)

13.30 A merchant's sale ('00s tons) of ordinary coal over a period were as shown below:

Year	Quarters			
	I	II	III	IV
1996	118	87	47	53
1997	94	73	41	68
1998	73	61	36	56

- a) Construct seasonal indices, using the percentage centred moving-average method.
- b) Construct quarterly seasonal index values by the ratio-to-moving average method and use them to deseasonalize the 1997 values.

13.31 Quarterly sales of a certain fertilizer over the period 1994-98 in thousand of tons, were as follows:

Quarter	Year				
	1994	1995	1996	1997	1998
1st		50	68	93	84
2nd	52	46	34	56	61
3rd	16	22	26	16	29
4th	35	40	35	45	48

By means of centred moving-averages, compute the trend and estimate the seasonal indices, and hence forecast sales for each quarter of 1999.

13.32 Compute seasonal indices for the four quarters by the ratio-to-trend method for the data in question 13.29.

13.33 Compute the indices of seasonal variation by ratio-to-trend method from the following data:

Year	Summer	Autumn	Winter	Spring
1982	70	81	89	115
1983	75	93	108	152
1984	80	105	150	195

Use the seasonal indices to deseasonalize the 1984 values.

13.34 Construct seasonal indices by ratio-to-trend method from the following data:

Quarter	Year			
	2000	2001	2002	2003
I	107	118	126	121
II	115	122	129	122
III	103	115	118	116
IV	98	104	107	103

13.35 For the time series data in exercise 13.28(b),

- determine the trend line by the least-squares method;
- assuming the multiplicative model, compute the following:
 - seasonal indices for the four quarters;
 - desasonalized values.

13.36 Calculate the indices of seasonal variation by *link-relative method* from the following data:

Year	Quarters			
	I	II	III	IV
1970	112	123	119	110
1971	119	147	115	
1972	111	142	150	118
1973	128	151	162	125

(P.U., B.A./B.Sc. 1988)

13.37 a) What is forecasting? Write the different methods of forecasting?

- b) The following table gives the numbers of deep freezers sold by a certain company:

Year	Quarters			
	I	II	III	IV
2001	42	45	33	36
2002	38	45	40	65
2003	51	60	65	94

Use the trend equation and the seasonal index to forecast the sales for each quarter of 2004.

- c) Calculate the exponentially smoothed series, using $w=0.1$ and $w=0.3$. Which will produce smoother trend?

13.38 a) Describe the residual method in time series analysis.

- b) Use the data of exercise 13.29 to calculate the cyclical irregular and cyclical relatives.

13.39 The following data pertain to the rainfall in inches in England and Wales:

Months Year	Jan	Feb	Mar	April	May	June	July	Aug	Sep	Oct	Nov	Dec
2000	6.2	1.8	0.9	1.4	3.2	2.3	2.2	3.2	3.4	3.4	2.7	2.1
2001	3.3	1.7	0.5	2.2	1.5	2.5	2.8	3.2	4.2	4.5	6.1	2.8
2002	3.4	3.1	1.3	1.7	3.2	3.2	2.6	2.8	2.5	4.1	0.8	4.1
2003	3.4	3.4	1.5	1.8	3.0	3.4	3.1	5.4	4.9	1.5	6.2	4.0

- a) Determine the trend by the least-squares method.
- b) On the basis of the least-squares line and the multiplicative model, compute the following:
- i) seasonal indices for the twelve months;
 - ii) de-seasonalized values;
 - iii) cyclical and irregular variations.
- 13.40 a) Explain what you understand by serial correlation.
- b) The following noise measurements were recorded at an intersection in time order they were observed:
- 65, 64, 63, 61, 60, 58, 63, 64, 62, 64; 63, 63, 62, 61, 62, 64, 66, 68, 68, 69.
- i) Plot the scatter diagram for the pairs (x_{t-1}, x_t) .
 - ii) Calculate the first serial correlation coefficient r_1 , and the coefficient of auto-correlation of lag 2.

Contents

**ANSWERS TO
EXERCISES**

<https://stat9943.blogspot.com>

ANSWERS TO EXERCISES

Chapter 1, Pp. 11–14

OBJECTIVE

- | | | | |
|---|---|---|---|
| (i) F (are),
(v) F (descriptive),
(ix) F (attribute); | (ii) F (parameter),
(vi) F (population),
(x) T; | (iii) F (statistic),
(vii) F (continuous), | (iv) F (inferential),
(viii) F (discrete). |
|---|---|---|---|

SUBJECTIVE

- | | | | |
|--|---|---|--------------------------------------|
| 1.9 (b) (i) Discrete,
(v) Continuous,
(ix) Continuous, | (ii) Continuous,
(vi) Continuous,
(x) Discrete, | (iii) Discrete,
(vii) Discrete, | (iv) Discrete,
(viii) Continuous, |
| 1.10 (i) Qualitative,
(v) Quantitative, | (ii) Quantitative,
(vi) Quantitative, | (iii) Qualitative,
(vii) Qualitative, | (iv) Quantitative, |
| 1.11 (i) ratio-level,
(iv) ratio-level,
(vii) ordinal-level, | (ii) ordinal-level,
(v) nominal-level,
(viii) ratio or interval | (iii) interval-level,
(vi) ratio level,
(ix) nominal-level, | (x) ratio-level. |
| 1.12 (i) 32.22,
(v) 0.07000, | (ii) 937.1,
(vi) 22.26 | (iii) 0.003599, | (iv) 1.004, |

Chapter 2, Pp. 44–45

OBJECTIVE

- | | | | |
|---|---|--|----------------------|
| (a) (i) F (time series data),
(iv) F (mutually exclusive),
(vii) F (histogram),
(x) F (height),
(xiii) F (1), | (ii) T,
(v) F (can),
(viii) F (cannot),
(xi) F (mid points),
(xiv) F (qualitative), | (iii) F (does),
(vi) F (graphically),
(ix) F (one dimensional),
(xii) F (lowest),
(xv) F (same), | |
| (b) (i) b,
(v) d,
(ix) a, | (ii) d,
(vi) c,
(x) b, | (iii) c,
(vii) d,
(xi) b, | (iv) b,
(viii) d, |

Chapter 3, Pp. 77–86

SUBJECTIVE

- | | | |
|---|--|--|
| (a) (i) T,
(iv) F (ii),
(viii) F (mode),
(x) F (geometric mean),
(xiii) F (negatively), | (ii) F (median),
(v) T,
(viii) F (second),
(x) T,
(xii) F (right), | (iii) T,
(vi) F (mode),
(ix) T,
(xi) T,
(xv) F (negatively), |
|---|--|--|

- | | | | |
|------------|----------|-----------|----------|
| (b) (i) c, | (iii) b, | (vii) c, | (ix) c, |
| (v) c, | (vi) a, | (viii) c, | (xii) b, |
| (ix) b, | (x) d, | | |

SUBJECTIVE

- 3.14 (i) Median or Mode, (ii) Mode, (iii) Mean.
- 3.15 (c) 6.5
- 3.16 (c) 2.13
- 3.17 (i) 18; (ii) 17; (iii) 17.07
- 3.18 $\bar{X} = 111.60$; G.M. = 55.35; H.M. = 28.82. Here G.M. is the best average.
- 3.19 $\bar{X} = 1037.73$; G.M. = 377.2; H.M. = 186.7.
- 3.20 Rs.2.75 per hour; (b) 69.25% marks.
- 3.21 (i) $\bar{X} = \text{Rs.}10.41$; (ii) $\bar{X}_w = \text{Rs.}10.08$
- 3.22 Rs.24.50
- 3.23 (a) 14.07 years; (b) 13.82 years
- 3.24 $\bar{X} = \frac{1}{n+1} \{2^{n+1} - D\}$; G.M. = $2^{\frac{n+1}{2}}$; H.M. = $\frac{n+1}{2 \left(1 - \frac{1}{2^{n+1}}\right)}$
- 3.25 $\bar{X} = \frac{1}{2(n+1)} \{3^{n+1} - 1\}$; G.M. = $3^{\frac{n+1}{2}}$; H.M. = $\frac{n+1}{2 \left(1 - \frac{1}{3^{n+1}}\right)}$
- 3.27 18.15%
- 3.28 (b) H.M. = 40 miles per hour; (c) H.M. = 6.8 miles per hour.
- 3.29 (a) 22.56 k.p.h.; (b) (i) H.M. = 7.66; (ii) G.M. = 29.3%
- 3.30 G.M. = 49.18; H.M. = 48.47.
- 3.31 G.M. = 19.80; H.M. = 16.27.
- 3.32 (i) Median = Rs. 1500; (ii) $\bar{X} = 65.5$; (iii) Median = 18.
- 3.33 Median = 7; $Q_1 = 6\frac{1}{2}$; $Q_3 = 7\frac{1}{2}$; $D_2 = 7\frac{1}{2}$; $P_{34} = 7\frac{1}{2}$.
- 3.34 Mean = 3.78; Median = 3; Mode = 3.
- 3.35 (ii) Median = Rs.9.04.
- 3.36 Median = 48.44; $Q_1 = 34.34$; $Q_3 = 61.45$.

ANSWERS TO EXERCISES

3.37 Median = 67.72 inches; $Q_1 = 65.4"$; $Q_3 = 69.4"$.

3.38 Median = 138.63.

3.39 Mean = 24.96 years; Median = 23.36 years

3.40 Median = 41.4; $Q_1 = 32.46$; $Q_3 = 50.72$.

3.41 Mean = 79.57; Median = 76.77; $Q_1 = 57.9$; $Q_3 = 99.14$.

3.43 (a) Mean = 91.27 mg; Median = 96.58; Mode = 103.33.

(b) $Q_1 = 77.04$ mg; $Q_3 = 109.06$; $D_1 = 54.96$; $P_{50} = 93.68$.

3.44 Assuming a range of Rs 55.00 to Rs 105.00, the frequency distribution would be:

Groups			$f(\%)$	f	
Rs.55.00	and	under	Rs.60.00	4	20
Rs.60.00	and	under	Rs.62.50	11	55
Rs.62.50	and	under	Rs.72.75	10	50
Rs.72.75	and	under	Rs.78.75	16	75
Rs.78.75	and	under	Rs.82.25	10	50
Rs.82.25	and	under	Rs.85.25	10	50
Rs.85.25	and	under	Rs.90.00	15	75
Rs.90.00	and	under	Rs.96.00	10	50
Rs.95.00	and	under	Rs.100.00	10	50
Rs.100.00	and	under	Rs.105.00	5	25
Total			100	500	

The mean is approximately Rs 80.80.

(b) Mean = 146.975; Median = 146.75; Mode = 147.20

46 Mean = 11.19; Median = 11.07; Mode = 11.06.

47 Mode = Rs 32.48; Mean = 32.49.

(c) Mode = 27

48 (a) (i) Median or Mode, (ii) Mean, (iii) Mean or Median, (iv) Weighted Mean, (v) Mean, (vi) Median or Mean, (vii) Mode, (viii) Mean.

Chapter 4, Pp. 116–129

EXERCISE

- | | | |
|---|---|--|
| (b) (i) F (Dispersion),
(iv) T,
(vi) F (relative),
(a) F (standard deviation). | (ii) F (zero),
(v) F (range),
(viii) T, | (iii) T,
(vi) F (square root),
(ix) F (range), |
|---|---|--|

- | | | | |
|------------|----------|----------|-----------|
| (b) (i) c, | (ii) c, | (iii) c, | (iv) d, |
| (v) b, | (vi) b, | (vii) b, | (viii) a, |
| (ix) a, | (x) a, | (xi) c, | (xii) a, |
| (xiii) c, | (xiv) b, | (xv) d. | |

SUBJECTIVE

- 4.4 (b) Q.D. = Rs.9.78
- 4.5 Median = 152 lb; SIQR = 7.5 lb; Mean = 152.75 lb; S.D. = 10.52 lb. Mean of original data = 152.917 lb; S.D. of original data = 10.32 lb.
- 4.6 (b) $\bar{X} = 46.17$; M.D. = 11.28
- 4.7 Group A: Q.D. = 1.285; M.D. = 1.45; Co-efficient of Q.D. = 0.62;
Co-efficient of M.D. = 0.024.
Group B: Q.D. = 1.435; M.D. = 1.60; Co-efficient of Q.D. = 0.612;
Co-efficient of M.D. = 0.026.
- 4.8 (b) $\sigma^2 = 6.85$, $\sigma = 2.62$.
- 4.11 (b) Mean = 32; S = 5; (c) Mean = 74.1, S = 1.56
- 4.12 (c) (i) S = 2; (ii) S = 2; answers of (i) and (ii) coincide because standard deviation remains unaffected if a constant is added.
- 4.15 $\bar{X} = 33.9$ in.; S = 1.507 in.
- 4.16 30.886 and 36.914; Contains 190 observations.
- 4.17 Place A: $\bar{X} = \text{Rs.106.32}$; s = Rs.30.6.
Place B: $\bar{X} = \text{Rs.106.32}$; s = Rs.32.7.
- 4.18 $\bar{X} = \text{Rs.12.006}$; s = 2.626.
- 4.19 $\bar{X} = \text{Rs.90.15}$; $s = 15.39$; 65%; 95%; 100%.
- 4.20 $\bar{X} = 14.23$; s = 0.72 in. smallest size = 12.82 in.; largest size = 17.14 in.
- 4.21 Actual class-intervals are: 109.5 - 115.5; 115.5 - 121.5; 121.5 - 127.5; 127.5 - 133.5; 133.5 - 139.5; 139.5 - 145.5; 145.5 - 151.5; 151.5 - 157.5 (h = 6, P.M. = 136.5)
- 4.22 Source A: $\bar{X} = 1060$ hours; s = 21.1 hours
Source B: $\bar{X} = 1060$ hours; s = 22.2 hours
These data give a false impression as the distribution is U-shaped.
- 4.24 (b) 13.87%
- 4.25 (b) s = 8.3; C.V. (A) = 16.58%; C.V. (B) = 9.50%.
Locality A has a greater relative dispersion.

- 4.26 C.V. for $X = 19.25$; C.V. for $Y = 25.58$.
 Candidate X showed more consistent performance.
- 4.27 (a) s (corrected) = 2.79; C.V. = 4.14%.
 (b) C.V. for batsman $A = 117.67\%$; C.V. for batsman $B = 70.45\%$.
 Batsman A is better as a run getter but batsman B is the more consistent player.
- 4.28 (a) C.V. = 9.33, 9.17.
 (c) Tab A has a greater absolute dispersion. Tab A has a greater relative dispersion.
- 4.29 Town A: $s = 21.88$, C.V. = 21.15%.
 Town B: $s = 12.95$, C.V. = 23.59%.
- 4.30 (A) C.V. = 24.99; (B) C.V. = 23.54; (C) C.V. = 22.80
- 4.32 (a) $\bar{X} = 37.06$, $S = 8.75$.
 (b) $\bar{X} = 16$, $S = 7.2$, C.V. = 45.
 (c) 78, 15.
- 4.34 Average score of student $A = 59$; Average score of student $B = 54$.
- 4.35 (b) Trimmed: mean = 71.33, s.d. = 5.83.
 Winsorized: mean = 71.2, s.d. = 6.61.
- 4.39 (a) $m_1 = 0$, $m_2 = 1.5$, $m_3 = 0$, $m_4 = 6$,
 $m'_1 = 10.5$, $m'_2 = 40.5$, $m'_3 = 10.5$,
 $m'_4 = 18.75$, $m'_5 = 57.825$, $m'_6 = 0.04$, $b_2 = 2.76$.
- (b) $m_1 = 0$; $m_2 = 2.49$; $m_3 = -0.77$; $m_4 = 18.33$.
- 4.41 $m_1 = 0$; $m_2 = 6.314$; $m_3 = -5.125$; $m_4 = 82.58$;
 $b_1 = 0.104$; $b_2 = 2.071$
- 4.42 $m_1 = 0$, $m_2 = 13.76$, $m_3 = 3.16$, $m_4 = 528.06$;
 $b_1 = 0.004$; $b_2 = 2.79$.
 (b) $b_1 = 0.0003$; $b_2 = 2.75$.
 $b_1 = 0.0002$; $b_2 = 2.97$; b_1 (corrected) = 0.002;
 b_2 (corrected) = 2.66.
 (b) (i) Symmetrical; (ii) Negatively skewed; (iii) Positively skewed.

- 4.47 (i) 0.32; (ii) 0.06.
- 4.48 $m_1 = 1$; $m_2 = 2.2083$; $m_3 = 0.1949$; $m_4 = 12.9646$; $b_1 = 0.0035$; $b_2 = 2.66$. Hence the distribution is slightly positively skewed and is Platy-kurtic.
- 4.49 (i) (ii) is more consistent. (ii), (iii) is negatively skewed.
- (iii) Mode of the distribution is mesokurtic.
- 4.50 (a) .3; (b) $D_1 = 0.49$; $k_2 = 0.65$; Plat-y-kurtic.
- 4.51 (a) (i) Second, (ii) Neither, (iii) First;
 (b) (i) Greater than 1875, (ii) 1875, (iii) Less than 1875.
- 4.54 (i) mean = 80, s.d. = 8.944 (ii) mean = 75, s.d. = 11.18
- 4.55 (i) $\bar{x} = 0.413$, C.V. = 25.43%
 (ii) Mean = 1633.33, S.D. = 796.70
 (iii) Mean = 1446.6630, S.D. = 876.39

Chapter 5, Pp. 174-187

OBJECTIVE

- | | |
|---|--|
| (a) (i) F (un-weighted),
(iv) F (geometric mean),
(vii) F (Laspeyres),
(viii) F (two), | (ii) F (weighted price index),
(v) F (geometric mean),
(vi) F (33.3%),
(vii) F (183), |
| (iii) T,
(ix) T, | (viii) T, |
| (b) (i) a,
(iv) b,
(v) c,
(xiii) d, | (ii) d,
(vii) a,
(xi) d,
(xv) d, |
| (ii) c,
(viii) b,
(x) d, | (iii) c,
(vi) b,
(xii) d, |

SUBJECTIVE

- 5.20 (i) 100, 99.9, 101.0, 104.7, 108.9, 110.6.
 (ii) 95.5, 95.5, 96.5, 100, 104.0, 105.6.
- 5.21 (i) 100, 137.9, 155.3, 166.7, 181.2, 196.8, 200.0, 223.4, 234.8, 341.5.
- 5.22 100, 100.6, 89.5, 92.9, 118.3, 114.7, 100, 91.6, 82.1, 83.7, 110.7, 113.4.
- 5.23 (i) 94.38, 97.67; (ii) 94.65, 98.60; (iii) 94.14, 97.59.
- 5.24 100, 99.2, 74.4, 53.9.
- 5.25 88.5, 88.9, 94.2, 94.7.
- 5.26 100, 107.7, 109.5, 115.0, 114.4, 118.6.
- 5.27 (i) 100, 101.8, 109.8, 125.8, 130.0 (ii) 100, 101.8, 115.8, 132.6, 138.2

- 5.28 (i) 116.6.
 5.29 (a) 84.14, 84.12; (b) 85.22, 85.22.
 5.30 (i) 101.6, (ii) 106.4.
 5.31 (i) 116.30, (ii) 116.30.
 5.32 (i) 115.16, 116.70, (ii) 115.37, 107.28.
 5.33 P_{01} (Marshall-Edgeworth) = 86.5, P_{01} (Fisher's) = 85.8.
 5.34 (i) 49.4; (ii) 262.5.
 5.35 Index for 2004 = 92.68; index for 2005 = 99.01
 5.36 (i) 99.06, 103.84 (ii) 99.06, 103.92; (iii) 99.06, 103.88; (iv) 99.06, 103.89
 (v) 99.06, 103.88; (vi) 99.06, 104.02.
 5.37 (a) (i) 100, 97, 94, 82; (ii) 100, 101, 104, 101.
 (b) Index for 1961 = 118.19; Index for 1962 = 120.00.
 5.38 126.72.
 5.39 Quantity Index: for 2007 on 1997 = 129.8, Quantity Index for 1997 on 2007 = 76.3
 Price index for 2007 with 1997 = 118.0 and price index for 1997 with 2007 = \$3.9
 5.40 100.33 (in both cases).
 5.41 (i) 173.8; (ii) 70.0.
 5.42 (ii) 98.15. The prices in 1929 as compared with the prices in 1928 have fallen down.
 5.43 (i) 121.23; (ii) 121.22.
 5.44 (i) 116.4; (ii) 116.4.
 5.45 124.35.
 5.46 (b) (i) 130.48, (ii) 165.34.
 5.47 (i) 70.24, (ii) 114.18, (iii) 205.97.

Chapter 6, Pp. 233–243

OBJECTIVE

- | | | |
|---|--|---|
| (a) (i) F (a fraction),
(iv) T,
(v) F,
(vi) F (are not equal), | (ii) F (dependent),
(v) F (not mutually exclusive),
(vi) T,
(x) F ($P(A \cap B) = P(A) P(B)$) | (iii) F (usually likely),
(viii) T, |
| (b) (i) c,
(v) a,
(ix) b,
(xiii) a, | (ii) c,
(vi) c,
(x) b,
(xiv) b, | (iv) b,
(vii) a,
(xi) d,
(xv) c, |

SUBJECTIVE

- 6.1 {chair, student}, {chair, pen}, {student, pen}, {chair}, {student}, {pen}, Ø

- 6.3 (i) $\{(1, 1), (1, 2), (2, 1), (2, 2)\}$; (ii) $\{(1, 2), (1, 3), (2, 2), (2, 3)\}$; (iii) $\{(2, 1), (2, 2), (2, 3), (3, 2)\}$; (iv) $\{(1, 2), (1, 3), (2, 2), (2, 3), (3, 3)\}$; (v) $\{(2, 3)\}$.
- 6.4 (b) (i) {5}; (ii) {1, 3, 4, 5, 6, 7, 8, 9, 10}; (iv) {2, 3, 4, 5}; (vi) {1, 2, 5, 6, 7, 8, 9, 10}.
- 6.5 (i) {0, 2, 3, 4, 5, 6, 8}; (ii) φ; (iii) {0, 1, 6, 7, 8, 9}; (iv) {1, 3, 5, 6, 7, 9}; (v) {0, 1, 6, 7, 8, 9}; (vi) {2, 4}.
- 6.6 (a) $A = \{(1, 1), (1, 2), (3, 2), (2, 1), (3, 1), (2, 2)\}$
 $B = \{(1, 6), (2, 6), (3, 6), (4, 6), (5, 6), (6, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5)\}$
- 6.7 (a) (i) 4; (ii) 14; (b) (i) 2730; (ii) 455.
- 6.8 635, 013, 559, 600.
- 6.9 2520.
- 6.12 (a) The investment counsellor's claim is wrong, as the sum of the given three mutually exclusive events cannot exceed unity.
(b) The given statement is wrong, as the probability of each of the outcomes is not 1/3.
(c) The given statement is wrong, as the sum of the given three mutually exclusive events cannot exceed unity.
(d) Same remarks as for (c) above.
- 6.13 (i) $\frac{1}{2}$; (ii) $\frac{5}{36}$; (iii) $\frac{7}{8}$; (iv) $\frac{5}{26}$.
- 6.14 (b) (i) $\frac{1}{3}$; (ii) $\frac{3}{5}$; (iii) $\frac{11}{14}$; (iv) $\frac{2}{5}$; (v) $\frac{4}{5}$.
- 6.15 (a) $\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}, \frac{11}{12}$
(b) $\frac{1}{6}$.
- 6.16 (a) $\frac{1}{6}, \frac{23}{36}, \frac{1}{3}$; (b) $\frac{5}{36}, \frac{1}{6}, \frac{1}{36}, \frac{5}{18}$.
- 6.17 $\frac{1}{15}, \frac{1}{15}, \frac{2}{15}, \frac{2}{15}, \frac{3}{15}, \frac{2}{15}, \frac{2}{15}, \frac{1}{15}, \frac{1}{15}$.
- 6.18 (i) $\frac{1}{3}$; (ii) $\frac{15}{36}$.
- 6.19 (b) 0.5177; 0.4914; (c) $\frac{25}{216}, \frac{1}{8}$.

6.20 (a) $\frac{8}{663}$; (b) $\frac{5}{8}$.

6.21 (a) (i) $\frac{16}{33}$; (ii) $\frac{19}{33}$; (b) (i) 0.38, (ii) 0.62, (iii) 0.12

6.22 (a) $\frac{1}{2}, \frac{1}{5}$ (b) (i) $\frac{1}{2}$; (ii) $\frac{17}{19}$; (c) (i) $\frac{40}{143}$; (ii) $\frac{9}{143}$

6.23 (a) (i) 0.596, (ii) $1 - \left(\frac{5}{6}\right)^8$ (b) 26; (c) 5.

6.24 (i) $\frac{160}{1001}$; (ii) $\frac{109}{143}$

6.25 (a) (i) $\frac{1}{6}$; (ii) $\frac{1}{30}$; (iii) $\frac{5}{6}$ (iv) $\frac{49}{143}$

6.26 (i) $\frac{1}{30}$; (ii) $\frac{1}{6}$; (iii) $\frac{5}{6}$

6.27 (b) $\frac{2}{3}$ (c) $\frac{9}{20}$

6.28 (b) $\frac{6}{26}$

6.29 (c) $\frac{5}{8}$

6.30 (b) $\frac{2}{9}$; (c) 0.9; (ii) 0.6

6.31 (b) (i) $\frac{1}{3}$; (ii) $\frac{2}{3}$; (iii) $\frac{1}{12}$ (c) (i) $\frac{1}{8}$; (ii) $\frac{1}{4}$; (iii) $\frac{7}{8}$; (iv) $\frac{1}{4}$

6.33 (b) (i) $\frac{2}{15}$; (ii) $\frac{2}{15}$

6.34 (a) $\frac{5}{9}$; (b) $\frac{1}{3}$; (c) 0.25.

6.35 (i) 0.0779; (ii) 0.4062; (iii) 0.5.

(c) 0.60, 0.76, 0.60, 0.46, 0.60.

- 6.37 (b) (i) 0.3; (ii) 0.5; (c) (i) 0.10, (ii) 0.20; (iii) 0.17.
- 6.38 (a) True; (b) False, if A and B are independent, then $P(A/B) = P(A)$; (c) True. (d) False, independent does not mean that two events have equal probabilities.
- 6.39 (b) $P(A) = \frac{1}{2}$
- 6.40 (i) not independent; (ii) not independent; (iii) not independent.
- 6.41 (a) $\frac{37}{40}$; (b) $\frac{8}{5525}$
- 6.42 (i) $\frac{4}{63}$; (ii) $\frac{10}{63}$; (iii) $\frac{2}{9}$; (iv) $\frac{5}{9}$
- 6.43 (a) $\frac{65}{176}$ (b) $\frac{49}{80}$
- 6.44 $\frac{7}{429}$
- 6.45 (b) $\frac{25}{1218}$
- 6.47 (i) $\frac{3}{32}$; (ii) $\frac{13}{32}$
- 6.48 $\frac{2}{5}$ or $\frac{94}{175}$ (In this case all three also favour).
- 6.49 (a) $\frac{5}{6}$ (b) 0.012
- 6.50 (i) $\frac{31}{72}$; (ii) $\frac{6}{31}$
- 6.51 (i) 0.12; (ii) 0.88; (iii) 0.38; (iv) 0.38
- 6.52 (a) $\frac{1}{8}$; (b) $\frac{5}{72}$; (c) $\frac{5}{36}$; (d) $\frac{19}{27}$
- 6.53 $\frac{901}{1680}$.
- 6.54 0.586.
- 6.55 (a) 2 : 1.

6.56 (a) $\frac{15}{64}$, (b) (i) 0.196, (ii) $\frac{1}{15}$

6.57 (i) 0.31; (ii) 0.04
<https://stat9943.blogspot.com>

6.58 (b) $\frac{1}{3}$.

6.59 $\frac{9}{29}$.

6.60 $\frac{3}{11}$

6.61 $\frac{1000}{29}\%$

6.62 (i) 0.256 (ii) 0.415 (iii) 0.328

6.63 0.0020; 0.0096; 0.9883.

Chapter 7, Pp. 294-302

OBJECTIVE

- | | | | |
|--------------------------|--------------------------------|-------------------------------|-----------|
| (a) (i) F (any), | (ii) F (discrete), | (iii) F (continuous), | |
| (iv) F ($\sum xP(x)$), | F ($\sum (x - \mu)^2 P(x)$), | (vi) F (fixed set of values), | |
| (vii) F (constant), | (viii) F (independent), | (ix) F (one), | |
| (x) F (at least). | | | |
| (b) (i) d, | (ii) a, | (iii) d, | (iv) a, |
| (v) b, | (vi) a, | (vii) b, | (viii) b, |
| (x) c. | (x) c. | | |

JECTIVE

7.2 (b) (i) $\frac{1}{12}$; (ii) $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$

(c) 0.3038, 0.4389, 0.2135, 0.0412, 0.00264.

7.3 (b) $\frac{15}{210}, \frac{80}{210}, \frac{90}{210}, \frac{24}{210}, \frac{1}{210}$.

7.4 (a) $\frac{120}{455}, \frac{225}{455}, \frac{100}{455}, \frac{10}{455}$ (b) $\frac{1}{56}, \frac{15}{56}, \frac{30}{56}, \frac{10}{56}$

7.5 (b) $\frac{3}{8}$.

7.6 $\frac{1}{8}, 0, \frac{15}{32}, \frac{8}{8}$ <https://stat9943.blogspot.com>

7.7 (i) $\frac{3}{16}$ (ii) 0, (iii) $\frac{11}{16}$ (iv) $\frac{0}{16}$ (v) $\frac{5}{16}$

7.8 (ii) $3x^2 - 2x^3$; (iii) $\frac{13}{27}, \frac{1}{2}$; (iv) $b = 0.6130$.

7.9 (i) $\frac{1}{4}$; (ii) $\frac{3}{64}$.

7.10 (b) $g(x) = \frac{4}{15}, \frac{1}{3}, \frac{2}{5}; h(y) = \frac{2}{5}, \frac{1}{3}, \frac{4}{15}$

Conditional p.d. of X given that $Y = 1; f(x/1) = \frac{1}{2}, \frac{1}{3}, \frac{1}{6}$.

$$f(y/2) = \frac{2}{5}, \frac{1}{2}, \frac{1}{10}.$$

7.11 (a) $g(x) = \frac{5}{36}, \frac{19}{36}, \frac{1}{3}; h(y) = \frac{14}{4}, \frac{14}{45}, \frac{79}{180}$.

$$f(x/1) = \frac{1}{3}, \frac{2}{3}, 0; f(y/3) = 0, \frac{3}{5}, \frac{2}{5}.$$

(b) X and Y are independent.

7.12 (i) $g(x) = \frac{x}{11}, x = 2, 4, 5; h(y) = \frac{y}{6}, y = 1, 2, 3$.

X and Y are independent.

(ii) $g(x) = \frac{x}{6}, x = 1, 2, 3; h(y) = \frac{y^2}{5}, y = 1, 2$.

X and Y are independent.

7.13 (b) $g(x) = \frac{3}{2}x^2 + x; h(y) = y + \frac{3}{2}y^2; f(x/y) = \frac{3x(x+y)}{1+\frac{3}{2}y}$,

$$f(y/x) = \frac{3y(x+y)}{1+\frac{3}{2}x}.$$

7.14 (a) It is a p.d.f. (b) (i) $\frac{5}{6}$ (ii) $\frac{7}{24}$ (iii) $\frac{65}{72}$, (iv) $\frac{5}{32}$.

7.15 (b) $g(x) = \frac{3}{2}x^2 + x$; $h(y) = \frac{3}{3}y^2 + \frac{3}{3}y$; $f(x/y) = \frac{6x(x+y)}{1+4y}$.

$$f(y/x) = \frac{x+2y}{x+1}; \frac{5}{12}.$$

7.16 $g(x) = \frac{3}{2}x^2 + x$; $h(y) = y + \frac{3}{2}y^2$; $f(x/y) = \frac{6x(x+y)}{2+3y}$;

$$f(y/x) = \frac{6y(x+y)}{2+3x}; \frac{769}{2528}.$$

7.17 $g(x) = 12x^2(1-x)$; $f(y/x) = 2y$. The variables X and Y are independent.

7.19 (c) 7.

7.20 (a) $\frac{105}{56}$ (b) $E(Y)$ does not exist.

7.22 (a) $\frac{3}{4}, \frac{9}{8}$ (b) (i) 0.55; 1.35; (ii) 2.1; 5.4

7.23 (a) 20 (b) $\mu = \frac{2}{3}$, $\sigma^2 = \frac{2}{3}$ (c) $\frac{11}{64}$

7.24 (b) 7 (c) 2.75, $\frac{15}{16}$

7.25 (a) 7, $\frac{35}{6}$; (b) $\frac{n+1}{2}, \frac{n^2-1}{12}$.

7.26 (a) Rs.6 and Rs.5 (b) £64, £48, £36, £27 (c) Rs.9, Rs.6, Rs.3

7.27 (a) $\mu = 0.6$, $\sigma = 0.2$.

(b) (i) $\frac{2}{3}$; (ii) $\frac{14}{9}, \frac{13}{162}, \sqrt{\frac{13}{162}}$; (c) $\frac{1}{2}; 0.39$

7.28 (a) $\mu = \frac{2}{3}; \sigma^2 = \frac{2}{63}$; (b) $A = \frac{4}{81}, \mu = \frac{8}{5}, \sigma = 0.66$

7.29 $\mu = 0; \sigma = 1$

7.30 $k = \frac{2}{27}$; mode = $\frac{1}{2}$; $\mu = \frac{7}{6}$; $\sigma^2 = \frac{97}{180}$.

7.31 $k = \frac{2}{9}$; $\mu = \frac{7}{2}$; $\text{mode} = \frac{9}{20}$; $\text{M.D.} = \frac{7}{2}$

7.33 $\mu_1 = 0.001$.

7.34 (a) $\mu = 0$, $\sigma^2 = \frac{a^2}{3}$, M.D. = $\frac{a}{2}$.

(b) $\mu_1 = 0$, $\mu_2 = \frac{1}{12}$, $\mu_3 = 0$, $\mu_4 = \frac{1}{80}$, M.D. = $\frac{1}{4}$

7.35 $\mu = \frac{3}{2}$, $\sigma^2 = \frac{5}{12}$, $\beta_2 = 2.184$

7.36 $\mu_1 = 0$, $\mu_2 = 129$, $\mu_3 = 0$, $\mu_4 = 3.86$, $\beta_2 = 2.33$.

7.37 $g(x) = \frac{3}{5}, \frac{2}{5}; h(y) = \frac{1}{5}, \frac{1}{3}, \frac{7}{15}, \rho = 0.25$

7.38 -0.091

7.41 (a)

x_i	1	2	3	4	5	Total
$f(x_i)$	10/30	8/30	6/30	4/30	2/30	1

(b) Rs.4, Rs.3, Rs.2 and Rs.1.

7.42 (b) Mean = 12; variance = 36

(c) Expected value = 6,00,000, S.D. = 12,000

Chapter 8, Pp. 329-340

OBJECTIVE

- | | | | |
|-------------------------------|--------------------------|----------------------------|-----------|
| (a) (i) F, | (ii) F (two), | (iii) F (fixed specified), | |
| (iv) T, | (v) F (not equal), | (vi) T, | |
| (vii) F (equal), | (viii) F (small, large), | (ix) T, | |
| (x) F (one), | (xi) F (independent), | (xii) F (less), | |
| (xiii) F (positively skewed), | (xiv) T, | (xv) T. | |
| (b) (i) d, | (ii) b, | (iii) c, | (iv) b, |
| (v) a, | (vi) a, | (vii) b, | (viii) b, |
| (ix) d. | | | |

SUBJECTIVE

7.8.2 (b) 0, 0.29, 0.936, 0, 0.352.

8.3 (a) (i) 0.1317 , $\frac{131}{243}, \frac{200}{243}$ (b) (i) 0.28, (ii) 0.31, (iii) 0.23.

8.4 (a) (i) $\frac{15}{64}$, (ii) $\frac{57}{64}$ (b) (i) $\frac{57}{64}$, (ii) $\frac{21}{32}$.

8.5 (a) 0.512; (b) 0.1277, 0.024.

8.6 (a) (i) $\frac{2816}{3125}$; (ii) $\frac{53}{3125}$; (b) (i) 0.0055; (ii) 0.2903

8.7 (i) $\frac{32}{243}$; (ii) $\frac{192}{243}$; (iii) $\frac{40}{243}$; (iv) $\frac{11}{243}$

8.8 (a) 0.3134; (b) 0.10737, 0.99363

8.9 (a) 0.4374, 6.124, 35.72, 111.132, 194.48, 181.5, 70.6.

(b) Expected frequencies are: 3, 15, 30, 30, 15, 3.

8.10 (a) 27.31%; (b) 1 approximately.

8.11 52.08, 41.67, 12.50, 1.67, 0.08; Mean = $\frac{2}{3}$:

8.12 (c) 0.0073;

8.13 (a) $p = 0.36$, $n = 100$; (b) $p = 0.303$, $n = 41$; (c) No, it makes $q = 1.8$ which is wrong.

8.17 (b) Median = 5, Mode = 5.

8.18 (a) 0.65, (b) Theoretical frequencies are 8, 56, 155, 192, 89.

8.19 $p = 0.25$ and the expected frequencies are, 35.60, 71.19, 59.33, 26.37, 6.59, 0.88, 0.05.

8.20 $p = 0.32$ and the expected frequencies are: 32, 60, 43, 13 and 2.

8.21 $\bar{X} = 5.42$ and $s = 1.70$.

8.23 (b) $E(X) = 9$; $\text{Var}(X) = 2.25$; 0.3907.

8.26 $\frac{840}{462}; \frac{84}{121}$.

8.27 (a) $\frac{4}{20}, \frac{12}{20}, \frac{4}{20}$; (b) $\frac{3}{14}$.

8.28 0.004, 0.50.

8.29 (a) 0.3179; (b) 0.82.

8.31 (a) $\frac{\binom{30}{3} \binom{120}{1}}{\binom{150}{4}} = 0.0240$ (b) $b(3; 4, 0.2) = 0.0256.$

<https://stat9943.blogspot.com>

8.32 (e) 0.2019, 0.3230, 0.2584, 0.2167

8.33 (a) 0.04918, 0.1494, 0.2241, 0.2241, 0.1681.

(b) 0.2644, 0.1037 (c) 0.135, 0.857.

8.34 (b) (i) 0.99994 (ii) 0.9442

8.35 (b) (i) 0.1937 (ii) 0.1839

8.36 (a) (i) 0.9513 (ii) 0.9989; (b) 0.221.

8.37 (b) 0.2231, 0.1913.

8.38 0.5620

8.39 (b) 0.1839.

8.40 (a) The statement is wrong. (b) 5! and 11.

8.44 Expected frequencies are: 202.16, 137.96, 27.56, 10.72, 1.84, 0.24.

8.45 (b) (i) 0.642 (ii) 0.073.

8.46 123, 110, 49, 14, 3, 1, 0.

8.47 90.3, 108.4, 65.0, 26.0, 7.8, 1.4, 0.4, 0.1; 0.0341

8.48 (b) 0.3679.

8.49 (a) 0.5272, 0.8646; (b) 0.2231

8.50 (i) 0.2682; (ii) 0.0614

8.51 (i) 0.4380; (ii) 0.5620

8.54 (b) 0.10033

8.55 (b) no, (c) 0.1172.

8.56 (b) 0.0515.

8.58 (c) 1/8.

8.59 (b) 0.0129

8.60 (a) 0.09; (b) (i) 0.135, (ii) 0.081, (iii) 0.081.

8.61 (b) (i) 0.132; (ii) 0.0682

(c) 0.09

8.62 (b) (i) 0.5217, (ii) 0.0059, (iii) 0.0000061.

Chapter 9, Pp. 411–420

OBJECTIVE

- (a) (i) T, (ii) F (zero), (iii) F (all),
 (iv) T, (v) F ($0 < \mu_1 < \mu_2$), (vi) F (mean),
 (vii) F (equal), (viii) F (zero, one), (ix) F (one),
 (x) F (equal to), (xi) F (1), (xii) F (whole & fraction),
 (xiii) F (equal to), (xiv) F (will), (xv) F (not same).
 (b) (i) c, (ii) d, (iii) d, (iv) b,
 (v) c, (vi) d, (vii) b, (viii) d,
 (ix) c, (x) b.

SUBJECTIVE

- 9.1 (b) $\frac{1}{t}(e^{t/2} - e^{-t/2})$; $\mu_2 = \frac{1}{12}$; $\mu_3 = 0$, $\mu_4 = \frac{1}{80}$.
- 9.2 (b) 0.2231.
- 9.3 (a) $\mu = 2$; $\sigma^2 = 4$; $\frac{1}{1-2t} : 0.2231$; 0.2232.
 (b) 18.1%, 49.8%.
- 9.4 (b) m.g.f. = $\frac{1}{1-at}$; $\mu_1 = 0$, $\mu_2 = a^2$, $\mu_3 = 2a^3$, $\mu_4 = 9a^4$.
- 9.5 $\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 4$, $\mu_4 = 12$.
- 9.7 (a) $\mu = 0$, $\sigma^2 = 2$; $k = \frac{2}{\pi}$; 0.6014.
- 9.11 (a) The distribution is $\beta_2(m, n)$ -variate.
- 9.15 $k = \frac{1}{\sqrt{24\pi}}$; $\mu = 3$; $\sigma = \sqrt{12} = 3.464$.
- 9.18 (b) (i) 43.82%; (ii) 67.05
- 9.19 (a) 0.5000; 0.3694 (b) (i) 0.1672; (ii) 0.7492.
- 9.20 (a) 0.6147; (b) 0.4822; (c) 0.9973.
- 9.21 (a) 0.3085; (b) 0.6915; (c) 0.0548; (d) 0.1832; (e) 0.6898;
 (f) 0.9452.
- 9.22 (b) (i) 0.4052; (ii) 0.3745; (iii) 0.2358
- 9.23 (a) (i) 0.0663; (ii) 0.0062; (iii) 0.9198 (b) 69.15%.

- 9.24 (a) (i) 0.0918; (ii) 0.5375 (b) 0.3983; 0.4649.
- 9.25 (i) 6; (ii) 131; (iii) 880; (iv) 24.
- 9.26 (a) (i) 26,600; 22,660 (b) (i) 2 days (ii) 58 days (iii) 228 days.
- 9.27 (a) (i) 2.28; (ii) 15.87; (iii) 68.26; (iv) 13.36 (b) 3.37.
- 9.28 (i) The old route is better; (ii) the new route is better.
- 9.29 (a) 99.36; (b) 0.263.
- 9.30 (i) 0.4404; (ii) 12.43, 82.77; (iii) 26.86, 39.11, 85.28.
- 9.31 (a) (i) 570; (ii) 0.4052.
- 9.32 (a) 0.38% (b) 0.4514, (c) 23, (d) 189.95.
- 9.33 (a) 168 cm, (b) 6.24 years.
- 9.34 87 inches.
- 9.35 (a) $\mu = 12.5$; $\sigma = 6.67$ (b) 50; 10 (c) $\mu = 71.46$, $\sigma^2 = (13.0)^2$
- 9.36 40.37; 12.32.
- 9.37 $\mu = 1.7905$ m, $\sigma = 0.0706$ m, 2.009m.
- 9.38 (b) 0.0143.
- 9.39 (b) (i) 0.9962; (ii) 0.0681; (iii) 0.0558.
- 9.40 (a) (i) 0.1925; (ii) 0.2177 (b) (i) 0.392; (ii) 0.6970.
- 9.41 (a) 0.599 (b) 0.7469, (i) 0.624, (ii) 0.7462.
- 9.42 (a) (i) 0.3251, (ii) 0.1781, (iii) 0.2803, (iv) 0.0459.
- 9.43 (a) $\bar{X} = 67.9$; $s = 2.7$ approx. Proportions are 69%, 95% and 100%. The distribution is nearly normal.
 (b) Fitted frequencies by the area method are:
 2.55, 5.70, 15.30, 32.62, 59.25, 92.78, 116.18, 124.65, 114.82, 84.38, 54.52, 28.05, 12.52, 6.68.
- 9.44 The equation to the fitted normal distribution is
- $$f(x) = \frac{280}{9.50\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - 64.961}{9.50}\right)^2\right]$$
- Frequencies: 4.42, 10.00, 23.58, 41.61, 54.82, 57.20, 43.93, 26.80, 12.12, 5.52.
 Ordinates: 3.14, 9.86, 23.31, 41.31, 56.03, 57.50, 44.71, 26.25, 11.85, 3.98.
- 9.45 Frequencies: 6.8, 30.7, 100.4, 210.4, 269.6, 221.0, 114.6, 37.6, 8.9.

Chapter 10, Pp. 449-456

OBJECTIVE

- (a) (i) F (does not indicate),
 (iii) F (-1 and +1),
 (v) F (correlation analysis),
 (vii) T,
 (ix) F (is always positive),
 (xi) F (correlation analysis),
 (xii) F (must not both),
 (xiii) T,
 (x) F (zero).
 (b) (i) a, (ii) a, (iii) a, (iv) d,
 (v) c, (vi) b, (vii) a, (viii) a,
 (ix) b, (x) c.

SUBJECTIVE

- 10.6 (a) $\hat{Y} = 27.63 - 1.063X$ (b) 17.00, 15.94, 11.69, 9.56, 7.43, 6.37
 (c) $\sum(Y - \hat{Y})^2 = 12.995$, $s_{Y,X} = 1.802$
- 10.7 (b) $\hat{Y} = 2.00 + 0.387X$ (d) 5.87
- 10.8 (a) $\hat{Y} = 30 - X$; (b) $\hat{Y} = 364.68 + 0.095X$,
 (c) $\hat{Y} = -3.17 + 0.32X$, (d) $\hat{Y} = -75.9 + 0.89X$
 (e) $\hat{Y} = 53.68 + 3.53X$.
- 10.9 (b) $\hat{Y} = 130.62 - 0.57X$; (c) 122.
- 10.10 (a) $\hat{Y} = -0.43 + 2.06X$; (b) 0.41; (c) $\hat{X} = 0.27 + 0.47Y$; (d) 0.20.
- 10.11 (b) $r^2 = 0.95$.
- 10.12 (b) (i) 1234; (ii) 1168.45; (iii) 65.55; 95% variability is explained by regression model.
- 10.14 (b) $r = -0.92$ (c) $r = -0.92$.
- 10.15 (b) $r = 0.93$.
- 10.16 (a) (i) 0.7; (ii) -0.7 (b) $r = 0.70$.
- 10.17 (i) $r = 0.91$ (ii) $\hat{Y} = 33.3 + 1.43X$.
- 10.18 $r = 0.94$.
- 10.19 $r = -0.61$.
- 10.20 $r = -0.98$; $\hat{Y} = 31.55 - 1.94X$ and $\hat{X} = 16 - 0.5Y$.
- 10.21 (a) $r = 0.876$ (b) $r = 0.97$

10.22 $r = 0.898.$

10.23 $r = 0.67.$

10.24 (a) $r = -0.415;$ (b) $r = 0.75$ (c) $r = 0.54$; (d) $r = -0.94$

10.25 $r = 0.58;$ $\hat{Y} = 11.32 + 0.54X.$

10.26 $r = 0.39.$

10.27 (b) $r = 0.9883;$ $\hat{Y} = 0.33\hat{X} - 54.09;$ $\hat{X} = 3.03Y - 192.16$

10.28 $r = \frac{1}{2}.$

10.29 (b) $r_s = 0.80.$

10.30 (b) $r = -0.6.$

10.31 (b) $r_s = 0.8545.$

10.32 Denoting the judges by 1, 2, 3; $r_{12} = -0.21;$ $r_{23} = -0.70;$ $r_{13} = 0.64.$ This suggests that judges 2 and 3 have the nearest approach to common tastes.

10.33 $r_{xy} = 0.625;$ $r_{xz} = 0.503;$ $r_{yz} = 0.673.$ Pair (Z, X) has the nearest approach to common tastes.

10.34 $r_s = 0.33.$

10.35 $W = 0.21.$

Chapter 11, Pp. 478-484

OBJECTIVE

- | | | | |
|---|-------------------------------|----------|-----------|
| (a) (i) F (co-efficients of Multiple correlations), | (ii) F (Partial correlation). | | |
| (iii) F (sq-root of) | (iv) F (positive), | | |
| (vi) F (n) | (vii) F (n - k - 1), | | |
| (ix) T, | (x) F (change). | | |
| (b) (i) c, | (ii) b, | (iii) a, | (iv) c, |
| (v) b, | (vi) d, | (vii) c, | (viii) c, |
| (ix) a, | (x) b. | | |

SUBJECTIVE

11.2 $a = 3.88,$ $b_1 = 2.09,$ $b_2 = 2.65.$

11.3 (a) $\hat{Y} = 4.49 - 0.04X_1 + 0.64X_2.$

11.4 (a) $\hat{X}_1 = 0.04 + 0.21X_2 + 0.28X_3;$ (b) 0.21; (c) 0.9863; 0.99.

11.5 (a) $\hat{X}_3 = 61.40 - 3.65X_1 + 2.54X_2$ (b) 40; (c) 0.9927.

11.6 (a) $\hat{Y} = 20.1084 + 0.4136X_1 + 2.0253X_2$ (b) $\hat{Y} = 57.5$ kg

11.7 (b) $R_{1,23} = 0.77$

<https://stat9943.blogspot.com>

11.8 (a) $r_{12} = 0.952$, $r_{13} = -0.056$, $r_{23} = 0.304$; $R_{1,23} = 0.98$,
 $R_{2,31} = 0.98$; $R_{3,12} = 0.82$.

11.9 (b) $R_{2,13} = 0.83$ (c) $R_{2,13} = 0.68$

11.10 (a) $R_{1,23} = 0.80$ (b) $R_{2,13} = 0.66$; $r_{23,1} = 0.64$.

11.11 (b) 0.48.

11.12 $r_{12,3} = 0.759$; $r_{23,1} = -0.436$; $r_{13,2} = 0.097$.

$X_1 = 9.22 + 3.37X_2 + 0.0038X_3$.

11.13 $\hat{X}_1 = -26.77 + 0.39X_2 + 0.23X_3$;

$r_{12,3} = 0.55$; $r_{13,2} = 0.15$; $r_{23,1} = 0.21$.

11.14 $r_{12,3} = 0.63$; $r_{13,2} = 0.49$; $r_{23,1} = -0.035$.

11.15 (b) (i) Not possible; (ii) Not possible; (iii) Is possible.

11.16 $r_{13,2} = -0.586$; $r_{42} = 0.874$, $r_{43} = 0.836$, $r_{42,3} = -0.01$.

11.17 (c) $R_{2,13} = \sqrt{(r_{23}^2 - r_{13}^2)/(1 - r_{13}^2)}$, which is not necessarily zero.

11.18 (b) $r_{12,3} = r_{31,2} = r_{23,1} = -1$.

11.19 $\hat{Y} = 2.08 + 0.64X - 0.1X^2$; 0.13

Chapter 12, Pp. 503–510

OBJECTIVE

11.1 (c) $Y = 1.52 + 1.66X$

11.2 (c) $Y = 7.2 + 1.28X$. The calculated values are 7.20, 8.48, 9.76, 11.04, 12.32, 13.60, 14.88, 16.16, 17.44.

11.3 (b) $Y = 1.2 + 0.5X$; 5.0001.

11.4 (b) $Y = 11.25 + 0.70X$.

11.5 (a) $Y = 6.32 + 0.84X$ (b) $Y = 0.9 + 4.7X$ and the estimated value of $Y = 29.1$; and $X = 0.09 + 0.19Y$ and the fitted value of $Y = 31.1$.

11.6 (c) $Y = 0.83 + 1.60X - 0.71X^2$.

11.7 (b) $Y = 1.172 + 0.056X + 2.28X^2$ (c) $Y = 1.428 + 0.244X + 2.214X^2$

11.8 $Y = 1.04 - 0.20X + 0.24X^2$

12.9 $Y = -6.622 + 1.033X - 0.0056X^2$

12.11 (i) $Y = 123.2 + 8.29X; 9008.4$.

(ii) $Y = 244.04(1.22)^X$ and the estimated value of $Y = 244.13$.

12.12 (i) $Y = 0.32 + 1.13X; 5.819$.

(ii) $Y = 1.42 - 1.07X + 0.55X^2; 1.584$.

(iii) $Y = 1.03 + 1.725X - 1.40X^2 + 0.325X^3; 0.063$.

12.13 (b) $Y = 10 + 5X + 2X^2$.

12.14 (b) $Y = 74.04 (1.22)^X$ and the estimated value of $Y = 244.13$.

12.15 $Y = 32.15 (1.427)^X$ and the estimated value of $Y = 387.40$.

12.16 $Y = 0.8576 (1.393)^X$ or $Y = 3.22 (1.393)^X$ with origin at 4.

12.17 $Y = 8.478 (1.195)^X$.

12.18 $Y = 2.39 (1.19)^X$; 1954 for $X = 12$.

12.19 (a) $Y = 158.5 (2.17)^X$ (b) $Y = 2.50 + 3.50 \sqrt{X}$

12.20 $Y = 2033 (X)^{-1.7488}$.

12.21 $Y = 6483.1 (X)^{-1.04}$ and the estimated value of $Y = 12.76$.

12.22 $a = 5.703, n = 1.02$;

12.23 $v = 9.998 e^{-0.2t}$

12.24 (a) $Y = 9.88 e^{1.002X}$

12.25 $\log A = 12.1662; a = -2.3064$.

12.26 $\gamma = 1.42$.

12.27 (b) $\frac{1}{Y} = 0.10 + 0.025X$.

12.28 (b) $Y = 4.49 - 0.04X_1 + 0.64X_2$.

12.29 (b) $X = 0.034, Y = -0.305$

12.30 (a) $X = 1.162, Y = 2.262$; (b) $X = 1.02, Y = 2.50; \sum e_i^2 = 0.1292$

12.31 (a) $X = 0.9997, Y = 2.0010$; (b) $X = 2.31, Y = -1.02$.

12.32 $X = 1.17, Y = -0.75, Z = 2.08$.

12.33 (b) $X = 2.47, Y = 3.55, Z = 1.92$.

Chapter 13, Pp. 543–552

OBJECTIVE

- (a) (i) F (long-term), (ii) F (four), (iii) F (seasonal variations),
 (iv) F (constant), (v) F (harmonic), (vi) F (multiplication),
 (vii) F (secular), (viii) T, (ix) F (300,000),
 (x) T.
- (b) (i) a, (ii) d, (iii) c, (iv) c,
 (v) c, (vi) a, (vii) a, (viii) b,
 (ix) a, (x) d.

SUBJECTIVE

- 13.1 (b) (i) cyclical, (ii) long-term trend, (iii) seasonal, (iv) seasonal, (v) irregular.
- 13.2 (b) (i) Irregular, (ii) cyclical, (iii) seasonal, (iv) long-term trend, (v) seasonal, (vi) cyclical,
 (vii) seasonal, (viii) long-term trend, (ix) irregular, (x) long-term trend.
- 13.10 (ii) 132.5, 152.7, 172.9, 193.1, 213.3, 235.5, 253.7, 273.9, 294.1, 314.3, 334.5, 354.7.
- 13.11 127.6, 143.2, 160.8, 176.8, 192.8, 207.2.
- 13.12 (ii) 149.8, 142.0, 141.0, 139.8, 138.2, 140.0, 142.0, 147.6, 128.2, 126.2, 119.8, 113.0, 104.4,
 102.2.
- 13.13 (a) 46.3, 46.7, 46.3, 46.7, 46.0, 46.1, 46.3, 46.7
 $\hat{Y}_t = 46.57 + 1.28X$ with origin at midpoint of Saturday and Sunday.
 Trend values: 38, 40, 41, 42, 43, 45, 46, 47, 48, 50, 51, 52, 54, 55.
- 13.14 146.6, 138.0, 133.6, 130.7, 128.6, 129.6, 130.6, 133.3, 139.6, 145.1, 157.9, 176.3, 197.6,
 220.3, 243.4.
- 13.15 (a) 6.9, 6.7, 6.3, 6.6, 7.6, 8.6, 8.8, 8.7, 8.6, 8.2, 8.7, 9.7, 10.6, 10.7, 10.3, 10.0, 9.7, 10.0, 10.8,
 11.4.
 (b) 27.25, 29.88, 32.2, 33.50, 34.12, 34.88.
- 13.16 83.1, 83.0, 82.6, 81.5, 79.8, 78.9, 79.2, 80.1, 80.9, 81.2, 81.8, 82.1, 82.6, 83.2, 84.0.
- 13.17 82.4, 89.6, 97.2, 117.1, 153.2, 188.1, 220.5, 269.6, 327.1, 379.8, 438.0, 528.8.
- 13.18 (a) 2-year centred moving averages are: 50.8, 64.8, 65.8, 66.2, 70.2, 74.2, 81.0, 83.8.
 3-year weighted moving averages are: 50.8, 64.8, 65.8, 66.2, 70.2, 74.2, 81.0, 83.8.
 (b) 50.2, 67.8, 64.5, 65.8, 71.2, 72.8, 82.0, 84.5.
- 13.19 (i) 3.8, 4.6, 5.4, 6.2, 7.0, 7.8, 8.6, 9.4, 10.2., (ii) 4.0, 6.0, 7.0, 7.0, 8.0, 10.0.
 (iii) $\hat{Y}_t = 7 + X$ with origin at 1972.
- 13.20 (i) 24.0, 36.0, 42.0, 42.0, 42.0, 48.7, 59.7.
 (ii) $\hat{Y}_t = 41.89 + 5.95X$ with origin at 1974 and units of X are 1 year.

13.21 (a) $\hat{Y}_t = 99.6 - 12.4X$ with origin at 1997 and units of X are 1 year.

(b) $\hat{Y}_t = 39.9 - 0.7673X$ with origin at 1953.

Trend values are: 43.74, 42.97, 49.20, 41.43, 40.67, 39.90, 39.13, 38.37, 37.60, 36.83, 36.06.

13.22 (i) $\hat{Y}_t = 107.14 + 6.38X$ with origin at 1980; estimated value of $\hat{Y}_t = 139.04$.

(ii) $\hat{Y}_t = 103.24 + 6.38X + 0.976X^2$ with origin at 1980.

13.23 $\hat{Y}_t = 133.09 + 30.25X + 0.155X^2$ with origin at 1970-71; estimated value of $\hat{Y} = 385.01$.

13.24 $\hat{Y} = 31.57 + 5.46X + 0.155X^2$ with origin at 2000 and units of X are 1 year. Trend values are: 18.52, 18.85, 20.32, 22.93, 26.68, 31.57, 37.60, 44.77, 53.08, 62.53, 73.12.

13.25 $\hat{Y} = 351.1 + 13.188X + 0.311X^2$ with origin at Jan. 1, 2005 and
 X is measured in units of a half year.

13.26 $\hat{Y} = 192 - 20X + 1.332X^2$ with origin at 1924. Estimated value of $\hat{Y} = 133$.

13.27 $\hat{Y} = 5.35(1.345)^X$ and estimated value of $\hat{Y} = 130$.

13.28 (b) 118.30, 93.70, 78.75, 108.95.

13.29 101.5, 98.6, 100, 100.

13.30 (a) 135.44, 105.88, 59.75 and 98.

(b) 125.3, 104.6, 60.3, 109.7

Deseasonalization of Data for 1997: 75, 69.8, 68.0, 62.0

13.31 160.3, 105.8, 46.6, 87.2.

13.32 99.6, 100.9, 98.3, 101.2.

13.33 72.06, 85.76, 110.02, 132.18. Deseasonalization of Data for 1954: 111.02, 122.43, 136.38, 147.53,*

*13.34 104.13, 107.24, 98.87, 89.77.

13.35 (a) $\hat{Y} = 94.08 + 0.83X$.

(b) (i) 121.6, 90.2, 78.6, 109.5.

13.36 93.6, 106.6, 112.6, 87.2.

13.37 Seasonal Indices: 95.81, 109.46, 85.28, 109.45; and forecast: 73, 88, 72, 96.

13.38 Cyclical relatives: 97.70, 99.37, 99.47, 101.58, 103.61, 104.13, 103.42, 102.27, 100.19, 96.35, 95.41, 96.76, 100.33.

13.40 (b) $r_1 = 0.66$.

<https://stat9943.blogspot.com>

TABLES

<https://stat9943.blogspot.com>

TABLES

Table 1 – Logarithms

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170						1	2	13	17	21	26	30	34	38
11	0414	0453	0492	0531	0569						4	8	12	16	20	23	27	31	35
12	0792	0828	0864	0899	0934						3	7	11	15	18	22	26	29	33
13	1139	1173	1206	1239	1271						3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584						3	6	9	12	15	19	22	25	28
15	1761	1790	1818	1847	1875						3	6	9	11	14	17	20	23	26
16	2041	2068	2095	2122	2148						3	6	8	11	14	16	19	22	24
17	2304	2330	2355	2380	2405						3	5	8	10	13	15	18	20	23
18	2553	2577	2601	2625	2648						2	4	7	9	12	14	17	19	21
19	2788	2810	2833	2856	2878						2	4	6	8	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3561	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3910	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4395	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4684	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4780	4800	4817	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4957	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	5	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8

https://stat9943.blogspot.com

Table 1 - (Continued) Logarithms

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7252	7261	7270	7279	7288	7297	7306	7315	7324	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	2	3	3	4	5	5	6	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8383	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8501	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8863	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8944	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9048	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9148	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	4	4

https://stat9943.blogspot.com

Table 2 – Antilogarithms

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
.02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
.03	1072	1075	1077	1079	1082	1084	1087	1090	1093	1096	0	1	1	1	1	1	2	2	2
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	1	2	2	2
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	1	2	2	2
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	1	2	2	2
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	1	2	2	2
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	1	2	2	2
.09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	1	2	2	2
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	1	2	2	2
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	1	1	2	2	2
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	1	1	2	2	2
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	1	1	2	2	2
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	1	1	2	2	2
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	0	1	1	1	1	1	2	2	2
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	1	1	2	2	2
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	1	1	2	2	2
.18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	1	1	2	2	2
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	1	1	1	1	1	1	2	2	2
.20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	1	1	2	2	3
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	1	1	1	2	2	3
.22	1660	1663	1667	1671	1675	1679	1683	1687	1691	1694	0	1	1	1	1	1	2	2	3
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	1	1	1	2	2	3
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	1	1	1	2	2	3
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	1	1	1	2	2	3
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	1	1	1	2	2	3
.27	1862	1866	1871	1875	1879	1883	1888	1892	1897	1901	0	1	1	1	1	1	2	2	3
.28	1905	1910	1914	1919	1923	1927	1932	1936	1941	1945	0	1	1	1	1	1	2	2	3
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	1	1	1	2	2	3
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	1	1	1	2	2	3
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	1	1	1	2	2	3
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	1	1	1	2	2	3
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	1	1	1	2	2	3
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	1	1	1	1	2	2	3
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	1	1	1	1	2	2	3
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	1	1	1	1	2	2	3
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	1	1	1	1	2	2	3
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	1	1	1	1	2	2	3
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	1	1	1	1	2	2	3
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	1	1	1	1	2	2	3
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	1	1	1	1	2	2	3
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	1	1	1	1	2	2	3
.43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	1	1	1	1	2	2	3
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	1	1	1	1	2	2	3
.45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	1	1	1	1	2	2	3
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	1	1	1	1	2	2	3
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	1	1	1	1	2	2	3
.48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	1	1	1	1	2	2	3
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1	1	1	1	1	1	2	2	3

https://stat9943.blogspot.com

Table 2 – (Continued) Antilogarithms

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
.52	3311	3318	3327	3334	3342	3350	3358	3366	3373	3381	2	2	3	4	5	5	6	7	
.53	3388	3395	3404	3412	3420	3428	3436	3444	3452	3460	2	2	3	4	5	6	6	7	
.54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	2	3	4	5	6	6	7
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
.56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	4	5	5	6	7	8
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	2	3	4	5	7	8	9	10	
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	2	3	4	6	7	8	9	10	
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	4999	1	2	3	5	6	7	8	9	10
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
.71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
.72	5248	5260	5272	5284	5297	5309	5321	5334	5358	5370	1	2	4	5	6	7	9	10	11
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
.74	5495	5508	5521	5534	5546	5559	5571	5583	5596	5610	1	3	4	5	6	8	9	10	12
.75	5623	5636	5649	5662	5675	5688	5692	5715	5728	5741	1	3	4	5	7	8	9	10	12
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
.79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	6354	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
.82	6607	6622	6638	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
.85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20

https://stat9943.blogspot.com

Table 3 – Squares and Square Roots

<i>n</i>	<i>n</i> ²	\sqrt{n}	$\sqrt{10n}$	<i>n</i>	<i>n</i> ²	\sqrt{n}	$\sqrt{10n}$
1.0	1.00	1.00	3.16	5.5	30.25	2.345	7.416
1.1	1.21	1.05	3.31	6.0	31.00	2.300	7.483
1.2	1.44	1.095	3.464	6.7	32.49	2.387	7.550
1.3	1.69	1.140	3.606	7.8	33.64	2.408	7.616
1.4	1.96	1.183	3.742	8.9	34.81	2.429	7.681
1.5	2.25	1.225	3.873	6.0	36.00	2.449	7.746
1.6	2.56	1.265	4.000	6.1	37.21	2.470	7.810
1.7	2.89	1.304	4.123	6.2	38.44	2.490	7.874
1.8	3.24	1.342	4.243	6.3	39.69	2.510	7.937
1.9	3.61	1.378	4.359	6.4	40.96	2.530	8.000
2.0	4.00	1.414	4.472	6.5	42.25	2.550	8.062
2.1	4.41	1.449	4.583	6.6	43.56	2.569	8.124
2.2	4.84	1.483	4.690	6.7	44.89	2.558	8.185
2.3	5.29	1.517	4.796	6.8	46.24	2.608	8.246
2.4	5.76	1.549	48.99	6.9	47.61	2.627	8.307
2.5	6.25	1.581	5.000	7.0	49.00	2.646	8.367
2.6	6.76	1.612	5.099	7.1	50.41	2.665	8.426
2.7	7.29	1.643	5.196	7.2	51.84	2.683	8.485
2.8	7.84	1.673	5.292	7.3	53.29	2.702	8.544
2.9	8.41	1.703	5.385	7.4	54.76	2.720	8.602
3.0	9.00	1.732	5.477	7.5	56.25	2.739	8.660
3.1	9.61	1.761	5.568	7.6	57.76	2.757	8.718
3.2	10.24	1.789	5.657	7.7	59.29	2.775	8.775
3.3	10.89	1.817	5.745	7.8	60.84	2.793	8.832
3.4	11.56	1.844	5.833	7.9	62.41	2.811	8.888
3.5	12.25	1.871	5.920	8.0	64.00	2.828	8.944
3.6	12.96	1.897	6.000	8.1	65.61	2.846	9.000
3.7	13.69	1.924	6.083	8.2	67.24	2.864	9.055
3.8	14.44	1.949	6.164	8.3	68.89	2.881	9.110
3.9	15.21	1.974	6.245	8.4	70.56	2.898	9.165
4.0	16.00	2.000	6.325	8.5	72.25	2.915	9.220
4.1	16.81	2.025	6.403	8.6	73.96	2.933	9.274
4.2	17.64	2.049	6.481	8.7	75.69	2.950	9.327
4.3	18.49	2.074	6.557	8.8	77.44	2.966	9.381
4.4	19.36	2.098	6.633	8.9	79.21	2.983	9.434
4.5	20.25	2.121	6.708	9.0	81.00	3.000	9.487
4.6	21.16	2.145	6.782	9.1	82.81	3.017	9.539
4.7	22.09	2.168	6.856	9.2	84.64	3.033	9.592
4.8	23.04	2.191	6.928	9.3	86.49	3.050	9.644
4.9	24.01	2.214	7.000	9.4	88.36	3.066	9.695
5.0	25.00	2.236	7.071	9.5	90.25	3.082	9.747
5.1	26.01	2.258	7.141	9.6	92.16	3.098	9.798
5.2	27.04	2.280	7.211	9.7	94.09	3.114	9.849
5.3	28.09	2.302	7.280	9.8	96.04	3.130	9.899
5.4	29.16	2.324	7.348	9.9	98.01	3.146	9.950

<https://stat9943.blogspot.com>

INDEX

<https://stat9943.blogspot.com>

INDEX

A

Absolute error, 7
Accidental error, 7
Addition law of probability, 191
Aggregative expenditure method, 157
Aggregative index number: simple, 135
 weighted, 140
Algebra of sets, 177
Analysis of time series: (see time series)
Approximating Curves, 455
Arithmetical mean:
 definition of, 56
 from grouped data, 59
 merits & demerits of, 75, 134
 properties of, 57
 weighted, 57, 248
Attribute, 5
Auto-Correlation, 502
Average, definition of, 57
 criteria of, 57
 types of, 57
Axiomatic probability, 186

B

Bayes' theorem, 212
Binomial trials, 287
Beta distribution, 344, 349, 350
 properties of, 349, 351
Beta function, 346
Biased errors, 7
Binomial frequency
 distribution, 291
Binary probability
 distribution, 285
Binomial approximation, by
 normal, 374
By Poisson, 309
Derivation of, 385
Displaying of data by, 298
 m.g.f. and c.g.f. of, 299
 properties of, 292
 recurrence formula of, 297
Discrete distribution
 function, 237
 frequency table, 411
 probability
 function, 238

Boole's inequality, 193

C

Causation and Correlation, 408
Cartesian product set, 177
Census, 9
Centred moving averages, 484
Central Tendency,
 measures of, 55
Chain indices, 132, 133
Characteristic function, 273
Charlier check, 111
Chebyshev's, inequality, 268
 Rule, 97
Circular test, 155
Class: boundaries
 interval, 28
 limits, 20
 marks
Class intervals, 177
Classification:
 aims of, 15
 basic principles of, 15
 definition of, 15
Co-efficient of:
 autocorrelation, 502
 concordance, 417
 correlation, 263
 determination, 404
 dispersion, 87
 mean deviation, 90
 multiple determination, 434
 quartile deviation, 88
 rank correlation, 413
 skewness, 114
 standard deviation, 92
 variation, 98
Collection of data, 9
Combinations, 182
Complementation law of probability, 190
Composite index number, 131
Conditional distributions, 239
Conditional probability, 239
Consumer price index, 156
 Shortcomings of, 159
Continuity correlations, 375

Continuous:

- bivariate distribution, 243
- probability distribution, 233, 341
- random variable, 233

Correlation, 263, 395, 406

Correlation and causation, 406

Correlation co-efficient:

- for grouped data, 411
- multiple, 434, 438
- partial, 441
- properties of, 263, 408
- random variables, 263
- serial, 501
- simple linear, 406
- Spearman's rank, 414
- tied ranks, 416

Correlation table, 411

Cost of living index, 156

Counting sample points, 181

- combinations, 182
- multiplication, 181
- permutations, 182

Covariance of two r.v.'s, 263

Cumulants, 271

Cumulant generating function, 271

Cumulative frequency distribution, 26

Cumulative errors, 7

Cumulative frequency curve, 40

Curve-fitting, 455

- criteria for, 468

Curvi-linear regression, 441

Cyclical fluctuations, or

- variations, 479

- measurement of, 499

D

Data: collection of, 9

- deseasonalization of, 498
- editing of, 11
- primary, 9
- secondary, 9, 11

Deciles, 68

de Moivre-Laplace Theorem, 374

Dependent events, 205

Descriptive statistics, 3

Deterministic models, 395

Detrending, 491

Diagrams, 29, 30

- component bar, 31

- multiple bar, 30

- flow chart, 33

- pie, 34

- profit and loss, 35

- rectangles, 32

- scatter, 396

- sector, 34

- simple bar, 30

- sub-divided bar, 32

Discrete probability

- distribution, 228, 285

- binomial, 285

- geometric, 325

- hypergeometric, 302

- monomial, 328

- negative binomial, 321

- poisson, 308

- joint events, 180

Dispersion, absolute, 87

- co-efficient of, 87

- measures of, 87

- relative, 87

Distribution, 15

- function, 227

E

Error of measurement, 7

Event, 180

- compound, 180

- dependent, 205

- disjoint, 180

- equally likely, 181

- exhaustive, 181

- independent, 205

- mutually exclusive, 180

- simple, 180

- symbolic representation of, 181

Expectation (expected value):

- definition of, 248

- of a function, 250

- properties of, 256

Explained variation, 404

Exponential, curve, 462
 distribution, 342
 m.g.f. of, 344
 properties of, 343

<https://stat9943.blogspot.com>

F

factor reversal test, 153

Fisher's *ideal* index, 141, 147

fitting of data by,

binomial distribution, 298

criteria for, 468

exponential curve, 462

higher degree parabola, 460

normal distribution, 379

parabolic curve, 457

Poisson distribution, 316

straight line, 455

forecasting, 501

percentiles, 68

one-hand curve

method, 455, 481

frequency curve, 40

polygon, 39

types of, 41

frequency distribution, 20

construction a, 21

describing a, 116

mean and relation, 178

G

Gamma distribution, 344, 347

m.g.f. of, 348

properties of, 347

gamma function, 344

geometric distribution, 351

negative distribution, 325

deviation of, 325

m.g.f. of, 327

properties of, 326

geometric mean, definition

62

merits and demerits

75, 134

weighted, 63

geometric curve, 467

28, 35

cumulative frequency polygon, 40

frequency curve, 40

frequency polygon, 39

histogram, 37

one-hand, 36

Grouped data, 20

correlation for, 411

mean from, 59

variance of, 92

H

Harmonic analysis, 500

Harmonic mean, 64, 266

merits and demerits of, 75

Histogram, 37

Historigram, 36

Household budget method, 157

Hypergeometric distribution, 302

derivation of, 302

properties of, 305

I

Independence, 239

Independent events, 205

trials, 211

Index numbers, 131

chain indices, 132, 133

composite, 131

consumer price, 156

construction of, 131

cost of living, 156

Fisher's *ideal*, 141, 147

fixed based, 132

Laspeyres', 140, 145, 147

limitations of, 160

Lowe's, 142

Marshall-Edgeworth's, 141

Paasche's, 141, 145, 147

Palgrave's, 146

problems involve in, 133

quantity, 147

simple, 131

tests for, 151

unweighted, 135

uses of, 160

Walsh, 142

weighted, 140
 Inferential statistics, 3, 3
 Interquartile range, 88
 Intersection of sets, 175
 Inverse use of area table, 370
 Irregular or random variations, 480
 analyzing the, 500

J

Joint distributions, 237
 Joint event, 192
 Joint probability, 192

K

Kurtosis, 115

L

Laspeyres' inde, 140, 145, 147
 Laws of probability, 189
 addition, 192
 complementation, 190
 multiplication, 200
 Least-squares, principle of, 398, 487
 Lepto-kurtic, 115
 Linear regression,
 multiple, 429
 properties of, 402
 simple, 396
 Linearizing transformations, 463
 Link relatives, 133
 Link relative method, 497
 Logistic curve, 467
 Lowe's index, 142

M

Makeham curve, 467
 Margieal distribution, 238
 Marshall-Edgeworth index, 141
 Mathematical expectation (see (expectation))
 Mean, arithmetic, 56
 geometric, 62
 harmonic, 64
 trimmed, 104

weighted, 57
 Winsorized, 104
 Mean deviation, 89
 Measures of, central tendency, 55
 merits and demerits of, 75, 134
 skewness, 114
 Median, definition of, 67
 merits and demerits of, 75
 of continuous r.v., 266
 Mesokurtic, 115
 Method of least-squares, 398, 487
 Mode, definition of, 72
 merits and demerits of, 75
 of continuous r.v., 266
 Modified exponential curve, 465
 Moments, 251
 Moment-generating function, 269
 Moro-ratios, 108
 Moving average method, 483
 Multinomial distribution, 328
 derivation of, 328
 Multiple correlation, 429, 438
 Multiple regression, 395, 429, 437
 Multiplication law of probability, 201
 Mutually exclusive events, 180

N

Negative binomial distribution, 321
 derivation of, 322
 properties of, 324
 Normal approximation to:
 binomial, 374
 Poisson, 378
 Normal curve, 115
 Normal distribution, 351
 area of, 364
 continuity correction, 375
 fitting of data by, 379
 inverse use of area table, 370
 m.g.f. and c.g.f. of, 362
 ordinates of, 382
 properties of, 355
 standardized, 353

Normal equations: derivation of, 469
 for higher degree parabola, 460
 for least-squares line, 455
 for least-squares parabola, 456
 for simple linear regression, 398
 for multiple linear regression, 429
 Null set, 174

O

Observations, 5
 Ogive, 41
 Outlier, 104

P

Paasche's index, 141, 145, 147
 Palgrave's index, 146
 Parabolic curve, 457, 460
 Parameter, 3
 Partial correlation, 441
 Partition of sets, 177
 Pascal distribution, 322
 Pearson's coefficient of: correlation, 406
 Skewness, 114
 Percentage-of-annual average method, 492
 Percentage-error, 7
 Percentiles, 68
 Permutations, 182
 Pictograms, 33
 Pie-diagrams, 34
 Platy-kurtic, 115
 Poisson distribution, 308
 approximation by normal, 378
 derivation of, 309
 fitting of data by, 317
 frequency distribution, 312
 m.g.f. and c.g.f. of, 320
 properties of, 313
 recurrence formula of, 316
 Poisson process, 308, 318
 Population, 4
 Posterior probability, 185
 Primary data, 9
 Price index, (see index number)
 Probabilistic models, 395

Probability, 173
 axiomatic definition of, 186
 Bayes' theorem, 212
 classical definition of, 184
 laws of, 189
 of subevent, 191
 personalistic, 187
 posterior, or relative frequency definition of, 185
 Probability density function, 233
 Probability distributions, (see discrete probability distributions)

Q

Qualitative variables, 5
 Quantiles, 68
 Quantitative variables, 5
 Quantitative index, (see index numbers)
 Quotient deviation, 88
 Quartiles, 68, 266
 Questionnaire, 10

R

Random error, 7
 Random experiment, 179
 Random variable, 227
 continuous, 233
 discrete, 228
 Range, 87
 Rank correlation, 413
 derivation of, 414
 for tied ranks, 416
 Ratio charts, 43
 Ratio-to-moving average method, 493
 Ratio-to-trend method, 495
 Rectangular distribution, 341
 m.g.f. of, 342
 properties of, 341
 Regression:
 assumptions, 397
 curvilinear, 395
 definition of, 395
 multiple-linear, 395
 simple linear, 396
 properties of LS, 402

Regression co-efficient:

partial, 429

simple, 396

Relation and function, 178

Relative error, 7

Relative frequency, 26

Repeated trials, 211

Residual method, 499

Residuals, properties of, 436

Root-mean square deviation, 92

Rounding off numbers, 8

S

Sample, 4

Sample space, 179

Scatter diagram, 396

Seasonal indices, 492

Seasonally adjusted data, 498

Seasonal relative, 492

Seasonal variations, 479

analysing the, 492

Second degree curve, 457

Secondary data, 9

Secular trend, 478

analysing the, 481

Semi-average method, 481

Semi-interquartile range, 88

Semi-logarithmic graphs, 43

Serial correlation, 501

Sets, 173

algebra of, 177

cartesian product of, 177

complementation, 177

difference of, 176

disjoint, 180

empty or null set, 174

intersection of, 175

operations on, 175

partition of, 177

sub-sets, 174

union of, 175

unit or singleton set, 174

Sheppard's corrections, 108

Significant digits, 8

Simple aggregative index, 135

Simple average of relatives, 136

Simple index number, 131

Skewed distributions, 42

Skewness, 114

Spearman's rank correlation, 413

Spurious correlation, 408

Standard deviation,

definition of, 91, 251

interpretation of, 97

of population data, 91

properties of, 100

of regression, 402

of sample data, 91

trimmed, 104

Winsorized, 104

Standard error of

estimate, 402, 433

Standardized variables, 103

Standardized normal distribution, 353

Statistic, 1, 4

Statistics:

characteristics of, 2

descriptive, 3

importance of, 4

inferential, 3

meaning of, 1

Stem-and-leaf display, 27

Step function, 228

Stirling's formula, 183, 307

Straight line equation, 455

Subjective probability, 187

Subscript notation, 435

Subsets, 174

Symmetrical distributions, 41

T

Tabulation: definition of, 15

main parts of, 16

types of, 16

Tests for indices, 151

circular, 155

factor reversal, 153

time reversal, 151

Ties, 413

Tied and correlation, 416

Time series: analysis of, 477

components of, 478

decomposition, 480

definition of, 15, 477

<https://stat9943.blogspot.com>

- forecasting, 501
 Tree diagram, 178
 Trend, linear, 487
 quadratic, 488
 Trimmed: mean, 104
 standard deviation, 104
- U**
- Unbiased error, 7
 Unexplained variation, 404
 Uniform distribution, 341
 m.g.f. of, 342
 properties of, 341
 Union of sets, 175
 Unweighted indices, 135
 Uses of indices, 160
- V**
- Value index, 147
 Variables: continuous, 5
 discrete, 5
 qualitative or quantitative, 5
 random, 227
 Variance, of
 population, 91, 92, 251
 properties of, 100
- W**
- Walsh index, 142
 Weighted aggregative indices, 140
 Weighted average of relatives, 145
 Weighted mean:
 arithmetic, 57, 248
 geometric, 63
 harmonic, 65
 Weighted index, 140
 Wild observation, 104
 Winsorized
 mean, 104
 standard deviation, 104
- Y**
- Yule's notation, 435
- Z**
- Z-scores, 104

<https://stat9943.blogspot.com>