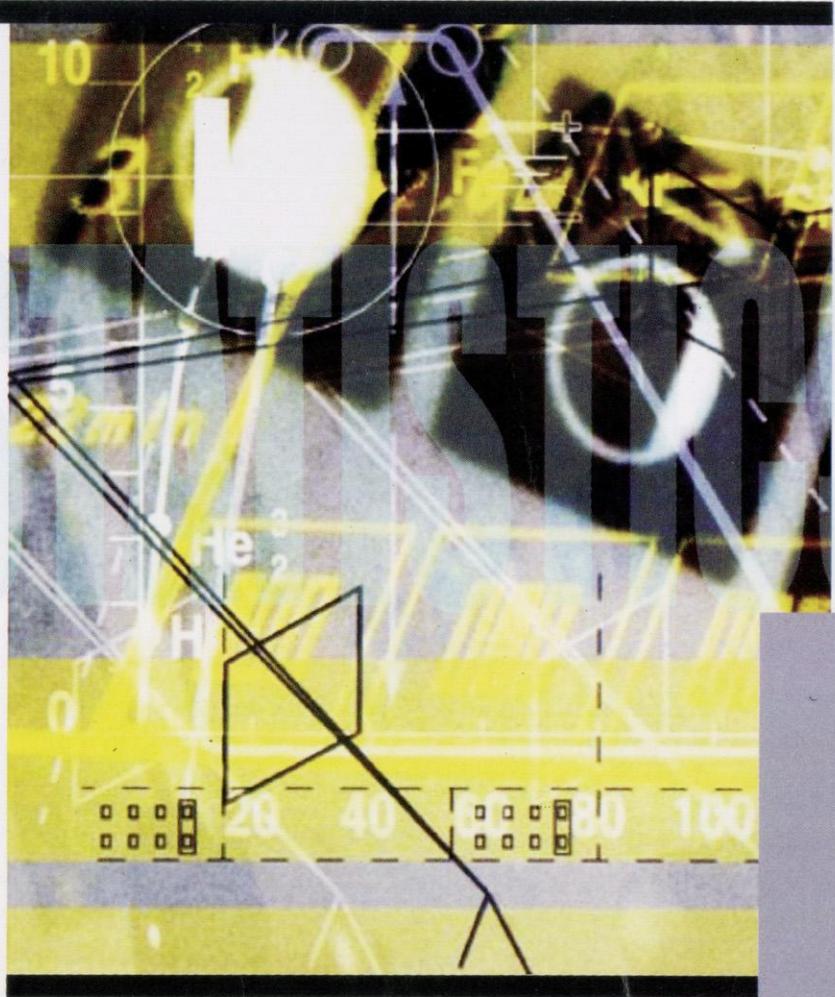


# Biostatistics



**P.N. Arora  
P.K. Malhan**

# **BIOSTATISTICS**

"This page is Intentionally Left Blank"

# BIOSTATISTICS

**DR. P.N. Arora**

M.A., Ph.D., (Delhi Univ.)

*Reader, Dyal Singh College  
(University of Delhi)  
New Delhi*

**DR. P.K. Malhan**

M.Sc. (Statistics), Ph.D.

*Formerly Deputy Director  
Association of Indian Universities  
New Delhi*



**Himalaya Publishing House**

• Mumbai • Delhi • Bangalore • Hyderabad • Chennai  
• Ernakulam • Nagpur • Pune • Ahmedabad • Lucknow

"This page is Intentionally Left Blank"

© No part of this book shall be reproduced, reprinted or translated for any purpose whatsoever without prior permission of the publisher in writing.

ISBN : 978-81-83186-91-9  
REVISED EDITION : 2010

---

**Published by :** Mrs. Meena Pandey  
for **HIMALAYA PUBLISHING HOUSE,**  
“Ramdoot”, Dr. Bhalerao Marg, Girgaon, Mumbai-400 004.  
Phones : 23860170/23863863 Fax : 022-23877178  
Email : hmpub@vsnl.com  
Website : www.himpub.com

**Branch Offices**

- Delhi** : “Pooja Apartments”, 4-B, Murari Lal Street, Ansari Road,  
Darya Ganj, New Delhi-110 002  
Phones : 23270392, 23278631 Reliance : 30180394/96  
Fax : 011-23256286 Email : hphdel@vsnl.com
- Nagpur** : Kundanlal Chandak Industrial Estate, Ghat Road,  
Nagpur-440 018  
Phone : 2721216, Telefax : 0712-2721215
- Bangalore** : No. 16/I (old 12/1), Ist floor, Next to Hotel Highland,  
Madhava Nagar, Race Course Road, Bangalore-560 001  
Phones : 22281541, 22385461 Fax : 080-2286611
- Hyderabad** : No. 3-4-184, Linampally, Besides Raghavendra Swamy Matham,  
Kachiguda, Hyderabad-500027  
Phone : 040-655501745, Fax : 040-27560041
- Chennai** : No. 2, Rama Krishna Street, North Usman Road,  
T-Nagar, Chennai-600 017  
Phone : 28144004, 28144005 Mobile : 09380460419
- Pune** : No. 527, “Laksha Apartment”, First Floor, Mehunpura,  
Shaniwarwadi, (Near Prabhat Theatre), Pune-411 030  
Phone : 020-24496333, 24496333, 24496323
- Lucknow** : C-43, Sector C, Ali Gunj, Lucknow - 226 024  
Phone : 0522-4047594
- Ahemdabad** : 114, Shail, 1st Floor, Opp. Madhu Sudan House,  
C G Road, Navrang Pura, Ahmedabad-380 009  
Mobile : 9327324149
- Eranakulam** : No. 39/104A, Lakshmi Apartment, Karikkamuri Cross Road  
Eranakulam, Cochin-622 011, Kerala  
Phone : 0484-2378012, 2378016
- Printed at** : A to Z Printers, Daryaganj, New Delhi-110002

"This page is Intentionally Left Blank"

# CONTENTS

<b>Chapter</b>	<b>Pages</b>
<b>1. INTRODUCTION TO BIOSTATISTICS</b>	<b>1-15</b>
1.1 Definition of Biostatistics	1
1.2 Development of Biostatistics	1
1.3 Application of Biostatistics	2
1.4 Role of Biostatistics	3
1.5 Definition of Statistics	3
1.6 Descriptive and Inferential Statistics	5
1.7 Some Definitions Concerning Statistics Inference	7
1.8 Data and its Collection	9
1.9 Classification of Data	9
1.10 Several Meaning of Statistics	10
1.11 Characteristic of Statistics	11
1.12 Importance and Usefulness of Statistics	11
1.13 Limitation of Statistics	11
1.14 Kinds of Statisticians	12
1.15 The Role of Applied Statistics	13
1.16 Some Popular Concepts and Words about Statistics	14
<b>2. PRELIMINARY CONCEPTS</b>	<b>16-30</b>
2.1 Variables and Constants	16
2.2 Population and Samples	16
2.3 Random Samples	17
2.4 Discrete and Continuous Variables	17
2.5 Relationship and Prediction	18
2.6 Variables in Biology	19
2.7 Derived Variables, Ratio, Index and Rates	20
2.8 Levels of Measurements of Biological Data	22
2.9 Parameter and Statistic	24
2.10 Accuracy and Precision	25
2.11 Accuracy in a Set of Observations	26
2.12 Levels of Measurement and Problems of Statistical Treatment	27
2.13 Units of Observations	28
2.14 The Summation Sign	28

<b>Chapter</b>		<b>Pages</b>
<b>3. TABULATION AND FREQUENCY DISTRIBUTION</b>		<b>31-39</b>
3.1 Tabulation		31
3.2 Frequency Table of Frequency Distribution		31
3.3 Preparation of a Frequency Table		33
3.4 Relative Frequency Distribution		35
3.5 Cumulative Frequency Distribution		36
<b>4. GRAPHICAL REPRESENTATION OF DATA</b>		<b>40-65</b>
4.1 Graphical Representation of Statistical Data		40
4.2 Types of Graphs		41
4.3 Modes of Graphical Representation of Data		41
4.4 Line Graph		42
4.5 Bar Diagram		43
4.6 Pie Chart or Circle Chart or Sector Chart		47
4.7 Pictograph or Pictogram		50
4.8 Graphical Representation of Grouped Data (Frequency Distribution)		51
4.9 Histogram		51
4.10 Frequency Polygon		54
4.11 Comparison between the Histogram and the Frequency Polygon		56
4.12 Frequency Curve		56
4.13 Cumulative Frequency Curve or Ogive		57
4.14 Proportional Change Diagram		59
4.15 Arithlog or Ratio Diagrams		60
<b>5. MEASURES OF CENTRAL TENDENCY</b>		<b>66-108</b>
5.1 Measures of Central Tendency or Average		66
5.2 Characteristics of an Ideal Measure of Central Tendency		66
5.3 Arithmetic Mean		67
5.4 Weighted Arithmetic Mean		73
5.5 Combined Mean		74
5.6 Corrected Mean		76
5.7 Merits, Demerits and Uses of arithmetic mean		76
5.8 Median		77
5.9 Calculation of Median		77
5.10 Calculation of Median for Grouped Data		78
5.11 Calculation of Median for Continuous Series		80
5.12 Merits, Demerits and Uses of Median		85
5.13 Mode		86
5.14 Types of Model Series		86
5.15 Computation of Mode for Individual Series		87
5.16 Computation of Mode by Grouping Method		87
5.17 Computation of Mode in a Continuous Frequency Distribution		90
5.18 Merits, Demerits and Uses of Mode		90
5.19 Empirical Relation between Mean, Median and Mode		91
5.20 Mid-Range		91
5.21 Geometric Mean		91

<b>Chapter</b>	<b>Pages</b>
5.22 Merits, Demerits and Uses of Geometric Mean	93
5.23 Harmonic Mean	94
5.24 Merits, Demerits and Uses of Harmonic Mean	95
5.25 Choice of an Average for Decisions Making	96
5.26 Comparison of the Mean, Median and Mode— Advantages and Disadvantages	97
5.27 Partition Values	98
5.28 Difference between Averages and Partition Values	99
5.29 Quartiles	99
5.30 Deciles	101
5.31 Percentiles	103
<b>6. MEASURES OF DISPERSION</b>	<b>109-130</b>
6.1 Variability	109
6.2 Range	110
6.3 Interquartile Range	110
6.4 Mean Deviation or Average Deviation	111
6.5 Coefficient of Mean Deviation	115
6.6 Standard Deviation	116
6.7 Merits, Demerits and Uses of Standard Deviation	117
6.8 Calculation of Standard Deviation—Individual Observations	118
6.9 Calculations of Standard Deviation —Discrete Series or Grouped Data	120
6.10 Calculation of Standard Deviation—Continuous Series	123
6.11 Limits of Variability	125
6.12 Empirical Relationships	126
6.13 Variance and Coefficient of Variation	126
<b>7. SKEWNESS, MOMENTS AND KURTOSIS</b>	<b>131-142</b>
7.1 Skewness	131
7.2 Definition of Skewness	131
7.3 Positively and Negatively Skewness	132
7.4 Purpose of Skewness	132
7.5 Difference between Dispersion and Skewness	132
7.6 Measures of Skewness	133
7.7 Relative Measures	133
7.8 Karl Pearson's Coefficient of Skewness	133
7.9 Bowley's Coefficient of Skewness	137
7.10 Kelly's Measure of Skewness	137
7.11 Co-efficient of Skewness Based on Moments	138
7.12 Role of Moments	139
<b>8. CORRELATION ANALYSIS</b>	<b>143-157</b>
8.1 Correlation	143
8.2 Covariance	144
8.3 Calculation of Covariance	145
8.4 Correlation Analysis	145

<b>Chapter</b>		<b>Pages</b>
8.5 Correlation Coefficient Calculated from Ungrouped Data		145
8.6 Spearson's Rank Correlation Coefficient		154
8.7 Scatter or Dot Diagram —Graphical Method		155
<b>9. REGRESSION ANALYSIS</b>	<b>158-168</b>	
9.1 Regression Analysis		158
9.2 Regression Coefficients		159
9.3 Properties of Regression Coefficients		159
9.4 Standard Error of Estimate or Prediction		159
9.5 Linear Regression Line or Equation		160
<b>10. PROBABILITY AND BAYE'S THEOREM</b>	<b>169-222</b>	
10.1 Introduction		169
10.2 Some Important Terms and Concepts		169
10.3 Definitions of Probability		172
10.4 Theorems on Probability		174
10.5 Complimentary Events		177
10.6 Happening of at Least One Event		177
10.7 Addition Theorem for Compatible Events		179
10.8 Definition of Probability in Terms of Odd in Favour or Odds Against the Event		181
10.9 Application of Permutation and Combination		184
10.10 Conditional Probability		191
10.11 Independent Events		192
10.12 Bayes' Theorem		205
10.13 Prior Probabilities (or Priori) and Posterior Probabilities		205
10.14 Inverse Probability		205
<b>11. BINOMIAL, POISSON AND NORMAL DISTRIBUTIONS</b>	<b>223-268</b>	
11.1 Theoretical Distributions		223
11.2 The binomial Distribution		223
11.3 Binomial Distribution Law		225
11.4 Mean and Variance of Binomial Distribution		226
11.5 Conditions for Application of Binomial Distribution		226
11.6 Pascal's Triangle		226
11.7 Characteristics of Binomial Distribution		226
11.8 Recursion Formula or Recurrence Relation for Binomial Distribution		227
11.9 Fitting of a Binomial Distribution		229
11.10 Some Remarks		234
11.11 Poisson Distribution		238
11.12 Conditions under which Poisson Distribution is Used		240
11.13 Characteristics of Poisson Distribution		240
11.14 Binomial Approximation to Poisson Distribution		240
11.15 Mean and Variance of Poisson Distribution		241
11.16 Recurrence Relation		241
11.17 Fitting of a Poisson Distribution		243

<b>Chapter</b>	<b>Pages</b>
11.18 Normal Distribution	250
11.19 Standard Normal Distribution	251
11.20 Properties of Normal Curve	251
11.21 Applications or Uses of Normal Distribution	253
11.22 Method to find the Probability when the Limits of Standard Normal Variates are given	253
11.23 Method to find the Probability when the Variate is Normally Distributed	255
<b>12. HYPOTHESIS TESTING AND LARGE SAMPLES TESTS</b>	<b>269-310</b>
12.1 Introduction	269
12.2 Population and Sample	269
12.3 Parameter and Statistic	270
12.4 Sampling	271
12.5 Sampling Theory	271
12.6 Sampling and Non-sampling Errors	271
12.7 Sampling Fluctuations	272
12.8 Sampling Distribution of a Statistic	273
12.9 Standard Error of a Statistic	274
12.10 Utility of Standard Error of a Statistic	275
12.11 Estimation Theory	275
12.12 Point Estimation	275
12.13 Interval Estimation	276
12.14 Method to Compute Confidence Limits for a Population Parameter $\theta$	277
12.15 Interval Estimation for Large Samples	277
12.16 Confidence Interval of the Mean	277
12.17 Testing of Hypothesis	280
12.18 Procedure of Testing a Hypothesis	281
12.19 The Relationship between Hypothesis Testing and Confidence Interval Estimation	285
12.20 Test of Significance of Mean—Large Sample	286
12.21 Test of Significance of Difference between two Means—Large Samples	292
12.22 Test of Significance for Difference of two Standard Deviations—Large Samples	296
12.23 Test of Significance for Single Proportion —Large Samples	301
12.24 Test of Significance of Difference between two Sample for Large Samples	303
<b>13. STUDENTS' T-TEST</b>	<b>311-357</b>
13.1 Introduction	311
13.2 Student's $t$ -Distribution	312
13.3 Assumptions for $t$ -test	312
13.4 Properties of $t$ -distribution	313
13.5 Application of $t$ -distribution	313
13.6 Interval estimate of Population Mean	313
13.7 Determination of Sample Size	314
13.8 Small (or Exact) Sample Tests	316
13.9 Computation of Test Statistic : $t$ -values	316

<b>Chapter</b>	<b>Pages</b>
13.10 Test of Significance of a Single Mean—Small Samples	317
13.11 Test of Significance of Difference between Two—Means Small Samples	326
13.12 Paired <i>t</i> -test for Difference of Means (when the Sample Observations are not Completely Independent)	338
<b>14. CHI-SQUARE TEST</b>	<b>358-385</b>
14.1 Chi-square Test	358
14.2 Degrees of Freedom	359
14.3 Chi-square Distribution	359
14.4 Properties of $\chi^2$ -distribution	360
14.5 $\chi^2$ -test	360
14.6 Uses of $\chi^2$ Test	361
14.7 Conditions for Using the Chi-square Test	362
14.8 $\chi^2$ -Test' for Goodness of Fit	362
14.9 $\chi^2$ Distribution of Sample Variance	378
14.10 Testing a Hypothesis About the Variance of a Normally Distributed Population	379
<b>15. F-TEST OR FISHER'S F-TEST</b>	<b>380-399</b>
15.1 <i>F</i> -test or Fisher's <i>F</i> -test	386
15.2 <i>F</i> -Statistic	386
15.3 Assumptions in <i>F</i> -test	387
15.4 Tests of Hypothesis About the Variance of Two Populations	387
<b>16. DEMOGRAPHY AND VITAL STATISTICS</b>	<b>400-423</b>
16.1 Demography	400
16.2 Vital Statistics	400
16.3 Methods of Collection of Vital Statistics	401
16.4 Rates and Ratios	402
16.5 Formulae for Calculation of Vital Statistics Rates	405
16.6 Adjustment of Rates	410
16.7 Follow-up Life Table	417
<b>17. NON-PARAMETRIC METHODS</b>	<b>424-476</b>
17.1 Introduction	424
17.2 Non-Parametric OR Distribution Free METHODS	424
17.3 Types of Non-parametric Tests	426
17.4 Advantages of Non-parametric Methods	427
17.5 Disadvantages of Non-parametric Methods	427
17.6 Uses of Non-parametric Methods	428
17.7 The Sign Test for Paired Data	429
17.8 One Sample Sign Test	431
17.9 Rank Sum Tests	435
17.10 Mann-Whitney U-Test	435
17.11 Kruskal-Wallis Test or H-Test	439
17.12 The One Sample Runs Test	446
17.13 Median Test for Randomness or Runs Above and Below the Median	451

<b>Chapter</b>	<b>Pages</b>
17.14 Spearman's Rank Correlation Test	455
17.15 Kolmogorov — Smirnov Test	458
17.16 Kendall Test of Concordance	461
17.17 Median Test for Two Independent Samples	464
17.18 Wilcoxon Signed-Rank Test	465
17.19 The Matched-pairs Sign Test	469
<b>18. FACTORIAL ANALYSIS</b>	<b>477-483</b>
18.1 The Factorial Principle	477
18.2 Basic Ideas and Notation in the $2^n$ Factorial	477
18.3 The Analysis of Variance for a $2^n$ Factorial	479
18.4 The Scope of More Advanced Designs	482
<b>19. CIRCULAR DISTRIBUTIONS AND DESCRIPTIVE STATISTICS</b>	<b>484-502</b>
19.1 Data on a Circular Scale	484
19.2 Graphical Presentation of Data	486
19.3 Sines and Cosines of Circular Data	488
19.4 The Mean Angle	490
19.5 Angular Dispersion	492
19.6 The Median and Modal Angles	493
19.7 Confidence Limits for the Mean and Median Angles	494
19.8 Diametrically Bimodal Distributions	494
19.9 Second-order Analysis: the Mean of Mean Angles	496
19.10 Confidence Limits for the Second-order Mean Angle	498
<b>20. TESTS OF SIGNIFICANCE FOR CIRCULAR DISTRIBUTIONS</b>	<b>503-514</b>
20.1 Goodness of Fit Testing	503
20.2 The Significance of the Mean and Median Angles	505
20.3 Parametric Two-sample and Multisample Testing of Angles	509
<b>21. ASSOCIATION OF ATTRIBUTES</b>	<b>515-536</b>
21.1 Attributes and Variables	515
21.2 Association of Attributes	515
21.3 Correlation and Association	515
21.4 Classification of Data	516
21.5 Terms and Notations	516
21.6 Order of Classes	517
21.7 Number of Classes	517
21.8 Ultimate Classes	517
21.9 Positive and Negative Classes	517
21.10 Nine Square Table	518
21.11 Consistency of Data	519
21.12 Association and Independence	520
21.13 Types of Association	522
21.14 Methods of Determining Association	522
21.15 Comparison of Observed and Expected Frequencies or Frequency Method	522
21.16 Comparison of Proportions or Proportion Method	524

<b>Chapter</b>		<b>Pages</b>
21.17 Yule's Co-efficient of Association		527
21.18 Yule's Co-efficient of Colligation		529
<b>22. MODELS OF DATA PRESENTATION WITH SPECIAL REFERENCE TO BIOLOGICAL SAMPLES</b>		<b>537-555</b>
22.1 Introduction		537
22.2 Hopkins Perspective (Model) on Biostatistics		537
22.3 Various Fields for Biostatistical Models		538
22.4 Covariance Models		539
22.5 Cock Ricing Model		540
22.6 Spatial Statistical Models		540
22.7 Multivariate Spatial Models		541
22.8 Markov Random Field Spatial Models		542
22.9 Gaussian Random Process Models		543
22.10 Non-Gaussian Random Process Models		544
22.11 Generalized Mixed Model Framework		545
22.12 Hierarchical Models		545
22.13 Spatiotemporal Models		546
22.14 Computers and Biology		546
22.15 High-level or Low-level Language and Models		547
22.16 Some Examples on Statistical Models		547

# 1

# *Introduction to Biostatistics*

## 1.1 DEFINITION OF BIOSTATISTICS

Biostatistics is the application of statistics to biology. It is frequently associated with applications to medicine and to agriculture.

*Biostatistics may be defined as the application of the statistical methods to the problems of biology, including human biology, medicine and public health.* It is also known as *Biometry* (literally meaning 'biological measurement's).

The terms biostatistics and biometry are sometimes used interchangeably, although biometry tends to connote a biological or agricultural, rather than a medical, application. (Note that there is a more recent meaning of biometrics).

Because research questions and data sets in biology and medicine are diverse Biostatistics has a broad meaning.

## 1.2 DEVELOPMENT OF BIOSTATISTICS

Perhaps the earliest important figure in Biostatistics thought was Adolphe Quetelet (1796-1874), a Belgian astronomer and mathematician, who in his work combined the theory and practical methods of statistics and applied them to the problems of biology, medicine and sociology. Francis Galton (1822-1911), a cousin of Charles Darwin has been called the **Father of biostatistics and eugenics**, the two subjects that he studied interrelatedly. The inadequacy of Darwin's genetic theories stimulated Galton to try to solve the problems of heredity. Galton's major contribution to biology is his application of statistical methodology to the analysis of biological variations, such as the analysis of variability and the study of regression and correlation in biological measurements. His hope of unrevealing the laws of genetics through these procedures was in vain. He started with the most difficult material and with the wrong assumptions. However, his methodology has become the foundation for the application of statistics to biology. Karl Pearson (1857-1936) at University College, London, became interested in the application of statistical methods to biology, particularly in the demonstration of natural selection, through the influence of W.F.R. Weldon, (1860-1906), a zoologist at the same institution. Weldon,

incidentally, is credited with coining the term biometry for the type of studies pursued by him. Pearson continued in the tradition of Galton and laid the foundation for much of descriptive and correlational statistics. The dominant figure in statistics and biometry in this century has been Ronald A. Fisher (1890-1962). His many contributions to statistical theory will become obvious even to the cursory reader of this book.

Biostatistical reasoning and modelling were also critical in formation of foundational theories of biology. After the 1900's and the rediscovery of Mendel's work there were conceptual gaps in understanding between genetics and evolutionary Darwinism. These gaps lead to several vigorous debates, including a struggle between the biometricalians (e.g., Walter Frank Raphael Weldon and Karl Pearson, and the Mendelists, e.g., Charles Benedict Davenport and William Bateson). Statisticians and models that exploited statistical reasoning helped to bridge this gap. Work at the intersection of genetics, evolution, and populations lead to a foundational advance in the history of biological thought, called the Neo-Darwinian Modern evolutionary synthesis. The prominent role that statistics played in this synthesis is shown by the fact that Sir Ronald Fisher was one of the major founders of this synthesis. Sir Ronald developed several basic statistical methods. However, he also wrote foundational and widely influential work in biology, such as *The Genetical Theory of Natural Selection*. Sewall G. Wright, a co-founder of this synthesis also used statistics in the development of modern population genetics. The mathematical biologist, J.B.S. Haldane is often credited as the third major founder of this synthesis. These individuals and the work of other biostatisticians, mathematical biologists, and statistically inclined geneticists helped bring together evolutionary biology and genetics into a consistent, coherent whole that could begin to be quantitatively modelled.

### 1.3 APPLICATION OF BIOSTATISTICS

The ever-increasing importance and application of statistics to biological data is evident even on cursory inspection of almost any biological journal. Why has there been such a marked increase in the use of statistics in biology? It has apparently come with the realisation that in biology the interplay of causal and response variables obeys laws that are not in the classic mold of 19th century physical science. In that century, biologists such as Robert Mayer, Helmholtz, and others, in trying to demonstrate that biological processes were nothing but physicochemical phenomena, helped create the impression that the experimental methods and natural philosophy that had led to such dramatic progress in the physical sciences should be imitated fully in biology. Regrettably, opposition to this point of view was confounded with vitalistic movement, which led to unproductive theorizing.

Thus, many biologists even to this day have retained the tradition of strictly mechanistic and deterministic concepts of thinking, while physicists, as their science became more refined and came to deal with ever more "elementary" particles, began to resort to statistical approaches. *In biology most phenomena are affected by many casual factors, uncontrollable in their variation and often unidentifiable. Statistics is needed to measure such variable phenomena with a predictable error and to ascertain the reality of minute but important differences.*

A Biostatistics Center could jointly organize working groups, the seminar series, computing infrastructure and possibly consulting and clinical trials coordinating center services.

The main objective of the centre would be to stimulate, collaborate on, and disseminate results of research in a particular subspecialty in the following areas:

- *Statistical methods for longitudinal studies;*
- *Statistical genetics;*
- *Foundations of inference;*
- *Environmental statistics;*
- *Bayesian biostatistics;*
- *Biostatistical practice and education.*

The most critical short-term problem faced in the field of Biostatistics is the information systems. We need to incorporate modern, web-based technologies into the everyday workings of the Department of Biostatistics. We need reliable and accessible systems that are competitive with those available to departments of statistics and biostatistics. We likely need to build collaborations with computer science students.

A related opportunity, is the need for quantifying uncertainty from sources beyond sampling variation. In most observational or experimental studies, there is uncertainty about key parameters derived from the sample begin different from the intended population, from the measured health outcomes being imperfect measures of the intended variable, from informative missing data, and other sources. By what process might these multiple sources of uncertainty be quantified, given a single study or many studies? What novel designs might improve our ability to quantify uncertainty?

## 1.4 ROLE OF BIOSTATISTICS

Discussions over the last year, culminating in the retreat, have identified the following priority missions for Biostatistics:

- I. Conduct of original research on important biostatistics problems across the spectrum: foundations ↔ methodology ↔ applications;
- II. Leadership of biostatistics education for public health/biomedical scientists and professionals;
- III. Participation in other current and future educational programs involving substantial statistical reasoning. Such as quantitative genetics, bioinformatics and clinical investigations and information technology;
- IV. Facilitation of biomedical and public health research that depends on statistical collaboration or consultation.
- V. Given continued growth and decentralization of biostatistics research and applications across the country, we have to discuss how best to organise this discipline and to make it maximally useful in advancing health.

## 1.5 DEFINITION OF STATISTICS

Statistics is a branch of applied mathematics with roots in a part of mathematics called probability theory, and is fundamental to all observational science: physical science, biological science and social science. Wherever observations are made, recorded in numerical form, and

analysed for the purpose of reaching scientific conclusions or judgements, statistical is a major partner.

Different authors have given different definitions of statistics. Some of the definitions of statistics describing it as quantitatively are:

*"Statistics are the classified facts representing the conditions of the people in a state especially those facts which can be stated in number or in a table of numbers or in any tabular or classified arrangement" — Webster.*

This definition is narrow as it is confined only to the collection of the people in a state.

Another definition is due to Bowley.

*"Statistics is a numerical statement of facts in any department of enquiry placed in relation to each other".*

This definition is also not clear and exhaustive. But the following definition given by Secrist is modern and convincing. It also brings out the major characteristics of statistical data.

*"By statistics we mean the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other". — Secrist.*

Statistics is also defined as *the scientific study of numerical data based on natural phenomena*. All parts of this definition are important and deserve emphasis.

**Scientific study :** We are concerned with the commonly accepted criteria of validity of scientific evidence. Objectively in presentation and in evaluation of data and the general ethical code of scientific methodology must constantly be in evidence if the old canard that "figures never lie, only statisticians do" is not be revived.

**Data :** Statistics generally deals with populations and groups of individuals; hence it deals with *quantities* of information, not with a single datum. Thus the measurement of a single animal or the response from a single biochemical test will generally not be of interest.

**Numerical :** Unless the data of a study can be quantified in one way or the other, they would not be amenable to statistical analysis. Numerical data can be measureable — the length or width of a structure or the amount of a chemical in a body fluid — or counts — the number of bristles or teeth.

**Natural phenomena :** We use this term in a wide sense, including all those events that happen in animate and inanimate nature not under the control of man, plus those evoked by the scientist and partly under his control, as in an experiment. Different biologists will concern themselves with different levels of natural phenomena; other kinds of scientists, with yet different ones. But all would agree that the chirping of crickets, the number of peas in a pod, and the age of maturity of a chicken are natural phenomena. The heartbeat of rats in response to adrenalin or the mutation rate in maize after irradiation may still be considered natural, even though man has interfered with the phenomenon through his experiment. However, the average biologist would not consider the number of stereo hi-fi sets bought by persons in different states in a given year to be natural phenomenon, although sociologists or human ecologists might so consider it and deem it worthy of study. The qualification "*natural phenomena*" is included in the definition

of statistics largely to make certain that the phenomena studied are not arbitrary ones that are entirely under the will and control of the researcher, such as the number of animals employed in an experiment.

Although no single definition of statistics is satisfactory for the purposes, the following statement will be useful:

*Statistics is the study of methods and procedures for collecting, classifying, summarizing and analysing data and for making scientific inferences from such data.*

Statistics, according to the definition given above, breaks naturally into two reasonably distinct subcategories: descriptive statistics and inferential statistics.

## 1.6 DESCRIPTIVE AND INFERNENTIAL STATISTICS

**Descriptive Statistics :** *Descriptive statistics serve as devices for organising data and bringing into focus their essential characteristics for the purpose of reaching conclusions at a later stage.*

The area of descriptive statistics involves, through the use of graphic, tabular, or numerical devices, the abstraction of various properties of sets of observations. Such properties include the frequency with which various values occur, the notion of a typical or usual value, the amount of variability in a set of observations, and the measurement of relationships between two or more variables.

Notice that the field of descriptive statistics is not concerned with the implications or conclusions that can be drawn from sets of data. The failure to choose appropriate descriptive statistics has often been responsible for faulty scientific inferences. *The primary function of descriptive statistics is to provide meaningful and convenient techniques for describing features of data that are of interest.*

**Inferential statistics :** A second branch of statistics practice, known as *inferential statistics*, provides the procedures to draw an inference about conditions that exist in a larger set of observations from study of a part of that set. This branch of statistics is also known as *sampling statistics*.

A basic characteristic of experimental science is the necessity for reaching conclusions on the basis of *incomplete information*. For example, Mendel, in studying the way pea plants differed from one another in height, colour of seeds, colour of pods, and colour of flowers, necessarily had to make his conclusions on the basis of relatively small group of plants compared with the entire population of pea plants of a particular type. In making a statement about, say, colour of flowers, Mendel's conclusions were dependent on the particular sample of plants available for study. He certainly did not expect colour variations found in the group of plants he studied to be identical to those in the universe of all possible pea plants of that type. However, he was able to reach general, genetically accurate conclusions about the inheritance of colour in the universe of pea plants.

Alternatively, suppose that a geneticist wonders whether heritability of colour trait in a newly discovered flower is the same as that noted in Mendel's pea plants. The geneticist knows the familiar 1 : 2 : 1 pattern developed by Mendel and wishes to see if the newly observed plants are the same. Again, only a limited number of plants of the new type will be available for study,

and there is thus the possibility of making an error whether concluding that the new type is the same or different from the old.

Although this example is chosen from the field of botany, it illustrates the situation confronting experimental biologists in any field of interest. In testing the efficacy of a new hypotensive drug, for example, the physician will have only a limited number of hypertensive patients with whom to work. It is unlikely that a second physician will be interested in the particular group of patients studied in the experiment rather than those patients who present themselves to his or her own clinic. Thus, the second physician is interested in the extent to which generalisations can be made from the published experimental results in a different group. Of course, this type of generalisation is necessarily inductive rather than deductive, in that we attempt to reach conclusions concerning a large group on the basis of studying only a small subset.

In statistics terminology this inductive procedure involves making inferences about an appropriate *population* or universe in light of a single subset or *sample*. *The population is the full set of individuals to whom we limit any discussion or inferences, while the sample is a subset or a part of that population.* We will define and expand these concepts later in the book.

*Statistical inference is concerned with the procedures whereby such generalisations can be made.* This concern is most often limited to the quantitative aspects of the generalisation, but more and more often in practice the biostatistician is asked to contribute to the process of reaching substantive conclusions as well. In many respects the biostatistician in this role functions as an applied philosopher of science. It should be pointed out clearly, however, that the statistician is not necessarily in as good a position to make substantive inferences as is the experimenter. In fact, often the reverse is true. However, study of the quantitative aspects of the inferential process provides a solid basis on which the more general substantive inference by virtue of experience with, and abstraction from, other similar problems.

Another application of inferential statistics is particularly suited to the evaluation of the outcome of an experiment. Is it possible that a certain drug has an effect on the speed of learning? Let us suppose that an investigator decides on the kind of subjects he wishes to study, selects at random two groups of 25 subjects each, and administers the drug to one of the groups. Both groups are given a learning task and are in all ways treated alike except for the drug. From the outcome of the study, he finds that the average learning score of the two groups differs by five points.

Now some difference between the groups would be expected even if they were treated alike, because of chance factors involved in the random selection of groups. The question faced by the experimenter is whether the observed difference is within the limits of expected variation. If certain preconditions have been met, statistical theory can provide the basis for an answer. If the experimenter finds that the obtained difference of five points is larger than can be accounted for by chance variation, he will infer that other factors must be at work. If examination of his experimental procedures reveals no reasonable cause for the difference other than the deliberate difference in experimental treatment, he may conclude that the drug is the responsible factor.

*The ability to make generalised conclusions inferring characteristics of the whole from characteristics of its parts, lies within the parts of inferential statistics.*

## 1.7 SOME DEFINITIONS CONCERNING STATISTICS INFERENCE

Some definitions concerning the place of statistics in observational and experimental science by presenting a series of informal definitions are given below. This approach is intended only as a brief overview. All the ideas presented here will be discussed in greater detail later.

1. **Unit** : The smallest object or individual that can be investigated, the source of the basic information. Units thus must have some observable characteristic in common. The examples of the units are, respectively, individual Indian women aged 40 to 49, small sub-areas of land and individual patients. In surveys, the units are often called sampling units; in experiments, *experimental units*.
2. **Study or investigation** : An organised scientific undertaking with a defined set of purposes or objectives.
3. **Experiment** : A study that alters existing conditions in a defined manner in order to assess the effect of one or more “treatments”. In an agronomic investigation a piece of land on which a crop has been planted is subdivided into small areas, and various fertiliser combinations are applied to the subareas to assess the differential effect on yield. In clinical pharmacology, a new medication is administered to one group of patients suffering from a disease and a placebo (inactive or inert substance) to another group in order to assess the comparative response. In these examples the state of nature is altered by the investigator to gain comparative information. Most experimental studies are comparative.
4. **Survey** : A study whose purpose is to assess conditions as they exist in nature, altering them as little as possible. For example, we may wish to know the average resting heart rate of Indian women aged 40 to 49. An investigation to estimate this average would be a survey. Such studies are often called *observational studies*.
5. **Analysis** : The procedures for summarising and extracting numerical information from the units selected for study.
6. **Sample** : A subset or a part of units in the underlying population or universe. The sample provides the actual numerical information used in making inferences about the population.
7. **Population or universe** : A very large (possibly infinite) group of units concerning which scientific inferences are to be made. In the above examples the populations or universes are: all Indian women aged 40 to 49, all possible pieces of land that could be studied, and all patients (past, present and future) suffering from the disease.
8. **Parameter** : A characteristic of a population or universe. In the examples: the average resting heart rate of all Indian women aged 40 to 49, the increased crop yield of all subareas of land that might have received fertiliser combination A compared to those receiving combination B, the average response for patients who might receive the drug minus the average for all receiving placebo. In practice, the actual numerical value of a parameter is unknown and is the subject of a statistical inference.
9. **Statistic** : A characteristic of a sample, used for making inferences about the parameter. In the examples, the statistic are: the average resting heart rate of women selected in the

sample, the increased crop yield for the subareas of land actually studied, the average response of patients scheduled to receive the drug minus the average for those scheduled to receive the placebo. A mnemonic device notes the correspondence between the initial letters: population parameter and sample statistic.

10. **Design** : The detailed specification of the procedures whereby information will be obtained. The design includes a statement of study objectives, a specific definition of units and of the population or universe, and a description of procedures for selecting units from the population. If the study is experimental, the design includes a description of procedures for assigning treatments to the selected units.
11. **Subject matter inference** : The inference from a sample of units for the populations of units. A statistical inference is based on numerical observations on the units selected into the sample and refers to the numerical observations or units in the population. In other words, inferences about Indian women aged 40 to 49 are subject matter inferences, whereas inferences about their heart rates as recorded in the study are statistical inferences. We proceed from a sample of observations to a population of observations by way of a statistical inference and to a population of units from a population of observations by way of a subject matter inference.
12. **Statistical inference** : A conclusion about a population or universe on the basis of information contained in a sample. For example, a statement about a parameter based on the observed value of the corresponding statistic. If our sample of Indian women aged 40 to 49 was selected in a particular way (involving probability and randomness), we may be able to conclude on the basis of an observed average resting sample heart rate of 71.13 (value of the statistic) that the average resting heart rate of the population is likely to lie between 70.13 and 72.13 (values of the parameter). Note that we never know population values exactly, since we have studied only a part of the population, i.e., a sample. Statistical inferences utilise the laws of probability.
13. **Variable** : The characteristic observable on the units. For example, resting heart rate, crop yield response to a medication. The characteristic is called a variable because it can vary. Different women may (and often do) have different heart rates. Characteristic that do not vary are called constants and are incredibly uninteresting. The business of science is the study of variation.
14. **Continuous variable** : A variable that can potentially take any value within a range. Examples are body temperature, blood pressure, serum cholesterol level, weight, height. Note that word ‘potentially’ in the definition Variables are never available in continuous form for analysis, but are recorded in discrete form, that is to the nearest degree or mm Hg or kilogram and so on. Although actual temperature exists to the nearest tenth, hundredth, thousandth, and so on, of a degree it is recorded to a determined accuracy of measurement. It is the actual rather than the recorded value that determines whether a variable is continuous or discrete. A count of 2 — there is no issue of measurement accuracy.
15. **Discrete variable** : A variable that is intrinsically gappy, in the sense that between any two potentially attainable values lies at least one unattainable value. The best example is

a count. Count take values 0, 1, 2, 3 .... A count of  $2 \frac{1}{2}$  is not possible. The opposite of a discrete variable is a continuous variable.

## 1.8 DATA AND ITS COLLECTION

**Data :** The information collected through census and surveys or in a routine manner or other sources is called a *raw data*. The word *data* means *information* (its literacy meaning is given facts). The adjective raw attached to data indicates that the information thus collected and recorded cannot be put to any use immediately and directly. It has to be converted into more suitable form or processed before it begins to make sense to be utilised gainfully. Raw data is like a raw rice. A raw rice has to be cooked properly and tastefully before it is eaten and digested. Similarly, a raw data has to be converted into proper form such as tabulation, frequency distribution form, etc., before any inference is drawn from it. In other words, *a Raw Data is a statistical data in original form before any statistical techniques are used to redefine, process or summarise.*

The important step in statistics is the collection of statistical data or simply data. This depends upon the purpose for which the statistical data is required. It is a most important step because it is the basis or foundation of statistical investigations. There are two types of statistical data:

1. *Primary data*

2. *Secondary data*

*Primary data* : It is the data collected by a particular person or organisation for his own use from the primary source.

*Secondary data* : It is the data collected by some other person or organisation for their own use but the investigator also gets if for his use.

In other words, the *Primary* data are those data which are collected by you to meet your own specific purpose whereas the *secondary* data are those data which are collected by somebody else. A data can be primary for one person and secondary for the other.

## 1.9 CLASSIFICATION OF DATA

When data is collected through primary methods, it is in the form of unarranged facts and figures, which practically give no information, whatsoever it may be. It must be sorted out, arranged and properly classified in such a manner which suits the purpose most. *The process of arranging things in groups of classes according to their common characteristics and affinities is called the classification of data.* In other words, a *classification* is the process of dividing things into different classes according to their resemblance.

*Basis of Classification* : There are four basis of classification of data.

(i) **Qualitative** : When the basis of classification is according to differences in quality, the classification is called *qualitative*. For example, rich and poor persons, educated and uneducated persons, intelligent and dull students, etc.

(ii) **Quantitative** : When the basis of classification is according to differences in quantity, the classification is called *quantitative*. For example, a class of students split up into groups according to their heights or ages.

- (iii) **Temporal** : When the basis of classification is according to differences in time, the classification is called temporal. For example, the students who got first division during the last three years are classified year-wise.
- (iv) **Spatial or geographical** : When the basis of classification is according to differences in geographical location or space, the classification is called spatial or geographical. For example, birth rate of India is divided statewise.

### 1.9.1 Kinds of Classification

Classification may be of two kinds:

- (i) **Simple classification** : In simple classification items are classified according to one attribute only. Here each class is divided into two sub-classes. For example, classification of persons according to their sex : males or females; according to education : educated or non-educated.
- (ii) **Manifold classification** : In manifold classification items are classified according to more than one attribute. Classification may give rise to several classes and sub-classes. For example, candidates taking up the higher secondary examination are classified as males and females, sub-divided into students from Government schools and aided schools, further sub-divided into pass and fail, and so on.

### 1.10 SEVERAL MEANING OF STATISTICS

The word *statistics* is used in two different senses — **Plural** and **Singular**. In its *plural* form it refers to the numerical data collected in a systematic manner with some definite aim or object in view such as the number of persons suffering from malaria in different colonies of Delhi or number of unemployed girls in different states of India and so on. In other words, it is used as the plural of the noun *statistic*, which refers to any one of many computed or estimated statistical quantities, such as the mean or the standard deviation, or the correlation coefficient. Each one of these is a statistic. In *singular* form, the word *statistics* means the science of statistics or the subject itself. It includes the methods and principles concerned with collection, analysis and interpretation of numerical data.

So far, the term *statistics* has been encountered in several contexts. It can mean *applied statistics*, the science of organising, describing, and analysing bodies of quantitative data. It can also mean *statistical theory*. In this sense it is best regarded as a branch of mathematics, owing much to the theory of probability. In a third meaning, statistics refers to a *set of indices*, such as averages, which are the outcome of statistical procedures. The general public often uses the word in this sense, as reflected in the request, “Give me the statistics”.

A fourth meaning, similar to the third, is of *significance test* to be discussed later in this book. In this sense, a statistics is an *index descriptive of a sample*. The same index, if descriptive of a population, is called a *parameter*. Thus, *the average or mean of a sample is a statistic; the average or mean of the population is a parameter*.

## 1.11 CHARACTERISTIC OF STATISTICS

According to Sechrist the following are the characteristics of statistics:

1. Statistics are the aggregate of facts.
2. Statistics are numerically expressed.
3. Statistics are affected to a marked extent by multiplicity of causes and not by a single cause.
4. Statistics should be collected in a systematic manner.
5. Statistics should be collected for a pre-determined purpose.
6. Statistics should be placed in relation to each other.
7. The reasonable standard of accuracy should be maintained in statistics.

## 1.12 IMPORTANCE AND USEFULNESS OF STATISTICS

1. Statistics help in presenting large quantity of data in a simple and classified form.
2. It gives the methods of comparison of data and it weights and judges them in the right perspective.
3. It enlarges individual mind.
4. It, when considered as the logic of figures, assists in arriving at correct views based on facts.
5. It helps in finding the conditions of relationship between the variables.
6. It tries to give material for the businessmen as well as the administrators so as to serve as a guide in planning and in shaping future policies and programmes.
7. It proves useful in a number of fields like Railways, Banks, Army etc.

## 1.13 LIMITATION OF STATISTICS

1. Statistical laws are held to be true on the average or in the long run. They are not exact laws which are true in every case. Statistics deals with phenomena affected by a large number of causes. Many of these causes have to be neglected or studied jointly with other more important causes. As such the conclusions arrived under similar conditions at all times, are not identical though they are of the same nature.
2. Statistics does not take cognisance of individual cases. It deals with aggregates though for purposes of analysis these aggregates are very often reduced to single figure. Thus the average income of a group of persons might have remained the same over two periods and yet many persons in the group might have become poorer than what they were before.
3. Statistical data must always be treated as approximations or estimates and not as precise measurement. This is particularly so in business or economic studies where experiments cannot be carried out under controlled conditions.
4. Statistical methods cannot be applied to facts that cannot be measured quantitatively such as, health, blindness, culture, intelligence, poverty etc. These subjective concepts will have to be related in an indirect fashion to numerical data before they can be studied statistically.

5. *Statistical results are ascertained by samples.* If the selection of samples is biased, errors will accumulate and results will not be reliable.
6. The greatest limitation of statistics is that *only one who has an expert knowledge of statistical methods can efficiently handle statistical data.* Statistics, like medicine in the hands of quacks are capable of being easily misused by the inexpert. Knowingly or unknowingly, such a person can draw faulty conclusions.
7. *Statistical results might lead to fallacious conclusions if they are quoted without their context.* Thus the average marks of two students for three examinations might be the same but one may be improving from bad preparation to good while the other may be deteriorating in his studies.

#### **1.14 KINDS OF STATISTICIANS**

Those who work with statistics might be divided into four classes:

- (i) Those who need to know statistics in order to appreciate reports of findings in their professional field.
- (ii) Those who must select and apply statistical treatment in the course of their own inquiry.
- (iii) Professional statisticians.
- (iv) Mathematical statisticians

The main interest of those in the first two classes is in their own subject matter. Statistics in an aid to them in organising and making meaningful the evidence that bears on questions that have been raised. Among their ranks are the biologist, educator, psychologist, engineer, census taker, medical researcher, geologist, agriculturist, physicist, personal officer, counselor, business person and city manager, all these and many more regularly find that statistical procedures can be of assistance. We might think of them as amateur statisticians, and like amateurs in most areas, their statistical knowledge may range from novice to expert.

At the next level, we have the professional statistician. In earlier years, he or she may have been trained in a university mathematics department. A more recent graduate is probably a product of a department of statistics and has undergone extensive training in statistical theory and relevant mathematics. The practising statistician acts as a "middleman" in the process of research, one who assists those with substantive question in finding and applying statistical models with which to examine evidence relative to their inquiry. The advantage of a professional statistician is that of expert knowledge of statistical theory and of its general applicability. However, lack of expertise in the field of application is a limitation.

The three types of persons discussed so far, all have in common a primary interest in applied statistics, although the latter two may indeed have interest in, and make contributions to statistical theory. The primary interest of a mathematical statistician is in pure statistics and probability theory. Professional statisticians have been heard to complain that mathematical statisticians think them too practical and that those whom they serve as consultants think them too theoretical. In fairness, it should be pointed out that the later view is less likely to emerge when those seeking advice have had some elementary education in statistics.

## 1.15 THE ROLE OF APPLIED STATISTICS

We know that an applied statistics is a tool and neither a beginning nor an end by itself. An investigator, poses a problem. He may turn to statistical procedures to find a convenient and meaningful way to organise and view the data that he proposes to collect in her quest. *Statistical procedures provide only a statistical answer; they do not by themselves provide a substantive answer.* They will give one view, a view characterised by certain properties. In choosing statistical techniques, one have to decide that technique which gives the views that is most illuminating, and he must keep in mind the limitations. To do this, he must know the properties of the technique. To interpret the data, statistical knowledge is necessary but not sufficient. Faced with the outcome of statistical analysis, he must then take into account the numerous factors peculiar conclusion.

*The use of statistical procedures is therefore always a middle step.* The typical steps are:

1. *The substantive question* is formulated and refined, and a plan is developed to obtain relevant evidence. A substantive question is a question of facts in a subject – matter area. Suppose one group of subjects learns a foreign language vocabulary by spending two hours a day for four days, while another learns by spending a half hour a day for sixteen days. The total practice time is the same for both groups, but will the amount retained be equal? Some of the aspects of this step will include selection of type of material to be learned, the kind of subjects to be studied, and the measure of retention to be used.
2. When appropriate, *a statistical model is chosen to assist in organising and analysing the data to be collected.* At this point, a statistical question may be developed, the answer to which may be expected to throw light on the substantive question. A statistical question differs from a substantive question in that it always concerns a *statistical* property of the data, such as the average of the set of measures. For example, in the problem above we may ask whether the average number of words retained differs so greatly under the two conditions of practice that chance variation cannot account for it. Often there are alternative statistical questions which are relevant to exploration of the substantive question. For instance, we might ask whether the *proportion* of subjects who retain three quarters or more of the words learned differs substantially under the two conditions of practice. Part of the study of statistics is to learn how to choose among alternative statistical approaches.
3. Upon applying the statistical procedure, one arrives at a *statistical conclusion.* For example, a possible outcome of the learning experiment is that the difference in average number of words retained under the two conditions is too great to be attributed to chance variation. Again, a statistical conclusion concerns a statistical property of the data: in this case, it is about averages.
4. Finally, *a substantive conclusion is drawn.* In the above example, the conclusion may be warranted that under the circumstances studied, retention is better when short practice periods are used. Although the substantive conclusion derives partly from the statistical conclusion, other factors must be considered. If average performance under the two conditions is so different that chance variation among the groups cannot be considered the sole factor, it will not be clear that the difference in method of learning is the responsible factor unless the experiment has been so conducted as to rule out other

factors. The investigator, therefore, must weigh both the statistical conclusion and the adequacy of the experimental design in arriving at a substantive conclusion.

## 1.16 SOME POPULAR CONCEPTS AND WORDS ABOUT STATISTICS

Many people think of statistics of beginning and ending with lengthy tabulations of figures and numerical data. Although numerical data form the grist for the statistician's mill. Consideration of problems of statistical inference immediately reveals that this is but a small part of the field of statistics.

1. Nearly everyone has heard the statement originally made by Disraeli, "*There are three kinds of lies: lies, damned lies and statistics, or the statement "figures don't lie, but liars figure," or, "statistics can prove anything".*" Certainly, the statistical method is a two-edged sword and can be incorrectly used. The field of marketing and advertising provide many examples of misuses. While the use of pseudo-statistics and pseudo-experiments in modern advertising for the purpose of motivating consumers is well understood by many educated person, there seems to be a less general awareness that similar methods can be unwittingly used by a well meaning investigator, and that equally erroneous conclusions may often be reached.
2. It is true that a statistical statement is usually one that is made about a group rather than about an individual. Consideration of the individual is certainly possible. However, a batting average refers to the performance of a single person. In study of groups there may be much that is significant for the individual. Consider a study in educational technique, in which elementary statistics is taught by "standard" methods and also by an experimental method. Suppose that, at the conclusion of the study, average performance is substantially higher for the experimental group, and that average level of student anxiety during the course is substantially lower for the same group. This out-come does not mean that every student will profit in these ways from the new technique; for some students the possibility exists that the new method is worse. However, in the absence of further information, this experiment says that the odds are in favour of students who have the opportunity to study under the new method. Thus, the "group" approach can often be turned to be probable advantage of an individual.
3. In a given state college, it may be that 80% of the total number of credit units of instruction were taught by teachers holding the doctorate, that 75% of the courses were taught by teachers holding the same qualification, and that 65% of teachers in the institution hold the doctorate. If a journalist's theme is the "rotten state of higher eduction". We can guess which figure is most likely to be mentioned. On the other hand, if an economy-minded representative of the taxpayers' association wishes to show that higher salaries are unnecessary to attract a fully qualified faculty, a different choice will be made. Unfortunately, if we are presented in a telling way with only one of these approaches, the possibility of others may never occur to us. We might remember the old saying: "*figures never lied, but liars figure*".

Some people complaints that statistical methodology of investigations, and that it is too mathematical for anyone but an expert to understand since there is some truth in each of these charges, we shall examine them.

A statistical result is simply a statement about the condition of data. This is bound to be dry unless the person is interested in the question to which the data relate and understands the significance of those findings for the question. *It is the conclusion that may be drawn from the data rather than the state of the data that is of possible interest.* If an increase of 851 children is expected in the public schools of your city next year, this information may not stir me greatly if your city is not mine. Even Sally Q. Citizen of your city may feel that she has been told more than she wants to know. Mentally translating the figure into "quite an increase", she thinks it would be wise to vote to approve the school bond issue in the coming election. However, the detailed nature of the number is of vital interest to the school superintendent because it tells how many new classrooms and teachers must be ready and how much adjustment will be needed in the budget.

### EXERCISE

1. Explain the role of statistics in Biology.
2. Discuss the scope, utility and limitation of statistics.
3. Comments on the following statements.
  - (a) "Statistics is the science of averages"
  - (b) "Statistics is the science of estimates and probabilities"
  - (c) "Statistics can prove anything"
  - (d) "Figures cannot lie"
4. Explain clearly what do you understand by the science of statistics. Discuss its scope and limitations.
5. Discuss the role of applied statistics.
6. State a few important definitions of statistics and state the merits and the demerits of each one of them.



# 2

# Preliminary Concepts

## 2.1 VARIABLES AND CONSTANTS

**Variable :** A variable is a characteristic that may take on different values. Typical examples are: intelligence test score, height, number of errors on a spelling test, position of a baseball team in the league standing, eye colour, marital status and sex etc. The concept of a variable does not imply each observation must differ from all the others. All that is necessary is that the possibility of difference exists. Thus, if a school teacher interested in the height of fifth grade students selects a sample of three for study and finds that they are all of the same height, it is still proper to refer to height as a variable since the possibility of getting students of different heights existed. On the other hand, his decision to study height only among the fifth grade students means that grade level is a constant, rather than a variable, in this inquiry.

**Constant :** When it is not possible for a characteristic to have other than a single value, that characteristic is referred to as a constant. In a particular study, there are often several variables and constants to which consideration must be given. Suppose, we are interested in religious attitude among students in a college, our prime concern is the variable of religious attitude. However, other variables such as age, sex and parental religious affiliation may have to be taken into account in interpreting our findings. At the same time, membership in the specific college and the time when the data were collected are constants.

## 2.2 POPULATION AND SAMPLES

In statistical work, you will find much talk about *populations* and *samples*. As a first approach to these important concepts, we find that *population refers to a group of persons (or objects) about which the investigator wishes to draw conclusions and that a sample consists of a part of that population*. If poll stars are trying to take the pulse of the nation prior to an election, their target population consists of those who will go to the polls and vote, whereas those whose opinions they actually obtain constitute a sample of that population. In most research in social science, the investigator hopes that findings can be generalised to persons who are like those used in the study. Thus a sample is studied in the hope that it will lead to conclusions about the larger “target” population.

The biological definition of population refers to all the individuals of a given species (perhaps of a given life-history stage or sex) found in a circumscribed area at a given time. In statistics population always means the totality of individual observations about which inferences are to be made, existing anywhere in the world or at least within a definitely specified sampling area limited in space and time. Some time, of course, the group at hand is the group that we want, and therefore is a population. If teacher wants to know how the present class performed on the first mid-term, the class members constitute the population and not a sample.

So far, we have used the word *population* which refer to a group of persons or objects. A somewhat different definition usually works better in statistics because it eliminates certain confusions. According to this definition, *population refers to the complete set of observations as measurements about which we would like to draw conclusions*. In this usage, the word refers not to people but rather to some observed characteristic. Thus in a study of individuals, rather than the individuals themselves, we shall call a single observation or measurement an *element*. Thus if our interest is in the mathematics achievement of all currently enrolled students in a school, the observation to be recorded is the achievement test score of each student. An element is the test score of a particular student, the scores of students of a class room is a sample, and the complete set of scores constitutes the population.

### 2.3 RANDOM SAMPLES

A *random sample* is a sample chosen in a very specific way. Infact, if a sample is selected at random, known principles of inference apply, and if it is not so chosen, they don't apply. *For a sample to be random, it must have been selected in such a that every element in the population had an equal opportunity of being included in the sample.*

An example may be helpful. Suppose a deck of 52 cards is thoroughly shuffled, and 4 cards are drawn. The hand thus obtained may be considered to be a random sample of the population, because every card in the deck had an equal opportunity of inclusion in the hand, and every possible set of four cards had an equal opportunity of selection.

There are two properties of random samples that we need to know now. First, if several random samples are drawn from the same population, the elements of the sample will differ, and therefore the statistical characteristics will change from sample to sample. Thus if two polling organizations each draw a random sample of voters from the voting population, the elements of their samples will not be identical, and so their projections of the final vote will not be exactly the same. Similarly, in dealing with the 52 playing cards from a standard deck to 4 players, we do not anticipate that each hand of cards will contain one king, one ace etc., nor do we suppose that the number of cards from each suit (clubs, diamonds, hearts, spades) will necessary be equally distributed in a given hand.

Second, the larger the sample, the less is the variation of characteristics of the sample from random sample to random sample.

### 2.4 DISCRETE AND CONTINUOUS VARIABLES

If we ask how many were present at a meeting, we might learn that 23 persons attended, or 24 but no value between these two figure is possible. Variables having this kind of property are called *discrete variables*. Values of such variables are stepwise in nature. Other examples of

such variables are the number of rooms in a school, kinds of occupations, number of cataract operations done in a day in a hospital etc. *A discrete variable is therefore one that can take only certain values, and none in between.*

The second kind is known as a *continuous variable*, despite the assertion of a student that it is called an “*indiscrete*” variable. Consider the question of length. It is possible for an object to be 3 ft 1 in. or 3 ft 2 in. in length, or any conceivable amount in between. The characteristic of a continuous variable is that within whatever limits its values may range, any value at all is possible. **The discrete variable has gaps in its scale; the continuous variable has none.** Examples of variables that are reasonably conceived as continuous variables are weight, age, temperature administration, and musical talent.

The difference between the two types of variables has doubtless been noticed by everyone, although perhaps not completely analyzed. Although we accept as normal a statement that males of given age average 5.3 ft in height, we may find it a bit odd to be told that in the principality of Ruritania, 60 year-old-males have had an average of 1.3 wives. The use of the average is, however entirely appropriate in both instances. All that is required is to recognize what an average is, and that it does not necessarily characterise an individual.

*Even though a variable is continuous in theory, the process of measurement always reduces it to a discrete one.* Suppose we are measuring length and decide to record our observations to the nearest ten-thousand of an inch. A particular observation will then be recorded as 1.3024 inches if the object appears to be closer in length to that figure than to 1.3023 or 1.3025. As a result, our *recorded* measurement will form a discrete scale in steps of one ten-thousandth of an inch.

## 2.5 RELATIONSHIP AND PREDICTION

Experience tells us that there is some relationship between intelligence of parents and their offspring, and yet it is not perfect. We expect that the parents of the brightest child in the class are also bright but would not expect that they are necessarily the most intelligent of all parents of children in this group. Can we describe with greater exactness the extent of this relationship?

The personnel office of an industrial concern gives an aptitude test to its prospective clerical employees. Is there really any relationship between the score on this test and subsequent level of job proficiency? How much? If there is a relationship, what per cent of applicants may be expected to have success on the job if only those who score above a certain point on the test are accepted for employment? How does this compare with what would happen if the test were eliminated from the company's procedures?

These are examples of the type of question that probes the existence and extent of *relationship* between two (or more) factors and explores the possibility of *prediction* of standing in one factor from knowledge of standing in the other. This kind of analysis is so frequently of interest that a considerable body of statistical techniques has been developed to deal with it. Because of the importance and distinctiveness of these two questions, they have been separately identified here. Nevertheless, both are problems in description and inference, the two major categories of statistical endeavour.

## 2.6 VARIABLES IN BIOLOGY

Each biology discipline has its own set of variables, which may include conventional morphological measurements such as: concentrations of chemicals in body fluids, rates of certain biological processes, frequencies of certain events as in genetics and radiation biology, physical readings of optical or electronic machinery used in biological research and many more.

We have already referred to biological variables in a general way, but we have not yet defined them. We shall define a *variable* as a *property with respect to which individuals in a sample differ in some ascertainable way*. Length, height, weight, number of teeth, vitamin C content, and genotypes are examples of variables in ordinary, genetically and phenotypically diverse groups of organisms. Warm bloodedness in a group of mammals is not variables, since they are all alike in this regard, although body temperature of individual mammals would of course, be a variable.

We can divide biological variables into the following three types:

1. *Measurement variables*
2. *Ranked variables*
3. *Attributes*

**1. Measurement variables :** *Measurement variables are all those whose differing states can be expressed in a numerically ordered fashion.* They are divisible into two kinds. The first of these are *continuous variables*, which at least theoretically can assume an infinite number of values between any two fixed points. For example, between the two length measurements 1.5 and 1.6 cm, there is an infinite number of lengths that could be measured if one were so inclined and had a precise enough method of calibration to obtain such measurements. Any given reading of continuous variable, such as a length of 1.57 mm, is therefore an approximation to the exact reading, which in practice is unknown. Many of the variables studied in biology are continuous variables. Examples are lengths, areas, volumes, weight, angles, temperatures, periods of time, percentages and rates.

Contrasted with continuous variables are the *discontinuous variables*, also known as *meristic or discrete variables*. These are variables that have only certain fixed numerical values, with no intermediate values possible in between. Thus the number of segments in a certain insect appendage may be 4 or 5 or 6 but never  $5\frac{1}{2}$  or 4.3. Examples of discontinuous variables are numbers of a certain structure (such as segments, bristles, teeth or glands), the numbers of offspring, the numbers of colonies of micro-organisms or animals or the numbers of plants in a given fixed space.

**2. Ranked variables :** Some variables cannot be measured but at least can be ordered if ranked by their magnitude. Thus in an experiment one might record the rank order of emergence of ten pupae without specifying the exact time at which each pupa emerged. In such cases we code the data as a *ranked variable*, the order of emergence. Special methods for dealing with such variables have been developed. By expressing a variable as a series of ranks, such as 1, 2, 3, 4, 5,... we do not imply that the difference in magnitude between, say, ranks 1 and 2 is identical to or even proportional to the difference between rank 2 and rank 3.

**3. Attributes :** Variables that cannot be measured quantitatively but can be expressed qualitatively are called attributes. These are all properties, such as black or white; pregnant or not pregnant; dead or alive; male or female. When such attributes are combined with frequencies, they can be treated statistically. Of 80 mice, we may, for instance, state that four were black, two agouti, and the rest gray. When attributes are combined with frequencies into table suitable for statistical analysis, they are referred to the enumeration data. Thus the enumeration data on colour in mice just mentioned would be arranged as follows:

Colour	Frequency
Black	4
Agouti	2
Gray	74
Total number of mice	80

In some cases attributes can be changed into variables if this is desired. Thus colours can be changed into wavelengths or colour chart values, which are measurable variables. Certain other attributes that can be ranked or ordered can be coded to become ranked variables. For example, three attributes referring to a structure as "poorly developed", "well-developed" and "hypertrophied" could conveniently be coded 1, 2 and 3.

A term that has not yet been explained is *variate*. In this book we shall use it as a single reading, score, or observation of a given variable. Thus, if we have measurements of the length of the tails of five mice, tail length will be a continuous variable, and each of the five readings of length will be a variate. In this text we identify variables by capital letters, the most common symbol being Y. Thus Y may stand for all length of mice. A variate will refer to a given length measurement;  $Y_i$ , is the measurement of tail length of the  $i^{\text{th}}$  mouse, and  $Y_4$  is the measurement of tail length of the fourth mouse in our sample.

## 2.7 DERIVED VARIABLES, RATIO, INDEX AND RATES

The majority of variables in biometric work are observations recorded as direct measurements or counts of biological material or as readings that are the output of various types of instruments. However, there is an important class of variables in biological research that we may call the *derived or computed variables* which are generally based on two or more independently measured variables whose relations are expressed in a certain way. We are referring to ratios, percentages indices rates and the like.

**Ratio :** A ratio expresses as a single value the relation that two variables have one to the other. In the simplest form it is expressed as in 64 : 24, which may represent the number of wild type versus mutant individuals or the number of males versus females or the proportion of parasitized individuals versus those not parasitized and so on. The above examples implied ratios based on counts; a ratio based on a continuous variables might be similarly expressed as 1.2 : 1.8, which may represent the ratio of width to length in a sclerite of an insect or the ratio between the concentrations of two minerals contained in water or soil. Ratios may also be

expressed as fractions; thus the two ratios above could be expressed      and      . However, for

expressed as fractions; thus the two ratios above could be expressed  $\frac{64}{24}$  and  $\frac{1.2}{1.8}$ . However, for computational purposes it is most useful to express the ratio as a quotient. The two ratios cited above would therefore be 2.666.... and 0.666, respectively. These are pure numbers, not expressed in measurement units of any kind. It is this form for ratios that we shall consider further below. *Percentages* are also a type of ratios. Ratios or percentages are basic quantities in biological research, widely used and genetically familiar.

**Index :** An *index* is the ratio of one anatomic variable divided by a larger, so called standard one. A well-known example of an index in this sense is the cephalic index in physical anthropology. Conceived in the wide sense, an index could be the average of two measurements — either simply, such as  $\frac{1}{2}$  (length of *A* + length of *B*), or in weighted fashion, such as  $\frac{1}{3}$  ( $2 \times$  length of *A*) + (length of *D*).

**Rates :** *Rates* will be important in many experiment fields of biology. The amount of a substance liberated per unit weight or volume of biological material, weight gain per unit time, reproductive rates per unit population size and time (birth rates) and death rates would fall in this category.

The use of ratios and percentages is deeply ingrained in scientific thought processes. Often ratios may be the only meaningful way to interpret and understand certain types of biological problems. If the biological process being investigated operates on the ratio of the variables studied, one must examine this ratio to understand the process.

There are several disadvantages of ratios : first is their relative inaccuracy. Let us return to the ratio  $\frac{1.2}{1.8}$  mentioned above and recall from the previous section that a measurement of 1.2 indicates a true range of measurement of variable from 1.15 to 1.25; similarly, a measurement of 1.8 implies a range from 1.75 to 1.85. We realize therefore that the true ratio may vary anywhere from  $\frac{1.15}{1.85}$  to  $\frac{1.25}{1.75}$  or 0.622 and 0.714 respectively. We note a possible maximal error of 4.2% if 1.2 were an original measurement:  $[(1.25 - 1.2)/1.2]$ ; the corresponding maximal error for the ratio is 7.0%  $[(0.714 - 0.667)/0.667]$ . Furthermore, the best estimate of a ratio is not usually the mid-point between its possible ranges. Thus in our example the midpoint between the implied limits is 0.668 and the true ratio is 0.666.... only a slight difference, which may, however, be greater in other instances.

Another drawback to ratios and percentages is that their distributions may be rather unusual and they may therefore not be more or less normally distributed as required by many statistical tests. This difficulty can frequently be overcome by the transformation of the variable. Another disadvantage of ratios is that they do not provide information on the relationship between the two

## 2.8 LEVELS OF MEASUREMENTS OF BIOLOGICAL DATA

### Variable in Biology

A characteristic that varies from one biological entity to another is termed a *variable* (*or variate*). Different sorts of variables may be encountered by biologists, which can be measured differently. It is desirable to distinguish them.

There are five types of measurements of biological data:

1. Data on an interval scale.
2. Data on ratio scale.
3. Date on an ordinal scale.
4. Data on nominal scale.
5. Discrete and continuous data.

**1. Data on an interval scale :** Some measured scales possess a constant interval size but not a true zero; they are called *interval scales*. An outstanding example is that of the two common temperature scales: Celsius (C) and Fahrenheit (F). We can see that the same difference exists between  $20^{\circ}\text{C}$  ( $68^{\circ}\text{F}$ ) and  $25^{\circ}\text{C}$  ( $77^{\circ}\text{F}$ ) as between  $5^{\circ}\text{C}$  ( $41^{\circ}\text{F}$ ) and  $10^{\circ}\text{C}$  ( $50^{\circ}\text{F}$ ); i.e., the measurement scale is composed of equal-sized intervals. But it cannot be said that a temperature of  $40^{\circ}\text{C}$  ( $104^{\circ}\text{F}$ ) is twice as hot as a temperature of  $20^{\circ}\text{C}$  ( $60^{\circ}\text{F}$ ); the zero point is arbitrary. [Temperature measurements on the absolute, or Kelvin (K), scale can be referred to a physically meaningful zero and thus constitute a ratio scale].

Some interval scale encountered in biological data collection are *circular scales*. Time of day and time of the year are examples of such scales. The interval between 2.00 p.m. (i.e., 1400 hr) and 3.30 p.m. (1530 hr) is the same as the interval between 8.00 a.m. (0800 hr) and 9.30 a.m. (0930 hr). But one cannot speak of ratios of times of day because the zero point (midnight) on the scale is arbitrary, in that one could just as well set up a scale for time of day which would have noon, or 3.00 p.m., or any other time as the zero point. Circular biological data are occasionally compass points, as if one records the compass direction in which an animal or plant is oriented. Since *the designation of north as  $0^{\circ}$  is arbitrary, this circular scale is a form of interval scale of measurement*.

**2. Data on a Ratio scale :** Consider that the heights of a group of plants constitute a variable of interest, and perhaps the number of leaves per plant is another variable under consideration. Thanks to measuring devices at the biologists disposal, it is possible to assign a numerical value to the height of each plant, and simply counting the leaves allows a numerical value to be assigned to the number of leaves on each plant. Regardless of whether the height measurements are recorded in centimeters, inches, or any other units, and regardless of whether the leaves are counted in a number system using base 10 or any other base, there are two fundamentally important characteristics of these data.

First, there is a constant size interval between any adjacent units on the measurement scale. That is, the difference in height between a 46 – cm and a 47 – cm plant is the same as the difference between a 49 – cm and a 50 – cm plant, and the difference between 8 to 10 leaves is equal to the differences between 9 to 11 leaves. (This may seem simple minded, but it is very important, as we shall see on examining the other scales of measurement).

Second, it is important that there exists a zero point on the measurement scale and that there is a physical significance to this zero. This enables us to say something meaningful about the ratio of measurements. We can say that a 30 – cm tall plant is half as tall as a 60 cm plant, and that a plant with 45 leaves has 3 times as many leaves as a plant with 15.

*Measurement scales having a constant interval size and a true zero point are said to be ratio scales* of measurement. Besides lengths and numbers of items, ratio scale includes weight (mg, lb, etc.), volumes (cc, cu, ft, etc.) capacities (ml, qt, etc.), rates (cm/sec, mph, mg/min, etc.) and lengths of time (hr, yr, etc.).

**3. Data on a nominal scale :** Sometimes the variables under study is classified by some quality it possesses rather than by a numerical measurement. In such cases, the variable is called an *attribute* and we are said to be using a *nominal scale* of measurement. Genetic phenotypes are commonly encountered biological attributes; the possible manifestation of an animal's eye colour may be blue or brown, and if human hair colour are the attribute of interest, we might record black, brown, blonde, or red. On a nominal scale ("nominal" is from the Latin word for "name"), animals might be classified as male and female, or as left – or right – handed. Or plants might be classified as dead or alive, or as with or without thorns. Taxonomic categories also form a nominal classification scheme (e.g., a plant might be classified as pine, spruce, or fir). Sometimes data from an ordinal, interval or ratio scale of measurement may be recorded in nominal scale categories. For example, heights may be recorded as tall or short or performances on an examination as pass or fail.

As will be seen, statistical methods useful with ratio, interval or ordinal data generally are not applicable to nominal data, and we must, therefore, be able to identify such situations when they occur.

**4. Data on an ordinal scale :** The preceding paragraphs on ratio and interval scales of measurement discussed data between which we know numerical differences. For example, if man *A* weights 60 kg and man *B* weights 70 kg, then man *A* is known to have a weight of 10 kg more than *B*. But our data may, instead, be a record only of the fact that man *A* weights more than man *B* (with no indication of how much more). Thus, we may be dealing with relative difference rather than with quantitative differences. Such data consist of an ordering or ranking of measurements and are said to be on *ordinal scale* of measurement ("ordinal" being from the Latin word for "order"). One may speak of one biological entity being shorter, darker, faster or more active than another; the sizes of five cell types might be labelled 1, 2, 3, 4 and 5, to denote their magnitudes relative to each other; or success in learning to run a maze may be recorded as *A*, *B* and *C*.

It is often true that biological data expressed on the ordinal scale could have been expressed on the interval or ratio scale had exact measurement been obtained (or obtainable). Sometimes data that were originally on interval or ratio scales will be changed to ranks; for example, examination grades of 95, 85, 73 and 56% (ratio scale) might be recorded as *A*, *B*, *C* and *D* (ordinal scale), respectively.

Ordinal scale data contain and convey less information than ratio or interval data, for only relative magnitudes are known. Consequently, quantitative comparisons are impossible (e.g., we cannot speak of a grade of *C* being half as good as a grade of *A* or of the difference

between cell sizes 1 and 2 being the same as the difference between sizes 3 and 4). However, we will see that a great many statistical procedures are, in fact, applicable to ordinal data.

**5. Discrete and continuous data :** However, when speaking of the number of leaves on a plant, we are dealing with a variable that can take on only certain values. It might be possible to observe 27 leaves, or 28 leaves, but 27.43 leaves and 27.9 leaves are values of the variable that are impossible to obtain. Such a variable is termed as *discrete or discontinuous variable* (also known as a *meristic variable*). The number of white blood cells in 1 mm<sup>3</sup> of blood, the number of giraffes visiting a water hole, and the number of eggs laid by a grasshopper are all discrete variables. The possible values of a discrete variable, generally, are consecutive integers.

**6. Continuous data :** When we speak of plant heights, we were dealing with a variable that could have any conceivable value within any observed range; this is referred to as a *continuous variable*. That is, if we measure a height of 35 cm and a height of 36 cm an infinite number of heights is possible in the range from 35 to 36 cm: a plant might be 35.07 cm tall or 35.988 cm tall, or 35.3263 cm tall, etc., although, of course, we do not have devices sensitive enough to detect this infinity of heights. A continuous variable is one for which there is a possible value between any other two possible values.

## 2.9 PARAMETER AND STATISTIC

Several measures help to describe or characterize a population. For example, generally, a preponderance of measurements occurs somewhere around the middle of the range of a population of measurement. Thus some indication of a population "average" would express a useful bit of descriptive information. Such information is called a *measure of central tendency*, and several such measures (e.g., the mean and the median).

It is also important to describe how dispersed the measurements are around the "average". This is, we can ask whether there is a wide spread of values in the population or whether the values are rather concentrated around the middle. Such a descriptive property is called a *measure of dispersion*, and several such measures (e.g., the range and the standard deviation).

*A quantity such as a measure of central tendency or a measure of dispersion is called a parameter when it describes or characterises a population.* Thus one rarely is able to calculate parameters. However, by random sampling of population parameters can be estimated very well. *An estimate of a population parameter is called a statistic.* It is biostatistical convention to represent population parameters by Greek letters and samples statistics by Latin letters.

The statistic one calculates will vary from sample to sample for samples taken from the same population. Since one uses sample statistic as estimates of population parameter, it allows the researcher to arrive at the "best" estimates possible. As for what properties to desire in a "good" estimate, consider the following.

First, it is desirable that if we take an indefinitely large number of samples from a population, the long run average of the statistic obtained will equal the parameter being estimated. That is, for some samples a statistic may underestimate the parameter of interest, and for others it may overestimate that parameter; but in the long run the estimates that are too low and those that are too high will "average out". If such a property is exhibited by a statistic, we say that we have an *unbiased statistic* or an *unbiased estimator*.

Second, not only should the deviations of the statistic from the parameter cancel out in the long run, but it is also desirable that a statistic obtained from any single sample from a population be very close to the value of the parameter being estimated. This property of a statistic is referred to as *precision, efficiency or reliability*. As one frequently secures only one sample from a population, it is important to arrive at a close estimate of a parameter from a single sample.

Third, keep in mind that one can take larger and larger samples from a population (the largest sample being the population itself). As the sample size increases, a *consistent* statistic will become a better estimate of the parameter it is estimating. Indeed, if the sample were the size of the population, the best estimate would equal the parameter itself.

## 2.10 ACCURACY AND PRECISION

“Accuracy” and “Precision” are used synonymously in everyday life, but in statistics we define them more rigorously. *Accuracy is the closeness of a measured or computed value to its true value. Precision is the closeness of repeated measurement of the same quantity.* A biased but sensitive scale might yield inaccurate but precise weight. By chance an insensitive scale might result in an accurate reading, which would however be imprecise, since a repeated weighting would be unlikely to yield an equally accurate weight. Unless there is bias in a measuring instrument, precision will lead to accuracy. We need, therefore, mainly be concerned with the former.

Precise variates are usually but not necessarily whole numbers. Thus, when we count four eggs in a nest, there is no doubt about the exact number of eggs in the nest if we have counted correctly; it is four, not three nor five and clearly it could not be four plus or minus a fractional part. Meristic variables are generally measured as exact numbers. Seemingly, continuous also be exact numbers. For instance, ratios between exact number are themselves also exact. If in a colony of animals there are 18 females and 12 males, the ratio of females to males is 1.5, a continuous variate but also an exact number.

For example, if we report that the hind leg of a frog is 8 cm long, we are stating the number 8 (a value of a continuous variable) as an estimate of the frog’s true leg length. This estimate was made using some sort of a measuring device. Had the device been capable of more accuracy, we might have concluded that the leg was 8.32 cm long. When recording values of continuous variables, it is important to designate the accuracy with which the measurements have been made. By convention the value 8 denotes a measurement in the range of 7.50000..... to 8.49999...., the value 8.3 designates a range of 8.25000.... to 8.34999.... and the value 8.32 implies that the true value lies within the range of 8.31500.... to 8.32499.... . That is, the reported value is the midpoint of the implied range, and the size of this range is designated by the last decimal place in the measurement. The value of 8 cm implies a range of accuracy of 1 cm, 8.3 cm implies a range of 0.1 cm and 8.2 cm implies a range of 0.01 cm. Thus, to record a value of 8.0 implies greater accuracy of measurement than does the recording of a value of 8; For in the first instance the true value is said to lie between 7.95000.... and 8.04999.... (e.g., within a range of 0.1 cm), whereas 8 implies a value between 7.50000.... and 8.49999.... (i.e., withing a range of 1 cm). To state 8.00 cm implies an accuracy in measurement which ascertains the frog’s limb length to lie between 7.99500.... and 8.00499.... cm (i.e., within a range of 0.01 cm). Those digits in a number that denote the accuracy of the measurement are referred to as *significant figures*. Thus, 8 has

one significant figure; 8.0 and 8.3 each have two significant figures; and 8.00 and 8.32 each have three significant figures.

In working with exact values of discrete variables, the preceding considerations do not apply. That is, it is sufficient to state that our frog has 4 limbs or that its left lung contains 13 flukes. The use of 4.0 or 13.00 would be inappropriate, for since the numbers involved are exactly 4 and 13, there is no question of accuracy or significant figures.

But there are instances where significant figures and implied accuracy come into play with discrete data. An entomologist may report that there are 72,000 moths in a particular forest area. In doing so, it is probably not being claimed that this is the exact number but an estimate of the exact number, perhaps accurate to 2 significant figures. In such a case 72,000 would imply a range of accuracy of 1000, so that the true value might lie anywhere from 71,500 to 72,500. If the entomologist wished to convey the fact that this estimate is believed to be accurate to the nearest 100 (i.e., to 3 significant figures), rather than to the nearest 1000, he had better present his data in the form of *scientific notation*, as follows: If the number  $7.2 \times 10^4$  (= 72,000) is written, a range of accuracy of  $0.1 \times 10^4$  (= 1000) is implied, and the true value is assumed to lie between 71,500 and 72,500. But if  $7.20 \times 10^4$  were written, a range of accuracy of  $0.01 \times 10^4$  (= 100) would be implied, and the true value would be assumed to be in the range of 71,950 to 72,050. Thus, the accuracy of large values (and this applies to continuous as well as discrete variables) can be expressed succinctly using scientific notation.

## 2.11 ACCURACY IN A SET OF OBSERVATIONS

Numbers in a discrete series may be *exact numbers*. If we perform an experiment using 15 subjects, we may speak of 15 as an exact number, since there is no margin of error. Numbers lacking this kind of accuracy are known as *approximate numbers*. There are two ways in which we may be confronted with a set of approximate numbers. First, our method of collecting data may have degraded the potential accuracy in a set of discrete values. This happens when we count the "house" at a theatre by estimating how many seats are filled, or in reporting the national debt Rs. 182,000,000,000, rather than Rs. 182,344,541,776,98. This may be potentially be achieved.

Second, we have seen that recorded measured obtained from a scale that is essentially continuous constitute a discrete scale. However, such recorded values are approximately numbers. If we are measuring height to be nearest half inch, recorded height of 5 ft  $4\frac{1}{2}$  in. means that the actual height lies somewhat within the range of 5 ft  $4\frac{1}{4}$  in. and 5 ft  $4\frac{3}{4}$  in. It should be clear that any measurement of a continuous variable must be treated as an approximate number.

In any event, it is up to the investigator to determine the degree of accuracy appropriate to his problem. A report of weight to the nearest pound will be adequate for determining your weight, but not be satisfactory for buying candy, much less gold! Once the desired degree of accuracy is determined, it falls in the computational domain to adjust procedures so that in the process of numerical manipulation we do not arrive at an outcome that pretends to greater accuracy than is warranted, or that unnecessarily loses desired accuracy. Soon you will ask, "How many decimal places should I keep?". The answer is, "Whatever seems sensible." We now know some of the factors involved, but there is one more to consider: the effect of sampling variation.

Very often we are interested in studying the characteristics of a sample not merely for its own sake, but because of the implication that the sample findings may have for population from which the sample was drawn. In this case, we must deal not only with accuracy in the data at hand, but with the fluctuation attributable to sampling. Of course, the characteristics that we obtain in a particular sample will not be exactly duplicated in a second sample. Although the observations taken might be exact numbers, they are subject to inaccuracy from the stand point of our primary interest in the characteristics of the population.

In general, it is good to keep a little more accuracy in the course of a series of statistical computations than that, we think is the minimum to which we are entitled, since in a sequence of computation it is possible to compound inaccuracy. Once a statistical computation is completed, we should round back to a figure that seems sensible. In so far as the factor of sampling variation is concerned, remember that an outcome based on a large sample has, other things being equal, greater stability than one based on a small sample.

## **2.12 LEVELS OF MEASUREMENT AND PROBLEMS OF STATISTICAL TREATMENT**

An understanding of the major types of measurement scales provides for a framework for understanding certain problems in the interpretation of data. In psychology and education, as well as in other behavioural sciences, most common measurements can not be demonstrated to have the full properties of interval or ratio scales. For example, a test of spelling, or any of the ordinary achievement test (including examinations prepared for college classes) almost certainly do not have an absolute zero. A score of zero in spelling means that the person could not answer the simplest question, but easier questions exist. At the same time, it is not clear that such tests have the property of measurement according to equal intervals. This would certainly be doubtful in the case of the spelling test if the words varied in difficulty.

We should, therefore, be alert to the necessity of resisting several erroneous but tempting propositions, such as the assertion that a person with an IQ of 100 is twice as bright as one with an IQ of 50, or that the difference between 15 and 25 points on a spelling test necessarily represents the same increment in spelling ability as the difference a score of 30 and 40 points on the same test. In psychological measurement, this problem may be particularly critical when a test does not have enough "top" or "bottom" to make adequate differentiation among the group measured. For example, imagine a test of ability that has maximum possible score of 50 points and that is too easy for the group measured. Between two persons who score 50 points, the score for one may indicate maximum level of attainment, but the second person with the same score may be capable of a much higher level of performance: the measuring instrument is simply incapable of showing it.

"Scale problems" can sometimes cause headache in the interpretation of research outcomes. Consider, for example, the problem of evaluating a method of teaching spelling. We want to know whether the method is of differential effectiveness for high- and low- achieving students. We select two such groups, measure their spelling performance before and after exposure to the teaching method, and consider comparing the average gain made by one groups with that made by the other. But, if the two groups, do not start at the same level (and they most likely will not in this study), we are in a poor position to compare the gains unless it is possible to assume that

a given gain at one point in the measurement scale represents the same increase in ability as an equal amount of gain in another part of the scale. In short, we must be able to assume an interval scale in order to be certain that we can make an entirely sensible interpretation of the comparison of gains. We are well advised to be alert to possible scale problems. Fortunately, the weight of the evidence suggests that in most situations, interpretability of statistical outcomes is not seriously incapacitated by uncertainty to the level of measurement achieved.

## 2.13 UNITS OF OBSERVATIONS

We know that a variable can take different values all of which may or not be actually observed in a given situation. Each observed or recorded value of a variable is associated with a place, person, object etc., and each recorded value of land area refers to a particular plot of land, each recorded value of a student refers to the marks he has secured or refers to the class in which he is studying. Thus in each case we have to clarify what each recorded value of a variable refers to or associated with. In order to overcome this difficulty, the term *unit of observation* or statistical units will be used to describe, what the values of a variable are attached to. A *statistical unit* is a well defined and identifiable object or group of objects with which the measurements or counts in any statistical investigation are associated. For example, in the results of a particular examination, the variable of observation could be the marks obtained and the corresponding units of observation would be the ‘student’. A very important step before the collection of data begins is to define clearly the statistical units on which the data are to be collected. In a number of situations the units are conventionally fixed like *physical units* of measurement such as meters, kilometres, kilograms, grams, quintals, hours, days, weeks, years, etc., which are well defined and need no elaboration or explanation.

However, in many statistical investigations, particularly relating to socio-economic studies, *arbitrary units* are used which must be clearly defined. The following points might serve as guidelines for deciding about the unit in any statistical investigations:

1. *It should be unambiguous and it must cover the entire population.*
2. *It should be specific.*
3. *It should be stable over a long period of time and also with respect to places.*
4. *It should be appropriate to enquiry.*
5. *It should be uniform and homogeneous throughout the investigation so that measurement obtained are comparable.*

The statistical units are classified as:

- (i) Units of collection
- (ii) Units of analysis and interpretation

## 2.14 THE SUMMATION SIGN

In statistics the operation of addition is used so frequently that a notational shortcut is needed to indicate summation. The Greek capital sigma ( $\Sigma$ ) is most often used as a *summation sign*. Appearing just to the left of an algebraic symbol this sign is read “the sum of”, for example  $\Sigma x_i$ , “the sum of  $x_i$ ’s”.

However, by itself the symbol  $\Sigma$  is of limited usefulness and can lead to ambiguities. For example, suppose that we want to sum only part of the  $x$ 's or suppose that there are several different kinds of  $x$ 's. Suppose, for instance that our  $x$ 's are heights of a group of college students and we want the sum of the heights of the sophomores only.

This leads us to introduce first some sort of tagging or lagging or labelling device. Subscripts are frequently used for this purpose. Suppose, for instance that our  $x$ 's heights of a group of college students and we want the sum of the heights of the sophomores only.

This leads us to introduce first some sort of tagging or lagging or labelling device. Subscripts are frequently used for this purpose. Suppose that we have measured the heights of  $n$  students. Let  $x_1$  be the height of the first student,  $x_2$  the height of the second student, and so on, letting  $x_n$  be the height of the  $n^{\text{th}}$  student. The letter  $i$  is frequently used as a roving index to help us keep track of the observations. Let  $x_i$  be the height of  $i^{\text{th}}$  student. It is important to realize that  $i$  identifies only the individual in whom we are interested and not the height.

Now we are in a position to introduce a precise notation for the sum of the heights of the  $n$  student. One way of writing this is  $x_1 + x_2 + x_3 + \dots + x_n$ . A shorter way, using our shorthand,

is  $\sum_{i=1}^n x_i$ . This symbol, which is now complete, is read "the sum of  $x_i$ , where  $i$  ranges from 1 to  $n$ ".

The  $i = 1$  written below the  $\Sigma$  tells us that this is the first value taken by  $i$  (we start with the first student). The  $n$  written above the  $\Sigma$  tells us that this is the last value taken by  $i$  (we finish with the last or  $n^{\text{th}}$  student). Now suppose that the sophomores in our sample are numbered from

1 through  $n$  inclusively. We can indicate the sum of their heights by  $\sum_{i=1}^n x_i$ .

The  $\Sigma$  is merely an operator which tells us what do with the  $x_i$ 's — namely, sum them. The  $i$  is an index or tag. The entries above and below the  $\Sigma$  tell us the range of the index over which we are to sum.

## EXERCISE

1. If we are interested in ascertaining the existence of a TV set in the residences of Area
  - (a) What is the observation to be recorded?
  - (b) What is an element?
  - (c) What is the population?
2. In some a school, suppose we are interested in the number of chairs in each room.
  - (a) What is the observation to be recorded?
  - (b) What is an element?
  - (c) What is the population?
3. We wish to set up experiment to test the relative effectiveness of morning hours and afternoon hours as a times for study. Identify several variables that it would be desirable to hold constant.

4. In a safety study, we might record the number of auto accidents incurred by each subject during a five year period. This variable could be considered as discrete or, from another point of view, as continuous. Explain.
5. The weights of nine children were recorded to the nearest pound and their average weight was found to be 118.555 lb. What do you think about reporting the average as:
  - (a) 118.555?
  - (b) 118.6?
  - (c) 120 (round to the nearest to 10 lb)?
6. Ram, is asked whether he would rather have a grade of *C* for sure, or a 50-50 chance of getting a *B* or a *D*. He replies that he would prefer the certain *C*. It must be that for him the psychological distance between a *B* and a *C* is less than the distance between a *C* and a *D*. Therefore, the grades *B*, *C* and *D* form a scale no higher in level of measurement than (which scale).
7. In one state, voters register as Congress, BJP, or independent. Can we consider this variable as one of the four “scales of measurement”? If so, which one?
8. Temperature in degrees Centigrade forms an interval scale. What kind of scale is formed by degrees Fahrenheit? Explain.
9. Lecture, assistant professor, associate professor, and professor form what kind of scale?
10. Assume the following series of numbers form an interval scale:  
0, 1, 2, 3, ..., 19, 20.
  - (a) Would it still be an interval scale if we added 10 points to each score? Explain.
  - (b) Would it still be an interval scale if we multiplied each score by 10? Explain.
11. In an interval scale, is it proper to consider that an increase of 20 points is twice as much as an increase of 10 points? Explain.
12. Explain the following terms:  
Population, precision and accuracy, data on a interval scale, Desired variables, attributes, ratio, index and rates.
13. Give the various levels of measurements of biological data.
14. Explain the various types of variables in biology.



# 3

# *Tabulation and Frequency Distribution*

## **3.1 TABULATION**

The last stage in the compilation of data is *tabulation*. After the data have been collected and classified. It is essential to put them in the form of tables with rows and columns. *Tabulation is a scientific process used in setting out the collected data in an understandable form.* It is used to make available answers to various questions concerning the enquiry at a glance. As classification is of simple and manifold form, so tabulations is simple and manifold. There are no hard and fast rules for preparing tables. But at the same time it is necessary that the tables should be prepared in such a way that they can be used to the best advantage with a minimum effort. The following general rules are to be born in mind for tabulation:

- (i) A rough draft of the table should be prepared first. Before drawing out the final table, rough draft should be examined very carefully.
- (ii) Figures to be compared should be placed as near to each other as possible.
- (iii) A suitable heading should be given to the table. It should be brief, comprehensive and self-explanatory.
- (iv) Heading of columns and rows should be brief and clear.
- (v) Columns and rows should be numbered to facilitate reference to the tables.
- (vi) The rows and columns should be arranged in a logical order.
- (vii) Explanatory notes should always be given as footnotes.
- (viii) The table should be self-explanatory and as simple as possible.
- (ix) Major items should be separated by bold or double lines.
- (x) The sources from which data are obtained should be given.

## **3.2 FREQUENCY TABLE OF FREQUENCY DISTRIBUTION**

Suppose there are 50 students in a class. Their heights in centimetres are given below:  
105, 101, 101, 109, 103, 122, 103, 104, 102, 101, 105, 103, 106, 119, 120, 116, 115, 118,

122, 109, 108, 107, 106, 105, 104, 103, 102, 106, 103, 109, 117, 114, 120, 122, 107, 116, 113, 119, 116, 101, 115, 110, 122, 107, 108, 105, 106, 101, 117, 109, 125.

The data given in the above form is **ungrouped data**. Therefore, to avoid confusion, we first of all write them in the ascending or the descending order. Arranging the above data in ascending order, we get,

101, 101, 101, 101, 101, 102, 102, 103, 103, 103, 103, 104, 104, 104, 105, 105, 105, 106, 106, 106, 107, 107, 107, 108, 108, 109, 109, 109, 110, 113, 114, 115, 115, 115, 116, 116, 116, 117, 117, 118, 119, 119, 120, 120, 122, 122, 122, 122.

We call this way of arrangement as **array** and the data are said to be **arrayed**. To make the work easier, we can further group these figures in the form of a table.

**Table 1**

Heights (in cm)	No. of students	Heights (in cm)	No. of students
101	5	110	1
102	2	113	1
103	5	114	1
104	2	115	2
105	4	116	4
106	3	117	2
107	3	118	1
108	2	119	3
109	3	120	2
		122	4

The method of arranging the given data in the above form is known as **frequency distribution**. Here height called **variate** and the corresponding number of students is called **frequency of the variate**. Thus frequency is the number of times a variate has been repeated. Though the frequency table given above is an improvement over the arranged data, yet we can further simplify it by classifying it into groups:

**Table II**

Heights (in cm)	No. of students
101 – 105	18
106 – 110	12
111 – 115	4
116 – 120	12
121 – 125	4
	<b>50</b>

Such an arrangement is called **grouped frequency distribution**. In the table the first class-interval is 101-105. In it 101 is called the **lower limit** and 105 the **upper limit** of the class-interval. All those students whose heights are between 101 cm and 105 (both inclusive) are

placed into this group. This group consists of 18 students. We do not feel any difficulty in making such frequency table, but sometimes we come across a difficulty. Let the height of a student be 105.5 cm. Now the question is whether we should place him in the class-interval (101-105) or (106-110). Obviously we cannot place him in the class-interval (101-105), because his height is 105.5 cm and the upper limit of the class interval is 105 cm. Similarly, we cannot place him in the class-interval (106-110), because his height is 105.5 cm and the lower limit of the class-interval is 106. Now where should we put him? In such cases, we form the class intervals as given below:

**Table III**  
**Heights in centimetres**

100 to below 105
105 to below 110
110 to below 115
115 to below 120
120 to below 125

Thus, we overcome the difficulty mentioned above, because now we put the student whose height is 105.5 cm in the second group. In practical use we shall write the above table as:

**Table IV**  
**Heights in cm**

100 – 105
105 – 110
110 – 115
115 – 120
120 – 125

Such an arrangement as given by Table III or IV is called continuous frequency distribution, while the frequency distribution given by Table I or II is known as discontinuous distribution.

### 3.3 PREPARATION OF A FREQUENCY TABLE

- (i) Arrange the scores in ascending order to form an array.
- (ii) Draw a table consisting of three columns:
  - (a) Class Interval,      (b) Tally,      (c) Frequency.
- (iii) Bearing in mind the lower and the upper limits, write down the class intervals or the variables in the first column.
- (iv) Against each interval or the variable, write down as many vertical lines in the “Tally column” as the number of scores it contains.
- (v) Count the number of vertical lines, crossing of 4 lines to be counted as 5 and put down the number in the ‘frequency column’.

**Note :** The total of the frequency column must be equal to the total number of items of the given data.

**Example 1 :** Form a frequency table for the following variables :

59, 52, 51, 60, 68, 63, 64, 65, 66, 67, 68, 52, 59, 58, 60, 51, 54, 55, 56, 61, 62, 69, 70, 58, 69, 65, 67, 66, 63, 63, 62, 61, 51, 55, 59, 63, 68, 67, 69, 53, 53, 51, 59, 56, 55, 70, 65, 62, 65, 66, 69, 70, 52, 55, 64, 65, 69, 70, 61, 63, 54, 64, 61, 61, 62, 51, 52, 52, 54, 55.

**Solution :** Arranging the given data in the form of an array we get:

51, 51, 51, 51, 51; 52, 52, 52, 52, 52; 53, 53, 54, 54, 54; 55, 55, 55, 55, 55; 56, 56; 58, 58; 59, 59, 59, 59; 60, 60; 61, 61, 61, 61, 61; 62, 62, 62, 62; 63, 63, 63, 63, 63; 64, 64, 64, 65, 65, 65, 65; 66, 66, 66; 67, 67, 67; 68, 68, 68; 69, 69, 69, 69, 69; 70, 70, 70, 70.

Values of variables (i)	Tally (ii)	Frequency (iii)
51		5
52		5
53		2
54		3
55		5
56		2
58		2
59		4
60		2
61		5
62		4
63		5
64		3
65		5
66		3
67		3
68		3
69		5
70		4
Total		70

**Example 2 :** Present the following data of the sample of frequency of 40 households taken from a factory in the form of a frequency table with 9 classes with class interval data 100.

200	120	350	550	400	140	350	85	180	110	110
600	350	500	450	200	170	90	170	800	190	700
630	170	210	185	250	120	180	350	110	250	430
140	300	400	200	400	210	305				

**Solution :** Arranging the given data in the ascending does order, we get:

85	90	110	110	110	120	120	140	140	170	170
170	180	180	185	190	200	200	200	210	210	250
250	300	305	350	350	350	350	400	400	400	430
450	500	550	600	630	700	800				

Taking a class interval of 100 starting 0 – 100 we have following frequency distribution table.

**Frequency Distribution Table**

Class Interval	Frequency	
0 – 100		2
100 – 200		14
200 – 300		7
300 – 400		6
400 – 500		5
500 – 600		2
600 – 700		2
700 – 800		1
800 – 900		1
	Total	40

### 3.4 RELATIVE FREQUENCY DISTRIBUTION

We know that the frequency is defined as the total number of data points that fall within that class. Frequency of each class can also be expressed in a fraction or percentage terms. These are known as **relative frequencies**. In other words, a *relative frequency* is the class frequency expressed as a rate of total frequency, i.e.,

$$\text{Relative Frequency} = \frac{\text{Class Frequency}}{\text{Total Frequency}}$$

A relative frequency distribution is given by the following table of marks secured by 25 students out of 100.

**TABLE : Relative Frequency Distribution for the Collection Days**

Class (marks)	Frequency (No. of students)	Relative Frequency
20 – 40	6	$6/25 = 0.24 = 24\%$
40 – 60	12	$12/25 = 0.48 = 48\%$
60 – 80	4	$4/25 = 0.16 = 16\%$
80 – 100	3	$3/25 = 0.12 = 12\%$
Total	25	1.00 or 100%

It may be observed that the sum of all relative frequencies is 1.000 or 100 per cent because the frequency of each class has been expressed as a percentage of the total frequencies. (or data).

The relative frequencies express the frequency of any class as a percentage of the total frequency. It is obtained by the formula.

$$\text{Relative frequency} = \frac{\text{Frequency of item in a group}}{\text{Total frequency of the group}}$$

The relative frequencies are used to compare two or more frequency distributions or two or more items in the same frequency distribution.

In the above example, we can compare the number of students securing the marks 20 – 40 and 80 – 100. As there are 24% students who secured marks between 20 and 40 and 12% students have secured marks between 80 and 100. In the same manner, we can compare the production of rice in Punjab and U.P. in 2005.

It is important to note that for comparing two frequency distributions by means of relative frequencies, the variates as well as the divisions of the values of variable into classes must be same for both the distributions. Relative frequencies are used in percentage sub-divided Bar diagrams and Pie diagrams.

### 3.5 CUMULATIVE FREQUENCY DISTRIBUTION

*Cumulative frequency corresponding to a class is the sum of all the frequencies up to and including that class.* In cumulative frequency distribution the frequency of a particular class is obtained by adding to the frequency of that class all the frequencies of its previous classes. Thus the cumulative frequency table is obtained from the ordinary frequency table by successively adding the several frequencies.

Cumulative frequency series are of two types:

- (i) “Less than series”    (ii) “More than series”.

Suppose we are given the following discrete series of marks obtained by 100 students. With the help of this series, we shall form the ‘Less than’ and ‘More than’ series.

Marks obtained	No. of students
30 – 40	8
40 – 50	12
50 – 60	20
60 – 70	25
70 – 80	18
80 – 90	17

#### Less than Cumulative Series

Marks obtained	No. of students
Less than 40	8
Less than 50	20

Less than 60	40
Less than 70	65
Less than 80	83
Less than 90	100

**More than Cumulative Series**

Marks obtained	No. of students
More than 30	100
More than 40	92
More than 50	80
More than 60	60
More than 70	35
More than 80	17

**Example 3 :** The marks obtained by 35 students of 11<sup>th</sup> class of a school are:

628, 665, 560, 328, 421, 525, 326, 480, 470, 405, 421, 664, 668, 620, 300, 305, 520, 420, 370, 326, 440, 328, 480, 565, 650, 480, 360, 325, 450, 360, 426, 440, 306.

Form a cumulative frequency table with class interval of 50.

**Solution :** Let us arrange the given data in the ascending order of the magnitudes.

300, 305, 306, 325, 326, 326, 328, 328, 360, 360, 360, 370, 405, 420, 420, 421, 421, 426, 440, 440, 450, 470, 480, 480, 480, 520, 525, 560, 565, 620, 628, 650, 664, 665, 668.

Let us now put them in a group of class interval 50 in the following cumulative frequency distribution form:

Class Interval	Frequency	Cumulative Frequency
300 – 350	8	8
351 – 400	4	12
401 – 450	9	21
451 – 500	4	25
501 – 550	2	27
551 – 600	2	29
601 – 650	3	32
651 – 700	3	35

**Example 4 :** The heights (in centimetres) of 40 persons are:

110, 112, 125, 135, 150, 152, 150, 155, 159, 130, 128, 138, 133, 143, 147, 151, 154, 156, 112, 116, 119, 111, 113, 115, 118, 121, 123, 120, 125, 121, 110, 113, 114, 149, 153, 155, 150, 156, 152, 111.

Array the data and form a cumulative frequency table with class interval of 10.

**Solution :** Arranging the given data in ascending order of magnitude, we get

110, 110, 111, 111, 112, 112, 113, 113, 114, 115, 116, 118, 119, 120, 121, 121, 123, 125, 125, 128, 130, 130, 133, 135, 138, 143, 147, 149, 150, 150, 150, 151, 152, 152, 153, 154, 155, 155, 156, 159.

Class interval (i)	Tally (ii)	Frequency (iii)	Cumulative frequency (iv)
110 – 120		5	5
120 – 130		6	11
130 – 140		4	15
140 – 150		3	18
150 – 160		5	23

**Example 5 :** Form a frequency table from the following data :

Age (in year)	No. of persons	Age (in year)	No. of persons
under 5	0	under 40	22
under 10	7	under 45	29
under 15	10	under 50	35
under 20	14	under 55	41
under 25	16	under 60	47
under 30	17	under 65	50
under 35	20		

**Solution :** Here cumulative frequencies of the variates have been given. So to get the frequency of a particular class-interval, subtract the cumulative frequency of the lower limit from that of the upper limit.

Age (in year)	No. of persons	Age (in years)	No. of persons
5 – 10	7	35 – 40	2
10 – 15	3	40 – 45	7
15 – 20	4	45 – 50	6
20 – 25	2	50 – 55	6
25 – 30	1	55 – 60	6
30 – 35	3	60 – 65	3

### EXERCISE

1. What are the functions and importance of tabulation in the scheme of investigation?
2. Define the following terms: Class interval, frequency, array, cumulative frequency, lower and upper limit of a class interval.

3. What are the different types of statistical data? Give the different methods of collecting the data.

4. Array the data and form a frequency table for the following:

240, 288, 298, 239, 231, 240, 230, 233, 239, 240, 230, 237, 236, 235, 239, 238, 232, 228, 236, 235, 237, 239, 235, 232, 239, 231, 230, 244, 239, 231, 245, 232, 239, 240, 237.

5. The marks obtained by 30 students, out of 10 are:

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

Array the data and form the frequency distribution.

6. Present the following data of the marks of 60 students in the form of a frequency table with 10 classes of equal intervals, the first interval being 40 – 49:

63, 63, 54, 92, 60, 58, 70, 6, 67, 82, 57, 49, 34, 73, 54, 63, 36, 52, 32, 75, 9, 79, 28, 30, 42, 93, 43, 80, 3, 32, 67, 24, 64, 63, 11, 35, 82, 10, 23, 0, 41, 32, 72, 53, 92, 88, 52, 55, 60, 33, 40, 57.

7. Form a frequency table from the following data:

Marks	No. of students	Marks	No. of students
Below 5	7	Below 30	70
Below 10	15	Below 35	75
Below 15	25	Below 40	95
Below 20	35	Below 45	100
Below 25	50	Below 50	120

8. Make a discrete series from the following data :

15, 18, 16, 20, 21, 25, 25, 15, 16, 18, 22, 23, 25, 20, 18, 22, 24, 25, 24, 25, 23, 20, 15, 16, 25, 19, 18, 22, 21, 18.

9. Following are the marks obtained by 40 students in mathematics:

75, 80, 83, 85, 90, 95, 96, 99, 100, 45, 72, 76, 86, 80, 95, 49, 53, 45, 40, 75, 79, 48, 38, 75, 80, 85, 87, 82, 86, 87, 91, 93, 83, 73, 63, 67, 53, 43, 91, 81.

Make a frequency distribution taking 10 as magnitude of class interval.

10. Here are given the following marks secured by 25 students in the examination :

23, 28, 30, 32, 35, 36, 40, 41, 43, 44, 45, 45, 48, 49, 52, 53, 54, 56, 56, 58, 61, 62, 65, 68.

(a) Arrange the above data on frequency distribution taking class intervals 20 – 29, 30 – 39, 40 – 49, 50 – 59, 60 – 69.

(b) Form the cumulative frequency distribution.

11. From a frequency distribution from the following data taking the appropriate magnitude of class intervals:

Weight of 40 students in pounds.

138, 164, 115, 150, 120, 102, 140, 150, 115, 125, 140, 147, 125, 142, 115, 103, 142, 155, 115, 125, 163, 119, 135, 111, 125, 107, 145, 165, 125, 130, 142, 135, 160, 119, 109, 110, 146, 166, 116, 136.

# 4

# *Graphical Representation of Data*

## **4.1 GRAPHICAL REPRESENTATION OF STATISTICAL DATA**

The representation of quantitative data suitably through charts and diagrams is known as **Graphical Representation of Statistical Data**. Graphs include both **charts** and **diagrams**.

*A graphic representation is the geometrical image of a set of data.* It is a mathematical picture. It enables us to think about a statistical problem in visual terms. A picture is said to be more effective than words for describing a particular thing or phenomenon. Consequently, the graphic representation of data proves quite an effective and economic device for presentation, understanding and interpretation of the collected statistical data.

### **4.1.1 Advantages of Graphical Representation**

1. *It is easily understood by all.*
2. *The data can be presented in a more attractive form appealing to the eye.*
3. *It shows relationship between two or more sets of figures.*
4. *It shows the trend and tendency of values of the variables.*
5. *It helps interpolation of the values of the variables.*
6. *It has a universal applicability.*
7. *Various valuable statistics like median, mode, quartiles, may be easily observed.*
8. *Comparative analysis and interpretation may be effectively and easily made by graphic representation.*
9. *It provides a more lasting effect on the brain. It is possible to have an immediate and meaningful grasp of large amounts of data through such presentation.*
10. *The real value of graphical representation lies in its economy and effectiveness. It carries a lot of communication power.*
11. *It may help in the proper estimation, evaluation and interpretation of the characteristics of items and individuals.*

12. The graphical representation helps in forecasting as it indicates the trend and movements of the data in the past.

#### 4.1.2 Disadvantages of Graphical Representation

1. It takes a lot of time to prepare a graph.
2. A graph does not show all the facts (data) in detail.
3. It depicts only approximate values.

#### 4.2 TYPES OF GRAPHS

There are various types of graphs in the form of charts and diagrams. Some of them are:

1. Line chart, 2. Ratio chart, 3. Bar chart, 4. Pie chart, 5. Pictograph, 6. Histogram, 7. Frequency polygon, 8. Cumulative frequency polygon or Ogive and so on.

Though there are many types of graphs but an understanding of a few general types will be suffice for most ordinary medical data. The choice of a particular form of graph to be used is often a matter of personal preference. There are, however, certain general principles that are commonly accepted as preferable. Some of the most important of these are:

1. Every graph should be completely self-explanatory. Therefore it should be correctly labelled as to title, source, scales, and explanatory keys or legends.
2. When more than one variable is shown on a graph, each should be clearly differentiated by means of legends or keys.
3. The simplest type of graph consistent with its purpose is the most effective. No more lines or symbols should be used in a single graph than the eye can easily follow.
4. The position of the title for a graph is one of personal choice. In published graphs, however, the title is commonly placed below the graph.
5. The diagram or graph generally precedes from left to right and from bottom to top. All writing should be placed therefore, so as to be read from the bottom or from the right-hand side of the page.
6. Frequency is generally represented on the vertical line, method of classification on the horizontal line.
7. Scale divisions should be clearly indicated as well as the units into which the scale is divided.
8. On an arithmetic scale equal increments on the scale must represent equal numerical units.
9. On the scale representing frequency, the numerical scale should start with zero. If this is not possible the position of the zero line should be shown with a break in the vertical line or horizontal line clearly indicating that a break has been made. The break is generally, shown as: ‘—W—’.

#### 4.3 MODES OF GRAPHICAL REPRESENTATION OF DATA

We know that the data in the form of raw scores is known as ungrouped data and when it is organised into a frequency distribution, then it referred to as grouped data. Separate modes

and methods are used to represent these two types of data — ungrouped and grouped. Let us discuss them under separate heads.

### Graphical Representation of Ungrouped Data

For the ungrouped data (data not grouped into a frequency distribution) we usually make use of the following graphical representation:

1. Line graphs
2. Bar graph or Bar diagrams
3. Circle graph or Pie diagrams
4. Pictograms

#### 4.4 LINE GRAPH

Line graphs are simple mathematical graphs that are drawn on the graph paper by plotting the data concerning one variable on the horizontal  $x$ -axis and other variable of data on the vertical  $y$ -axis. With the help of such graphs the effect of one variable upon another variable during an experimental or normative study may be clearly demonstrated. The construction of these graphs can be understood through the following example:

**Example 1.** A word-nonsense syllables association test was administered on a student of class X to demonstrate the effect of practice on learning. The data so obtained may be studied from the following table.

Trial No.	1	2	3	4	5	6	7	8	9	10	11	12
Score	4	5	8	8	10	13	12	12	14	16	16	16

Draw a line graph for the representation and interpretation of the above data.

**Solution :** Plot the points  $(1, 4)$ ,  $(2, 5)$ ,  $(3, 8)$ ,  $(4, 8)$ ,  $(5, 10)$ ,  $(6, 13)$ ,  $(7, 12)$ ,  $(8, 12)$ ,  $(9, 14)$ ,  $(10, 16)$ ,  $(11, 16)$  and  $(12, 16)$ .

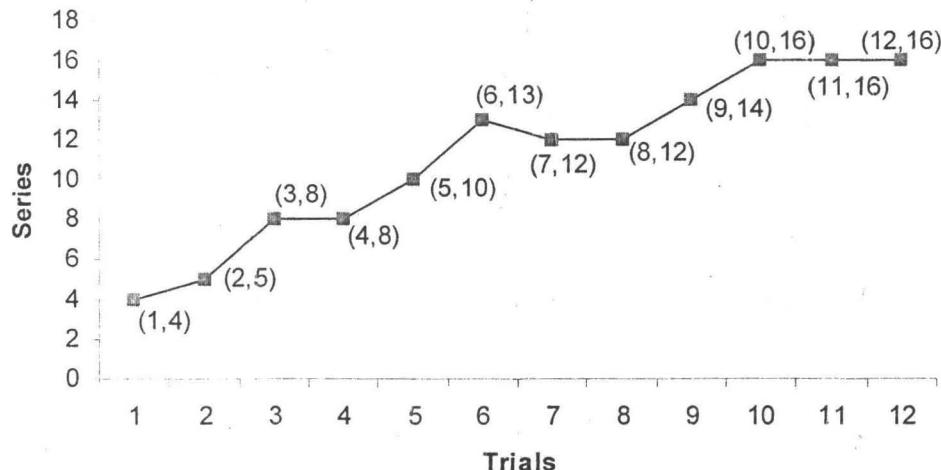
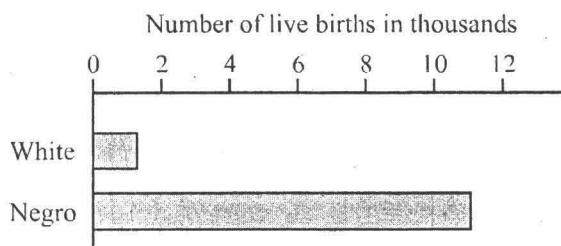


Fig. 4.1 : Line graph — The effect of practice on learning

#### 4.5 BAR DIAGRAM

The simplest type of graph that can be used is the bar diagram. It is especially useful in comparing qualitative data or quantitative data of discrete type. A bar diagram is a graph on which the data are represented in the form of bars. It consists of a numbers of bars or rectangles which are of uniform width with equal space between them on the  $x$ -axis. The length of the bar is proportional to frequency or the value it represents. It should be seen that the bars are neither too short nor too long. The scale should be clearly indicated and base line be clearly shown.



*Fig. 4.2 : Number of live births to white and Negro women.*

In this diagram the bars representing births to white and Negro women are drawn of equal width and with lengths equal to the respective number of births. Their areas, therefore, is proportional of their length. Hence, comparison of the lengths of the two bars gives a visual picture of the births at this hospital by race, showing that there were about ten times as many babies born to Negro mothers as to white mothers.

Bars may be drawn either horizontally or vertically. A good rule to use in determining the direction is that if the legend describing the bar can be written under the bars when drawn vertically, vertical bars should be used; when it cannot be, horizontal ones must be used. In this way the legends can be read without turning the graph. The descriptive legend should not be written at the end of the bars or within the bars, since such writing may distort the comparison. Usually the diagram will be more attractive if the bars are wider than the spaces between them.

The width of bars is not governed by any set rules. It is an arbitrary factor. Regarding the space between two bars, it is conventional to have a space about one half of the width of a bar.

The data capable of representation through bar diagrams, may be in the form of raw scores, total scores or frequencies, computed statistics and summarised figures like percentages and averages etc.

The bar diagram is generally used for comparison of quantitative data. It is also used in presenting data involving time factor. When two or more set of data over a certain period of time are to be compared a group bar diagram is prepared by placing the related data side by side in the shape of bars. The bars may be vertical or horizontal in a bar diagram. If the bars are placed horizontally, it is called a **Horizontal Bar Diagram**. When the bars are placed vertically, it is called a **Vertical Bar Diagram**. There are four types of Bar diagram: (i) **Simple Bar diagram**, (ii) **Multiple or Grouped Bar diagram**, (iii) **Subdivided or Component Bar diagram**, (iv) **Percentage Subdivided Bar diagram**.

(i) **Simple Bar Diagram** : It is used to compare two or more items related to a variable. In this case the data are presented with the help of bars. These bars are usually arranged according

to relative magnitude of bars. The length of a bar is determined by the value or the amount of the variables. A limitation of **Simple Bar diagram** is that only one variable can be represented on it. The method is illustrated by the following example:

**Example 2.** Draw a Bar Chart of the procurement of rice (in tons) in an Indian state:

Year :	1978	1979	1980	1981	1982	1983
Rice (in tons) :	4500	5700	6100	6500	4300	7800

**Solution :** The given data is represented by the **Simple Bar Diagram** as only one variable is to be represented.

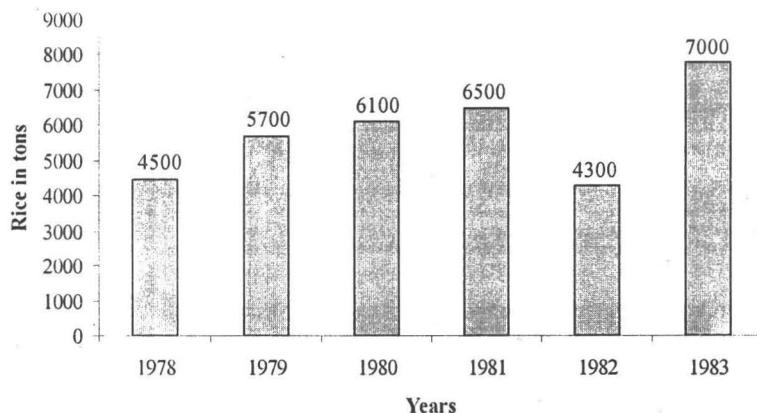


Fig. 4.3

Here we represent years on the  $x$ -axis and procurement of rice on the  $y$ -axis.

(ii) **Multiple or Grouped Bar Diagram** : A multiple or grouped bar diagram is used when a number of items are to be compared in respect of two, three or more values. In this case, the numerical values of major categories are arranged in ascending or descending order so that the categories can be readily distinguished. Different shades or colours are used for each category.

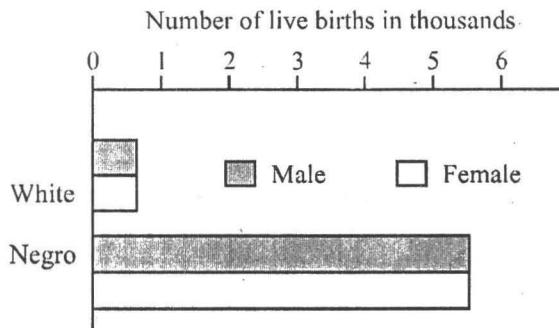


Fig. 4.4 : Number of live births by sex to white and Negro women

Sub classification of qualitative data may be shown by the use of multiple bars (Fig. 4.4). When using multiple bars there are two alternatives: either the bars representing the subclassification

may be drawn contiguous to one another as in Figure 4.4 or they may be separated by a small space. If the latter method is used the space between the bars in a single group should be narrower than the space between the group of bars. In all cases when multiple bars are used a key is necessary for differentiating the various subclassifications.

**Example 3.** Represent the following data by a suitable diagram showing the difference between proceeds and costs:

Proceeds and costs of a firm (in thousands of rupees).

Year	Total Proceeds	Total costs
2000	22.0	19.5
2001	27.3	21.7
2002	28.2	30.0
2003	30.3	25.6
2004	32.7	26.1
2005	33.3	34.2

**Solution :** In this case the two types of information — proceeds and costs — are shown on a diagram indicating the difference between them. The two types of information for each year are placed in such a way that a comparison may be made between them. So a grouped bar diagram is drawn in order to represent the given data.

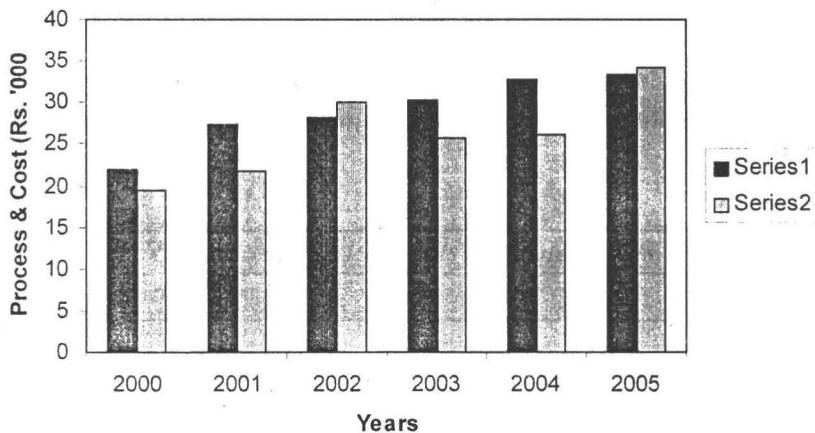


Fig. 4.5 : Grouped Bar Diagram showing Proceeds and Costs

(iii) **Sub-divided or Component Bar Diagram :** A component bar diagram is one which is formed by dividing a single bar into several component parts. A single bar represents the aggregate value whereas the component parts represent the component values of the aggregate value. It shows the relationship among the different parts and also between the different parts and the main bar. The design and procedure of constructing it is similar to simple bar diagrams except that in this form of presentation each bar is subdivided into its components. Different shades or colours or designs or different types or cross hatchings are used to distinguish the various components. The method is clear by the following example. ●

**Example 4.** Represent the following data of the development expenditure of Central Governments in India during 1977-78, 1978-79, 1979-80 by bar diagram.

Year	Loans and advances	Capital	Revenue	Total
2001-02	8,601	3,787	3,477	15,865
2002-03	10,535	4,456	4,036	18,827
2003-04	11,549	4,803	3,709	20,061

**Solution :** In this case we use subdivided bar diagram. Here the data for different years is represented by various parts of the single bar for the year. Different parts are indicating loans and advances capital and revenues for each year as shown in the figure.

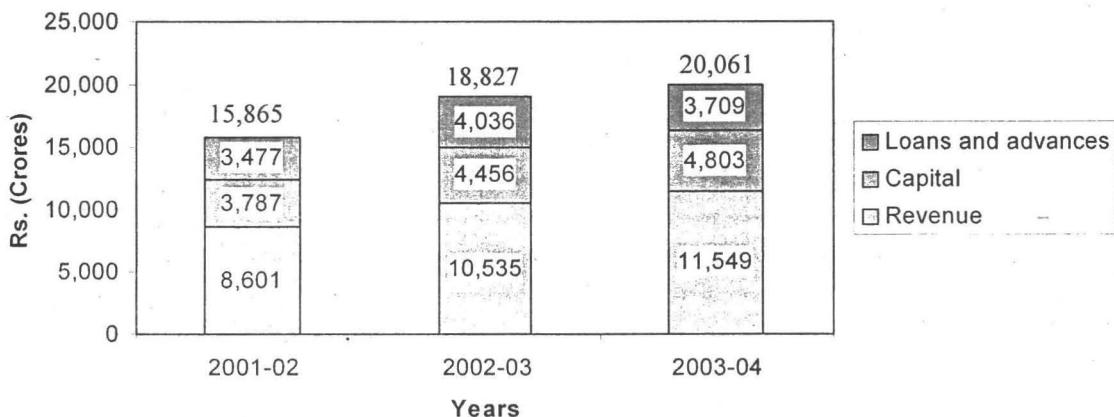


Fig. 4.6 : Development Expenditure of Central and State Government in India.

**(iv) Percentage Sub-divided Bar Diagram :** It consists of one or more than one bars where each bar totals 100%. Its construction is similar to the subdivided bar diagram with the only difference that whereas in the sub-divided bar diagram segments are used in absolute quantities, in the percentage bar diagram the quantities are transformed into percentages. Its construction is based on the following steps:

- Step I.** Convert the quantities in each case into percentage of the whole.
- Step II.** Take the cumulative percentages.
- Step III.** Represent each category of items in different shades or colours or different types of cross-hatchings.
- Step IV.** In case two diagrams showing different periods are given, then show each category of item with the same shade or colour or cross-hatchings.

**Example 5.** Following are the heads of income of Railway during 2004 and 2005.

	2004 (in crores of rupees)	2005 (in crores of rupees)
Coaching	26	31
Goods	40	39
Others	4.50	3.50

Represent the above data by a bar chart.

**Solution :** We are given three types of information – coaching, goods and other for two years. In order to facilitate comparison among them and also between the two years, a component bar chart is drawn to represent the given data. The graph has been drawn on percentage figures. Percentage of each item has been expressed on the total income of respective years by the formula.

$$\text{Percentage} = \frac{\text{Income of the item}}{\text{Total income of respective year}} \times 100$$

TABLE : Calculation of Percentage

Items	2004		2005	
	Income	Percentage	Income	Percentage
Coaching	26	36.88	31	42.18
Goods	40	56.74	39	53.06
Other	4.50	6.38	3.50	4.76
	70.50	100.00	73.50	100.00

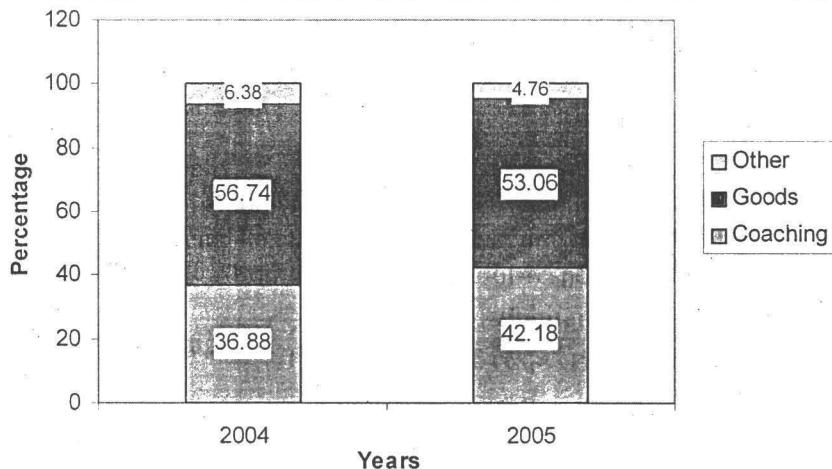


Fig. 4.7 : Percentage subdivided bar chart showing incomes of Railways.

#### 4.6 PIE CHART OR CIRCLE CHART OR SECTOR CHART

A pie chart is a circular graph which represents the total value with its components. The area of a circle represents the total value and the different sectors of the circle represent the different parts. The circle is divided into sectors by radii and the areas of the sectors are proportional to the angles at the centre. It is generally used for comparing the relation between various components of a value and between components and the total value. In pie chart, the data is expressed as percentage. Each component is expressed as percentage of the total value. A pie chart is also known as **circular chart or sector chart**.

The name pie diagram is given to a circle diagram because in determining the circumference of a circle we have to take into consideration a quantity known as ‘pie’ (written as  $\pi$ ).

*Method of construction :* The surface area of a circle is known to cover  $2\pi$  radians or 360 degrees. The data to be represented through a circle diagram may therefore be presented

through 360 degrees, parts or sections of a circle. The total frequencies or value is equated to  $360^\circ$  and then the angles corresponding to component parts are calculated (or the component parts are expressed as percentages of the total and then multiplied by  $360/100$  or  $3.6$ ). After determining these angles, the required sectors in the circle are drawn.

### WORKING RULE

- Step I.** Plot a circle of appropriate size with protractor, pencil and compass. The angles of a circle at the centre is  $360^\circ$ .
- Step II.** Convert the given value of the components of an item in percentage of the total value of the item.
- Step III.** In laying out the sector for a pie chart it is logical to adopt the common procedure to arrange sectors according to size with the largest at the top and the others in sequence running clockwise.
- Step IV.** Transpose the various component values correspond to the degrees on the circle. Since 100% is represented by  $360^\circ$ , the angle at the centre of the circle, therefore, 1% value is represented by  $360/100 = 3.6^\circ$ . If 8 be the percentage of a certain component, the angle which represent the percentage of such component is  $(3.6 \times 8)$  degrees.
- Step V.** Measure with protractor the points on a circle representing the size of each sector.
- Step VI.** Label each sector for identification.

The method is illustrated by the following example.

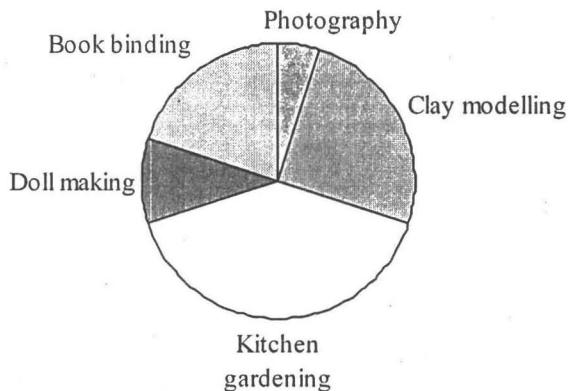
**Example 6.** 120 class XII students of a school were asked to opt for different work experiences. The details of these options are as under.

Areas of work experience	No. of students
Photography	6
Clay modelling	30
Kitchen gardening	48
Doll making	12
Book binding	24

Represent the above data through a pie chart.

**Solution :** The numerical data may be converted into the angles of the circle as given below:

Areas of work experience	No. of students	Angle of the circle
Photography	6	$6/120 \times 360 = 18^\circ$
Clay modelling	30	$30/120 \times 360 = 90^\circ$
Kitchen gardening	48	$48/120 \times 360 = 144^\circ$
Doll making	12	$12/120 \times 360 = 36^\circ$
Book binding	24	$24/120 \times 360 = 72^\circ$
Total	120	$360^\circ$



*Fig. 4.8 : Pie diagram – Area of work experience opted for students of class XII.*

**Example 7.** Draw a pie chart to represent the following data on the proposed outlay during the Fourth Five-Year Plan.

Item	Agriculture	Industrial and Minerals	Irrigation and Power	Communication	Miscellaneous
Rs. (in crores)	6,000	4,000	2,500	4,500	3,000

**Solution :** In constructing a pie chart, it is necessary to convert the percentages into angles of different degrees.

#### Calculation for Pie chart

Items	Amount (Rs. in crores)	Percentage on Total (%)	Angle for each percentage ( $360^\circ \times \frac{100}{100}$ )	Angle for each item at the centre of the Pie chart
Agriculture	6,000	$\frac{6,000 \times 100}{20,000} = 30$	3.6	$30 \times 3.6 = 108$
Industries and Minerals	4,000	$\frac{4,000 \times 100}{20,000} = 20$	3.6	$20 \times 3.6 = 72$
Irrigation and Power	2,500	$\frac{2,500 \times 100}{20,000} = 12.5$	3.6	$12.5 \times 3.6 = 45$
Communications	4,500	$\frac{4,500 \times 100}{20,000} = 22.5$	3.6	$22.5 \times 3.6 = 81$
Miscellaneous	3,000	$\frac{3,000 \times 100}{20,000} = 15$	3.6	$15 \times 3.6 = 54$
Total	20,000	100		360

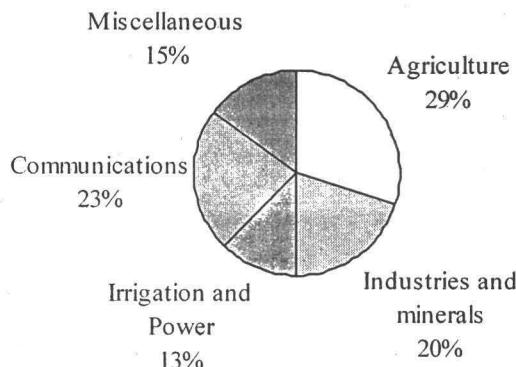


Fig. 4.9.

#### 4.7 PICTOGRAPH OR PICTOGRAM

It is a device of representing statistical data in pictures. In pictograph, a number of pictures of same size and equal in value are drawn. Each picture represents a number of units. Pictures are generally drawn side by side horizontally or vertically. It is widely used by government organisations as well as by private institutions. The chief advantage of this method is its attraction. In this method of representation each picture is assigned some numerical value which can be expressed either by denoting it or writing it on the edge of each picture. While drawing pictograph, it should be borne in mind that the pictures drawn are simple, clear and easy to understand, and the number of units which each picture represents should be clearly stated. It is illustrated by the following example:

Year	Strength of Students
2002 – 03	 650
2003 – 04	 500
2004 – 05	 800

Seats  
= 100

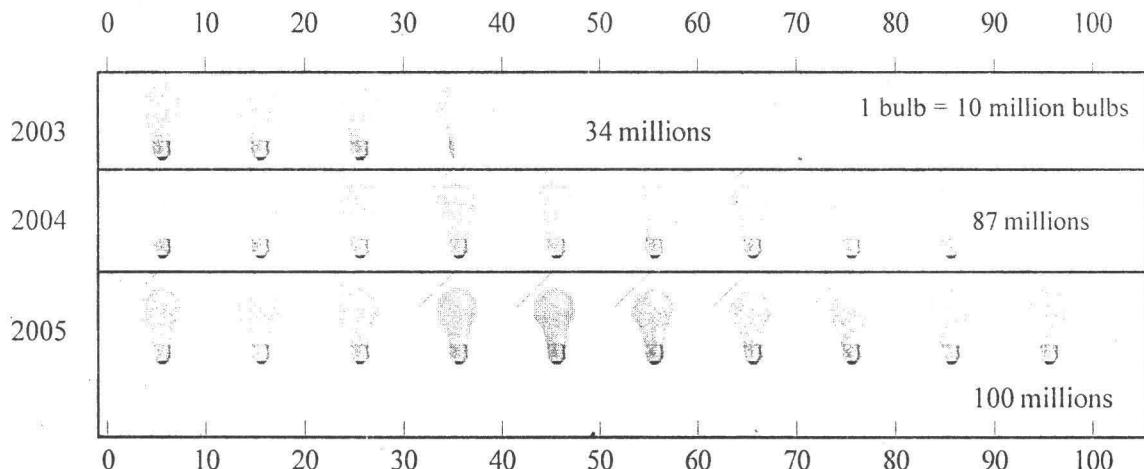
Note : Each student represents a strength of 100.

Fig. 4.10 : Strength of students in different years at a Govt. Boy's High School.

**Example 8.** Represent the following data of the production of electric bulbs of a factory by pictogram:

Year	2003	2004	2005
Production of bulbs in Million	34	67	100

**Solution :** The given data is represented by pictogram as shown in figure:



## 4.8 GRAPHICAL REPRESENTATION OF GROUPED DATA (FREQUENCY DISTRIBUTION)

There are four methods of representing a frequency distribution graphically:

1. Histogram
2. Frequency polygon
3. Cumulative change diagram
4. Proportional change diagram
5. Ratio diagram or Arithlog

## 4.9 HISTOGRAM

A Histogram is a graph containing a set of rectangles, each being constructed to represent the size of the class interval by its width and the frequency in each class-interval by its height. The area of each rectangle is proportional to the frequency in the respective classinterval and the total area of the histogram is proportional to the total frequency. A histogram is used to depict a frequency distribution.

In constructing a histogram, the class-boundaries are considered to be very important; and they are located on the horizontal axis of the graph. The vertical rectangles are erected side by side on the basis of the frequencies over the classintervals which are extended to their class boundaries.

When the class-intervals are unequal, difficulty arises, because the rectangles so formed will also of unequal width. So in order to remove this difficulty, the heights of rectangles are made proportional not to the class frequencies, but to the frequency densities.

This type of diagram is used exclusively for showing frequency distributions of quantitative data that are continuous in nature. It is essentially an area diagram composed of a series of adjacent rectangles. Hence, the areas used to represent the frequency in groups or class intervals when added together will give the composite area for the entire group.

#### 4.9.1 Construction of Histogram

A histogram or column diagram is essentially a bar graph of a frequency distribution. The following points are to be kept in mind while constructing the histogram for a frequency distribution.

1. Convert the data in the exclusive series if it is given in the inclusive series. For example, if it is given in the form. For example.

Age (years) :	10 – 19	20 – 29	20 – 39	40 – 49
No. of cases :	1	0	1	10

Here the given data is in inclusive series. Transform it into exclusive series of the form:

Age (years) :	9.5 – 19.5	19.5 – 29.5	29.5 – 39.5	39.5 – 49.5
No. of cases :	1	0	1	10

2. The scores in the form of actual class limits as 9.5 19.5, 19.5 29.5, etc., are taken in the construction of a histogram rather than the written class limits as 10, 19, 20, 29.
3. It is customary to take two extra intervals (classes) one below and other above the given grouped intervals or classes (with zero frequency). In the case of the frequency distribution of above data, we take 0.5 – 9.5 and 49.5 – 59.5 as two required class intervals — with zero frequency.
4. Now we take the actual lower limits of all the class intervals (including the extra intervals) and plot them on the  $x$ -axis. The lower limit of the lowest interval (one of the extra intervals) is taken at the intersecting point of  $x$ -axis and  $y$ -axis.
5. Each class or interval with its specific frequency is represented by a separate rectangle. The base of each rectangle is the width of the class interval and the height is the respective frequency of that class or interval.
6. Frequencies of the distribution are plotted on the  $y$ -axis.
7. Care should be taken to select the appropriate units of representation along the  $x$ -axis and  $y$ -axis. Both  $x$ -axis as well as  $y$ -axis should not be too short or too long.

#### 4.9.2 Types of Histograms

There are two types of Histograms

- (a) Histograms with equal class intervals
- (b) Histograms with unequal class intervals

**Histogram with equal class intervals :** The sizes of class intervals are drawn on  $X$ -axis with equal distances and their respective frequencies on  $Y$ -axis. Class and its frequency taken together form a rectangle. The graph of rectangles is known as histogram.

**Example 9.** The monthly profits in rupees of 100 shops distributed as follows:

Profit per shop	0 – 100	100 – 200	200 – 300	300 – 400	400 – 500	500 – 600
No. of shops	12	18	27	20	17	6

Draw a histogram of the above data.

**Solution :** This is the case of Histogram with equal frequencies.

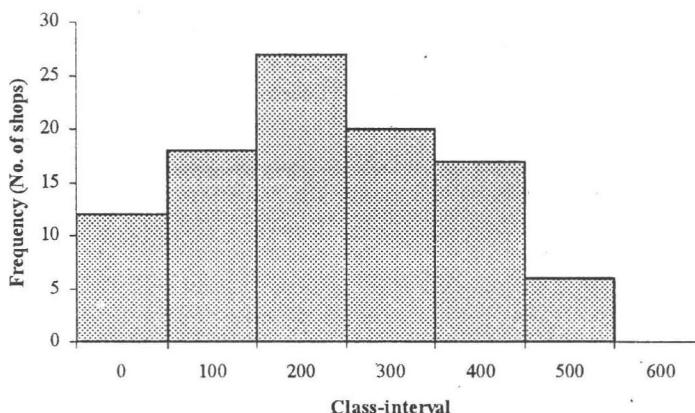


Fig. 4.12 : Histogram showing Monthly Profits

**Example 10.** Draw the histogram of the following frequency distribution and show the area on your graph which represents the total number of wage-earners in the age-group 19 – 32 years:

Age group	14 – 15	16 – 17	18 – 20	21 – 24	25 – 29	30 – 34	35 – 39
No. of wage-earners	60	140	150	110	110	100	90

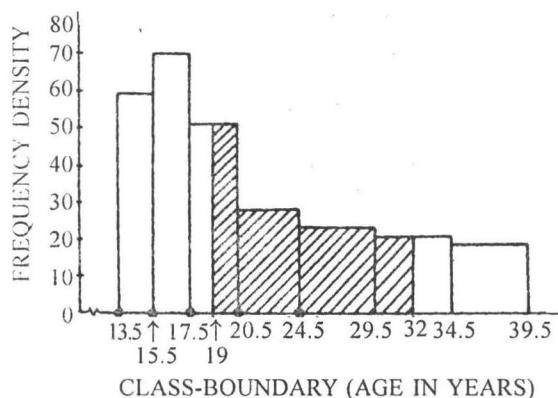
**Solution :** Here the class intervals have been marked by class-limits. As a result the upper limit of one class does not coincide with the lower limit of the next class. In order to draw a histogram the upper limit of one class must coincide with the lower limit of the next class. To draw a histogram in such case where the upper limit of one class does not coincide with the lower limit of the next class, class limits of all the classes should be extended to their class-boundaries. This will help the drawing of a histogram from a frequency distribution in, which class intervals are marked by class-limits.

In this example, the classes are not of equal width. Some have less width and some have more width. So the histogram should be drawn on the basis of frequency density and not on the basis of frequency.

### Calculation of Histogram

Class-interval (Age group)	Class-boundary (Age group)	Class-width	Frequency (No. of wage-earners)	Frequency density
14 – 15	13.5 – 15.5	2	60	80
16 – 17	15.5 – 17.5	2	140	70
18 – 20	17.5 – 20.5	3	150	50
21 – 24	20.5 – 24.5	4	110	27.5
25 – 29	24.5 – 29.5	5	110	22
30 – 34	29.5 – 34.5	5	100	20
35 – 39	34.5 – 39.5	5	90	18

The graph has been shaded showing the area which is represented by the total number of wage-earners in the age-group 19 – 32 years.



*Fig. 4.13 : Histogram showing the number of Wage-earners*

### 4.10 FREQUENCY POLYGON

*It is the curve obtained by joining the middle points of the tops of the rectangles in a histogram by straight lines.* It is generally used in a frequency distribution in which the class-intervals are equal and have a common width. It is particularly useful in representing a sample and ungrouped frequency distribution of discrete variable. It has no special significance but is gives a fair idea about the shape of the frequency distribution. In order to complete the drawing of a frequency polygon, the two end points of it are joined to the base line at the mid-points of the empty class at both ends of the frequency distribution.

**Example 11.** Construct a histogram and frequency polygon for the following data:

100 – 150      150 – 200      200 – 250      250 – 300      300 – 350

4

6

13

5

2

**Solution :** We have the case of equal class intervals.

Class interval	Frequency	c.f.
100 – 150	4	4
150 – 200	6	10
200 – 250	13	23
250 – 300	5	28
300 – 350	2	30

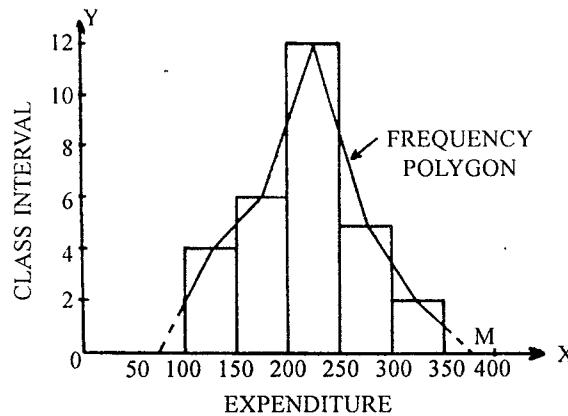


Fig. 4.14

**Example 12.** Construct a histogram and a frequency polygon for the following data:

Output (units (per worker)	500–509	510–519	520–529	530–539	540–549	550–559	560–569
No. of workers	8	18	23	37	47	26	16

**Solution :** Here we notice that the upper limit of one class does not coincide with the lower limit of the next class. In order to avoid such a difficulty and to facilitate the construction of histogram along with the construction of frequency polygon, class boundary for each class interval are determined. On the basis of these class boundaries, the histogram along with the frequency polygon has been drawn with the help of following table:

Class Interval	Class boundary	Frequency (No. of workers)
550 – 509	499.5 – 509.5	8
510 – 519	509.5 – 519.5	18
520 – 529	519.5 – 529.5	23
530 – 539	529.5 – 539.5	37
540 – 549	539.5 – 549.5	47
550 – 559	549.5 – 559.5	26
560 – 569	559.5 – 569.5	16

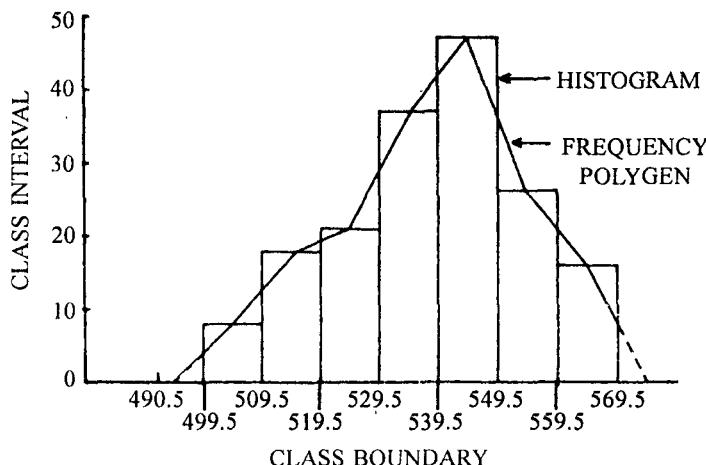


Fig. 4.15

#### 4.11 COMPARISON BETWEEN THE HISTOGRAM AND THE FREQUENCY POLYGON

Although both histogram and frequency polygon are used for the graphic representation of frequency distribution and are alike in many respects, yet they possess points of difference. Some of these difference are cited below :

- (i) Where histogram is essentially the bar graph of the given frequency distribution, the frequency polygon is a line graph of the distribution.
- (ii) In comparison to the histogram, frequency polygon gives a much better concept of the contours of the distribution. With a part of the polygon curve, it is easy to know the trend of the distribution but a histogram is unable to tell such a thing.
- (iii) In comparing two or more distributions by plotting two or more graphs on the same axis, frequency polygon is more useful and practicable than the histogram.
- (iv) In frequency polygon, it is assumed that the frequencies are concentrated at mid-points of class-intervals. It points out merely the graphical relationship between mid-points and frequencies and thus is unable to show the distribution of frequencies within each class interval. But the histogram gives a very clear as well as accurate picture of the relative proportions of frequency from interval to interval.

#### 4.12 FREQUENCY CURVE

It is drawn by the free hand through the various points of the polygon in such a way that the area included is just the same as that of the polygon. The basic object of drawing a frequency curve is to present graphically, the area covered by histogram in a more symmetrical manner. It is also known as or **smooth frequency curve**.

**Example 13.** Draw a histogram, a frequency polygon and a smoothed frequency curve from the following data relating to marks secured in Economics by students of school in a house-test.

Marks	No. of students	Marks	No. of students
0 – 5	3	25 – 30	14
5 – 10	7	30 – 35	10
10 – 15	13	35 – 40	7
15 – 20	25	40 – 45	4
20 – 25	40	45 – 50	2

Solution : The required curves are shown in Fig. 4.16.

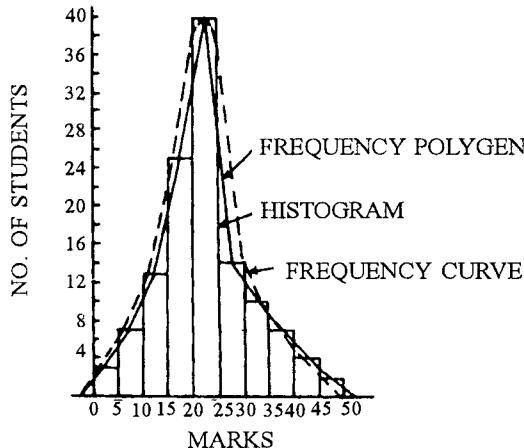


Fig. 4.16

#### 4.13 CUMULATIVE FREQUENCY CURVE OR OGIVE

It is a graph which represent the data of the cumulative frequency distribution. An Ogive, when it is drawn, is in the form of a curve. *The curve drawn from the data cumulated downward is known as less than ogive and the curve drawn from the data cumulated upward as more than ogive. An ogive is used to find median, quartiles, deciles and percentiles etc.*

It is also used to find the number of observations which are expected to lie between two given values.

A cumulative frequency curve or Ogive is obtained by plotting the successive cumulative frequencies on the graph paper and joining them by a free hand. An Ogive is drawn with the following points in mind:

- (i) The upper limits of the classes are represented along  $x$ -axis.
- (ii) The cumulative frequency of a particular class is taken along the  $y$ -axis.
- (iii) The points corresponding to cumulative frequency at each upper limit of the classes are joined by a free hand curve. This curve is called a **cumulative frequency curve or an Ogive**. The Ogive is always monotonically increasing because cumulative frequency increases as we move from one class to the next class.

*The main difference between the construction of a frequency graph and the Ogive is that in the case of former the frequencies must be plotted at the mid-points of the class but in the case of an Ogive, the cumulative frequency is plotted at the upper limit of the class.*

**Example 14.** Draw a cumulative frequency graph and estimate the number of persons between the ages 32 – 42 in the following table :

Age	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50	50 – 55	55 – 60
No. of persons	50	70	100	180	150	120	70	59

**Solution :** A cumulative frequency distribution table, showing cumulative frequency “less than ogive”, has been prepared and on the basis of the same a “less than ogive” cumulative frequency graph has been drawn as shown in the Fig. 4.17.

Class-interval (Age)	Class-boundary (Age)	Frequency (No. of persons)	Cumulative Frequency (Less than Ogive)
	20	0	0
20 – 25	25	50	50
25 – 30	30	70	120
30 – 35	35	100	220
35 – 40	40	180	400
40 – 45	45	150	550
45 – 50	50	120	670
50 – 55	55	70	740
55 – 60	60	59	799
		Total = 799	

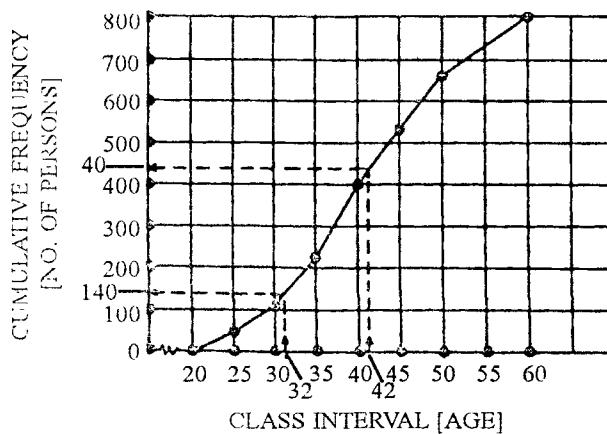


Fig. 4.17 : Cumulative Frequency Graph (Less than Ogive)

The graph has been drawn showing the age on the horizontal line and the number of persons on the vertical line. It is seen from the graph that the number of persons upto the age of 32 years is 140 and the number of persons upto the age of 42 is 440 as is disclosed on the vertical line of the graph. So the estimated number of persons between the ages 32 – 42 is  $440 - 140 = 300$ .

**Example 15.** Plot less than Ogive and more than Ogive for the following data:

Cost of production	4 – 6	6 – 8	8 – 10	10 – 12	12 – 14	14 – 16
No. of farms	13	111	182	105	19	7

**Solution :** We prepare a cumulative frequency distribution table showing cumulative frequency 'less than Ogive' and 'more than Ogive' as shown below:

Table : Computation of Ogive

Class interval (Cost of production)	Class boundary	Frequency (No. of farms)	Cumulative Frequency	
			(Less than Ogive)	(More than Ogive)
	4	0	0	437
4 – 6	6	13	13	424
6 – 8	8	111	124	313
8 – 10	10	182	306	131
10 – 12	12	105	411	26
12 – 14	14	19	430	7
14 – 16	16	7	437	0
<b>Total = 437</b>				

By plotting the points  $(4, 0), (6, 13), (8, 124), (10, 306), (12, 411), (14, 430), (16, 437)$  and joining them by a free hand, we get the **less than Ogive**. Again by plotting the points  $(16, 0), (14, 7), (12, 26), (10, 131), (8, 313), (6, 424), (4, 437)$  then by a free hand, we get the **more than Ogive**. Both the curves are given by Figure. 4.18.

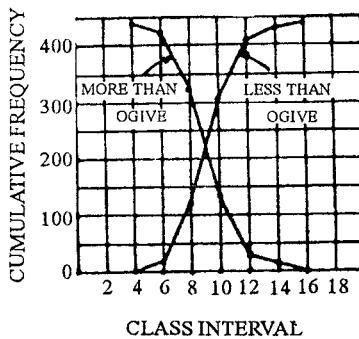
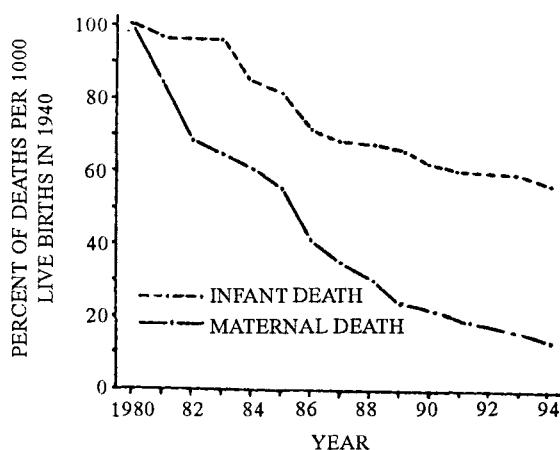


Fig. 4.18 : Cumulative Frequency Diagram.

#### 4.14 PROPORTIONAL CHANGE DIAGRAM

When comparing two or more sets of figures, it is in many instances more important to show the relative change in the variables than the absolute change. This is especially true when the figures to be compared are markedly different in magnitude. This may be done by two different techniques. The first of these is easily understood since it involves a proportional

change concept. The value for any given year, usually the first year in the series, is selected as a base. All other values are then expressed as a percentage of this value. The value of 47.0 deaths per 1,000 live birth for the year 1980 was selected as a base. The value of 45.3 deaths per 1,000 live birth for the year 1981 was then calculated to the 96.4 per cent of this value by the following method:  $45.3 \times 47.0 \times 100$ . This is repeated for each year. The resulting percentages are then plotted on the vertical scale in place of the absolute values (see Fig. 4.19). This scale must be labelled, therefore, "Per cent of deaths per 1,000 live births in base period" – in this case 1940. If the change is a decrease in the rate it is common to start the vertical scale with 0. If, however, the change is an increase the zero of the scale should be shown, although the values between 0 and 100 may be omitted by indicating a break in the scale.



**Fig. 4.19 : Percentage change in number of infant and maternal deaths per 1,000 live births in a city, 1980 – 1994**

There are two main objections to this type of diagram. First, both lines start at the same point 100 per cent. This may give the reader the first impression that the values for the two or more variables for this particular year are the same even though a study of the scales will show this not be true. Secondly, the process of calculating the necessary series of percentages may become very time consuming, if one does not have mechanical equipment available.

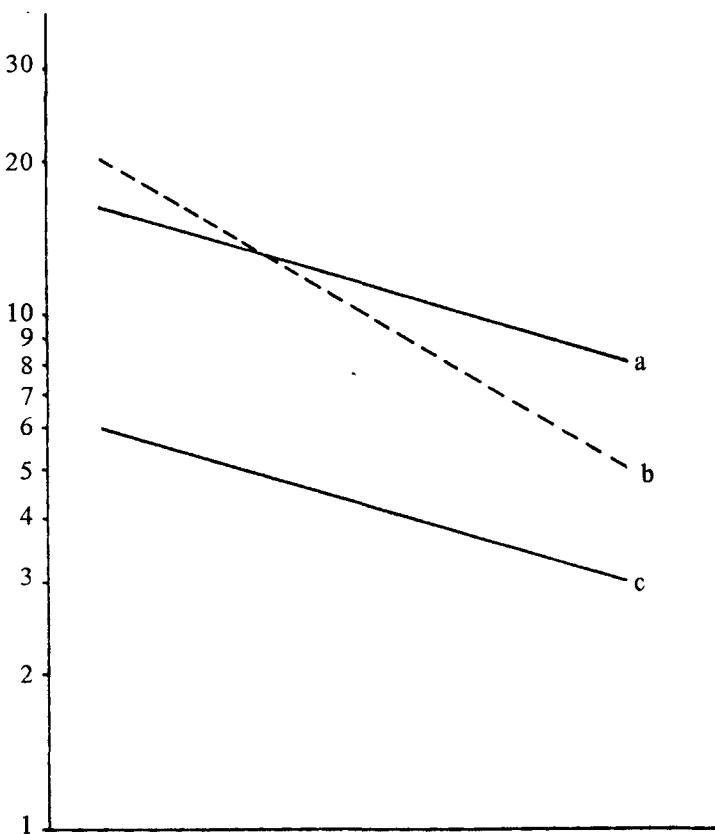
#### 4.15 ARITHLOG OR RATIO DIAGRAMS

This diagram has the added advantage that both relative and absolute changes can be determined from it. It is not necessary to calculate percentages. The basis of the construction is the use of a special paper known as arithlog or ratio paper which has scales so drawn that the same ratio between two numbers will always be represented by the same increment on the vertical scale between 8 and 4 is the same as that between 16 and 8 or between 80 and 40, since in each case the first number is twice the second. It will be seen that line *a*, which starts at 16 and ends at 8, is parallel to line *c*, which starts at 6 and ends at 3. This indicates that the same proportional change has occurred in the variable represented by line *a* as has occurred in the variable represented by line *c*. Line *b*, which shows a change from 20 to 5, has a much sharper decline showing that the proportional change in the variable represented has been much greater

than in the variables of lines *a* and *c*. This is true since the numbers indicate not a two fold decrease as in the first lines but a fourfold decrease.

This arithlog or ratio paper is provided in various sizes, depending on the multiples of ten that are shown. Each multiple of ten is commonly known as a cycle. Hence, if the data to be presented include numbers in the tens, hundreds, and thousands, a paper with three cycles would be used. The tens would be plotted in the first cycle, the hundreds in the second cycle, and the thousands in the third cycle. The vertical scale would be labelled the same as in the ordinary arithmetic line diagram.

The horizontal scale on this type of paper is the same as on arithmetic paper, that is, equal increments represent equal units of measure. Therefore, the same labelling of the horizontal scale will be used. The diagram is plotted by the same method as is the arithmetic line diagram.



*Fig. 4.20 : Illustration of Arithlog Diagram*

A comparison of identical data plotted by three types of line diagrams. At first glance, it shows that the infant deaths per 1,000 live births have decreased more rapidly than have the maternal deaths per 1,000 live births. In absolute numbers this is true as there has been an absolute decrease of 20.4 infant deaths per 1,000 live births as compared with an absolute decrease of 3.2 maternal deaths per 1,000 live births. The diagram on arithmetic paper should be used only when comparisons of absolute number are to be made. Either the proportional

change diagram or the diagram on ratio paper should be used if the relative change is of importance. This will be true usually when comparing variables in which the measurements are of different magnitudes. The ratio paper is similar to use but is not so easily understood by the untrained person.

## EXERCISE

1. What are the various types of statistical diagrams? Explain each in brief.
2. Write short note on the following:
  - (a) Bar diagram
  - (b) Relative frequency distribution
  - (c) Pie diagram
  - (d) Component or sub-divided bar diagram.
3. Show by suitable bar diagrams the absolute as well as relative changes in the students population of colleges A and B in the different faculties from 1988 to 1994.

<i>Faculties</i>	<i>College A</i>		<i>College B</i>	
	<i>2002</i>	<i>2003</i>	<i>2004</i>	<i>2005</i>
Arts	300	350	100	200
Science	120	500	150	250
Commerce	200	650	130	150
Law	100	300	100	120

4. Draw a pie diagram to represent the following data which gives the distribution of outlay in a five year plan of India under the major heads of development expenditure in crores of rupees.

<i>Head</i>	<i>Expenditure (Rs. in crores)</i>
Agriculture and Community Development	8,000
Irrigation and Power	4,000
Industry and Mining	7,000
Transport and Communication	5,500
Total	27,000

5. What do you understand by the term 'graphical representation of data'? Enumerate its advantages.
6. What are circle graphs or pie diagrams? Illustrate their construction through an example.
7. What are the various modes used for the graphical representation of ungrouped data? Discuss in brief.
8. What are line graphs? Discuss their utility in the presentation of statistical data.
9. What are pictograms? How can statistical data be represented through such diagrams? Illustrate with an example.

10. What is a histogram? How does it differ from a frequency polygon and frequency curve? Present the following data in Histogram:

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Frequency	4	10	16	22	20	18	8

11. Plot a frequency curve and an ogive for the following data showing the parental income of 122 school students of a class:

Income (Rupees)	No. of students
Under 250	0
250 to less than 500	5
500 to less than 750	20
750 to less than 1000	43
1000 to less than 1250	37
1250 to less than 1500	10
1500 to less than 1750	5
1750 to less than 2000	2
Total	122

12. Point out the various methods for the graphic representation of grouped data (frequency distribution). Discuss them in brief.
13. What is a cumulative percentage frequency curve or ogive? Illustrate, how it is constructed. Enumerate its different uses.
14. The data below gives the weekly earnings of 430 workers in a floor mill. Draw a frequency curve, an ogive, a histogram and a frequency polygon.

Earnings (in Rs.)	No. of workers	Earnings (in Rs.)	No. of workers
80 – 100	18	200 – 220	65
100 – 120	30	220 – 240	35
120 – 140	20	240 – 260	38
140 – 160	40	260 – 280	20
160 – 180	90	280 – 300	12
180 – 200	70		

15. The marks obtained out of 40, by a group of 100 candidates is given below. Plot the Ogive.

Marks	No. of candidates	Cumulative frequency
5 – 10	9	9
10 – 15	10	19
15 – 20	21	75
20 – 25	35	75
25 – 30	18	93
30 – 35	5	98
35 – 40	2	100

16. Draw a histogram to represent the following data of earnings of workers.

<i>Monthly earning</i>	<i>No. of workers</i>	<i>Monthly earning</i>	<i>No. of workers</i>
80 – 120	4	200 – 240	8
120 – 160	7	240 – 280	5
160 – 200	13	280 – 320	2

17. Draw a histogram of the following data:

Weekly wages in (Rs.) :	1 – 10	11 – 20	21 – 30	31 – 41	41 – 50
No. of workers :	14	28	36	12	10

18. Draw a histogram and the frequency polygon in the same diagram to represent the following data:

<i>Weight (in Kg)</i>	<i>No. of persons</i>	<i>Weight (in Kg)</i>	<i>No. of person</i>
50 – 55	12	70 – 75	5
55 – 65	8	75 – 80	7
60 – 65	5	80 – 85	6
65 – 70	4	85 – 90	3

19. The distribution of heights in cm of 100 people is given below:

<i>Height (in cm)</i>	<i>Frequency</i>	<i>Height (in cm)</i>	<i>Frequency</i>
145 – 155	3	175 – 185	15
155 – 165	35	185 – 195	20
165 – 175	25	195 – 205	2

Construct a histogram and polygon in the same diagram.

20. In a study of diabetic patients, the following data were obtained.

<i>Age (in years)</i>	<i>Number of persons</i>	<i>Age (in years)</i>	<i>Number of persons</i>
10 – 20	3	40 – 50	9
20 – 30	6	50 – 60	5
30 – 40	14	60 – 70	2

Represent the above data by a frequency polygon.

21. Following is the distribution of scores of 70 students in a certain school test:

<i>Marks upto:</i>	10	20	30	40	50
<i>No. of students</i>	3	8	17	20	22

Represent the above data by a frequency polygon.

22. Draw ogive for the following data.

<i>Class interval</i>	200–220	220–240	240–260	260–280	280–300	300–320
<i>Frequency</i>	5	4	6	8	0	4

23. Draw the ogive for the following distributions:

Age in years	0 – 10	10 – 20	20 – 30	30 – 40
No. of persons	15	12	8	20

24. From the following frequency distribution table, prepare a table showing cumulative frequency less than the upper class limit and hence draw the ogive scale.

(Scale = 1 cm = 5 limits)

Class	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50	50 – 55	55 – 60
Frequency	2	4	6	9	13	15	5



# 5

# Measures of Central Tendency

## 5.1 MEASURES OF CENTRAL TENDENCY OR AVERAGE

If we consider the scores of the students of a class and arrange them in a frequency distribution, we will find that there are very few students who either score very high or very low. The marks of most of the students lie somewhere between the highest and the lowest scores of the whole class. This tendency of a group about distribution is named as **central tendency**. The typical score lying between the extremes and shared by most of the students is referred to as a **measure of central tendency**. The measure of central tendency is defined as :

*"It is a sort of average or typical value of the items in the series and its function is to summarise the series in terms of this average value".*

Thus a single expression representing the whole group is selected which may convey a fairly adequate idea about the whole group. This single expression in statistics is known as the **average**. Averages, are generally, the central part of the distribution and therefore they are also called the **measures of central tendency**.

The most common measures of central tendency are :

1. Arithmetic mean or Mean
2. Median
3. Mode.

Each of them, in its own way, can be called a representative of the characteristics of the whole group and thus the performance of the group as a whole can be described by the single value which each of these measures gives. The values of mean, median or mode also help us in comparing two or more groups or frequency distributions in terms of typical or characteristic performance.

## 5.2 CHARACTERISTICS OF AN IDEAL MEASURE OF CENTRAL TENDENCY

According to Professor G.U. Yule, a good average must have the following characteristics:

1. *It should be rigidly defined so that different persons may not interpret it differently.*

2. It should be easy to understand and easy to calculate.
3. It should be based on all the observations of the data.
4. It should be easily subjected to further mathematical calculations.
5. It should be least affected by the fluctuations of the sampling.
6. It should not be unduly affected by the extreme values.
7. It should be easy to interpret.

### 5.3 ARITHMETIC MEAN

We have the following technique to calculate the arithmetic mean of a given data.

#### 5.3.1 Mean of Raw Data

We know that in an ungrouped raw data, we are given individual items. We also know that the average of  $n$  numbers is obtained by finding their sum (by adding and then dividing it by  $n$ ), let  $x_1, x_2, \dots, x_n$  be  $n$  numbers, then their average is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

**Example 1.** Find the arithmetic mean of the tryglycerides present in 10 patients in their blood samples in Escort Hospital:

$$25, 30, 21, 55, 47, 10, 15, 17, 45, 35$$

**Solution :** Let  $\bar{x}$  be the average tryglyceride value.

$$\therefore \text{Sum of all the observations} = 25 + 30 + 21 + 55 + 47 + 10 + 15 + 17 + 45 + 35 = 300.$$

$$\therefore \text{Number of patients} = 10.$$

$$\therefore \text{Arithmetic mean} = \frac{\text{Sum of observations}}{\text{Number of patients}} = \frac{300}{10} = 30.$$

#### 5.3.2 Mean of Grouped Data

Let  $x_1, x_2, x_3, \dots, x_n$  be the variates and let  $f_1, f_2, f_3, \dots, f_n$  be their corresponding frequencies, then their mean  $\bar{x}$  is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \dots + f_nx_n}{f_1 + f_2 + f_3 + \dots + f_n} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n f_i x_i}{N} .$$

$$\text{where } N = f_1 + f_2 + \dots + f_n.$$

**Example 2.** Find the Arithmetic Mean of a sample of reported cases of mumps in school children of the following data.

Blood LDL	52	58	60	65	68	70	75
No. of Patients	7	5	4	6	3	3	2

**Solution :** Let  $\bar{x}$  be the blood LDL and  $f$  be the frequency so that we have the following table.

$x$	$f$	$f \times x$
52	7	364
58	5	290
60	4	240
65	6	390
68	3	204
70	3	210
75	2	150
<b>Total</b>	<b>30</b>	<b>1848</b>

Here,  $N = \sum f = 30$  and  $\sum fx = 1848$

$$\therefore \text{Mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{1848}{30} = \frac{616}{10} = 61.6.$$

**Example 3.** Calculate the Arithmetic Mean of the number of florets on sunflower as given below:

Class interval	Frequency	Class interval	Frequency
10 – 20	2	60 – 70	10
20 – 30	7	70 – 80	3
30 – 40	17	80 – 90	2
40 – 50	29	90 – 100	1
50 – 60	29		

**Solution :** While calculating the arithmetic mean for such a tabular data, it is assumed that all the observations in any particular class interval have the same value. This value is the middle value or midpoint of the class interval, i.e., we replace the classes by the midvalues and proceed as above.

Class interval	Midvalues $x$	Frequency $f$	$f \times x$
10 – 20	15	2	30
20 – 30	25	7	175
30 – 40	35	17	595
40 – 50	45	29	1305
50 – 60	55	29	1595
60 – 70	65	10	650
70 – 80	75	3	225
80 – 90	85	2	170
90 – 100	95	1	95
	<b>Total</b>	<b>100</b>	<b>4840</b>

Here,  $N = \Sigma f = 100$  and  $\Sigma fx = 4840$

$$\therefore \text{Mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{4840}{100} = 48.4$$

### 5.3.3 Short-cut Method

For mean of ungrouped data, short-cut method is applied when the frequencies and the values of the variables are quite large and it becomes very difficult to compute the arithmetic mean as in the case of above example. In a frequency table of such a type a **provisional mean** is taken as that values of  $x$  (mid value of the class interval) which corresponds to the highest frequency or which comes near the middle value of the frequency distribution. This number is called the **Provisional Mean or Assumed Mean**. Also find the deviations of the variates from this provisional mean. Then the arithmetic mean is given by the formula:

(i) In the case of **ungrouped data**

$$\bar{x} = a + \frac{\sum d}{n},$$

where

$a$  = assumed mean,  $n$  = number of items,

$d = x - a$  = deviations of any variate from  $a$ .

#### Working Rule for short cut method for ungrouped data:

**Step I.** Denote the variable of the discrete series by  $x$  or  $X$ .

**Step II.** Take any item of series, preferably the middle one, and denote it by  $a$ . This number  $a$  is called the **assumed mean or provisional mean**.

**Step III.** Take the difference  $(x - a)$  and denote it by  $d$  or  $dx$  or  $d = (x - a)$ , where  $d$  is the deviation of any variable from ' $a$ '.

**Step IV.** Find the sum of  $\Sigma d$ .

**Step V.** Use the following formula to calculate the arithmetic mean.

$$\bar{x} = a + \frac{\sum d}{n}$$

(i) In the case of **grouped data**

$$\bar{x} = a + \frac{\sum fd}{N}$$

where,  $fd$  = product of the frequency and the corresponding deviation

$N = \Sigma f$  = the sum of all the frequencies.

#### Working Rule for short cut method for grouped data:

**Step I.** In the case of discrete series, denote the variable by  $x$  of  $X$  and the corresponding frequency by  $f$ . (But in the case of continuous series  $x$  is the mid value of the interval and  $f$ , the frequency corresponding to that interval).

- Step II.** Take any item  $x$  series, preferably the middle one and denote it by ' $a$ '. This number ' $a$ ' is called the **assumed mean or provisional mean**.
- Step III.** Take the difference  $x - a$  and denote it by  $d$  or  $dx$  or  $d = x - a$  = deviation of any variate  $x$  from  $a$ , the assumed mean.
- Step IV.** Multiply the respective  $f$  and  $d$  and denote the product under the column  $fd$ .
- Step V.** Find  $\Sigma fd$ .
- Step VI.** Use the following formula to calculate the arithmetic mean.

$$\bar{x} = a + \frac{\sum fd}{\sum f}.$$

**Example 4.** Find by **short-cut method**, the mean height of the following 8 students whose height in centimetre are

59, 65, 71, 67, 61, 63, 69, 73.

**Solution :** Let us take 65 as assumed mean. i.e.,  $a = 65$ . Let us prepare the following table:

$x$	$d = x - 65$
59	-6
65	0
71	+6
67	+2
61	-4
63	-2
69	+4
73	+8
	$\Sigma fd = 8$

Total deviation =  $\Sigma fd = 8$

Here,  $a = 65$ ,  $n = 8$ ,  $\Sigma fd = 8$

$$\therefore \text{Mean} = \bar{x} = a + \frac{\sum fd}{n} = 65 + \frac{8}{8} = 66 \text{ cm.}$$

**Example 5.** Ten patients were examined for uric acid test. The operation was performed 1050 times and the frequencies thus obtained for different number of patients ( $x$ ) are shown in the following table. Calculate the arithmetic mean by the short-cut method.

$x$	:	0	1	2	3	4	5	6	7	8	9	10
$f$	:	2	8	43	133	207	260	213	120	54	9	1

**Solution :** Let 5 be the assumed mean. i.e.,  $a = 5$ . Let us prepare the following table in order to calculate the arithmetic mean.

$x$	$f$	$d = x - 5$	$fd$
0	2	-5	-10
1	8	-4	-32
2	43	-3	-129
3	133	-2	-266
4	207	-1	-207
5	260	0	0
6	213	1	+213
7	120	2	+240
8	54	3	+162
9	9	4	+36
10	1	5	+5
	$\Sigma f = 1050$		$\Sigma fd = 12$

$$\therefore \text{Arithmetic Mean} = \bar{x} = a + \frac{\sum fd}{\sum f} = 5 + \frac{17}{1050} = 5 + 0.0114 = 5.0114 \text{ cm.}$$

Hence the average for uric acid is 5.0114.

#### 5.3.4 Step Deviation Method

When the class intervals in a grouped data are equal, then the calculations can be simplified further by taking out the common factor from the deviations. This common deviation of variates  $x$  from the assumed mean ' $a$ ' (i.e.,  $d = x - a$ ) are divided by the common factor. The arithmetic mean is then obtained by the following method :

$$\bar{x} = a + \frac{\sum fd}{N} \times i$$

where  $a$  = assumed mean or provisional mean

$$d = \frac{x - a}{i} = \text{the deviation of any variate from } a$$

$i$  = the width of the class interval

$N$  = the number of observations.

**Example 6.** Calculate by step deviation method, the arithmetic mean of the blood test for triglyceride of 384 patients.

Triglyceride	:	5	10	15	20	25	30	35	40	45	50
No. of Patients	:	20	43	75	67	72	45	39	9	8	6

**Solution :** Let  $a = 20$  and  $i = 5$ .

$x$	$f$	$dx = x - a$	$d' = d/i$	$f \times d'$
5	20	-25	-5	-100
10	43	-20	-4	-172
15	75	-15	-3	-225
20	67	-10	-2	-134
25	72	-5	-1	-72
30	45	0	0	0
35	39	5	1	39
40	9	10	2	18
45	8	15	3	24
50	6	20	4	20
		$\sum dx = 384$		$\sum f d' x' = -598$

Now,

$$\bar{x} = a + \frac{\sum f d' x'}{\sum f} \times i$$

$$= 30 - \frac{598}{384} \times 5 = 30 - 1.56 \times 5 = 30 - 7.80 = 22.2.$$

**Example 7.** Given below is the data on the height of plants grown under normal light. Calculate its mean.

Height	:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
No. of plants	:	42	44	58	35	26	15

**Solution :**

Height	Mid value $x$	$d = \frac{x - 35}{10}$	No. of plants ( $f$ )	$fd$
0 – 10	5	-3	42	-126
10 – 20	15	-2	44	-88
20 – 30	25	-1	58	-58
30 – 40	35	0	35	0
40 – 50	45	1	26	26
50 – 60	55	2	15	30
			$N = 220$	$\Sigma fd = -216$

Here  $a = 35$ ,  $N = 220$ ,  $\Sigma fd = -216$ ,  $i = 10$ 

**Mean :** 
$$\bar{x} = a + \frac{\sum fd}{N} \times i$$

$$\therefore \bar{x} = 35 + \left( \frac{-216}{220} \right) \times 10 = 35 - 9.8 = 25.2$$

**Example 8.** In a study on patients of typhoid fever the following data are obtained. Find the arithmetic mean.

Age (in years) :	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89
No. of cases :	1	0	1	10	17	38	9	3

**Solution :** The data is presented in the form of an inclusive series. We have to transform the inclusive series in an exclusive series. It can be transformed as follows :

We measure the distance between the lower limit of the second class interval and the upper limit of the first class-interval. This is equal to  $20 - 19 = 1$ . We subtract  $\frac{1}{2}$  of this distance (i.e., 0.5) from the lower limit and add it to the upper limit. The new classes will be formed as follows:

$$10 - 0.5 = 9.5 ; \quad 19 + 0.5 = 19.5$$

The new data will be as follows :

Age	f	Midvalue (x)	$d = \frac{x - 44.5}{10}$	fd
9.5 – 19.5	1	14.5	-3	-3
19.5 – 29.5	0	24.5	-2	0
29.5 – 39.5	1	34.5	-1	-1
39.5 – 49.5	10	44.5	0	0
49.5 – 59.5	17	54.5	1	17
59.5 – 69.5	38	64.5	2	76
69.5 – 79.5	9	74.5	3	27
79.5 – 89.5	3	84.5	4	12
	$N = 79$			$\Sigma fd = 128$

Now

$$\bar{x} = a + \frac{\sum fd}{N} \times i$$

$\therefore$

$$\bar{x} = 44.5 + \frac{128}{79} \times 10 = 44.5 + 16.2 = 60.7.$$

## 5.4 WEIGHTED ARITHMETIC MEAN

If  $w_1, w_2, w_3, \dots, w_n$  are the weight assigned to the values  $x_1, x_2, x_3, \dots, x_n$  be respectively, then the weighted average is defined as :

$$\text{Weighted Arithmetic Mean} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots + w_n x_n}{w_1 + w_2 + w_3 + \dots + w_n} = \frac{\sum w x}{\sum w}$$

**Example 9.** The following table gives the platelets count (in lakh/cmm) from the analysis of the blood samples on five different days in a pathology laboratory. Find the average platelets count per patient.

Day :	1	2	3	4	5
Platelets count : (in lakh/cmm)	0.50	0.75	1.00	1.50	2.00
No. of patients :	65	80	95	90	70

**Solution :** Here the values of  $x$  are 65, 80, 95, 90, 70 and their corresponding weights  $w$ 's are 0.50, 0.75, 1.00, 1.50, 2.00 respectively.

$$\therefore \text{Weighted arithmetic mean} = \frac{\sum wx}{\sum w}$$

$$= \frac{65 \times 0.50 + 80 \times 0.75 + 95 \times 1.00 + 90 \times 1.50 + 70 \times 2.00}{65 + 80 + 95 + 90 + 70}$$

$$= \frac{32.5 + 60.0 + 95.0 + 135 + 140}{400} = \frac{462.5}{400} = \text{Rs. } 1.156$$

Hence, the average platelets per patient are 1.15 lakh/cmm.

**Example 10.** A man travelled by motor car for 3 days. He covered 960 kms each day. He drove the first day 10 hours at 96 kms per hour, the second day 12 hours at 80 kms per hour and the third day 15 hours, at 64 kms per hour. What was his average speed?

**Solution :**

#### Calculation of Weighted Mean

Speed (kms per hour $X$ )	Hour $W$	$WX$
96	10	960
80	12	960
64	15	960
Total	$\Sigma W = 37$	$\Sigma WX = 2,880$

$$\text{Average speed} = \frac{\sum WX}{\sum W} = \frac{2880}{37} = 77.8 \text{ kms per hour}$$

## 5.5 COMBINED MEAN

If we are given the mean of two series and their sizes, then the combined mean for the resultant series can be obtained by the formula :

$$\text{Combined Mean} : \bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

where  $\bar{x}_{12}$  = Combined mean of the two series;

$\bar{x}_1$  = Mean of the first series ;  $n_1$  = size of the first series;

$\bar{x}_2$  = Mean of the second series ;  $n_2$  = size of the second series.

**Example 11.** An average monthly production of a certain factory for the first 9 months is 2,584 units and for the remaining three months it is 2,416 units. Calculate the average monthly production for the year.

**Solution :** Here,  $N_1 = 9$ ,  $\bar{X}_1 = 2,584$ ,  $N_2 = 3$ ,  $\bar{X}_2 = 2,416$

$$\text{Combined mean : } \bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

$$\Rightarrow \bar{X}_{12} = \frac{(9 \times 2,584) + (3 \times 2,416)}{9 + 3} = \frac{23256 + 7248}{12} = \frac{30504}{12} = 2542.$$

Hence, the average monthly production for the year is 2,542 units.

**Example 12.** There are two branches of a company, employing 280 and 320 persons respectively. If the arithmetic mean of the monthly salaries paid by the two companies are Rs. 750 and Rs. 937.5 respectively, find the arithmetic mean of the salaries of the employees of the companies as a whole.

**Solution :** Here,  $N_1 = 280$ ,  $\bar{X}_1 = 750$ ,  $N_2 = 320$ ,  $\bar{X}_2 = 937.50$ .

$$\begin{aligned} \therefore \text{ Combined Mean : } \bar{X}_{12} &= \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} \\ &= \frac{280 \times 750 + 320 \times 937.5}{280 + 320} \\ &= \frac{210000 + 300000}{600} = \frac{510000}{600} = \text{Rs. 850.} \end{aligned}$$

Hence, the average of the employees of the companies as a whole is Rs. 850.

**Example 13.** The mean wage of 100 labourers working in a factory, running two shifts of 60 and 40 workers respectively, is Rs. 38. The mean wage of 60 labourers working in the morning shift is Rs. 40. Find the mean wage of 40 labourers working in the evening shift.

**Solution :** Morning shift workers : Let  $N_1 = 60$ ,  $\bar{X}_1 = 40$ ,

Evening shift workers : Let  $N_2 = 40$ ,  $\bar{X}_2 = ?$

Let mean wage of 100 workers :  $\bar{X}_{12} = 38$

$$\text{Combined Mean : } \bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

$$\Rightarrow 38 = \frac{60 \times 40 + 40 \times \bar{X}_2}{60 + 40} = \frac{2400 + 40\bar{X}_2}{100}$$

$$\Rightarrow 3800 = 2400 + 40\bar{X}_2 \Rightarrow 40\bar{X}_2 = 1400 \Rightarrow \bar{X}_2 = \frac{1400}{40} = 35.$$

Hence, the mean wage of 40 labourers is Rs. 35.

## 5.6 CORRECTED MEAN

**Example 14.** Mean of 25 observations was found to be 78.4. But it was found that 96 was misread as 69. Find the correct mean.

**Solution :** We know that the mean is given by

$$\bar{x} = \frac{\sum x}{n} \text{ or } \Sigma x = n\bar{x}$$

Here,  $\bar{x} = 78.4$ ,  $n = 25$ ,  $\therefore \Sigma x = 25 \times 78.4 = 1960$

But this  $\Sigma x$  is incorrect as 96 was misread as 69.

$$\therefore \text{Correct } \Sigma x = 1960 - 69 + 96 = 1987$$

$$\text{Correct mean} = \frac{\text{Correct } \sum x}{n} = \frac{1987}{25} = 79.48.$$

**Example 15.** A firm of ready-made garments make both men's and women's shirts. Its profit average is 6% of sales. Its profits in men's shirts average 8% of sales; and women's shirts comprise 60% of output. What is the average profit per sales rupee in women's shirts.

**Solution :** Here  $\bar{X} = 6$ ,  $\bar{x}_1 = 8$ ,  $n_1 = 40$ ,  $n_2 = 60$ . Assuming that the total output is 100, we are required to find out  $\bar{x}_2$ . We know that

$$\bar{X} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{40 \times 8 + 60 \bar{x}_2}{40 + 60}$$

$$\bar{x}_2 = \frac{320 + 60 \bar{x}_2}{100}$$

$$\bar{x}_2 = \frac{600 - 320}{60} = \frac{280}{60} = \frac{10}{3} = 4.66$$

Thus, the average profits in women's shirt is Rs.4.66 per cent of sales, and Rs. 0.0466 per sale rupee.

## 5.7 MERITS, DEMERITS AND USES OF ARITHMETIC MEAN

**Merits:**

1. It can be easily calculated.
2. Its calculation is based on all the observations.
3. It is easy to understand.
4. It is rightly defined by the mathematical formula.
5. It is least affected by sampling fluctuations.
6. It is the best measure to compare two or more series (datas).
7. It is the average obtained by calculations and it does not depend upon any position.

**Demerits:**

1. It may not be represented in actual data so it is theoretical
2. The extreme values have greater affect on mean.
3. It cannot be calculated if all the values are not known.
4. It can not be determined for the qualitative data such as love, beauty, honesty etc.
5. Mean may lead to fallacious conditions in the absence of original observations.

**Use of Arithmetic Mean:**

1. A common man uses mean for calculating average marks obtained by a student.
2. It is extremely used in practical statistics.
3. Estimates are always obtained by mean.
4. Businessman uses it to find out the operation cost, profit per unit of article, output per man and per machine, average monthly income and expenditure, etc.

**5.8 MEDIAN**

*Median is defined as the middle most or the central value of the variable in a set of observations, when the observations are arranged either in ascending or in descending order of their magnitudes.* It divides the arranged series in two equal parts. Median is a position average, whereas the arithmetic mean is a calculated average. When a series consists of an even number of terms, median is the arithmetic mean of the two central items. It is generally denoted by M.

**5.9 CALCULATION OF MEDIAN**

(a) **When the data is ungrouped.** Arrange the  $n$  values of the variable in ascending (or descending) order of magnitudes.

**Case 1. When  $n$  is odd.** In this case  $\frac{n+1}{2}$ th value is the median.

$$\text{i.e., } M_d = \frac{n+1}{2} \text{th term}$$

**Case II. When  $n$  is even.** In this case there are two middle terms  $\frac{n}{2}$ th and  $\left(\frac{n}{2} + 1\right)$ th. The median is the average of these two terms, i.e.,

$$M_d = \frac{\frac{n}{2} + \left(\frac{n}{2} + 1\right)}{2}$$

**Example 16.** The number of blood LDL (in mg/dl) present in the blood samples of 11 patients are : 5, 19, 42, 11, 50, 30, 21, 0, 52, 36, 27

Find the median.

**Solution :** Let us arrange the values in ascending order:

$$0, 5, 11, 19, 21, 27, 30, 36, 42, 50, 52$$

.... (1)

In this data the number of items is  $n + 11$ , which is an odd number.

$$\therefore \text{Median} = M_d = \left( \frac{n+1}{2} \right) \text{th value} = \left( \frac{11+1}{2} \right) \text{th} = 6\text{th value.}$$

Now the 6th value in the data (1) is 27.

$$\therefore \text{Median} = 27 \text{ mg/dl.}$$

**Example 17.** Find the median of the following items :

$$6, 10, 4, 3, 9, 11, 22, 18$$

**Solution :** Let us arrange the values in ascending order.

$$3, 4, 6, 9, 10, 11, 18, 22:$$

In this data the number of items in  $n = 8$ , which is even.

$$\begin{aligned}\therefore \text{Median} = M_d &= \text{Average of } \left( \frac{n}{2} \right) \text{th and } \left( \frac{n}{2} + 1 \right) \text{th terms} \\ &= \text{Average of } \left( \frac{8}{2} \right) \text{th and } \left( \frac{8}{2} + 1 \right) \text{th items} \\ &= \text{Average of 4th and 5th items} \\ &= \frac{9+10}{2} = \frac{19}{2} = 9.5.\end{aligned}$$

**Example 18.** The weights (in kilogram) of 15 students are as follows :

$$31, 35, 27, 29, 32, 43, 37, 41, 34, 28, 36, 44, 45, 42, 30$$

Find the median.

**Solution :** Weights of 15 students (in kg) in ascending order is

$$27, 28, 29, 30, 31, 32, 34, 35, 36, 37, 41, 42, 43, 44, 45$$

$$\text{Median} = \left( \frac{n+1}{2} \right) \text{th item} = \frac{15+1}{2} = 8\text{th item} = 35 \text{ kg}$$

Hence, the median weight is 35 kg.

## 5.10 CALCULATION OF MEDIAN FOR GROUPED DATA

**Case 1 : When the series is discrete:**

In this case, the values of the variable are arranged in ascending to descending order of magnitudes. A table is prepared showing the corresponding frequencies and cumulative frequencies.

$$\text{Median} : M = \frac{(n+1)}{2} \text{th value}$$

If  $x$  is variable which takes the value  $x_1, x_2, x_3, \dots, x_n$  with respective frequencies  $f_1, f_2, f_3, \dots, f_n$  then the median of the given data is calculated by the following.

### Working Rule

**Step I.** Arrange the values of the variable in ascending or descending order of magnitude.

**Step II.** Find the cumulative frequency (c.f.).

**Step III.** Find  $\frac{N}{2}$ , when  $N = \sum f_i$ .

**Step IV.** Find the cumulative frequency just greater than  $N/2$  and determine the corresponding value of the variable.

**Step V.** The value obtained in Step IV above is the required median.

**Example 19.** Calculate median for the blood samples off or HDL of 43 patients.

No. of Patients	6	4	16	7	8	2
HDL (in mg/dl)	20	9	25	50	40	80

**Solution :** Arranging the marks in ascending order and preparing the following table.

HDL (in mg/dl)	No. of Patients	Cumulative frequency (C)
9	4	4
20	6	10
25	16	26
40	8	34
50	7	41
80	2	
$n = \sum f = 43$		

Here  $n = 43$   $\therefore$  Median =  $M = \left(\frac{n+1}{2}\right)$ th value =  $\left(\frac{43+1}{2}\right)$ th value = **22nd value**

The above table shows that all items from 11 to 26 have their values 25. Since **22nd item lies. This interval, therefore, it value is 25.**

Hence, Median = 25 mg/dl

**Example 20.** Find the median of the following frequency distribution.

x : 5	7	9	12	14	17	19	21
f : 6	5	3	6	5	3	2	4

**Solution :**

x : 5	7	9	12	14	17	19	21
f : 6	5	3	6	5	3	2	4
c.f : 6	11	14	20	25	28	30	34

Median = Average of  $\frac{n}{2}$  and  $\left(\frac{n}{2} + 1\right)$ th items ( $\because$  Here  $n = 34$  is an even number)

$$\text{Median} = \text{Average of 17th and 18th items} = \frac{12 + 12}{2} = 12$$

The above table shows that all items from 12 to 14 have their values 9. Since 12th item lies in this interval therefore its value is 9.

Hence, median = 9.

**Example 21.** The median of the observations 8, 11, 13, 15,  $x + 1$ ,  $x + 3$ , 30, 35, 40, 43 arranged in ascending order is 22. Find  $x$ .

**Solution :** The observations arranged in ascending order are :

$$8, 11, 13, 15, x + 1, x + 3, 30, 35, 40, 43$$

Total number of observations = 10 (even)

$$\therefore \text{Median} = \frac{\frac{n}{2}\text{th term} + \left(\frac{n}{2} + 1\right)\text{th term}}{2} = \frac{5\text{th item} + 6\text{th item}}{2}$$

$$\Rightarrow 22 = \frac{x + 1 + x + 3}{2}$$

$$\Rightarrow 22 = \frac{2x + 4}{2} \Rightarrow 2x + 4 = 44 \Rightarrow x = 20.$$

## 5.11 CALCULATION OF MEDIAN FOR CONTINUOUS SERIES

### Case II. When the series is continuous:

In this case the data is given in the form of a frequency table with class-interval, etc., and the following formula is used to calculate the Median.

$$M = L + \frac{\frac{n}{2} - C}{f} \times i$$

where,  $L$  = lower limit of the class in which the median lies,

$n$  = total number of frequencies, i.e.,  $n = \sum f$ ,

$f$  = frequency of the class in which the median lies,

$C$  = cumulative frequency of the class preceding the median class,

$i$  = width of the class-interval of the class in which the median lies.

**Example 22.** Find the median of blood samples for tryglyceride test of the following data is

Glyceride (in mg/dl) : 20 – 30    30 – 40    40 – 50    50 – 60    60 – 70

Patients	:	5	10	20	5	3
----------	---	---	----	----	---	---

**Solution :** Let us prepare the following computation table. Let  $x$  = the range of tryglycerides and  $f$  = number of patients.

## Computation of Median

$x$	$f$	c.f.
20 – 30	3	3
30 – 40	5	8
40 – 50	20	28
50 – 60	10	38
60 – 70	5	43

Here,  $N = 43$ .

$$\text{Median} = \text{Size of } \left(\frac{N}{2}\right)\text{th item} = \text{Size of } \frac{43}{2}\text{th item}$$

= 21.5th item, which lies in the class 40 – 50

$\therefore$  Median class = 40 – 50,  $L = 40$ ,  $N/2 = 21.5$ , c.f. = 8,  $f = 10$ ,  $i = 10$

$$\text{Median} = L + \frac{\frac{N}{2} - \text{c.f.}}{f} \times i$$

$$= 40 + \frac{21.5 - 8}{20} \times 10 = 40 + \frac{135}{20} = 40 + 6.75 = 46.75 \text{ mg/dl}$$

**Example 23.** The following table gives the marks obtained by 80 students in biology. Find the median.

Marks	No. of students	Marks	No. of students
10 – 15	4	30 – 35	7
15 – 20	6	35 – 40	3
20 – 25	10	40 – 45	9
25 – 30	5	45 – 50	6

**Solution :** Let us prepare the following table showing the frequencies and cumulative frequencies:

Marks	Frequency	Cumulative Frequency (c.f.)
10 – 15	4	4
15 – 20	6	10
20 – 25	10	20
<b>25 – 30</b>	<b>5</b>	<b>25</b>
30 – 35	7	32
35 – 40	3	35
40 – 45	9	44
45 – 50	6	50

Here,  $n = 50$ .  $\therefore \frac{n}{2} = 25$ th item. It lies in the class  $25 - 30$

$\therefore$  Median class =  $25 - 29$

$$\therefore \text{Median} = L + \frac{\frac{n}{2} - C}{f} \times i = 25 + \frac{25 - 20}{5} \times 5 = 25 + 5 = 30$$

Here,  $L$  = Lower limit of the median class = 25

$C$  = Cumulative frequency of class preceding the median class = 20.

$f$  = Frequency of the median class = 5

$i$  = Class interval of the median class = 5.

**Example 24.** In a survey of 950 families in a village, the following distribution of numbers of children was obtained.

No. of children	0 - 2	2 - 4	4 - 6	6 - 8	8 - 10	10 - 12
No. of families	272	328	205	120	15	10

Find the mean and median of the above distribution.

**Solution :** Let us prepare the following table by taking 7 as assume mean, i.e.,  $a = 7$ .

Class	$x_i$	$f_i$	c.f.	$d_i = \frac{x_i - 7}{2}$	$f_i d_i$
0 - 2	1	272	272	-3	-819
2 - 4	3	328	600	-2	-656
4 - 6	5	205	805	-1	-205
6 - 8	7	120	925	0	0
8 - 10	9	15	940	1	15
10 - 12	11	10	950	2	20
Total		$\Sigma f_i = 950$		$\Sigma f_i d_i' = -1642$	

$$\text{Now } \text{Mean} = a + \frac{\sum_{i=1}^n f_i d_i'}{\sum_{i=1}^n f_i} \times i = 7 + \frac{-1642}{950} \times 2 = 7 - \frac{3284}{950} = 7 - 3.46 = 3.54$$

$$\text{For Median, } \frac{N}{2} = \frac{950}{2} = 475. \text{ If lies in the interval } 2 - 4$$

$\therefore$  Median class =  $2 - 4$ ;  $L = 2$ ;  $C = 272$ ,  $i = 2$ .

$$\therefore \text{Median} = L + \frac{\frac{N}{2} - C}{f} \times i = 2 + \frac{475 - 272}{328} \times 2 \\ = 2 + \frac{203}{328} \times 2 = 2 + \frac{406}{328} = 2 + 1.24 = 3.24$$

**Example 25.** The following table gives the marks obtained by 80 students in Economics. Find the median.

Marks	No. of students	Marks	No. of students
10 – 14	4	30 – 34	7
15 – 19	6	35 – 39	3
20 – 24	10	40 – 44	9
25 – 29	5	45 – 49	6

**Solution :** The given series is an inclusive one. For finding the median, we have to make it a continuous series.

Let us prepare the following table showing the frequencies and cumulative frequencies:

Marks	Frequency	Cumulative Frequency
9.5 – 14.5	4	4
15.5 – 19.5	6	10
20.5 – 24.5	10	20
<b>24.5 – 29.5</b>	<b>5</b>	<b>25</b>
29.5 – 34.5	7	32
34.5 – 39.5	3	35
39.5 – 44.5	9	44
44.5 – 49.5	6	50

Also,  $n = 50$ .  $\therefore \frac{n}{2} = 25$ . The median lies the class interval 24.5 – 29.5.

Here,  $L$  = Lower limit of the median class = 24.5.

$C$  = Cumulative frequency of class preceding the median class = 20.

$f$  = Frequency of the median class = 5

$i$  = Class-interval of the median class = 5

$$\therefore \text{Median} = 24.5 + \frac{25 - 20}{5} \times 5 = 24.5 + 5 = 29.5.$$

**Example 26.** The following table gives the weekly expenditure of 100 families. Find the median weekly expenditure.

Weekly Expenditure (in Rs.)	Number of Families
0 – 10	14
10 – 20	23
20 – 30	27
30 – 40	21
40 – 50	15

**Solution :** Let us prepare a table which gives the frequency and cumulative frequency.

Weekly expenditure (in Rs.)	Number of families (frequency)	Cumulative frequency
0 – 10	14	14
10 – 20	23	37
20 – 30	27	64
30 – 40	21	85
40 – 50	15	100

Here  $n = \Sigma f = 100$

$$\therefore \text{Median} = \left(\frac{n}{2}\right)\text{th value} = \left(\frac{100}{2}\right)\text{th value} = 50\text{th value.}$$

Since the 50th value lies in the class 20 – 30, so

**Median class = 20 – 30.**

$$\text{Here } \frac{n}{2} = 50, \quad L = 20, \quad f = 27, \quad C = 37, \quad i = 10.$$

$$\therefore \text{Median} = L + \frac{\frac{N}{2} - C}{f} \times i = 20 + \frac{50 - 37}{27} \times 10 = 24.81.$$

**Example 27.** The frequency distribution of weight in grams of mangoes of a given variety is given below. Calculate the arithmetic mean and the median.

Weight in grams	No. of mangoes	Weight in grams	No. of mangoes
410 – 419	14	450 – 459	45
420 – 429	20	460 – 469	18
440 – 449	42	470 – 479	7
440 – 449	54		

**Solution :** The given series is an inclusive class interval series. Since the interpolation formula for median is based on continuous frequency distribution, so, we shall first convert the given series into exclusive class interval series.

TABLE : Calculations for mean and median

Weight in grams	No. of mangoes ( $f$ )	Mid value ( $X$ )	$d = \frac{X - 444.5}{10}$	$fd$	Less than (c.f.)
409.5 – 419.5	14	414.5	-3	-42	14
419.5 – 429.5	20	424.5	-2	-40	34
429.5 – 439.5	42	434.5	-1	-42	76
<b>439.5 – 449.5</b>	<b>54</b>	<b>444.5</b>	<b>0</b>	<b>0</b>	<b>130</b>
449.5 – 459.5	45	454.5	1	45	175
459.5 – 469.5	18	464.5	2	36	193
469.5 – 479.5	7	474.5	3	21	200
<b>Total</b>	$\Sigma f = 200 = N$			$\Sigma fd = -12$	

$$\text{Mean } (\bar{X}) = A + \frac{h \Sigma fd}{N} = 444.5 + \frac{10 \times (-12)}{200}$$

$$= 444.5 - 0.6 = 443.9 \text{ gms.}$$

Here  $N/2 = 100$ . The cumulative frequency just greater than 100 is 130.

Hence the corresponding class 439.5 – 449.5 is the median class.

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

$$= 439.5 + \frac{10}{54} (100 - 76) = 439.5 + \frac{10 \times 24}{54}$$

$$= 439.50 + 4.44 = 443.91 \text{ gms.}$$

## 5.12 MERITS, DEMERITS AND USES OF MEDIAN

### Merits :

1. It is easily understood.
2. It is not affected by extreme values.
3. It can be located graphically.
4. It is the best measure for qualitative data such as beauty, intelligence, honesty, etc.
5. It can be easily located even if the class-intervals in the series are unequal.
6. It can be determined even by inspection in many cases.

### Demerits:

1. It is not subject to algebraic treatments.
2. It cannot represent the irregular distribution series.
3. It is a positional average and is based on the middle item.

4. It does not have sampling stability.
5. It is an estimate in case of a series containing even number of items.
6. It does not take into account the values of all the items in the series.
7. It is not suitable in those cases where due importance and weight should be given to extreme values.

Uses :

1. It is useful in those cases where numerical measurements are not possible.
2. It is also useful in those cases where mathematical calculations cannot be made in order to obtain the mean.
3. It is generally used in studying phenomena like skill, honesty, intelligence etc.

### 5.13 MODE

Mode is defined as that value in a series which occurs most frequently. In a frequency distribution mode is that variate which has the maximum frequency. In other words, mode represents that value which is most frequent or typical predominant.

For example, in the series, 6, 5, 3, 4, 3, 7, 8, 5, 9, 5, 4; we notice that 5 occurs most frequently, therefore, 5 is the mode. Mode is also known as Norm.

**Example 28.** A shoe shop in Delhi had sold 100 pairs of shoes of a particular brand on a certain day with the following distribution:

Size of Shoe	4	5	6	7	8	9	10
No. of Pairs	10	15	20	35	16	3	1

Find the mode of the distribution.

**Solution :** Let us prepare the table showing the frequency.

Size of Shoe	4	5	6	7	8	9	10
No. of Pairs	10	15	20	35	16	3	1

In the above table, we notice that the size 7 has the maximum frequency, viz., 35.

Therefore, 7 is the mode of the distribution.

### 5.14 TYPES OF MODEL SERIES

A series of observation may have one or more modes.

**Unimodal series.** *The series of observations which contains only one mode, is called Unimodal series.*

**Bimodal series.** *The series of observations which contains two modes is called a bimodal series. In this series, the two modes are the same value of greatest density.*

**Trimodal series.** *The series of observations which contains three modes is called a trimodal series. In this series, the three modes are of same value of greatest density and highest concentration of observations.*

**Ill-defined Mode.** *If a series of observations has more than one mode then the mode is said to be ill defined.*

## 5.15 COMPUTATION OF MODE FOR INDIVIDUAL SERIES

### 5.15.1 Simple Series

In the case of simple series, the value which is repeated maximum number of times is the mode of the series.

**Example 29.** In Rajdhani Rubber Industry, Tilak Nagar, New Delhi seven labourers are receiving the daily wages of Rs. 5, 6, 6, 8, 8, 8 and 10. Find the modal wage.

**Solution :** In the series 5, 6, 6, 8, 8, 8, 10; since 8 occurs thrice and no other item occurs three times or more than three times and hence the modal wage is **Rs. 8**.

### 5.15.2 Discrete Frequency Distribution Series

In the case of discrete frequency distribution, mode is the value of the variable corresponding to the maximum frequency.

**Example 30.** A set of number consists of four 4's, five 5's, six 6's and nine 9's. What is the mode?

**Solution :** Let us prepare the following table.

Size of item :	4	5	6	9
Frequency :	4	5	6	9

Since 9 has the maximum frequency, viz., 9, therefore, 9 is the mode.

## 5.16 COMPUTATION OF MODE BY GROUPING METHOD

We notice that, the discrete series, mode is determined by inspection and therefore, an error of judgement is possible in those cases where the difference between the maximum frequency and the frequency preceding or succeeding it is very small and the items are heavily concentrated on either side. Under such circumstances the value of mode is determined by preparing a grouping table and analysis table.

### GROUPING TABLE

A grouping table has the following six columns.

- Column I.** *It has original frequencies and the maximum frequency is marked by bold type.*
- Column II.** *In this column the frequencies of column I are combined 'two by two'. (1 and 2; 3 and 4; 5 and 6 and so on). Here also the maximum frequency is marked by bold type.*
- Column III.** *Here, we leave the first frequency of the column I and combine the others in 'two by two' (2 and 3; 4 and 5; 6 and 7 and so on). Again the maximum frequency is marked by bold type.*

- Column IV.** *In this column the frequencies of the column I are combined (grouped) in 'three by three' (1, 2 and 3; 4, 5 and 6; 7, 8 and 9 and so on). And again the maximum frequency is marked by bold type.*
- Column V.** *Here we leave the first frequency of the column I and group the others in 'three by three' (2, 3 and 4; 5, 6 and 7; 8, 9 and 10 and so on). Again mark the maximum frequency by bold type.*
- Column VI.** *Now leave the first two frequencies of column I and combine the others in 'three by three'. (3, 4 and 5; 6, 7 and 8; 9, 10 and 11 and so on). Mark the maximum frequency by bold type.*

## ANALYSIS TABLE

After preparing the grouping table, we prepare the analysis table. While preparing this table we put the *column numbers on the left-hand side and the various probable value of the mode on the right-hand side. The values against which frequencies are maximum marked in the grouping table and are entered by means of a bar in the relevant 'box' corresponding to the values they represent.*

The procedure of preparing a grouping table and analysis table shall be clear from the following example.

**Example 31.** Calculate the mode from the following frequency distribution:

Size ( $x$ )	4	5	6	7	8	9	10	11	12	13
Frequency ( $f$ )	2	5	8	9	12	14	14	15	11	13

**Solution :** We find that the value 11 of the variable  $x$  occurs maximum numbers of times, i.e., 15. But we also notice that the difference between the frequencies of the values of the variable, on both sides of 15, which are very close to 11 is very small. This shows that the values of the variable  $x$  are heavily concentrated on either side of 11. Therefore, if we find mode just by inspection, an error is possible.

This problem is solved by the method of grouping as it is an irregular distribution in the sense that the difference between maximum frequency 15 and frequency preceding it is very small. Let us prepare the grouping and analysis table.

## GROUPING TABLE

Size (x)	Frequency (f) (I)	Grouping				
		Col. of two (II)	Col. of two leaving the first (III)	Col. of three (IV)	Col. of three leaving the first (V)	Col. of three leaving the first two (VI)
4	2					
5	5					
6	8					
7	9					
8	12					
9	14					
10	14					
11	15					
12	11					
13	13					

Let us now prepare the Analysis Table.

## ANALYSIS TABLE

X→	4	5	6	7	8	9	10	11	12	13
↓ Col. No.								I		
I								I		
II							I	I		
III						I	I			
IV							I	I	I	
V					I	I	I			
VI						I	I	I		
Total frequency					1	3	5	4	1	

From the analysis table, it is clear that the value 10 has the maximum number of bars, i.e., maximum frequency, viz., 5. Hence, the modal value is 10.

**Remark :** But by inspection one is likely to say that the modal value is 11, since it occurs the maximum numbers of times, i.e., 15, which is incorrect as revealed by analysis and grouping table which gives the correct modal value as 10. (though it occurs 15 times).

## 5.17 COMPUTATION OF MODE IN A CONTINUOUS FREQUENCY DISTRIBUTION

**Modal class :** It is that class in a grouped frequency distribution in which the mode lies. The modal class can be determined either by inspection or with the help of grouping table. After finding the modal class, we calculate the mode by the following formula:

$$\text{Mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i,$$

where  $l$  = the lower limit of the modal class

$i$  = the width of the modal class

$f_1$  = the frequency of class preceding the modal class

$f_m$  = the frequency of the modal class

$f_2$  = the frequency of the class succeeding the modal class.

Sometimes, it so happened that the above formula fails to give the mode. In this case, the modal value lies in a class other than the one containing maximum frequency. In such cases, we take the help of the following formula:

$$\text{Mode} = l + \frac{\Delta_2}{\Delta_1 + \Delta_2} \times i$$

where,  $l$  = width of interval;  $\Delta_1 = (f_m - f_1)$ ;  $\Delta_2 = (f_m - f_2)$

The procedure of finding the mode by the above method shall be clear by the following examples.

**Example 32.** Find the mode for the following data:

Height of plants	1 – 5	6 – 10	11 – 15	16 – 20	21 – 25
No. of plants	7	10	16	32	24

**Solution :** From the above table, it is clear that the maximum frequency is 32 and it lies in the class 16 – 20. Thus, the modal class is 16 – 20.

Here  $l = 16$ ,  $f_m = 32$ ,  $f_1 = 16$ ,  $f_2 = 24$ ,  $i = 5$

$$\begin{aligned} \text{Mode} &= l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i = 16 + \frac{32 - 16}{64 - 24 - 16} \times 5 \\ &= 16 + \frac{16}{24} \times 5 = 16 + \frac{16}{3} = 16 + 3.33 = 19.33 \end{aligned}$$

## 5.18 MERITS, DEMERITS AND USES OF MODE

**Merits:**

- It can be easily understood.

2. It can be located in some cases by inspection.
3. It is capable of being ascertained graphically.
4. It is not affected by extreme values.
5. It represents the most frequent value and hence it is very useful in practice.
6. The arrangement of data is not necessary if the items are a few.

**Demerits:**

1. There are different formulae for its calculations which ordinarily give different answers.
2. Mode is determinate. Some series have two or more than two modes.
3. It cannot be subjected to algebraic treatments. For example, the combined mode cannot be calculated for the modes of two series.
4. It is an unstable measure as it is affected more by sampling fluctuations.
5. Mode for the series with unequal class-intervals cannot be calculated.

**Uses:**

1. It is used for the study of most popular fashion.
2. It is extensively used by businessmen and commercial managements.

**5.19 EMPIRICAL RELATION BETWEEN MEAN, MEDIAN AND MODE**

A distribution in which mean, median and mode coincide is called a symmetrical distribution. If the distribution is moderately asymmetrical, then mean, median and mode are connected by the formula:

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

or

$$\text{Median} = \frac{1}{3} (\text{Mode} + 2 \text{ Mean})$$

**Example 33.** If the value of mode and mean is 60 and 66 respectively, find the value of median.

**Solution :** We know that

$$\text{Median} = \frac{1}{3} (\text{Mode} + 2 \text{ Mean}) = \frac{1}{3} (60 + 2 \times 66) = 64.$$

**OTHER MEASURES OF CENTRAL TENDENCY****5.20 MID-RANGE**

The mid-range is the value midway between the smallest and largest values in the sample, that is, the arithmetic mean of the largest and the smallest values. For example, in a set of radiologic counts 4, 5, 9, 1, 2 the mid-range is  $(9 + 1)/2$  or 5. It is clear that the mid-range will be influenced by extreme values.

**5.21 GEOMETRIC MEAN**

The geometric mean of a set of observations is the  $n$ th root of their product. The computation of the geometric mean requires that all observations be positive, that is greater than zero. For

example, the geometric mean of the radiologic counts 4 and 9 is  $\sqrt{4 \times 9} = \sqrt{36}$  or 6. A general formula for computing the geometric mean of the set of observations  $x_1, x_2, \dots, x_n$  is  $\sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$ . The geometric mean is sometimes denoted by  $\bar{x}_g$ . An interesting property is that the logarithm of the geometric mean is the arithmetic mean of the logarithms of the individual observations. This result is expressed by the formula

$$\log \bar{x}_g = \frac{\sum_{i=1}^n \log x_i}{n} \Rightarrow \bar{x}_g = \text{antilog} \left( \frac{\sum_{i=1}^n \log x_i}{n} \right)$$

In addition to other applications, the geometric mean is used in microbiology for computing average dilution titers.

**Example 34.** The turnover of a business in three years is doubled, trebled and quadrupled respectively. Find the average rate of increase in the turnover.

**Solution :** The average rate of the increase in turnover is the geometric mean of 2, 3, 4.

i.e.,  $\therefore \text{G.M.} = (2 \times 3 \times 4)^{1/3} = 2.88 \text{ per year.}$

### Geometric Mean (Definition)

If  $x_1, x_2, x_3, \dots, x_n$  are  $n$  values of a variate  $x$ , none of them being zero, then the geometric mean  $G$  is defined as

$$G = (x_1 \cdot x_2 \cdot x_3 \dots x_n)^{1/n} \quad \dots (1)$$

In particular, the geometric mean of 3, 9 and 27 is

$$= (3 \times 9 \times 27)^{1/3} = 9.$$

The difficulty of calculating the  $n$ th root is overpowered with the help of logarithms. Now taking logarithms of both sides of (1), we get

$$\log G = \log (x_1 \cdot x_2 \cdot x_3 \dots x_n)^{1/n} = \frac{1}{n} \log (x_1 \cdot x_2 \dots x_n)$$

or 
$$\log G = \frac{\log x_1 + \log x_2 + \dots + \log x_n}{n}$$

$$\therefore G = \text{antilog of} \left( \frac{\log x_1 + \log x_2 + \dots + \log x_n}{n} \right)$$

In the case of a frequency distribution, Geometric mean of  $n$  values  $x_1, x_2, \dots, x_n$  of a variate  $x$  occurring with frequency  $f_1, f_2, \dots, f_n$  respectively is given by

$$G = \left[ x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \dots x_n^{f_n} \right]^{1/n}$$

or

$$\log G = \frac{f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n}{n}$$

or

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

Thus, geometric mean is the antilog of weighted mean of the different values of  $\log x_i$  whose weights are the frequencies  $f_i$ .

In the case of continuous or grouped frequency distribution, the values of the variate  $x$  are taken to be the values corresponding to the mid-points of the class-intervals.

## 5.22 MERITS, DEMERITS AND USES OF GEOMETRIC MEAN

### Merits:

1. It is the only average that can be used to indicate the rate of change or ratios.
2. It is very simple and lends itself to algebraic treatment.
3. It is useful in the construction of index numbers.
4. It is not much affected by fluctuations of sampling.
5. It is based on all the observations.
6. It gives less weight to large items and large weights to small items.

### Demerits:

1. It cannot be easily understood.
2. It is relatively difficult to compute as it requires some special knowledge of logarithms.
3. It cannot be calculated when any value is zero or negative.
4. It may be a value which does not correspond to actual value.
5. It cannot be obtained by inspection.

### Uses:

It is used in those cases where it is necessary to average ratios which express rate of change. It is also used for the construction of index numbers.

**Example 35.** Compute the geometric mean for the following data:

10, 110, 120, 50, 52, 80, 37, 60

**Solution :** Let us prepare the following table:

Size of item ( $x$ )	$\log x$	Value of $\log x$
10	$\log 10$	1.000
110	$\log 110$	2.0414
120	$\log 120$	2.0792
50	$\log 50$	1.6990
52	$\log 52$	1.7160
80	$\log 80$	1.9031
37	$\log 37$	1.5682
60	$\log 60$	1.7782
$n = 8$		$\Sigma \log x = 13.7851$

Now  $\log G = \frac{1}{n} \sum \log x = \frac{13.7851}{8} = 1.723$

Taking antilog of both sides, we get

$$G = \text{antilog } 1.723 = 52.84.$$

### 5.23 HARMONIC MEAN

The harmonic mean of a set of observations is the reciprocal of the arithmetic mean of the reciprocals of the observations. That is, if the observations are  $x_1, x_2, \dots, x_n$ , then the harmonic mean is

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

The harmonic mean is often denoted by  $\bar{x}_h$ . Freund gives an interesting example of the usefulness of the harmonic mean. He suggests the problem of determining the average velocity of a car that has travelled the first 10 miles of a trip at 30 miles per hour and the second 10 miles at 60 miles per hour. At first glance the average velocity would seem to be the simple average of 30 and 60, that is, 45 miles per hour. However, this kind of average is usually defined to be total distance divided by total time. Here the total distance is 20 miles, whereas the total time  $\frac{1}{3}$  hour plus  $\frac{1}{6}$  hour of  $\frac{1}{2}$  hour, producing an average velocity of 40 miles per hour rather than 45 miles per hour. It is interesting to note that this average would be available as the harmonic mean of the velocities 30 and 60: that is 2 divided by  $\frac{1}{30} + \frac{1}{60} = 40$ .

**Definition :** The harmonic mean of  $n$  items  $x_1, x_2, x_3, \dots, x_n$  is defined as:

$$\text{Harmonic mean} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}.$$

For example, the harmonic mean of 2, 4 and 5 is  $\frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{5}} = \frac{60}{19} = 3.16$ .

H a r m o n i c M e a n      Let  $x_1, x_2, x_3, \dots, x_n$  be  $n$  items which occur with frequencies  $f_1, f_2, f_3, \dots, f_n$  respectively. Then their Harmonic Mean is given by:

$$\text{Harmonic mean} = \frac{\frac{f_1 + f_2 + f_3 + \dots + f_n}{\left( \frac{f_1}{x_1} + \frac{f_2}{x_2} + \frac{f_3}{x_3} + \dots + \frac{f_n}{x_n} \right)}}{\Sigma f_i \times \frac{1}{x_i}}.$$

## 5.24 MERITS, DEMERITS AND USES OF HARMONIC MEAN

### Merits:

1. It is easy to calculate.
2. It is rigidly defined.
3. It gives largest weight to the smallest items and can be used whenever so desired.
4. It is a useful average when we deal with average of rates.

### Demerits:

*It cannot be located by inspection.*

### Uses:

1. The harmonic mean is especially useful in averaging time rates, in finding the average price per unit when the data gives the amount of commodity for a given price and in the development of index numbers.
2. It is used when rates are expressed as  $x$  per  $y$ , where  $x$  is constant.

**Example 36.** The interest paid on each of three different sums of money yielding 3%, 4% and 5% simple interest p.a. respectively is the same. What is the average yield per cent on the total sum invested.

**Solution :** 
$$\text{H.M.} = \frac{3}{\frac{1}{3} + \frac{1}{4} + \frac{1}{5}} = \frac{3}{\frac{20 + 15 + 12}{60}} = \frac{3 \times 60}{47} = \frac{180}{47} = 3.38\%$$
.

**Example 37.** A train travels first 300 kilometres at an average of 30 k.p.h. and further travels the same distance at an average rate of 40 k.p.h. What is the average speed over the whole distance?

**Solution :** The average speed of train over the whole distance shall be the weighted harmonic mean of the speeds of 30 kilometres per hour over 300 kilometres and 40 kilometres per hour over 300 kilometres.

Here the weights  $w_1 = 300$ ,  $w_2 = 300$ .

We know that 
$$\frac{1}{H.M.} = \frac{\sum \frac{w}{x}}{\sum w} = \frac{1}{\text{average speed}} = \left( \frac{300}{30} + \frac{300}{40} \right) \times \frac{1}{(300 + 300)}$$

or 
$$\text{Average speed} = \frac{600}{10 + 75} = \frac{600}{17.5} = 34 \frac{2}{7} \text{ k.p.h.}$$

**Example 38.** The consumption of petrol by a motor was a gallon for 20 miles while going up from plains to hill station and a gallon for 24 miles while coming down. What particular average would you consider appropriate for finding the average consumption in miles per gallon for up and down journey, and why?

**Solution :** The average consumption of petrol in miles per gallon for up and down journey by the harmonic mean ( $H$ ) of 20 and 24, viz.,

$$\begin{aligned}\frac{1}{H} &= \frac{1}{n} [\Sigma(1/x)] = \frac{1}{2} \left[ \frac{1}{20} + \frac{1}{24} \right] = \frac{6+5}{2 \times 120} \\ \Rightarrow H &= \frac{240}{11} = 21.82 \text{ miles per gallon.}\end{aligned}$$

## 5.25 CHOICE OF AN AVERAGE FOR DECISIONS MAKING

We have studied the various kinds of averages such as **Arithmetic mean, Median, Mode, Geometric mean and Harmonic mean**. It is important to know **when and how to use which average?**. Thus averages cannot be used indiscriminately. A judicial selection of averages for sound statistical analysis depends upon the following factors:

- (i) The nature of the variable involved.
- (ii) The purpose of analysis.
- (iii) The system of classification adopted.
- (iv) The quality, nature and availability of data.
- (v) The study of average for further statistical computation required for the enquiry in mind.

We give below the suitability of some of the averages.

**Arithmetic Mean:** It is generally, used in business. Whenever, we talk of average cost of production or sale or average wages, we use arithmetic mean. It is also used for further statistical calculations such as **standard deviation**. Arithmetic mean is not recommended while dealing with frequency distribution with extreme observations or open end classes.

**Median:** It is to be used for finding the average when the data is qualitative, i.e., for finding average of intelligence, honesty, beauty etc., median is the only average to be used. It is a **positional average** as it divides the entire series in two equal parts, 50% of actual values will be below and 50% will be above it. It is suitable when there are **open extreme classes** or where there are **extreme values**. It is commonly used for **average wages of worker** as it would avoid the influence of a few very high or very low wage rates.

**Mode:** It is a positional average. It is to be used while dealing with open end classes. It is particularly used in business, when the businessman is not interested in the magnitude but only in the most common or fashionable value.

**Geometric and Harmonic Means:** Geometric mean and Harmonic mean are known as ratio averages as they are most appropriate where the data comprise rates, ratios or percentages instead of actual quantities. Geometric mean is to be used while dealing with rates and ratios. Harmonic mean is to be used in compiling special types of average rates or ratios, where time factor is variable and the act being performed, e.g., distance is constant.

## 5.26 COMPARISON OF THE MEAN, MEDIAN AND MODE — ADVANTAGES AND DISADVANTAGES

Relative to the five radiologic counts 1, 2, 4, 5, 9 we observe that the arithmetic mean 4.2 is larger than the median 4. This distinction points out a useful contrast between the mean and the median – the mean is sensitive to extremes, whereas the median is not. The extreme value 9 in this instance has increased the arithmetic mean, whereas the median, looking only at the middle value in the ordered set of observations, is not affected. Further insight to this distinction is provided by a physical characterisation of the arithmetic mean. Consider a weightless bar marked off in units and imagine each observation in the group as contributing an identical unit of weight that can be hung from its appropriate point on the bar. For the set of observations above, weights would hang from the points labelled 1, 2, 4, 5, 9. The point at which the bar would balance is the arithmetic mean. **Thus the mean is a kind of fulcrum or centre of gravity of a set of observations.** In this instance the bar would balance around the point 4.2. From this representation it is clear just why the arithmetic mean is sensitive to extreme values. If the largest value were increased beyond 9, say to 15, the arithmetic mean would follow that increase to maintain balance. However, as noted earlier, the median would be unaffected.

For some purposes we may wish to take into account extreme values. In these cases, the arithmetic mean is useful. However, in other instances it can give a very misleading picture of a set of observations. An interesting example concerns the ages of the five guests at Mary's party: 6, 6, 6, 6, 71. You see Grandpa was visiting the family and Mary wanted to be sure that he was also invited to her party. The mean age of the guests was

$$\frac{6 + 6 + 6 + 6 + 71}{5} = \frac{95}{5} = 19.$$

Now if we told you about a birthday party at which the average age of the guests was 19 and asked you to describe, for example, the types of recreational activity, you probably would not describe Mary's actual party very well. In fact, the arithmetic mean does not describe the ages very well. No one at the party was even remotely close to 19 years old. What is the problem? Grandpa has unintentionally (he might prefer to be 19) influenced the arithmetic mean. How about the median age? The median is 6 and is a much better measure in this case, as is the mode, which is also 6.

The median appears in many cases to ignore useful information. Although it takes all the observations into account, at least by virtue of nothing their relative positions, the only values directly influencing the computation of the median are the two middle observations

in the case of an even set or the single middle observation in the case of an odd set. This can result in loss of information. The mode has one possible advantage relative to the mean and median it does describe an individual, in fact several individual. By definition the mode must be attained by more than one observation in the sample. However, this advantage is of little value, since our purpose is to characterise groups rather than individuals. Furthermore, as we have seen, the same set of observations might well have more than one mode. Any set of observations has only one arithmetic mean and one median.

**Table 5.1 : Comparison Among Mean, Median and Mode**

	Mean	Median	Mode
Average	It is calculated average	It is positional average.	It is positional average.
Calculation	It is based on all the observation.	It is middle most value which divides the series into two equal parts.	It is the value around which the items of the series tend to concentrate densely.
Treatment	It is capable of mathematical treatments.	It is not capable of mathematical treatments.	It is not capable of mathematical treatments.
Item	It involves all the items for calculations.	Does not consider all the items.	Does not consider all the items.
Array	Does not require arraying.	Arraying of the values of the items in the series is essential.	Arraying of the values of the items in the series is essential.
Extreme values	It is affected by extreme and abnormal values of the items in the series.	It not affected by the extreme values.	It not affected by the extreme values.
Result	There is only one mean.	There is only one median.	In a series there may be one Mode or more than one Mode or no Mode.
Reliability	Most reliable measure.	Less reliable.	Less reliable.
Use	It is simple and widely used in statistical treatment and interpretation.	Not popular and is used only in appropriate cases.	Not popular and is used only in appropriate cases.

## 5.27 PARTITION VALUES

Median divides an arrayed series ascending or descending series into two equal parts. *When we are required to divide a series (array) into more than two equal parts, the dividing places are known as partition values.*

We know that one point divide a series into two equal parts called **halves**. Similarly, **three points** divides the arrayed series into four points called **quartiles**, **nine points** divide it into ten parts called **deciles** and **ninety-nine** points divide it into one hundred parts called **percentile**.

**Quartiles, deciles and percentiles are called partition values.**

For getting partition values, the most **important rule is that the values must be arranged in ascending order only**. In case of finding the median, we could arrange the data either in

ascending or descending order but here there is no choice – only ascending order is possible for calculating partition values.

## 5.28 DIFFERENCE BETWEEN AVERAGES AND PARTITION VALUES

An average (mean, median, mode, G.M., H.M.) is the representative of whole series, but quartiles, deciles, percentiles *are averages of parts of the distribution (series)*. For example, third quartile ( $Q_3$ ) is the average of the second half of the series, first decile ( $D_1$ ) is the average of first ten observations of the series and 35th percentile ( $P_{35}$ ) is the average of first 35 observations of the series.

Thus, quartiles, deciles, percentiles are not averages like, mean, median and mode. Partition values help us in understanding how various values are scattered around median.

Thus, partition values are used to study the scattardness of the values of the variable in relation to the median. Therefore, the special use of partition values is to study *dispersion* of items in relation to the median, i.e., it helps us in understanding the composition of a series.

## 5.29 QUARTILES

**Quartiles:** *Quartiles are those values of the variate which divides a series (in array) into four equal parts.*

Each portion contains equal number of items. The first, second and third points are termed as first quartiles ( $Q_1$ ), second quartile (better named as median) and third quartile ( $Q_3$ ). *The first quartile ( $Q_1$ ) or lower quartile, has 25% of the items of the distribution below it and 75% of the items are greater than it.  $Q_2$  (Median) the second quartile or median has 50% of the observations above it and 50% of the observations below it. The upper quartile or third quartile ( $Q_3$ ) has 75% of the items of the distribution below it and 25% of the items are above it.*

**Note :** It must be noted that  $Q_1 < Q_2 < Q_3$ .

## COMPUTATION OF QUARTILES

### Case I. Computation of Quartile for Individual Series

Let  $x_1, x_2, x_3, \dots, x_n$  be  $n$  values of a variate  $X$ . We than compute *quartiles* of these values by the following method.

### METHOD

**Step I.** Arrange the given data in ascending order of magnitude.

**Step II.** Find the total number of observation. Let it be  $N$ .

**Step III.** Calculate the three quartiles  $Q_1, Q_2, Q_3$  by the following formulae.

$$Q_1 = \text{Value of } \left( \frac{N+1}{4} \right) \text{th observation in the arrayed series.}$$

$$Q_2 = \text{Value of } \left( \frac{N+1}{2} \right) \text{th observation in the arrayed series if } n \text{ is odd.}$$

$Q_2$  = Mean of the values of  $\left(\frac{N}{2}\right)$ th and  $\left(\frac{N}{2} + 1\right)$ th observation if  $n$  is even.

$Q_3$  = Value of 3  $\left(\frac{N+1}{4}\right)$ th observation in the arrayed series.

**Example 39.** Compute  $Q_1$ ,  $Q_2$  and  $Q_3$  on 13 yellow flowered plants in a park.

13, 14, 7, 12, 17, 8, 10, 6, 15, 18, 21, 20

**Solution :** Arranging the given data in **ascending order**, we get

6, 7, 8, 9, 10, 12, 13, 14, 15, 17, 18, 20, 21

Here  $n = 13$ .

$$\begin{aligned} Q_1 &= \text{Value of } \left(\frac{N+1}{4}\right)^{\text{th}} \text{ term} = \text{Value of } \left(\frac{13+1}{4}\right)^{\text{th}}, \text{ i.e., } 3.5^{\text{th}} \text{ term} \\ &= \text{Value of } 3^{\text{rd}} \text{ term} + \frac{1}{2} (\text{value of } 4^{\text{th}} \text{ term} - \text{value of } 3^{\text{rd}} \text{ term}) \\ &= 8 + \frac{1}{2} (9 - 8) = 8 + \frac{1}{2} = 8.5. \end{aligned}$$

$$Q_2 = \text{Median} = \text{Value of } \left(\frac{N+1}{2}\right)^{\text{th}} \text{ term} = \text{Value of } \left(\frac{13+1}{2}\right)^{\text{th}}, \text{ i.e., } 7^{\text{th}} \text{ term} = 13.$$

$$\begin{aligned} Q_3 &= \text{Value of } 3 \left(\frac{N+1}{4}\right)^{\text{th}} = \text{Value of } 3 \left(\frac{13+1}{4}\right)^{\text{th}}, \text{ i.e., } 10.5^{\text{th}} \text{ term} \\ &= \text{Value of } 10^{\text{th}} \text{ term} + \frac{1}{2} (\text{value of } 11^{\text{th}} \text{ term} - \text{value of } 10^{\text{th}} \text{ term}) \\ &= 17 + \frac{1}{2} (18 - 17) = 17 + 0.5 = 17.5. \end{aligned}$$

Hence,  $Q_1 = 8.5$ ,  $Q_2 = 13$  and  $Q_3 = 17.5$ .

## Case II : Computation of Quartiles for a Discrete Frequency Distribution

### METHOD

**Step I.** Compute the cumulative frequencies.

**Step II.** Find  $N = \sum_{i=1}^n f_i$ .

**Step III.** Calculate the quartiles as under:

$Q_1$  : Find the cumulative frequency just greater than  $\frac{N}{4}$ . Determine the corresponding value of the variable  $X$ . This value is the lower quartile  $Q_1$ .

$Q_2$  : Find the cumulative frequency just greater than  $\frac{N}{2}$ . Determine the corresponding value of the variable  $X$ . **This value is the middle quartile  $Q_2$ , i.e., the median.**

$Q_3$  : Find the cumulative frequency just greater than  $\frac{3N}{4}$ . Determine the value of the variable  $X$ . **This value is the third quartile  $Q_3$ .**

### Case III. Computation of Quartiles for a Frequency Distribution with Class-Intervals

The following method is used to compute quartiles  $Q_1, Q_2, Q_3$  for a continuous distribution.

#### METHOD

**Step I.** Compute the cumulative frequency of the given distribution. Let  $N = \sum_{i=1}^n f_i$ .

**Step II.** Compute  $\frac{iN}{4}$ , where  $i = 1$  for lower quartile  $Q_1$ ,

$i = 2$  for middle quartile  $Q_2$  or median,  $i = 3$  for upper quartile  $Q_3$ .

**Step III.** Find the cumulative frequency just greater than  $\frac{iN}{4}$  and the corresponding class.

This class is called the quartile class.

**Step IV.** Use the following formula to calculate  $Q_1, Q_2$  or  $Q_3$ .

$$Q_i = L + \frac{(i \times N)/4}{f} \times h, \quad i = 1, 2, 3.$$

where

$L$  = Lower limit of the class in which a particular quartile lies.

$f$  = Frequency of the class-interval in which a particular quartile lies.

$h$  = Class-interval of the class in which a particular quartile lies.

$C$  = Cumulative frequency of the class preceding the class in which the quartile lies.

### 5.30 DECILES

*The value of the variable which divides the series, when arranged in ascending order, into 10 equal parts is called a decile.* The nine points which divide the given series (when arranged in ascending order) into ten parts are called deciles.

Deciles are denoted by  $D_1, D_2, D_3, \dots, D_9$ . The first decile,  $D_1$ , is the value of the variable such that it exceeds 10% of the observations and is exceeded by 90% of the observations. Similarly,  $D_6$ , the sixth decile, has 60% observations before it and 40% observations after it.

The fifth decile,  $D_5$ , is the median of the given data.

### 5.30.1 Computation of Deciles

#### Case I : Computation of Deciles for Individual Series

In this case, the  $k^{\text{th}}$  decile is given by

$$D_k = \text{Value of } k \left( \frac{n+1}{10} \right)^{\text{th}} \text{ term}, \quad k = 1, 2, 3, 4, \dots, 9.$$

when the series is arranged in ascending order.

**Example 40.** Compute third decile  $D_3$  of the following marks obtained by M.B.B.S. student in biostatistics. The marks are:

13, 14, 7, 12, 9, 17, 8, 10, 6, 15, 18, 21, 20.

**Solution :** Arranging the given data in ascending order, we get:

6, 7, 8, 9, 10, 12, 13, 14, 15, 17, 18, 20, 21.

Here  $n = 13$ .

$$\begin{aligned} D_3 &= \text{Value of } 3 \left( \frac{n+1}{10} \right)^{\text{th}} \text{ term} = \text{Value of } 3 \left( \frac{13+1}{10} \right)^{\text{th}} \text{ term} \\ &= \text{Value of } 4.2^{\text{th}} \text{ term} = \text{Value of } 4^{\text{th}} \text{ term} + \frac{1}{5} (\text{value of } 5^{\text{th}} \text{ term} - \text{value of } 4^{\text{th}} \text{ term}) \\ &= 9 + \frac{1}{5} (10 - 9) = 9.2 \end{aligned}$$

Thus, 3<sup>rd</sup> decile =  $D_3 = 9.2$ .

#### Case II : Computation of Deciles for a Frequency Distribution with Class-Intervals

#### METHOD

**Step I.** Compute the cumulative table. Let  $N = \Sigma f_i$ .

**Step II.** Compute  $\frac{iN}{10}$  to find  $D_i$ , the  $i^{\text{th}}$  decile  $i = 1, 2, 3, \dots, 9$ .

**Step III.** Find the cumulative frequency just greater than  $\frac{iN}{10}$  and the corresponding class.

This class is called the decile class.

**Step IV.** Use the formula

$$D_i = L + \frac{\frac{iN}{10} - C}{f} \times h.$$

where,  $L$  = Lower limit of the decile class

$C$  = Cumulative frequency of the class preceding the decile class

$f$  = frequency of the  $i^{\text{th}}$  decile class.

$h$  = width of the  $i^{\text{th}}$  decile class.

### 5.31 PERCENTILES

The *percentiles* of a set of observations divide the total frequency into hundredths. That is, the 30<sup>th</sup> percentile is that value of the variable below which 30 per cent of the observations lie. Below the 90<sup>th</sup> percentile lie 90 per cent of the observations and so on. If the vertical scale of a cumulative frequency polygon is changed to include percentage of total frequency, the percentiles can be determined directly by finding the appropriate percentage on the vertical scale reading across until the cumulative polygon is intersected and then reading the percentile value as the value on the horizontal axis just below the point of intersection.

In addition to percentiles, which divide the total frequency into thousands there is sometimes a need for quantities dividing total frequency into larger equal parts, for example, thirds, fourths, fifths or tenths. The division points for these various partitions are called *tertiles*, *quartiles*, *quintiles* and *deciles* respectively. Thus the first quartile is the same as the 25<sup>th</sup> percentile and the second quintile is the 40<sup>th</sup> percentile, the seventh decile is the 70<sup>th</sup> percentile, and so on. The second quartile and the fifth decile are simply other terms for the median.

The term *quantile* is a generic term of a division point relative to any partition. That is, percentiles, tertiles, quintiles and deciles are all examples of quantiles.

**Definition :** *The value of the variable which divides the series, when arranged in ascending or descending order, into 100 equal parts are called percentiles. There are 99 percentiles denoted by  $P_1, P_2, P_3, P_4, \dots, P_{99}$  respectively.*

The ninety-nine points which divide the given data, when arranged in ascending order, into hundred equal parts are called percentiles of the data.

$P_{30}$  is the value of variable such that it exceeds 30% of the observations and is exceeded by 70%. Similarly,  $P_{50}$  is the median.

The percentile rank of a particular numerical value of the observed variable is a percentage P such that the specified value of the variable is the P<sup>th</sup> percentile of the set of observations. Thus, the percentile rank of the median is 50. In the kidney weights we may read approximate percentile ranks, the percentile rank of 325 g is 48, the percentile rank of 370 g is 86, and so on. The vertical scale again being adjusted to show percentage, not frequency.

Be sure to note that a percentile, or any quantile, for that matter, is a value of the observed variable, whereas a percentile rank is a percentage.

Percentiles are widely used (and misused) in health applications. In constructing growth charts, for example, pediatricians have studied large samples of children at various ages, noting percentiles of body weight at 1 week, 1 month, 2 months and so on. At each age the 5<sup>th</sup> percentile and 95<sup>th</sup> percentile, say, are noted. Thus, a child whose weight falls between these two percentiles is exhibiting a weight comparable to the majority (90%) of children. However, a problem can arise in interpreting weights either below the 5<sup>th</sup> or above the 95<sup>th</sup> percentile as indicating some growth pathology. In any group of healthy children there must be a lightest 5 per cent and a heaviest 5 per cent, and extremes, although unusual, do not necessarily indicate pathology. Such pathology can best be determined by conducting additional tests or by longitudinal follow-up studies of such children.

In cardiovascular disease epidemiology, there is another application of percentiles to a phenomenon called “tracking”. In follow-up examinations, it may be important to obtain the earliest possible diagnosis of impending disease in order to initiate preventive measures. For example, if individuals whose blood pressure is above the upper quintile (upper 20 per cent of the population) at one examination tend to remain there with the passage of time, blood pressure will be said to exhibit tracking and such individuals will be at higher risk of developing subsequent blood pressures elevated to treatable levels. This is obviously a large and complex topic. For us, the major point is the utility and applicability of percentiles.

### 5.31.1 Computation of Percentiles

#### Case I : Computation of Percentile of Individual Series

In this case, the  $k^{\text{th}}$  percentile is given by

$$P_k = \text{Value} \left[ k \left( \frac{n+1}{100} \right) \right]^{\text{th}} \text{ term, when}$$

arranged in ascending order,  $k = 1, 2, 3, \dots, 100$ .

#### Case II : Computation of Percentiles for Discrete Frequency Distribution

**Step I.** Arrange the given data in ascending order.

**Step II.** Compute the cumulative frequencies.

**Step III.** Find  $N = \sum f_i$ .

**Step IV.** Compute  $\frac{iN}{100}$  to compute  $P_i$ , the  $i^{\text{th}}$  percentile,  $i = 1, 2, 3, \dots, 100$ .

**Step V.** Find the cumulative frequency just greater than  $\frac{iN}{100}$  and the corresponding value of the variable. This value is the  $D_i$ , the  $i^{\text{th}}$  percentile of the given data.

#### Case III : Computation of Percentiles for a Frequency Distribution with Class-Intervals.

**Step I.** Compute the cumulative frequency table. Let  $N = \sum f_i$ .

**Step II.** Compute  $\frac{iN}{100}$  for  $P_i$ , the  $i^{\text{th}}$  decile,  $i = 1, 2, 3, \dots, 99$ .

**Step III.** Find the cumulative frequency just greater than  $\frac{iN}{100}$  and the corresponding class.

This class is called  $i^{\text{th}}$  percentile class.

**Step IV.** Use the formula

$$P_i = L + \frac{\frac{iN}{100} - C}{f} \times h,$$

$C = 1, 2, 3, \dots, 99, 100$ .

where  $L$  = lower limit of percentile class

$C$  = cumulative frequency of the class preceding the percentile class.

$f$  = frequency of the percentile class.

$h$  = width of the percentile class.

**Example 41.** Find the 45<sup>th</sup> and 57<sup>th</sup> percentiles of the following data on marks obtained by 100 students in biostatistics.

Marks :	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50
No. of students :	10	20	20	15	15	20

**Solution :** We prepare the following table showing the frequency and cumulative frequency.

Marks	No. of students ( $f$ )	Cumulative frequency (c.f.)
20 – 25	10	10
25 – 30	20	10 + 20 = 30
30 – 35	20	30 + 20 = 50
35 – 40	15	50 + 15 = 65
40 – 45	15	65 + 15 = 80
45 – 50	20	80 + 20 = 100

Here  $n = 100$ ,  $P_{45} = \frac{45 \times n}{100} = \frac{45 \times 100}{100} = 45$ . Similarly,  $P_{57} = 57$ .

**45<sup>th</sup> Percentile:**

$P_{45}$  class is 30 – 35.  $\therefore L = 30$ ,  $C = 30$ ,  $i = 5$ ,  $f = 20$ .

$$\therefore P_{45} = 30 + \frac{45 - 30}{20} \times 5 = 30 - 1.25 = 28.75.$$

**57<sup>th</sup> Percentile :**

In this case  $P_{57} = 57$ , which lies in the class 35 – 40.

$$\therefore C = 50, \quad C = 5, \quad i = 15, \quad L = 35.$$

$$P_{57} = L + \frac{\frac{57n}{100} - C}{f} \times i = 35 + \frac{57 - 50}{15} \times 15 = 35 - 2.33 = 32.6\%.$$

### EXERCISE

- Define arithmetic mean and give its merits and demerits.
- What are the measures of central tendency and give their relative merits and demerits?
- Calculate the arithmetic mean of the heights in cms of plants of the following data:

25, 30, 21, 55, 47, 10, 15, 17, 45, 35.

4. Find the arithmetic mean of the data on oxygen percentages in respirometer: 11, 11, 12, 12, 13, 14, 13, 14, 13, 13, 15, 16, 16, 15, 17, 17, 15 and 18.

[Hint : Put the data in the frequency table:

$x :$	11	12	13	14	15	16	17	18
$f :$	2	3	5	2	3	2	2	1

$$\text{Mean : } \bar{x} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + \dots + f_n x_n}{f_1 + f_2 + f_3 + \dots + f_n = N} = \frac{280}{20} = 14]$$

5. Given below are the mean lengths measured by each group. Find the average length of scoliodon.

Group :	I	II	III	IV	V	VI
Mean length (mean) :	15	18	12	20	17	16
No. of Fishes :	8	6	9	7	5	3

[Hint: Weighted mean

$$= \frac{\sum f \times x}{\sum n} = \frac{(15 \times 8 + 18 \times 6 + 12 \times 9 + 20 \times 7 + 17 \times 5 + 16 \times 3)}{\sum f = 38} = 16.02]$$

6. The number of flower on 10 red flowered plants are:

6, 2, 3, 4, 10, 5, 8, 16, 14, 12. Find its mean.

7. Find the median height (in cm) of the plants of the following height (in cm).

4, 5, 8, 9, 7, 13, 10

[Hint : Arrange the plants in ascending order their heights 4, 5, 7, 8, 9, 10, 13.

$$\text{Median} = \text{Size of } \frac{N+1}{2}^{\text{th}} \text{ plant} = \text{size of } \frac{8}{2}^{\text{th}} \text{ plant} = \text{size of the plant} = 8]$$

8. When should the arithmetic mean be used in preference to other averages?

9. Compute the mean from the frequency table in the heights (in inches) of 30 students of a class:

Height	70	50	60	52	65	75	68
No. of students	3	5	7	6	2	3	3

10. Find the mean and median height from the following frequency distribution:

Heights in cms	150	160	158	155	164	168
No. of students	10	14	8	15	7	16

11. For a frequency distribution of marks in Biology for 100 students, the arithmetic average was found to be 50. Later on it was discovered that '48' was misread as '84'. Find the correct mean.

12. Define median. What are its merits and demerits? Also give its uses.

13. The mean height of 15 students is 154 cm. It was discovered later on that while calculating the mean the reading 175 cm was wrongly read as 145 cm. Find the correct mean height.

[Hint : Total height of 15 students =  $\Sigma x = 154 \times 15 = 2310$  cm.]

It was found that 175 cm was wrongly read as 145.

Correct sum =  $2310 - 145 + 175 = 2340$  cm.

$$\text{Correct mean} = \frac{2340}{15} = 156 \text{ cm}]$$

14. Calculate the median for the following data on the heights (in cm) of the plant of rice in different pools.

<i>Height :</i>	40	42	45	50	60	65	68
<i>No. of plants :</i>	4	5	8	10	12	10	0

25. The following tables gives the age distribution of patients of a certain disease reported in a hospital during a particular year. Find the median.

<i>Age group :</i>	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–50
<i>No. of patients :</i>	2	11	26	17	8	6	3	2	1

16. Following all the scores in Biostatistics paper of 12 students in a class test.

15, 9, 18, 20, 21, 26, 14, 13, 27, 22, 16, 28

Find  $D_7$  and  $P_{33}$ .

17. The heights of 8 boys in a class (in cm) are :

135, 133, 160, 141, 155, 146, 158, 149

Find the 20<sup>th</sup> and 61<sup>st</sup> percentile.

18. Calculate the median,  $Q_3$ ,  $D_7$  and  $P_{70}$  of the following data of marks obtained in biostatistics paper.

<i>Marks :</i>	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
<i>No. of students :</i>	3	10	17	7	6	4	2	1

19. Find the  $Q_1$ ,  $Q_3$ ,  $P_{40}$  and  $P_{84}$  from the following data of marks obtained in biology paper.

<i>Marks :</i>	0 – 7	7 – 14	14 – 21	21 – 28	28 – 35	35 – 42	42 – 49
<i>No. of students :</i>	3	4	7	11	2	14	9

**ANSWERS**

3. 30.                  4. 14.                  5. 16.02.                  6. 7.  
7. 8.                  9. Mean = 58.37, Median = 155                  10. 159.5.  
11. 49.64.                  13. 156 cm.                  14. Median = 60  
15. Median = 19.31.                  16.  $D_4 = 24$  ;  $P_{33} = 15.5$ .  
17.  $P_{20} = 139.8$  cm ;  $P_{61} = 153.98$  cm.  
18. Median = 20.07 marks ;  $Q_3 = 40.83$  marks;  $D_7 = 37.14$  marks ;  $P_{70} = 37.14$  marks.  
19.  $Q_1 = 19.5$  ;  $Q_3 = 40.5$  ;  $P_{40} = 25$  ;  $P_{64} = 37.5$ .



# 6

# *Measures of Dispersion*

## 6.1 VARIABILITY

You have just returned from an exhausting hike and your feet are tired and score. You locate two buckets, one for each foot, and proceed to soak the abused pedal extremities. The water in each bucket is a soothing, warm 78°F (26°C). Now consider a case in which the water in one bucket (left foot) is at 34°F (1°C) and that in the second (right foot) 122°F (50°C). In other words, your left foot is practically freezing and your right practically boiling. But, on the average you should be comfortable. After all the mean temperature is

$$\frac{34 + 122}{2} = 78^{\circ}\text{F}.$$

In fact, the median is also 78°F, so you should have no complaint. In further, in both these situations the mean and median are identical (78°F), but the situations are clearly very different. In fact, first case there is no variability in temperature, but in the second there is a great deal of temperature variation between the buckets.

Again considering radiologic counts, we examine the two sets: 1, 4, 4, 4, 7 and 4, 4, 4, 4, 4. These sets have much in common. They have the same arithmetic mean, median, mode and midrange. In other words, they are similar in central tendency. However, the first set of observations is more variable than the second. Variability and its measurements are of fundamental importance in the biological sciences.

In examining these two sets of radiologic count, we notice that the first set varies from a low value of 1 to a high value of 7, whereas the second set has 4 as both its lowest and highest value. The first sets shows some variability, whereas the second set shows none. In attempting to define a measure of variability it seems reasonable to require that it take the value zero if and only if the observations show no variability and that it takes a positive value for observations showing some variability. Furthermore, the greater the degree of variation in a set of observations, the greater the positive value of the measure of variability. Using these standards, we can proceed now to judge several proposed measures of variability.

In order to describe a frequency distribution adequately it is necessary not only to know the centre of the distribution but to have some idea of the variability in the measurements. Measures commonly used to describe such variability are the

1. Range
2. The interquartile range
3. Average deviation or Mean deviation
4. Standard deviation.

## 6.2 RANGE

A reasonable measure of variability, the range, might be obtained by subtracting the lowest value in a set of observations from the highest value. For the first set of radiologic counts this  $7 - 1 = 6$  and for the second set  $4 - 4 = 0$ . Considering the criteria we have established for measuring variability, we note that the range fills the bill, at least in the example. It is positive for the first set of observations, which shows variability, and zero for the second set. A common mistake in using the range is to assign two numbers instead of one; that is, we are tempted to say that the range of the first set is from 1 to 7. This is incorrect statistical usage, since the range and indeed any reasonable measures of variability should be a single number, not two numbers.

**Advantages :** The range has several advantages.

1. It is easy to compute.
2. Its units are the same as the units of the variable being measured.

**Disadvantages:**

1. The range does not take into account the number of observations in the sample, only takes into consideration the largest observation and the smallest observation, whatever they may be. Because we expect large samples to include occasional extreme values, we expect a large range. A measure of variability should depend on the number of observations.
2. A second disadvantage of the range is that it makes no direct use of many of the observations in the sample. Observations between the smallest and largest in a set are used only to determine which observations are smallest and largest. Some use of the actual values of intervening observations seems desirable.
3. The range also suffers from dependence upon extreme observations.
4. Range cannot be computed in case of open-end distributions.

## 6.3 INTERQUARTILE RANGE

This measure gives a little more knowledge about the distribution which the range does not give. It includes the middle half of the measurements in the series, the distribution being first divided into quarters, using the same principle as is used in calculation of the median. Consider the three series A, B and C. Thus, of the 16 numbers in Series A, half of them fall between 4 and 96; in Series B, half are between 47 and 53; and in Series C, half of them fall between 21 and 78. The interquartile range, therefore, gives more knowledge as to the distribution of the individual of the individual observations by showing the data are concentrated at the extremes;

in Series B, at the centre; and in Series C, fairly evenly distributed over the entire range of values.

Series A	Series B	Series C
1	1	1
1	44	8
2	45	11
3	46	14
5	48	28
6	48	30
6	49	37
$\frac{7}{93}$ median, 50	$\frac{50}{50}$ median, 50	$\frac{48}{52}$ median, 50
94	51	62
94	52	70
95	52	72
97	54	84
98	55	91
98	55	92
100	100	100
$800/16 = 50$	$800/16 = 50$	$800/16 = 50$

## 6.4 MEAN DEVIATION OR AVERAGE DEVIATION

*Mean deviation of a set of observations of a series is the arithmetic mean of all the deviations, without their algebraic signs, taken from its central value. (mean or median or mode).* In other words, *it is in the average of the modulus of the deviations of the observations in a series taken from mean or median or mode.* Mean deviation is one of the calculated measure in which all the values are considered in their calculations. It has a precise significance as it is an arithmetic average of the variations of the value of individual items in the series from their central tendency. While calculating it, we will come across the following two problems:

### 1. What average should be taken as central value?

The solution to it is that the central value may be any one of the averages – **mean, median or mode**. But, generally, arithmetic mean is taken as the central value.

### 2. What should be the algebraic signs of the deviations?

While calculating the mean deviations, the algebraic sign of the deviation is always taken as **positive**, because the sum of deviations with their algebraic signs, + and –, from the arithmetic mean is always zero.

**Illustration 1 :** Let us have a sample of six observations 3, 5, 13, 14, 15, 16. The mean of these observations is

$$\bar{x} = \frac{3 + 5 + 13 + 14 + 15 + 16}{6} = \frac{66}{6} = 11.$$

The deviation of the items from the mean 11 are  $(3 - 11)$ ,  $(8 - 11)$ ,  $(13 - 11)$ ,  $(14 - 11)$ ,  $(15 - 11)$ ,  $(16 - 11)$ , i.e.,  $-8, -6, +2, +3, +4, +5$ . The sum of all these deviations is  $-8 - 6 + 2 + 3 + 4 + 5 = 0$ . Thus the summation of deviations from the mean in the given series is zero and this will be so in all the other series. To avoid such a situation, we have the following rule:

- (i) **Signs (plus and minus) of deviations are disregarded and absolute values of the deviations are summed up.** Symbolically, we use  $|x_i - \bar{x}|$ , which means the deviation of the  $i^{\text{th}}$  observation of  $x$  from the central value  $\bar{x}$ , (which may be mean or median or mode) with positive sign. Here the vertical lines stand for positive value. Now

**add up all the  $n$  observations to get  $\sum_{i=1}^n |x_i - \bar{x}|$ . Then**

$$\text{Mean deviation } MD_v = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|, \text{ where } \bar{x} \text{ is the arithmetic mean.}$$

Similarly, Mean deviation from the median  $M_d$  for ungrouped data is denoted by  $MD_{\text{med}}$  or  $d_{\text{med}}$ , where

$$MD_{\text{med}} = \frac{1}{n} \sum |x_i - M_d|,$$

#### (a) Mean deviation from the mode $Z$ for ungrouped data

$$d_v \text{ or } MD_v = \frac{1}{n} \sum |x_i - Z|$$

But in practice, we generally use arithmetic mean as it is amenable to algebraic treatments. Median at times is not an actual quantity while mode of a series may not exist, i.e., it may be ill defined series.

**Example 1 :** Calculate the mean deviation about the mean for the following data of Urea (in mg/dl) present in the blood samples of 11 patients in a hospital.

$$15, 20, 17, 19, 21, 13, 12, 10, 17, 9, 12.$$

**Solution :** Here  $n = 11$ , and therefore

$$\text{Mean} = \frac{15 + 20 + 17 + 19 + 21 + 13 + 12 + 10 + 17 + 9 + 12}{11} = \frac{165}{11} = 15 \text{ mg/dl}$$

Now Let  $M = 11$

TABLE : Computation of Mean Deviation

$x$	$d = x - M$	$ d $
15	0	0
20	5	5
17	2	2
19	4	4
21	6	6
13	-2	2
12	-3	3
10	-5	5
17	2	2
9	-6	6
12	-3	3
		$\Sigma  d  = 38$

$$\therefore \text{Mean Deviation} = \frac{\Sigma |d|}{N} = \frac{38}{11} = 3.45 \text{ mg/dl.}$$

**Example 2 :** Calculate the mean deviation about the mean for the following data of Tryglyceride present in the blood sample of 50 patients tested in Pathology Laboratory on a certain day.

Tryglycerides :	5	15	25	35	45	55	65
No. of patients :	8	12	10	8	3	2	7

**Solution :**

TABLE : Computation of Mean Deviation

$x$	$f$	$d = \frac{x - 35}{10}$	$f.d.$	$ x_i - \bar{x} $	$f x_i - \bar{x} $
5	8	-3	-24	24	192
15	12	-2	-24	14	168
25	10	-1	-10	4	40
A = 35	8	0	0	6	48
45	3	1	3	16	48
55	2	2	4	26	52
65	7	3	21	36	252
	$N = \sum f = 50$		$\Sigma fd = -30$		$\Sigma f(x - \bar{x}) = 800$

Let the assumed mean  $A = 35$ .

$$\text{Mean} : \bar{x} = A + \frac{\sum fd}{f} \times i, \text{ where } d = \frac{x - A}{i}$$

$$= 35 + \left( \frac{-30}{50} \right) \times 10 = 299 \text{ mg/dl.}$$

$$\text{Mean Deviation} = \frac{1}{N} \sum f |x - \bar{x}| = \frac{1}{50} = (800) = 16 \text{ mg/dl}$$

**Example 3 :** The marks obtained in Biostatistics paper by 10 students in an examination were as follows: 70, 65, 68, 70, 75, 73, 80, 70, 83, 86.

Find the mode, median and mean deviation about the mean.

**Solution :** Arranging the data in the ascending order we have;

65, 68, 70, 70, 70, 73, 75, 80, 83, 86 ( $\because$  70 occurs maximum number of times)

Mode = 70

$$\text{Median} = \text{Average of } 5^{\text{th}} \text{ and } 6^{\text{th}} \text{ item} = \frac{70 + 73}{2} = 71.5$$

$$\text{Mean} = \frac{\sum X}{n} = \frac{65 + 68 + 70 + 70 + 70 + 73 + 75 + 80 + 83 + 86}{10} = \frac{740}{10} = 74.$$

$$\begin{aligned} \text{Mean Deviation about Mean} &= \frac{\sum |X - \bar{X}|}{n} \\ &= \frac{9 + 6 + 4 + 4 + 4 + 1 + 1 + 6 + 9 + 12}{10} = 5.6 \text{ marks.} \end{aligned}$$

**Example 4 :** Find the mean deviation about the mean for the sample observations on the weight (lbs) of a new born baby : 9, 12, 10, 11, 8, 13, 11, 12, 10, 11, 12, 12, 8, 11, 16.

**Solution (a) :**

Table : Computation of Mean and Mean Deviation

Weight (lbs) (x)	Tally Bars	No. of Babies : f	$f \times x$	$ x - 11 $	$f  x - 11 $
8		2	16	3	2
9		1	9	2	2
10		2	20	1	2
11		6	66	0	0
12		3	36	1	3
13		1	13	2	2
16		1	16	5	5
Total		$\Sigma f = 16$	176		20

$$\text{Mean} = \frac{\sum f x}{\sum f} = \frac{176}{16} = 11 \text{ lbs}$$

$$\text{Mean deviation} = \frac{\sum f |x - \text{Mean}|}{\sum f} = \frac{\sum |X - 11|}{16} = \frac{20}{16} = 1.25 \text{ lbs.}$$

(b) **Mean Deviation for Grouped data.** Let  $x_1, x_2, x_3, \dots, x_n$  occur with frequencies  $f_1, f_2, f_3, \dots, f_n$  respectively and let  $\sum f = n$  and  $M$  can be either Mean or Median or Mode, then the mean deviation is given by the formula.

$$\text{Mean deviation} = \frac{\sum f |x - M|}{\sum f} = \frac{\sum f |d|}{n},$$

where  $d = |x - M|$  and  $\sum f = n$

**Example 5 :** Find the mean deviation from mean for the following data:

Marks :	20	18	16	14	12	10	8	6
No. of Students :	2	4	9	18	27	25	14	1

**Solution : Mean deviation from mean :** Let us calculate the mean of the given data by forming the following table:

Marks (x)	No. of students (f)	$f \times x$	$ d  =  x - 12 $	$f \times  d $
6	1	6	6	6
8	14	112	4	56
10	25	250	2	50
12	27	324	0	0
14	18	252	2	36
16	9	144	4	36
18	4	72	6	24
20	2	40	8	16
	$\sum f = 100$	$\sum f \times x = 1200$		$\sum f  d  = 224$

$$\therefore \text{Mean} = \frac{\sum f x}{\sum f} = \frac{1200}{100} = 12$$

$$\therefore \text{Mean deviation (about means)} = \frac{\sum f |d|}{\sum f} = \frac{224}{100} = 2.24.$$

## 6.5 COEFFICIENT OF MEAN DEVIATION

The relative measure of dispersion corresponding to mean deviation is the coefficient of mean deviation. It is given by the following formula.

**Coefficient of Mean deviation** =  $\frac{dx}{\bar{x}}$  or  $\frac{MD_x}{\bar{x}}$ , where  $\bar{x}$  is the arithmetic mean.

Similarly, the coefficient of mean deviations based on median and mode can be found out by  $\frac{MD_{med}}{M_d}$  and  $\frac{MD_v}{d_v}$  respectively.

### Merits, Demerits and Uses of Mean Deviation

#### Merits :

1. It is easy to understand and compute.
2. Mean deviation is less affected by the extreme values as compared to range or standard deviation.
3. Mean deviation about an arbitrary point is least when the point is median.

#### Demerits :

1. In mean deviation the signs of all deviations are taken as positive and therefore, it is not suitable for further algebraic treatments.
2. It is rarely used in social sciences.
3. It does not give accurate results because the mean deviation from the median is least but median itself is not considered a satisfactory average when the variations in the series is large.
4. It is often not useful for statistical inferences.

#### Uses :

Mean deviation and its coefficient are used in studying economic problems such as distribution of income and wealth in a society.

### 6.6 STANDARD DEVIATION

Standard deviation is the most important and commonly used measure of dispersion. It measures the absolute dispersion or variability of a distribution. A small standard deviation means a high degree of uniformity of the observations as well as homogeneity in the series. It is extremely useful in judging the representativeness of the mean.

**Definition :** Standard deviation is the positive square root of the average of squared deviation taken from arithmetic mean. It is, generally, denoted by the Greek alphabet  $\sigma$  or by S.D. or s.d. Let  $x$  be a random variate which takes on  $n$  values, viz.,  $x_1, x_2, x_3, \dots, x_n$ , then the standard deviation of these  $n$  observations is given by.

$$\sigma = \sqrt{\sum \frac{(x - \bar{x})^2}{n}}, \quad \text{where}$$

$$\bar{x} = \frac{\Sigma x}{n} \quad \text{is the mean of these observations}$$

Alternatively,

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

But the formula

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2},$$

is used when the items are very small. On the other hand the relevance of this method is particularly useful when computers are used where by the values, even when they are of high magnitude, can be used directly for calculating  $\sigma$ .

Standard deviation is also known as **Root mean square deviation**.

$$\text{Coefficient of Standard Deviation} = \frac{\sigma}{\bar{x}}$$

Where  $\bar{x}$  is the arithmetic mean of the given series. It is a **relative measure** of standard deviation. This method is illustrated below.

**Example 6 :** Find the standard deviation of the numbers 3, 4, 5, 6.

**Solution :** Here  $n = 4$ ,  $\sum x = 3 + 4 + 5 + 6 = 18$ .

$$\sum x^2 = 3^2 + 4^2 + 5^2 + 6^2 = 9 + 16 + 25 + 36 = 86$$

$$\begin{aligned} \therefore \sigma &= \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\frac{86}{4} - \left(\frac{18}{4}\right)^2} \\ &= \sqrt{21.5 - (4.5)^2} = \sqrt{21.5 - 20.25} = \sqrt{1.25} = 1.12 \text{ nearly.} \end{aligned}$$

## 6.7 MERITS, DEMERITS AND USES OF STANDARD DEVIATION

**Merits :**

1. It is based on all the observations.
2. It is rigidly defined.
3. It has greater mathematical significance and is capable of further mathematical treatments.
4. It represents the true measurement of dispersion of a series.
5. It is least affected by fluctuations of sampling.
6. It is reliable and dependable measure of dispersion.
7. It is extremely useful in correlation etc.

**Demerits :**

1. It is difficult to compute unlike other measures of dispersion.
2. It is not simple to understand.
3. It gives more weightage to extreme values.
4. It consumes much time and labour while computing it.

Uses :

1. It is widely used in biological studies.
2. It is used in fitting a normal curve to a frequency distribution.
3. It is most widely used measure of dispersion.

## 6.8 CALCULATION OF STANDARD DEVIATION – INDIVIDUAL OBSERVATIONS

When the data under consideration consists of individual observations, the standard deviation may be computed by any of the following two methods.

- (a) By taking deviations of the items from the actual mean.
- (b) By taking deviations of the items form an assumed mean.

### Case I. When the deviations are taken from the actual arithmetic mean

This method is known as **Direct method**

#### DIRECT METHOD

In case of simple series, the standard deviation can be obtained by the formula

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

or  $\sigma = \sqrt{\frac{\sum d^2}{n}}, \text{ where } d = x_i - \bar{x}$

and  $x_i$  = value of the variable of observation,  $\bar{x}$  = arithmetic mean and  
 $n$  = total number of observations.

#### Working rule

**Step I.** Calculate the arithmetic mean  $\bar{x}$ .

**Step II.** Take the deviations of the items from the mean, i.e., calculate  $d = x_i - \bar{x}$ .

**Step III.** Take the sum of the quare of all these deviations, i.e., find  $\sum d^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ .

**Step IV.** Find the mean of the squared deviations obtained in **Step III**, i.e.,  $\frac{\sum d^2}{n}$ , where  
 $n$  is the total number of observations. It is known as **variance**.

**Step V.** Take the under root of variance to get the desired **standard deviation**.

The above method is explained by the following example.

**Example 7 :** Find the mean respiration rate per minute and its standard deviation when in 4 cases the rate was found to be: 16, 13, 17 and 22.

**Solution :** Here Mean =  $\bar{x} = \frac{16 + 13 + 17 + 22}{4} = \frac{68}{4} = 17$ .

Let us prepare the following table in order to calculate the standard deviation.

**Table : Computation of Standard Deviation**

(x)	$d = x - \bar{x} = x - 17$	$d^2 = (x - \bar{x})^2$
16	- 1	1
13	- 4	16
17	0	0
22	5	25
		$\Sigma d^2 = 42$

$$\text{Standard Deviation : } \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum d^2}{n}} = \sqrt{\frac{42}{4}} = 3.2.$$

### **Case II. When the deviations are taken from the Assumed Mean**

This method is also known as **Short-cut Method**.

#### **Short-Cut Method**

This method is applied to calculate standard deviation, when the mean of the data comes out to be a fraction. In that case it is very difficult and tedious to find the deviations of all observations from the mean by the above method. The formula used is

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left( \frac{\sum d}{n} \right)^2},$$

where,  $d = x - A$ ,  $A$  = assumed mean,  $n$  = total number of observations.

#### **Working Rule**

- Step I.** Take any arbitrary number as the assumed mean  $A$ .
- Step II.** Take the deviations from the assumed mean and denote it by  $d$ , i.e.,  $d = x - A$ . Take the total of these deviations, i.e., obtain  $\sum d$ .
- Step III.** Square these deviations and find  $\sum d^2$ .
- Step IV.** Calculate  $\frac{\sum d}{n}$ ,  $\left( \frac{\sum d}{n} \right)^2$  and  $\frac{\sum d^2}{n}$ , where  $n$  is the total number of observations.
- Step V.** Find  $\frac{\sum d^2}{n} - \left( \frac{\sum d}{n} \right)^2$ . Take its square root to get the standard deviation of the given data.

**Example 8 :** Find the standard deviation of the (ESR) erythrocyte sedimentation rate found to be: 48, 43, 65, 57, 31, 48, 59, 78 in 10 normal cases.

**Solution :** Let us prepare the following table in order to calculate the value of S.D. by assuming assumed mean  $A = 50$ .

TABLE : Computation of Standard Deviation

Value ( $x$ )	$d = x - A$	$d^2$
48	- 2	4
43	- 7	49
65	15	225
57	7	49
31	- 19	361
60	10	100
37	- 13	169
48	- 2	4
59	9	81
78	28	784
$n = 10$	$\Sigma d = 26$	$\Sigma d^2 = 1826$

Here  $\bar{x} = A + \frac{\Sigma d}{n} = 50 + \frac{26}{10} = 52.6$

which is a fraction. Let us apply the short-cut formula in order to calculate S.D.

$$\therefore \text{S.D.} = \sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{1826}{10} - \left(\frac{26}{10}\right)^2}$$

$$= \sqrt{182.60 - 6.76} = \sqrt{175.84} = 13.26.$$

## 6.9 CALCULATIONS OF STANDARD DEVIATION – DISCRETE SERIES OR GROUPED DATA

The standard deviation of a discrete series or grouped data can be calculated by any one of the following three methods.

- (a) Actual Mean Method or Direct Method
- (b) Assumed Mean Method or Short-cut Method
- (c) Step Deviation Method

### (a) Actual Mean Method or Direct Method

The standard deviation for the discrete series is given by the formula.

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}},$$

where  $\bar{x}$  is the arithmetic mean,  $x$  is the size of item,  $f$  is the corresponding frequency and  $n = \sum f$ .

However, in practice, this method is rarely used because if the arithmetic mean is in fraction, the calculations take a lot of time and are cumbersome.

### (b) Assumed Mean Method or Short-cut Method

In this method we use the following to calculate the standard deviation  $\sigma$

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2},$$

where  $A$  is the assumed mean,  $d = x - A$ , and  $n = \sum f$ .

#### Working Rule

- Step I.** Take a suitable item of the given series as assumed  $A$ .
- Step II.** Take the deviations of the items from the mean  $A$  and denote it by  $d$ .
- Step III.** Multiply the deviations by the respective frequency and denote it by  $fd$ . Obtain the total  $\sum fd$ .
- Step IV.** Calculate  $d^2$ , where  $d$ 's are obtained in Step II.
- Step V.** Multiply the squared deviations by respective frequencies to get  $\sum fd^2$ .
- Step VI.** Find the value of  $\sigma^2 = \frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2$ .
- Step VII.** Take the square root of  $\sigma^2$  obtained in Step VI to get the value of standard deviation  $\sigma$ .

**Example 9 :** Find the standard deviation of incubation period smallpox in 50 patients of the following data.

Period :	10	11	12	13	14	15	16
No. of patients :	2	7	11	15	10	4	1

**Solution :** Let us prepare following table.

Period (x)	No. of patients (f)	$d = x - A, A = 13$	$fd$	$d^2$	$fd^2$
10	2	-3	-6	9	18
11	7	-2	-14	4	28
12	11	-1	-11	1	11
13	15	0	0	0	0
14	10	1	10	1	10
15	4	2	8	4	16
16	1	3	3	9	9
Total	$n = \sum f = 50$		$\sum fd = -10$		$\sum fd^2 = 92$

Now 
$$\text{Mean} = \bar{x} = A + \frac{\sum d}{n} = 13 + \frac{(-10)}{50} = 12.8$$

Here  $\bar{x} = 12.8$  is a fraction.

$$\therefore \text{Standard deviation} = \sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$$

$$= \sqrt{\frac{92}{50} - \left(\frac{-10}{50}\right)^2} = \sqrt{1.84 - 0.04} = \sqrt{1.80} = 1.342.$$

### (c) Step Deviation Method

In this method we divide the deviations by a common class interval and use the following formula for computing standard deviation

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times i$$

where  $i$  = common class interval,  $d = \frac{x - A}{i}$ ,  $A$  is assumed mean,  $f$  is the respective frequency.

The method is illustrated by the following example.

**Example 10 :** Daily high blood pressure of a patient on 100 days are given below:

B.P. (mmHg) :	102	106	110	114	118	122	126
No. of days :	3	9	25	35	17	10	1

Calculate the mean and the standard deviation of the above data blood pressure of the patient.

**Solution :** Let us take the assumed mean  $A = 114$ . Let  $d = (x - A)/i = (x - 114)/4$ ,  $i = 4$ .

TABLE : Computation of Mean and Standard Deviation

B.P (mmHg)	No. of days ( $f$ )	$d = \frac{x - 114}{4}$	$fd$	$fd^2$
102	3	-3	-9	27
106	9	-2	-18	36
110	25	-1	-25	25
114	35	0	0	0
118	17	1	17	17
122	10	2	20	40
126	1	3	3	9
$N = 100$			$\Sigma fd = -12$	$\Sigma fd^2 = 154$

**Arithmetic Mean :**  $\bar{x} = A + \frac{\sum fd}{N} \times i$

$$A = 114 ; \Sigma fd = -12 ; n = 100 ; i = 4$$

$$\therefore \bar{x} = 114 - \frac{12}{100} \times 4 = 114 - 0.48 = 113.52 \text{ mm Hg.}$$

$$\therefore \sigma = \sqrt{\frac{\sum f d^2}{n} - \left(\frac{\sum f d}{n}\right)^2} \times i = \sqrt{\frac{154}{100} - \left(\frac{-12}{100}\right)^2} \times 4 = \sqrt{154 - (-0.12)^2} \times 4 \\ = \sqrt{154 - 0.0144} \times 4 = 1.235 \times 4 = 4.94 \text{ mmHg.}$$

## 6.10 CALCULATION OF STANDARD DEVIATION - CONTINUOUS SERIES

The standard deviation of a continuous series can be calculated by any one of the methods discussed for discrete frequency distribution. However, the practice only **Step Deviation Method** is mostly used. In this method the formula used is

$$\sigma = \sqrt{\frac{\sum f d^2}{n} - \left(\frac{\sum f d}{n}\right)^2} \times i,$$

where  $d = \frac{m - A}{i}$ ,  $i$  = class interval (or the common factor in case the class intervals are unequal),  $m$  is the mid-value of the intervals,  $A$  is the assumed mean.

### Working Rule

- Step I.** Find the mid-values or mid-points of the various classes and denote it by  $m$ .
- Step II.** Take any one of the values of  $m$ 's as the assumed mean  $A$  (Generally middle of the values is taken as  $A$ ).
- Step III.** Take the deviations of the mid-points from the assumed mean  $A$  and divide it by class interval or common factor  $i$ . Denote it by  $d$ .
- Step IV.** Multiply the respective frequencies  $f$  with the corresponding deviation  $d$  and obtain  $\sum f d$ .
- Step V.** Square the deviations  $d$  and multiply it with their respective frequencies. Obtain  $\sum f d^2$ .
- Step VI.** Substitute the values of  $\sum f d$ ,  $\sum f d^2$ ,  $i$  in the formula

$$\sigma = \sqrt{\frac{\sum f d^2}{n} - \left(\frac{\sum f d}{n}\right)^2} \times i$$

where  $n = \sum f$ . We get the desired standard deviation  $\sigma$ .

The above method is explained by the following examples.

**Example 11 :** Find the standard deviation of intelligence quotient (I.Q) of 68 students of the following data:

I.Q.	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
No. of students :	5	12	15	20	10	4	2

**Solution :** Let us prepare the following table in order to calculate the standard deviation, by assuming  $A = 45$ .

I.Q. (Class interval)	No. of Students (f)	Mid-value (x)	$d = \frac{x - 45}{10}$	$fd$	$fd^2$
10 – 20	5	15	-3	-15	45
20 – 30	12	25	-2	-24	48
30 – 40	15	35	-1	-15	15
40 – 50	20	45	0	0	0
50 – 60	10	55	1	10	10
60 – 70	4	65	2	8	16
70 – 80	2	75	3	6	18
Total	$\Sigma f = n = 68$			$\Sigma fd = -30$	$\Sigma fd^2 = 152$

$$\therefore \sigma = i \times \sqrt{\frac{\Sigma fd^2}{n} - \left(\frac{\Sigma fd}{n}\right)^2} = 10 \times \sqrt{\frac{152}{68} - \left(\frac{-30}{68}\right)^2} = 14.3 \text{ approx.}$$

**Example 12 :** Find the standard deviation by the step deviation method for the following data on the age of patients suffering from pulmonary disease.

Age in years :	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of patients :	6	14	10	8	1	3	8

**Solution :** Let the assumed mean be  $A = 35$ ,  $d = \frac{x - A}{c}$ , where  $c = 10$ .

TABLE : Computation of Standard Deviation

Age in years	No. of Patients (f)	Mid-value (x)	$d = \frac{x - A}{c}$ $= (x - 35)/100$	$d^2$	$fd$	$fd^2$
0 – 10	6	5	-3	9	-18	54
10 – 20	14	15	-2	4	-28	56
20 – 30	10	25	-1	1	-10	10
30 – 40	8	35	0	0	0	0
40 – 50	1	45	1	1	1	1
50 – 60	3	55	2	4	6	12
60 – 70	8	65	3	9	24	72
Total	50				$\Sigma fd = -25$	$\Sigma fd^2 = 205$

$$\text{Standard deviation : } \sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times c = \sqrt{\frac{205}{50} - \left(\frac{-25}{50}\right)^2} \times 10$$

$$\therefore \sigma = \sqrt{4.1 - 0.25} \times 10 = \sqrt{3.85} \times 10 = 19.62.$$

**Example 13 :** In a study on patients, the following data was obtained. Find the standard deviation of the data.

Age (in years) :	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89
Number of Cases:	1	0	1	10	17	38	9	3

**Solution :**

TABLE : Computation of Standard Deviation

Age (in years)	No. of cases (f)	Mid-value (x)	$d = \frac{x - 44.5}{10}$	$fd$	$fd^2$
10 – 19	1	14.5	-3	-3	9
20 – 29	0	24.5	-2	0	0
30 – 39	1	34.5	-1	-1	1
40 – 49	10	44.5	0	0	0
50 – 59	17	54.5	1	17	17
60 – 69	38	64.5	2	76	152
70 – 79	9	74.5	3	27	81
80 – 89	3	84.5	4	12	48
<b>Total</b>	$n = 79$			$\Sigma fd = 128$	$\Sigma fd^2 = 308$

Here  $= i = 10$

$$\text{Standard deviation} = i \times \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} = 10 \times \sqrt{\frac{308}{79} - \left(\frac{128}{79}\right)^2}$$

$$= \frac{10}{79} \sqrt{24332 - 16384} = \frac{10}{79} \times 89.15 = 11.28.$$

## 6.11 LIMITS OF VARIABILITY

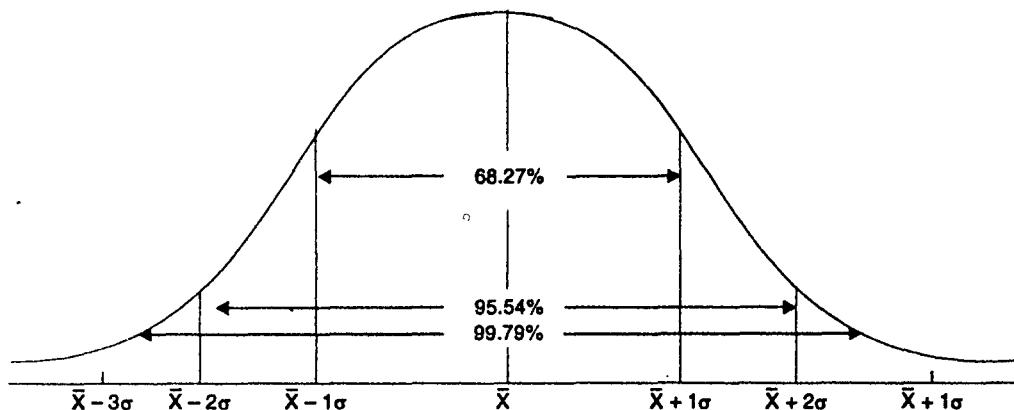
Standard deviation shows the limits of variability by which the individual item in a distribution will vary from the mean. For a symmetrical distribution with mean  $\bar{X}$ , the following area relationship holds good:

$\bar{X} \pm \sigma$  covers 68.27% items.

$\bar{X} \pm 2\sigma$  covers 95.45% items.

$\bar{X} \pm 3\sigma$  covers 99.73% items.

These limits are illustrated by the following curve known as Normal Curve.



## 6.12 EMPIRICAL RELATIONSHIPS

If the data is moderately non-symmetrical, then we have the following empirical relationships.

$$\text{Mean Deviation} = \frac{4}{5} \times \text{Standard deviation} = \frac{4}{5} \sigma$$

$$\text{Semi-Inter-quartile Range} = \frac{2}{3} \times \text{Standard deviation} = \frac{2}{3} \sigma$$

$$\text{Probable Error of Standard Deviation} = \frac{2}{3} \sigma = \text{Semi-inter-quartile Range}.$$

$$\text{Quartile Deviation} = \frac{5}{6} \text{ of Mean deviation.}$$

From these relationships, we have

$$4 \text{ Standard Deviation} = 5 \text{ Mean Deviation} = 6 \text{ Quartile Deviation}$$

## 6.13 VARIANCE AND COEFFICIENT OF VARIATION

**Variance :** The variance is the square of standard deviation and is denoted by  $\sigma^2$ . The methods for calculating variance are the same as for the standard deviation.

**Coefficient of Variation :** It is a relative measures of dispersion. It is, generally, denoted by C.V. and is given by the formula

$$\text{Coefficient of Variation or C.V.} = \frac{\sigma}{\bar{x}} \times 100$$

where  $\sigma$  = s.d. and  $\bar{x}$  = the mean of the given series.

It is important to note that the coefficient of variation is always a percentage.

*The coefficient of variation has great practical significance and is the best measure of comparing the variability of the two series. The series or group for which the coefficient of variation is greater is said to be more variable (less consistent). On the other hand, the series for which the variation is less is said be less variable or (more consistent).* Coefficient of variation can be employed for comparing the relative consistency of the prices of shares of two

or more companies. It will help a genuine investor (in shares) in selecting share, the price of which is relatively more stable. Thus the shares which are more consistent in the fluctuation of prices will be preferred by him.

**Example 14 :** In a series of adults, the mean blood pressure was 135 mmHg with S.D. as 10 mmHg. In the same series the mean height was 170 with S.D. as 6. Find which character shows greater variations.

**Solution :** C.V. (of B.P.) =  $\frac{\sigma}{\bar{X}} \times 100 = \frac{10}{135} \times 100 = 7.40\%$

[Here mean  $(\bar{X}) = 135$  mmHg and  $\sigma = 10$  mmHg]

$$\text{C.V. (of height)} = \frac{\sigma}{\bar{X}} \times 100 = \frac{10}{170} \times 6 = 3.52\%.$$

Since C.V. (B.P.) is more than the C.V. (height), so systolic blood pressure is more variable character.

**Example 15 :** A researcher collects data on the weight and length of fishes and is interested to find out which of the two characters is more variable. The data are:

Fish	Mean	Standard deviation
Weight	350 gms	12 gms
Lengths	16 inches	1.5 inch

**Solution :** Coefficient of variation for weights =  $\frac{\sigma}{\bar{X}} \times 100$ .

or C.V. (Weight) =  $\frac{12}{350} \times 100 = 3.43\%$ .

Coefficient of variation for heights =  $\frac{\sigma}{\bar{X}} \times 100$

or C.V. (length) =  $\frac{1.5}{16} \times 100 = 9.375$ .

Since C.V (weight) is more than C.V. (length). So there is a greater variability in the lengths of the fishes than their weights.

### EXERCISE

- The size 5 red fishes (in cm) is 1, 2, 5, 4, 3. Find its mean and standard deviation.

[Hint : Mean =  $\frac{\Sigma x}{n} = \frac{1+2+5+4+3}{5} = 3$  cm =  $\Sigma x^2 - \left( \frac{(\Sigma x)^2}{n} \right) = 55 - \frac{(15)^2}{5} = 10$  cm]

- Find the standard deviation of erythrocyte sedimentation rate (ESR) of the data: 3, 4, 5, 4, 2, 4, 5, 3 found in 8 normal persons.

[Hint : Here  $\Sigma x = 3 + 4 + 5 + 4 + 2 + 4 + 5 + 3 = 30$ . Similarly,  $\Sigma x^2 = 120$

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{120}{8} - \frac{900}{64}} = \sqrt{15 - 14.06} = \sqrt{0.94} = 0.97$$

3. Calculate the mean and standard deviation of following data on the length of fishes (in cm) reared under central conditions.

Length :	10	20	30	40	50	60	70
No. of Fishes :	4	6	10	15	20	15	10

[Hint :  $A = 40$ ,  $l = 10$ , then  $\Sigma fd = 46$ ,  $\Sigma fd^2 = 240$

$$\therefore \sigma = \sqrt{\frac{\Sigma fd^2}{n} - \left(\frac{\Sigma f}{n}\right)^2} \times l = \left[ \sqrt{\frac{240}{100}} - \frac{46}{100} \right] \times 10 = (\sqrt{2.4} - 0.21) \times 10 = 14.79 \text{ cm}$$

4. Calculate the mean and standard deviation from the following data of presence of urea in the blood samples of 520 patients in a hospital.

Range of (mg/dl) :	20–25	25–30	30–35	35–40	40–45	45–50	50–55	55–60
No. of patients :	20	44	60	101	109	84	66	10

[Hint : Take  $d = (m - 37.5) / 5$ ,  $C = 5$ , then

mid-point ( $m$ ) : 17.5 22.5 27.5 32.5 37.5 42.5 47.5 52.5 57.5

$d$ :	-4	-3	-2	-1	0	1	2	3	4
$f$ :	20	26	44	60	101	109	84	66	10
$fd$ :	-80	-78	-88	-60	0	109	168	198	40
$fd^2$ :	320	234	176	60	0	109	336	594	160

Then  $\Sigma fd = 209$ ,  $\Sigma fd^2 = 1989$ ;  $N = \Sigma f = 520$ .

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times C = \sqrt{\frac{1989}{520} - \left(\frac{209}{520}\right)^2} \times 5 = 9.58.$$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times C = 37.5 + \frac{209}{520} \times 5 = 39.5$$

5. Calculate the standard deviation for data of hypo B.P. (in mmHg).

B.P. ( $X$ ) :	50	60	70	80	90	100	110	120
No. of Patients ( $f$ ) :	14	40	54	46	26	12	6	2

[Hint : Take  $A = 80$ ,  $C = 10$ ,  $d = (x - 80)/10$ , then  $\Sigma fd = -100$ ;  $\Sigma d^2 = 500$ .

$$\text{Standard Deviation : } \sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times C = \sqrt{\frac{500}{200} - \left(\frac{-100}{200}\right)^2} \times 10 = 15 \text{ mmHg}$$

6. In two series of adults age 25 years and children 7 months old the following values were obtained for the height. Find which series shows a greater variations?

Person	Mean height	Standard Deviation
Adults	160 cm	10 cm
Children	60 cm	5 cm

$$[\text{Hint : } \text{C.V (adults)} = \frac{\sigma}{\bar{X}} \times 100 = \frac{10}{160} \times 100 = 6.25\%]$$

$$\text{C.V (children)} = \frac{\sigma}{\bar{X}} \times 100 = \frac{5}{60} \times 100 = 8.33\%$$

Since C.V (children) > C.V. (adults), so these is a greater variations in the heights of children than that of adults.]

7. In a series of boys, the mean systolic blood pressure was 120 mmHg and S.D. was 10. In the same series mean heights and S.D. were 160 and 5 cm respectively. Find which character shows greater variations?
8. A hospital carry out experiments on 10 patients for the effect of two medicine *A* and *B* on to reduce the total cholesterol level in their blood. The following results were obtained. Find which medicine has more variable effect.

Medicine	Mean	Standard Deviation
<i>A</i>	157 mg/dl	2.6 mg/dl
<i>B</i>	175 mg/dl	3.1 mg/dl

$$[\text{Hint : } \text{C.V. (A)} = \frac{\sigma}{\bar{X}} \times 100 = \frac{2.6}{157} \times 100 = 1.66\%]$$

$$\text{C.V. (B)} = \frac{\sigma}{\bar{X}} \times 100 = \frac{3.1}{175} \times 100 = 1.77\%$$

Since C.V. (B) > C.V. (A)  $\Rightarrow$  medicine *A* has more variable effect]

9. Compute the S.D. for the following frequency distribution of wage earners in a factory:

Wages per hour (in Rs.) ( <i>X</i> ) :	9	12	15	18	21	24	27	30
No. of Wage Earners. ( <i>f</i> ) :	20	60	150	250	200	120	50	40

[Hint : Take *A* = 18, *C* = 3, and *d* =  $(x - 18)/3$ ; *N* = 890. Then  $\Sigma fd = 420$ ,  $\Sigma fd^2 = 2340$ .

$$\sigma = \left[ \sqrt{\frac{\sum fd^2}{N}} - \left( \frac{\sum fd}{N} \right)^2 \right] \times C = \left[ \sqrt{\frac{2340}{890}} - \frac{420}{890} \right] \times 3 = 4.65]$$

10. The following table gives the marks obtained by 10 B.Sc. (Zoology Honours) students in biostatistics papers. Calculate its standard deviation.

R.No. :	1	2	3	4	5	6	7	8	9	10
Marks. :	43	48	65	57	31	60	37	48	78	59

[Hints : Take  $A = 50$ ,  $d = x - 50$ , then  $\sum d = 26$ ,  $\sum d^2 = 1826$ ,

$$\therefore \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{1826}{10} - \left(\frac{26}{10}\right)^2} = 13.26]$$

11. In an experiment to study the effectiveness of a new variety of seeds, yields per hectare (in quintals) were recorded from 50 fields with the following data:

Compute the mean and standard deviation.

Size of Field (in hectare) :	6	7	8	9	10	11	12
No. of Fields :	3	6	9	13	8	5	4

[Hint : Take  $A = 9$ ,  $d = x - A (= x - 9)$ , then  $\sum fd = 0$ ,  $\sum fd^2 = 124$ .

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{124}{48} - \left(\frac{0}{48}\right)^2} = 1.6]$$

12. Suppose that the following represent the number of children for 10 physicians on a particular hospital staff : 3, 2, 0, 1, 4, 7, 3, 2, 4, 2. Find the following descriptive measures:

- (a) The arithmetic mean; (b) The median ; (c) The mode ; (d) The variance ;  
 (e) The standard deviation.

13. Thirteen sheep were fed pingue (a toxin-producing weed of the south-western United States) as part of an experiment by Aanes (1) and died as a result. The time of death in hours after the administering of pingue for each sheep was as follows: 44, 27, 24, 24, 36, 36, 44, 44, 120, 29, 36, 36, 36. For these data, compute

- (a) The arithmetic mean ; (b) The median ; (c) The mode  
 (d) The variance ; (e) The standard deviation.

## ANSWERS

1.  $\bar{X} = 3$  cm,  $\sigma = 3.0$  cm.
2.  $\sigma = 0.97$ .
3.  $\sigma = 14.79$  cm.
4. 39.5
5.  $\sigma = 15$  mmHg.
6. Greater variations in the heights of children than in adults.
7. B.P. is found to be more variable character than height.
8. Medicine A has more variable effect.
9.  $\sigma = \text{Rs. } 4.65$ .
10.  $\sigma = 13.26$  marks.
11.  $\sigma = 1.6$  hectare.



# 7

# Skewness, Moments and Kurtosis

## 7.1 SKEWNESS

The literally meaning of ‘skewness’ is ‘lack of symmetry’. *Skewness is opposite to symmetry* and its presence tells us that a particular distribution is not symmetrical.

A distribution is said to be symmetrical when mean, median and mode are identical or coincide.

A symmetrical distribution when plotted on a graph give a perfectly bell-shaped curve as shown in the figure. A distribution said to be skewed if its frequency curve is not symmetrical but it is stretched more on one side than of skewness to the other side.

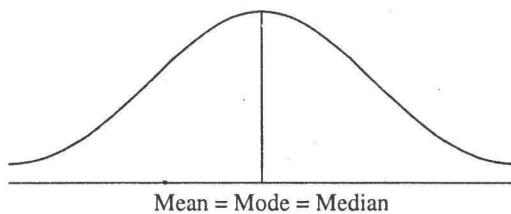


Fig. 7.1 : Symmetrical curve

## 7.2 DEFINITION OF SKEWNESS

Various statisticians defined skewness in the following various ways :

“Skewness is the lack of symmetry. When a frequency distribution is plotted on a chart, skewness present in the series tends to be dispersed more on one side of the mean than on the other”. — Riggleman and Frisbee

“Skewness or asymmetry is the attribute of a frequency distribution that extends further on one side of the class with the highest frequency than on the other”. — Simpson and Kafka

“When a series is not symmetrical it is said to be asymmetrical or skewed”.

— Croxton and Cowden

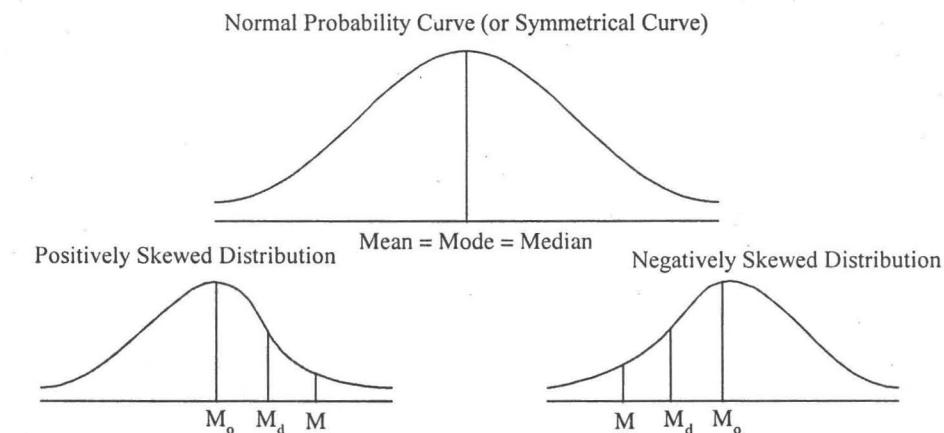
A distribution having 'mode' can be divided into *three parts* (i) *the left-tail*; (ii) *the middle part*; and (iii) *the right tail*. In the case of symmetrical distribution, the two tails (Right and left tails) are of equal length. But in asymmetrical distribution (or skewed distribution) one tail is longer than the other.

### 7.3 POSITIVELY AND NEGATIVELY SKEWNESS

A distribution which is not symmetrical is called a **skewed** distribution and in such distributions, the Mean, the Median. The Mode will not coincide, but the values are pulled a part.

**Positive skewness :** If the curve of the distribution has a longer tail towards the right, it is said to skewness. In this case **Mean > Median > Mode**.

**Negative skewness :** If the curve has a longer tail towards the left, it is said to be negative skewness. In this case **Mean < Median < Mode**.



### 7.4 PURPOSE OF SKEWNESS

1. To know whether the distribution is normal many statistical measures are based on the normal distribution (i.e., bell shaped curve).
2. To find out the nature and degree of concentration of items (or observations) of the distribution.

### 7.5 DIFFERENCE BETWEEN DISPERSION AND SKEWNESS

Dispersion	Skewness
<ol style="list-style-type: none"> <li>1. It shows us the spread of individual values about the central value. i.e., mean.</li> <li>2. It is a type of averages of deviation-average of the second order.</li> <li>3. It judges the truthfulness of the central tendency.</li> <li>4. It shows the degree of variability.</li> </ol>	<ol style="list-style-type: none"> <li>1. It shows us departure from symmetry, i.e., direction of variations.</li> <li>2. It is not an average, but is measured by the use of the mean, median and mode.</li> <li>3. It judges the differences between the central tendencies.</li> <li>4. It shows whether the concentration is in higher or lower values.</li> </ol>

## 7.5 MEASURES OF SKEWNESS

**Skewness** is used to find out the extent of asymmetry (i.e., departure from symmetry and direction in a series).

The measures of asymmetry are usually called measures of skewness. Measures of skewness indicate not only the extent of skewness (in numerical expressions), but also the direction; i.e., the manner in which the deviations are distributed. These measures can be absolute or relative.

### Absolute Measures

The absolute measures are also known as measures of skewness. *The relative measures are known as the coefficient skewness. The absolute measure tells us the extent of asymmetry, whether it is positive or negative.*

It is based on the difference between mean and mode.

$$\text{Absolute Skewness} = \text{Mean} - \text{Mode}$$

*In a symmetrical distribution absolute skewness will be zero because in this case Mean = Mode.*

*In a positively skewed frequency distribution absolute skewness will be because in this case "Mean is greater than Mode". Similarly, in a negatively skew distribution absolute skewness will be negative because in this "Mean less than Mode".*

The absolute measure of skewness is not very useful measure because it cannot be effectively used to compare the two or more distributions. Moreover absolute measure is expressed in the units of the original data and therefore, cannot be used for the comparison of skewness in two different distributions if they are in different units.

Thus for comparison purpose, we use relative measure of skewness known as **coefficient of skewness**.

## 7.7 RELATIVE MEASURES

There are four important measures relative skewness.

1. Karl Pearson's coefficient of skewness.
2. Bowley's coefficient of skewness.
3. Kelly's co-efficient of skewness.
4. Measures of coefficient based on moments.

## 7.8 KARL PEARSON'S CO-EFFICIENT OF SKEWNESS

Karl Pearson's coefficient of skewness enables us to find out the direction as well as **extent of skewness**. This method is based on the fact that in an asymmetrical distribution mean and mode pull apart from one another and that the greater the distance between the two, the greater is the degree of skewness. The formula is:

$$Sk = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

Suppose, if the mode is **ill-defined**, then co-efficient of skewness is determined by the following formula:

$$Sk = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

### Properties of Karl Pearson's coefficient of skewness

1. It's value usually lies between  $\pm 1$ .
2. When it's value is zero, there is no skewness, i.e., the distribution is symmetrical.
3. When its value is negative, the distribution is negatively skewed.
4. When its value is positive, the distribution is positively skewed.

**Example 1 :** Calculate Karl Pearson's coefficient of skewness for the following data on the number of red flowers on a plant 12, 18, 35, 22 and 18.

**Solution :**

#### Calculation of Mean and Standard Deviation

Sl.No.	No. of Red flowers ( $X$ )	$(X - \bar{X})$	$X$
1	12	-9	81
2	18	-3	9
3	35	14	196
4	22	1	1
5	18	-3	9
$N = 5$	$\Sigma X = 105$		$\Sigma X^2 = 296$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{105}{5} = 21$$

$$\sigma = \sqrt{\frac{\Sigma X^2}{N}} = \sqrt{\frac{296}{5}} = \sqrt{59.2} = 7.7$$

Mode = 18, because it occurs maximum number of times in the series.

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{21 - 18}{7.7} = \frac{3}{7.7} = 0.4.$$

**Example 2 :** 120 patients were tested their blood for total cholesterol from two pathology laboratories A and B. The following results were obtained in mg/dl).

**Solution :**

Laboratory A : Mean = 46.83 ; Mode = 51.67 ; S.D. = 14.8

Mean = 47.83 ; Mode = 47.07 ; S.D. = 14.8.

Determine the results of which laboratory is more skewed.

$$\text{Laboratory A : } Sk_A = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{46.83 - 51.67}{14.8} = \frac{-484}{14.8} = -0.327$$

$$\text{Laboratory } B : Sk_B = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{47.83 - 47.07}{14.8} = \frac{0.76}{14.8} = 0.0514.$$

Thus we find that  $|Sk_A| = 0.327$  is greater than  $|Sk_B| = 0.0514$ , so the results of the pathology laboratory *A* are more skewed.

**Example 3 :** Consider the following distribution of blood test for fasting sugar of 100 persons in two pathology laboratory.

	Laboratory A	Laboratory B
Mean	100	90
Median	90	80
Standard Deviation	10	10

Both the results have the same degrees of skewness. True/False?

**Solution :** Karl Pearson's co-efficient of skewness is:

$$Sk = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Skewness for the laboratories *A* and *B*.

$$\text{Laboratory } A : \text{Coefficient of skewness} : Sk(A) = \frac{3(100 - 90)}{10} = 3$$

$$\text{Laboratory } B : \text{Coefficient of skewness} : Sk(B) = \frac{3(90 - 80)}{10} = 3$$

Since  $Sk(A)$  and  $Sk(B) = 3$ , the statement that both the laboratories have the same degree of skewness is true.

**Example 4 :** From a moderately skewed distribution of retail prices for men's shoes, it is found that the mean price is Rs. 20 and the median price is Rs. 17. If the coefficient of variation is 20%, find the Pearsonian coefficient of skewness of the distribution.

**Solution :** We are given that:

Mean = 20 and Median = 17. To find the coefficient of skewness, we need standard deviation.

$$C.V. = \frac{\text{Standard deviation} \times 100}{\text{Mean}}$$

$$\text{or } 20 = \frac{\sigma}{20} \times 100 \Rightarrow 5\sigma = 20 \text{ or } \sigma = 4.$$

$$Sk = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}} = \frac{3(20 - 17)}{4} = \frac{3 \times 3}{4} = 2.25.$$

**Example 5 :** From the following data find out Karl Pearson's coefficient of skewness:

Measurement : 10      11      12      13      14      15

Frequency : 2      4      10      8      5      1

(Guru Nanak Dev. Uni. B. Com. II. Sept, 1982)

**Solution :**  $X$ : Measurement ;  $f$ : Frequency.

**Computation of Mean and Mode s.d.**

$x$	$f$	$fx$	$fx^2$
10	2	20	200
11	4	44	484
12	10	120	1440
13	8	104	1352
14	5	70	980
15	1	15	225
<b>Total</b>	$N = 30$	$\Sigma fx = 373$	$\Sigma fx^2 = 4681$

$$\text{Mean } (M) = \frac{\Sigma fx}{N} = \frac{373}{30} = 12.43$$

$$\begin{aligned}\text{S.D. } (\sigma) &= \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2} \\ &= \sqrt{\frac{4681}{30} - \left(\frac{373}{30}\right)^2} = \sqrt{156.0333 - 154.5049} = \sqrt{1.5284} = 1.2363\end{aligned}$$

Mode is the value of  $x$  corresponding to the maximum frequency, viz., 10.

∴ Mode = 12

Karl Pearson's coefficient of skewness is given by

$$Sk = \frac{M - M_0}{\sigma} = \frac{12.43 - 12.00}{1.2363} = 0.3473.$$

**Example 6 :** Find the coefficient of skewness for the following:

<i>Wages (Rs.)</i>	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5
<i>No. of workers</i>	35	40	48	100	125	87	43	22

**Solution :**

**Calculation of Mean, Mode and Standard Deviation**

<i>Wages (Rs.)</i> $X$	<i>No. of workers</i> $f$	$(X - 8.5)$ $d$	$fd$	$fd^2$
4.5	35	-4	-140	560
5.5	40	-3	-120	360
6.5	48	-2	-96	192
7.5	100	-1	-100	100
8.5	125	0	0	0
9.5	87	1	87	87
10.5	43	2	86	172
11.5	22	3	66	198
	$N = 500$		$\Sigma fd = -217$	$\Sigma fd^2 = 1669$

$$\bar{X} = A + \frac{\sum fd}{N} = 8.5 + \frac{-217}{500} = 8.5 - 0.43 = 8.07$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \\ &= \sqrt{\frac{1669}{500} - \left(\frac{-217}{500}\right)^2} = \sqrt{3.34 - 0.19} = \sqrt{3.15} = 1.77\end{aligned}$$

Mode = 8.5, because it occurs maximum no. of times.

$$\text{Coefficient of skewness} = \frac{\bar{X} - \text{Mode}}{\sigma} = \frac{8.07 - 8.5}{1.77} = \frac{-0.43}{1.77} = -0.24.$$

## 7.9 BOWLEY'S COEFFICIENT OF SKEWNESS

Prof. A.L. Bowley's coefficient of skewness is based on the quartiles and is given by

$$\text{Bowley's coefficient of skewness : } Sk = \frac{Q_3 - Q_1 - 2 \text{ Median}}{Q_3 - Q_1}.$$

where,  $Q_1$  = First quartile,  $Q_3$  = Third quartile

Limits for Bowley's coefficient of skewness : It ranges from -1 to 1

i.e.,  $-1 \leq Sk \text{ (Bowley)} \leq 1$ .

**Example 7 :** A distribution had  $Q_1 = 31.3$ , median = 35, and  $Q_3 = 36.4$ . Calculate the coefficient of skewness.

**Solution :** Here  $Q_3 = 36.4$  ;  $Q_1 = 31.3$  ; med = 35.

$$\text{Co-eff. skewness} = \frac{Q_3 - Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

$$\text{or } Sk = \frac{36.4 + 31.3 - 2 \times 35}{36.6 - 31.3} = \frac{-2.3}{5.3} = -0.43.$$

Hence the distribution is negatively skewed.

## 7.10 KELLY'S MEASURE OF SKEWNESS

Bowley's co-efficient of skewness ignores 50% of the data towards the extremes. This can partially removed by taking two deciles or percentiles equidistant from the median values. This refinement was suggested by Kelly. Kelly has suggested the following formula for measuring skewness upon 10th and 90th percentiles.

$$\text{Kelly's coefficient of skewness} = \frac{P_{10} + P_{90} - 2 \text{ median}}{P_{90} - P_{10}}$$

$$\text{Kelly's coefficient of skewness} = \frac{D_1 + D_9 - 2 \text{ median}}{D_9 - D_1}$$

**Remark :** This method is primarily of theoretical importance only and is seldom used in practice.

**Example 8 :** Calculate percentile co-efficient of skewness from the following positional measures given below:

$$P_{90} = 101 ; P_{10} = 58.12 ; P_{50} = 79.06.$$

**Solution :**

$$\text{Kelly's coefficient of skewness} = \frac{P_{10} + P_{90} - 2 \text{ median}}{P_{90} - P_{10}}$$

$$\text{or } Sk(\text{kelly}) = \frac{101 + 58.12 - 2(79.06)}{101 - 58.12} = \frac{159.12 - 158.12}{42.88} = \frac{1}{42.88} = 0.02$$

Hence the distribution is positively skewed.

## 7.11 CO-EFFICIENT OF SKEWNESS BASED ON MOMENTS

The term ‘moment’ in mechanics refers to the turning or the rotating effect of a force. In statistics, it is used to describe the peculiarities of a frequency distribution. Using moments, we can measure the central tendency of set of observations, their scatter, asymmetry and the peakedness of the curve. Deviations of items are taken from the arithmetic mean of the distribution. The arithmetic mean of the various powers of the deviations will give the required moments of the distribution. The moments about the actual mean is denoted by the Greek letter  $\mu(\mu)$ .

rth order moment about the mean:

$$\mu_r = \frac{(X_1 - \bar{X}) + (X_2 - \bar{X}) + \dots + (X_n - \bar{X})}{N} = \frac{\sum(X - \bar{X})}{N},$$

where  $\bar{X}$  = is the mean of items  $X_1, X_2, \dots, X_n$  and

$$\bar{X} = \frac{\sum X}{N}$$

The first four moments about arithmetic mean are called central moments and are given by the following formulae.

Moments	Individual Series	Discrete Series
First moment : $\mu_1$	$\frac{\sum(X - \bar{X})}{N}$	$\frac{\sum f(X - \bar{X})}{N}$
Second moment : $\mu_2$	$\frac{\sum(X - \bar{X})^2}{N}$	$\frac{\sum f(X - \bar{X})^2}{N}$
Third moment : $\mu_3$	$\frac{\sum(X - \bar{X})^3}{N}$	$\frac{\sum f(X - \bar{X})^3}{N}$
Fourth moment : $\mu_4$	$\frac{\sum(X - \bar{X})^4}{N}$	$\frac{\sum f(X - \bar{X})^4}{N}$

Where  $N = \Sigma f$ .

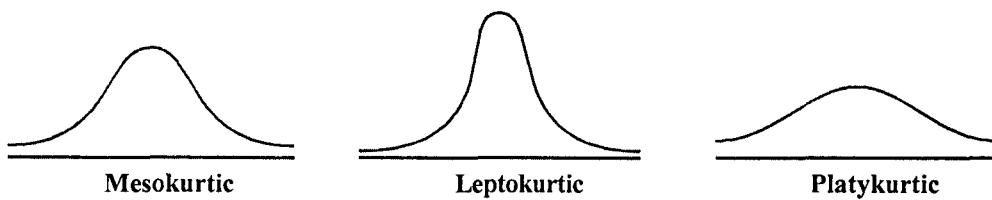
## 7.12 ROLE OF MOMENTS

1. The first moment ( $\mu_1$ ) of a frequency distribution is always zero, i.e.,  $\mu_1 = 0$ . **It measures mean of the distribution, i.e.,  $\mu_1 = \bar{X} = 0$ .**
2. The second moment ( $\mu_2$ ) of a frequency distribution about the mean is the variance of the distribution, i.e.,  $\mu_2 = \sigma^2$ . **It measures variance i.e., the spread of the different terms in a distribution.**
3. The third moment ( $\mu_3$ ) gives an **idea about the degree of skewness present in a series**.
4. The fourth moment ( $\mu_4$ ) throws light on the height of a frequency distribution, i.e., whether it is more peaked or more flat topped than the normal curve. **It measures Kurtosis.**
5. Co-efficient of skewness is given by  $\beta_1$ , where  $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ .
6. Kurtosis is measured by  $\beta_2$ , where  $\beta_2 = \frac{\mu_4}{\mu_2^2}$ .

## 7.13 KURTOSIS

Kurtosis enables us to have an idea about the shape and nature of the hump (middle part) of a frequency distribution. It is concerned with the flatness or peakedness of the frequency curve. According to C.M. Mayers "**Kurtosis refers to the degree of peakedness of the hump of the distribution**".

Karl Pearson called it a "**Measures of Convexity**" of the curve. He introduced three broad patterns of peakedness which are illustrated in the following diagram.



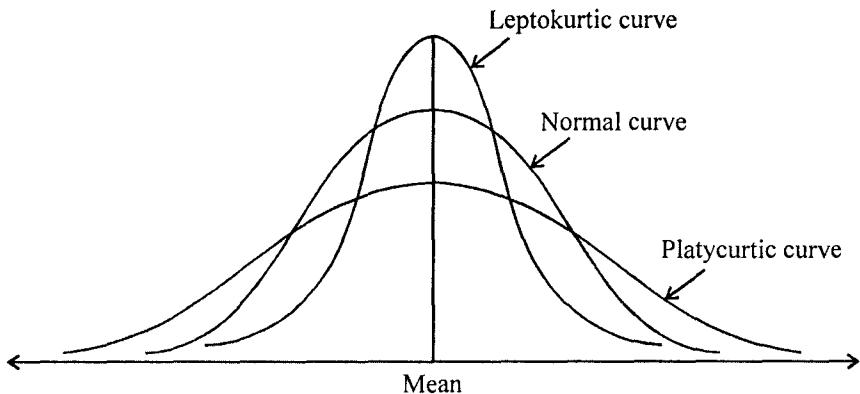
The curve which is neither flat nor peaked is known as **normal curve or mesokurtic**. A curve which is more peaked than the normal curve is known as **Leptokurtic** and the curve which is flatter than the normal curve is called the **Platykurtic**.

**Measures of Kurtosis :** As a measure of Kurtosis, Karl Pearson gave the coefficient of Kurtosis as co-efficient of Beta two ( $\beta_2$ ) and its next derivative as  $r_2$ . The measures are defined as:

$$\beta^2 = \frac{\mu^4}{\mu_2^2} = \frac{\mu^4}{\sigma^4}$$

$$r_2 = \beta_2 - 3 = \frac{\mu^4}{\sigma^4} - 3 = \frac{\mu^4 - 3\sigma^4}{\sigma^4}$$

When the value of  $\beta_2 > 3$ , it is a leptokurtic.



I. When the value of  $\beta_2 < 3$ , it is a platykurtic.

II. When the value of  $\beta_2 = 3$ , it is a mesokurtic.

**Example 9 :** The number of bacteria in 1 ml of blood from 5 persons are 2, 3, 7, 8, 10. Calculate the first, second, third and fourth moments about the mean.

Also find skewness and Kurtosis.

**Solution :**

Table : Calculation of moments

$X$	$(X - \bar{X})$ $x$	$(X - \bar{X})^2$ $x^2$	$(X - \bar{X})^3$ $x^3$	$(X - \bar{X})^4$ $x^4$
2	-4	16	-64	256
3	-3	9	-27	81
7	1	1	1	1
8	2	4	8	16
10	4	16	64	256
$N = 5$	$\Sigma x = 0$	$\Sigma x^2 = 46$	$\Sigma x^3 = -18$	$\Sigma x^4 = 610$

$$\mu_1 = \frac{\Sigma X}{N} ; \quad \mu_2 = \frac{\Sigma X^2}{N} ; \quad \mu_3 = \frac{\Sigma X^3}{N} ; \quad \mu_4 = \frac{\Sigma X^4}{N}$$

$$\mu_1 = \frac{0}{5} = 0 ; \quad \mu_2 = \frac{46}{5} = 9.2 ; \quad \mu_3 = \frac{-18}{5} = -3.6 ; \quad \mu_4 = \frac{610}{5} = 122.$$

$$\text{Skewness } (\beta_1) = \frac{\mu_3^2}{\mu_2^3} = \frac{12.96}{778.688} = 0.0166.$$

$$\text{Kurtosis } (\beta_2) = \frac{\mu_4^4}{\mu_2^2} = \frac{122}{(9.2)^2} = 1.4.$$

As the Kurtosis is less than 3, i.e., 1.4, the distribution is platykurtic.

**Example 10 :** The first four central moment of a distribution are 0, 2.5, 0.7 and 18.75. Test the skewness and kurtosis of the distribution.

**Solution :**

$$\begin{aligned}\text{Skewness } \beta_1 &= \frac{\mu_3^2}{\mu_2^3} \\ &= \frac{(0.7)^2}{(2.5)^3} = \frac{0.49}{15.625} = + 0.031\end{aligned}$$

The distribution is not perfectly symmetrical as  $\beta_1 = +0.03$ .

$$\begin{aligned}\text{Kurtosis } \beta_2 &= \frac{\mu_4}{\mu_2^2} \\ &= \frac{18.75}{(2.5)^2} = \frac{18.75}{6.25} = 3.\end{aligned}$$

The distribution is mesokurtic as  $\beta_2 = 3$ .

### EXERCISE

1. Calculate Karl Pearson's co-efficient of skewness for the following data of blood samples of 9 patients for the triglycerides (in mg/dl) present in their blood.

25, 15, 23, 40, 27, 25, 23, 25, 20.

2. From the information given below, calculate Karl Pearson's coefficient of skewness and also quartile coefficient of skewness.

Measure	Firm A	Firm B
Mean	150	140
Median	142	155
Standard deviation	30	55
Third quartile	195	260
First quartile	62	80

3. In a distribution, the difference between two quartiles is 30 and their sum is 70 and median is 40. Find the coefficient of skewness.

[Hint :  $Q_1 + Q_3 = 70$  and  $Q_3 - Q_1 = 30$ , and median = 40]

$$\therefore Sk = \frac{Q_3 + Q_1 - 2 \text{ median}}{Q_3 - Q_1} = \frac{70 - 80}{30} = -0.33]$$

4. The first four central moments of a distribution blood samples of 100 patients for their lipid profile are 0, 2.3, 0.9 and 15.65. Test the skewness and kurtosis of the distribution.
5. In a distribution the difference of two quartiles is 2.03 and their sum is 72.67 and the median is 36.8. Find the coefficient of skewness.
6. For a group of 10 rats,  $\Sigma X = 452$ ,  $\Sigma X^2 = 24270$  and Mode = 43.7. Find their Karl Pearson's coefficient of skewness.

7. For the data set  $x_1 = -3, x_2 = 5, x_3 = 40$ , find the first four central moments. Also find the skewness and kurtosis.

[Hint :  $\bar{X} = \frac{42}{3} = 14$ ;  $\Sigma(X - \bar{X}) = 0$ ,  $\Sigma(X - \bar{X})^2 = 1046$ ;

$$\Sigma(X - \bar{X})^3 = 11934, \quad \Sigma(X - \bar{X})^4 = 547058, N = 3]$$

8. Find the measures of skewness for the following data of red flowers on a plant : 20, 3, 10, 5 and 2.  
 9. The data on the number of peas in a pod is 4, 5, 2, 7, 2. Find its  
     (i) first four moments about the mean ;      (ii) skewness and ;      (iii) Kurtosis.  
 10. Calculate any measure of skewness for the following :

No. of Flowers on a Plant	0	1	2	3	4	5	6	7
No. of Plants	12	17	29	19	8	4	1	0

### ANSWERS

1.  $-0.03$       2.  $Sk$  (Karl Pearson)  $= -0.2$  ;  $Sk$  (Bowley)  $= 0.17$ .      3.  $Sk = -0.33$   
 4.  $\beta_1 = 0.07$  ;  $\beta_2 = 2.96$  ; the curve is Platykurtic.      5.  $Sk = 0.153$   
 6.  $0.007$   
 7.  $\mu_1 = 0, \mu_2 = 348.66, \mu_3 = 3978, \mu_4 = 182352.66; Sk = 0.61$  ; Kurtosis  $= -1.50$   
 8.  $Sk = 0.95$   
 9. (i)  $\mu_1 = 0, \mu_2 = 3.60, \mu_3 = 2.40, \mu_4 = 22.80$ .  
     (ii)  $Sk = 0.35$  ;      (iii) Kurtosis  $= -1.24$ .  
 10.  $Sk = 0.08$ .



# 8

# Correlation Analysis

## 8.1 CORRELATION

We know that the area  $A$  of a circle of radius  $r$  is given by  $A = \pi r^2$ . This gives us that a circle with larger radius will always have a larger area than a circle with a smaller radius. Thus there exists a functional relationship between  $A$  and  $r$ . Now consider the two variables  $h$  and  $w$ , the height and weight of a given group of people. We know that in most of the cases the taller people will have higher weight as compared to persons of shorter height. But we also come across, at the same time, a short fat person who may have a higher weight as compared to a lean tall person. Thus we cannot determine the weight of person from his height though we could try to make reasonable guess about his weight, i.e., we cannot put the statement. "A tall person is more likely to be heavier in weight than a short person", as an exact rule to be true in all cases.

In this chapter we shall study the relationship of the type between height and weight or price and demand or wage and price index, etc. Such a relationship is called *statistical relationship*. Special methods have been developed to discover the existence or *statistical relationship* between the two variable from the bivariable data. When both variables in the bivariate data are quantitative, we use the term **Correlation Analysis** to describe the methods designed to find out if the statistical relationship between the two variables exists or not. When the variables are of such a character that they may be measured and the result expressed in quantitative units and paired measurements for the two variables are available for a group of individuals, it is possible under certain conditions to calculate a constant known as the **correlation coefficient**, which will express the degree of relationship. For example, suppose that the temperature and pulse had been taken at the same time on a group of five patients. For these patients the corresponding measurement were :

Patient	Temperature (F)		Pulse
	x	y	
A	102	100	
B	101	90	
C	100	80	
D	99	70	
E	98	60	

An a first step the data have been plotted on a scatter diagram (Fig. 8.1) in order to determine whether there is evidence of association or correlation. Examination of this figure shows that the points representing the five individual all fall on a straight line, and that this line has an upward or positive slope. That is to say that as the temperature rises, the pulse increase. In this instance the rise is uniform; for an increase of one degree in temperature the pulse increased 10 beats. How would the degree of this relationship be expressed statistically?

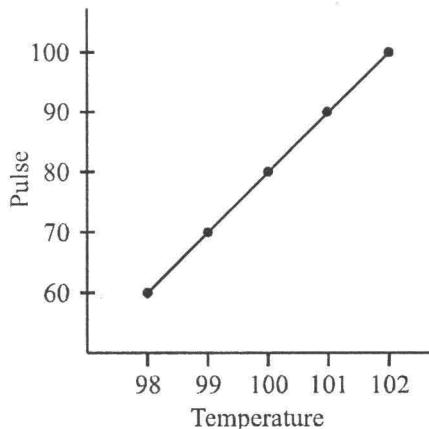


Fig. 8.1 : Relationship between pulse and temperature for five individuals.

We shall also study some methods of correlation and regression analysis when both the variables are quantitative. Croxton and Cowden defined the correlation as :

**"The relationship of quantitative nature. The appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation."**

## 8.2 COVARIANCE

Before we study the correlation analysis we introduce the concept of covariance between two quantitative variable  $X$  and  $Y$ . Let the corresponding values of the two variables  $X$  and  $Y$  on the given set of  $n$  units of observations be given by the ordered pairs.

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$$

Then the covariance between  $X$  and  $Y$  is denoted by Cov. ( $X, Y$ ). It is defined as

$$\text{Cov. } (X, Y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

$$= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \dots (1)$$

where  $\bar{x}$  and  $\bar{y}$  are the means of  $X$  and  $Y$  respectively.

i.e.,  $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}, \quad \bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$

### 8.3 CALCULATION OF COVARIANCE

The above formula for the calculation of covariance is complicated and may have more chances of the occurrence of error. We now give below which makes the calculations easier to carry out and which also reduces the chances of error.

The formula is

$$\text{Cov. } (X, Y) = \frac{1}{n} \left\{ \sum_{i=1}^n x_i y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right) \right\} \quad \dots (II)$$

$$\text{Cov. } (X, Y) = E(XY) - E(X) E(Y) \quad \dots (III)$$

where  $E(X)$ ,  $E(Y)$ ,  $E(XY)$  are the expectations of  $X$ ,  $Y$  and  $XY$  respectively.

### 8.4 CORRELATION ANALYSIS

#### 8.4.1 Correlation and Co-efficient of Correlation

**Correlation :** Correlation may be defined as a tendency towards interrelation variation and the **Coefficient of correlation** is a measure of such a tendency, i.e., the degree to which the two variables are interrelated is measured by a coefficient which is called the **Coefficient of correlation**. It gives the degree of correlation.

**Definition :** The relationship between two variables such that a change in one variable results in a positive or negative change in the other and also a greater change in one variable results in corresponding greater or smaller change in the other variable is known as **correlation**.

The coefficient of correlation between the two variables  $x$ ,  $y$  is generally denoted by  $r$  or  $r_{xy}$  or  $\rho(x, y)$  or  $\rho$ .

#### 8.4.2 Properties of Co-efficient of Correlation

1. It is a measure of the closeness of a fit in a relative sense.
2. Correlation coefficient lies between  $-1$  and  $+1$ , i.e.,  $-1 \leq r \leq 1$ .
3. The correlation is perfect and positive if  $r = 1$  and it is perfect and negative if  $r = -1$ .
4. If  $r = 0$ , then there is no correlation between the two variables and thus the variables are said to be independent.

### 8.5 CORRELATION COEFFICIENT CALCULATED FROM UNGROUPED DATA

A constant known as the correlation coefficient ( $r$ ) may be used to measure the degree of this relationship. This is calculated from the following formula:

$$r = \frac{\sum \left( \frac{x - \bar{x}}{\sigma_x} \right) \left( \frac{y - \bar{y}}{\sigma_y} \right)}{n}$$

where  $x$  and  $y$  represent measurements for the two variables taken at the same time  $\bar{x}$  and  $\bar{y}$  means of the two distributions of measurements and standard deviations of the measurements. For the series of five patients given below, if  $x$  represents the temperature and  $y$  the pulse for individual patients:

Patient	Temperature $x$	Pulse $y$	$\left( \frac{x - \bar{x}}{\sigma_x} \right)$	$\left( \frac{y - \bar{y}}{\sigma_y} \right)$	$\left( \frac{x - \bar{x}}{\sigma_x} \right) \left( \frac{y - \bar{y}}{\sigma_y} \right)$
A	102	100	$2/\sqrt{2}$	$20/\sqrt{200}$	$40/\sqrt{400}$
B	101	90	$1/\sqrt{2}$	$10/\sqrt{200}$	$10/\sqrt{400}$
C	100	80	$0/\sqrt{2}$	$0/\sqrt{200}$	$0/\sqrt{400}$
D	99	70	$-1/\sqrt{2}$	$-10/\sqrt{200}$	$10/\sqrt{400}$
E	98	60	$-2/\sqrt{2}$	$-20/\sqrt{200}$	$40/\sqrt{400}$
Total	500	400	0	0	$+100/\sqrt{400}$

$$\bar{x} = 500/5 = 100 \quad \sigma_x = \sqrt{2}$$

$$\bar{y} = 400/5 = 80 \quad \sigma_y = \sqrt{200}$$

$$r = \frac{+100 / \sqrt{400}}{5} = \frac{5}{5} = +1.$$

The correlation is, for +1. This indicates first that there is positive correlation; that is, the variables change in the same direction; second that the correlation is perfect.

Suppose the measurements on the five individuals had been:

Individual	Temperature	Pulse
A	102	60
B	101	70
C	100	80
D	99	90
E	98	100

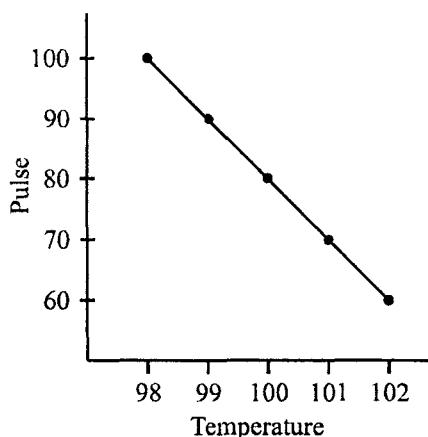


Fig. 8.2 : Relationship between pulse and temperature for five individuals.

Plotting these data on a scatter diagram gives the picture in Figure 8.2. Again all the points fall on a straight line, except that now the line has a downward or negative slope. This indicates an inverse relationship, that is, as one variable increases, the pulse rate decreases. The change again is uniform; for an increase of one degree of temperature the pulse decrease 10 beats.

What would be the correlation coefficient for these data? Using the same method the calculations are:

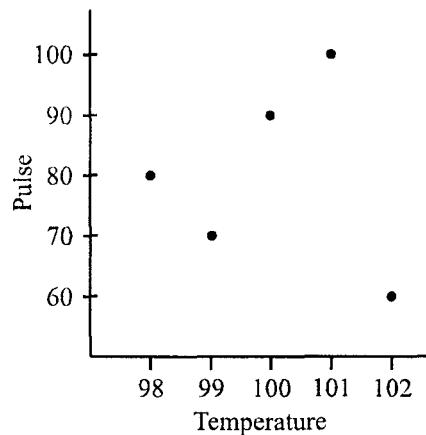
Patient	Temperature $x$	Pulse $y$	$\left( \frac{x - \bar{x}}{\sigma_x} \right)$	$\left( \frac{y - \bar{y}}{\sigma_y} \right)$	$\left( \frac{x - \bar{x}}{\sigma_x} \right) \left( \frac{y - \bar{y}}{\sigma_y} \right)$
A	102	60	$2\sqrt{2}$	$-20/\sqrt{200}$	$-40/\sqrt{400}$
B	101	70	$1/\sqrt{2}$	$-10/\sqrt{200}$	$-10/\sqrt{400}$
C	100	80	$0/\sqrt{2}$	$0/\sqrt{200}$	0
D	99	90	$-1/\sqrt{2}$	$10/\sqrt{200}$	$-10/\sqrt{400}$
E	98	100	$-2\sqrt{2}$	$20/\sqrt{200}$	$-40/\sqrt{400}$
Total	500	400	0	0	$+100/\sqrt{400}$

$$\bar{x} = 100 \quad \sigma_x = \sqrt{2}$$

$$\bar{y} = 80 \quad \sigma_y = \sqrt{200}. \quad \therefore r = \frac{-100/\sqrt{400}}{5} = \frac{-5}{5} = -1.$$

Note that the only difference here is that the product terms  $\left( \frac{x - \bar{x}}{\sigma_x} \right) \left( \frac{y - \bar{y}}{\sigma_y} \right)$  are all negative.

This indicates that for these data there is perfect negative correlation, shown by the fact that  $r = -1$ .



**Fig. 8.3 : Relation between pulse and temperature for five individuals.**

Now suppose the paired values had been :

Individual	Temperature	Pulse
A	102	60
B	101	100
C	100	90
D	99	70
E	98	80

These data are plotted in Figure 8.3. It can easily be seen that no straight line can be drawn on which all these points will fall. What will be the correlation coefficient expressing the degree on this relationship? Using the same method as before the calculations are:

Patient	Temperature $x$	Pulse $y$	$\left( \frac{x - \bar{x}}{\sigma_x} \right)$	$\left( \frac{y - \bar{y}}{\sigma_y} \right)$	$\left( \frac{x - \bar{x}}{\sigma_x} \right) \left( \frac{y - \bar{y}}{\sigma_y} \right)$
A	102	60	$2/\sqrt{2}$	$-20/\sqrt{200}$	$-40/\sqrt{400}$
B	101	100	$1/\sqrt{2}$	$+20/\sqrt{200}$	$+20/\sqrt{400}$
C	100	90	$0/\sqrt{2}$	$0/\sqrt{200}$	0
D	99	70	$-1/\sqrt{2}$	$-10/\sqrt{200}$	$+10/\sqrt{400}$
E	98	80	$-2/\sqrt{2}$	$0/\sqrt{200}$	0
Total	500	400	0	0	$-10/\sqrt{400}$

$$\bar{x} = 100 \quad \sigma_x = \sqrt{2}$$

$$\bar{y} = 80 \quad \sigma_y = \sqrt{200}$$

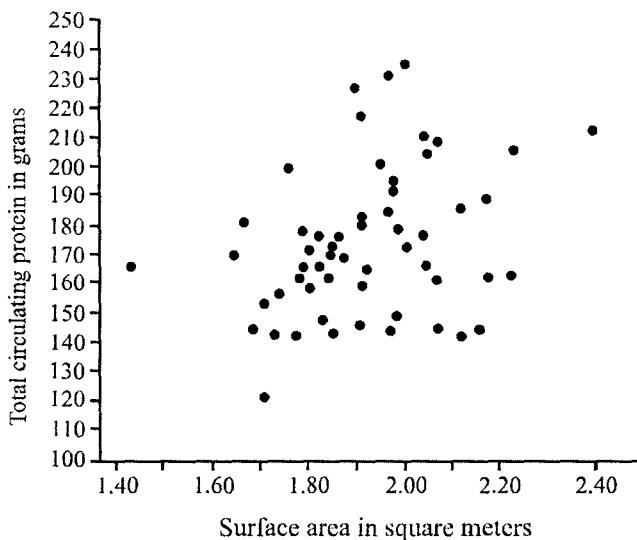
$$r = \frac{-10/\sqrt{400}}{5} = \frac{-10/20}{5} = -0.1.$$

This value of  $r$  indicates first, that what correlation is present in the sample is negative and second, that the correlation is of a very low degree. This substantiates what the diagram shows.

In general then, **correlation may be either positive or negative or zero**. Positive correlation indicates that the change in two variables is in the same direction, that is, as one increase the other increases; or if one decreases the other decreases. Negative correlation indicates that the variables change in opposite directions, that is, as one increases the other decreases. The value of the correlation coefficient must lie between 0 and +1 or 0 and -1. A coefficient of 0 indicates no correlation; a coefficient of 1 (either positive or negative) indicates perfect correlation. Obviously the closer the value of the coefficient is to one, the greater is the degree of intensity of association.

For sake of simplicity, the concept of correlation was presented from fictitious data and, as stated earlier, with a number of observations much too small to justify the determination of a correlation coefficient of association.

For an example from real data, consider the following information in regard to the relationship between surface area of the body and total circulating protein in the blood for a group of 62 normal males. The entire group of 62 pairs has been plotted in a scatter diagram in figure 8.4.



*Fig. 8.4 : Relationship between total circulating protein and surface area.*

Examination of this graph shows considerable scatter of the individuals but indicates that there is a relationship which is essentially linear, a necessary condition for the calculation of the correlation coefficient. A few representative pairs of values are given below:

<i>Individual No.</i>	<i>Surface area (sq. metre)</i> <i>x</i>	<i>Total circulating protein (gm)</i> <i>y</i>
1	1.75	170
2	2.18	182
3	1.74	160
4	1.77	174
:		
:		
60	2.00	188
61	2.12	199
62	1.67	144

The method used in the three previous examples could be used in this instance but the calculations would be very laborious.

The formula for the calculations of  $r$  originally given as:

$$r = \frac{\sum \left( \frac{x - \bar{x}}{s_x} \right) \left( \frac{y - \bar{y}}{s_y} \right)}{n}$$

can be written also in the following form which is easier to use for computation.

$$r = \frac{\frac{\sum xy}{n} - \frac{\sum x}{n} \frac{\sum y}{n}}{\sqrt{\frac{\sum x^2}{n} - \left( \frac{\sum x}{n} \right)^2} \sqrt{\frac{\sum y^2}{n} - \left( \frac{\sum y}{n} \right)^2}}$$

or simplifying,

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

For the above 62 observations the following sums have been computed:

$$\sum x = 117.76$$

$$\sum x^2 = 225.4290$$

$$\sum y = 11016$$

$$\sum y^2 = 2000748$$

$$\sum xy = 21050.38$$

$$n = 62$$

Substituting the appropriate values in the last formula gives :

$$r = \frac{62(21050.38) - 117.76 \times 11016}{\sqrt{62 \times 225.4290 - (117.76)^2} \sqrt{62 \times 2000748 - (11016)^2}}$$

$$r = +0.46$$

The above example shows that the computation is not simple.

We give the following direct method:

### 8.5.1 Direct Method

If  $X$  and  $Y$  are two variates having their means  $\bar{x}$  and  $\bar{y}$  respectively, then

$$\rho(X, Y) = \frac{\sum dxdy}{\sqrt{\sum dx^2 \times \sum dy^2}},$$

Where  $dx = x_i - \bar{x}$ ,  $dy = y_i - \bar{y}$ ,  $dx^2 = (x_i - \bar{x})^2$ ,  $dy^2 = (y_i - \bar{y})^2$ .

It can also be written as  $r_{xy} = \frac{\sum dxdy}{n\sigma_x \times \sigma_y}$

where  $n$  is the number of observations in  $X$  or  $Y$  series,  $\sigma_x$ ,  $\sigma_y$  are standard deviation of  $X$  and  $Y$  respectively.

The following formula can also be deduced from above:

$$\rho(X, Y) = \frac{\sum dxdy - \left( \frac{\sum dx \sum dy}{n} \right)}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{n}} \times \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{n}}}$$

where  $n$  is the number of observations.

### 8.5.2 Working Rule

The coefficient of correlation is calculated by the following steps :

- Step I.** Denote one series by  $x$  and the other series by  $y$ .
- Step II.** Calculate  $\bar{x}$  and  $\bar{y}$  of the  $x$  and  $y$  series respectively.
- Step III.** Take the deviations of the observations in  $x$  series from  $\bar{x}$  and write it under the column headed by  $dx = x - \bar{x}$ . Take the deviations of the observations in  $y$  series from  $\bar{y}$  and write it in a column headed by  $dy = y - \bar{y}$ .
- Step IV.** Square these deviations and write them under the columns headed by  $dx^2$  and  $dy^2$ .
- Step V.** Multiply the respective  $dx$  and  $dy$  and write it under the column headed by  $dxdy$ .
- Step VI.** Apply the following formula to calculate  $r$  or  $r_{xy}$ , the coefficient of correlation.

$$r = \frac{\sum dxdy}{\sqrt{\sum dx^2 \times \sum dy^2}}, \text{ or } r = \frac{\sum dxdy}{n\sigma_x \times \sigma_y},$$

where  $n$  is the number of observations in  $x$  or  $y$  series,  $\sigma_x$  and  $\sigma_y$  are the standard deviations of  $x$  and  $y$ . The above method is illustrated by the following example.

**Example 1 :** Find the coefficient of correlation between the heights of fathers and sons from the following data:

Heights of fathers (in inches) (x)	65	66	67	68	69	70	71
Heights of sons (in inches) (y)	67	68	66	69	72	72	69

**Solution :** Let the heights of fathers be denoted by  $x$  and that of sons be  $y$ , then

$$\bar{x} = \frac{65 + 66 + 67 + 68 + 69 + 70 + 71}{7} = 68.$$

$$\bar{y} = \frac{67 + 68 + 66 + 69 + 72 + 72 + 69}{7} = 69.$$

Let us prepare the following table:

$x$	$dx = (x - 68)$	$dx^2 = (x - 68)^2$	$y$	$dy = (y - 68)$	$dy^2 = (y - 68)^2$	$dxdy$
65	-3	9	67	-2	4	6
66	-2	4	68	-1	1	2
67	-1	1	66	-3	9	3
68	0	0	69	0	0	0
69	1	1	72	3	9	3
70	2	4	72	3	9	6
71	3	9	69	0	0	0
	$\Sigma dx = 0$	$\Sigma dx^2 = 28$		$\Sigma dy = 0$	$\Sigma dy^2 = 32$	$\Sigma dxdy = 20$

Now

$$r = \frac{\Sigma dxdy}{\sqrt{\Sigma dx^2 \times \Sigma dy^2}} = \frac{20}{\sqrt{(28 \times 32)}} = \frac{5}{7.5} = 0.67 \text{ (approx.)}.$$

**Example 2 :** Calculate the correlation coefficient between  $X$  and  $Y$  from the following data:

$X$	5	9	13	17	21
$Y$	12	20	25	33	35

**Solution :** Here  $\bar{X} = \frac{65}{5} = 13$ ,  $\bar{Y} = \frac{125}{5} = 25$ .

$X$	$dX = (X - 13)$	$dX^2 = (X - 13)^2$	$Y$	$dY = (Y - 25)$	$dY^2 = (Y - 25)^2$	$dXdY$
5	-8	64	12	-13	169	104
9	-4	16	20	-5	25	20
13	0	0	25	0	0	0
17	4	16	33	8	64	32
21	8	64	35	10	100	80
	$\Sigma dX = 0$	$\Sigma dX^2 = 168$		$\Sigma dY = 0$	$\Sigma dY^2 = 358$	$\Sigma dXdY = 236$

$$r = \frac{\sum dXdY}{\sqrt{\sum dX^2 \times \sum dY^2}} = \frac{236}{\sqrt{160 \times 358}} = \frac{236}{239.33} = 0.986.$$

### 8.5.3 Short-cut Method

The above direct method for calculating  $r$  is not convenient when (i) the terms of the series  $x$  and  $y$  are big and the calculation of  $\bar{x}$  and  $\bar{y}$  becomes difficult or (ii) the means  $\bar{x}$  and  $\bar{y}$  are not integers. In these cases we apply the following formula of assumed mean:

$$r_{xy} = \frac{\sum dxdy - \left( \frac{\sum dx \sum dy}{n} \right)}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{n}} \times \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{n}}}$$

where  $dx = x - a$ ,  $a$  is the assumed mean of  $x$  series,  $dy = y - b$ ,  $b$  is the assumed mean of  $y$  series and  $n$  is the number of observations of  $x$  or  $y$ .

#### Working Rule

- Step I.** Take any term  $a$  (preferably the middle one) of  $x$  series as assumed mean and any term  $b$  (preferable middle one) as assumed mean for  $y$  series.
- Step II.** Take deviations of the observations in  $x$  series from  $a$ , i.e.,  $dx = x - a$ . Take deviations of the observations in  $y$  series from  $b$ , i.e.,  $dy = y - b$  and write it under columns  $dx$  and  $dy$ .
- Step III.** Find  $dx^2$  and  $dy^2$  and write it under columns  $dx^2$  and  $dy^2$ .
- Step IV.** Find  $dxdy$  and write it under the column  $dxdy$ .
- Step V.** Apply the formula

$$r = \frac{\sum dxdy - \left( \frac{\sum dx \sum dy}{n} \right)}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{n}} \times \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{n}}}$$

The method is illustrated by the following example.

**Example 3 :** Calculate the coefficient of correlation between  $x$  and  $y$  for the following data:

$x :$	1	3	4	5	7	8	10
$y :$	2	6	8	10	14	16	20

**Solution :** Let 5 be the assumed mean for the values of  $x$  and 14 be assumed mean for the values of  $y$ .

Let  $dx = x - 5$ ,  $dx^2 = (x - 5)^2$ ,  $dy = (y - 14)$ ,  $dy^2 = (y - 14)^2$ . We have the following table :

$x$	$y$	$dx$	$dy$	$dx^2$	$dy^2$	$dxdy$
1	2	-4	-12	16	144	48
3	6	-2	-8	4	64	16
4	8	-1	-6	1	36	6
5	10	0	-4	0	16	0
7	14	2	0	4	0	0
8	16	3	2	9	4	6
10	20	5	6	25	36	30
$n = 7$	$n = 7$	$\Sigma dx = 3$	$\Sigma dy = 22$	$\Sigma dx^2 = 59$	$\Sigma dy^2 = 300$	$\Sigma dxdy = 106$

Since we have taken the deviations from assumed means, so we shall apply the following formula of correlation for assumed mean :

$$\begin{aligned}
 r_{xy} &= \frac{\Sigma dxdy - \left( \frac{\Sigma dx \Sigma dy}{n} \right)}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{n}} \times \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{n}}} \\
 &= \frac{106 + \frac{66}{7}}{\sqrt{\left( 59 - \frac{9}{7} \right)} \times \sqrt{\left( 300 - \frac{484}{7} \right)}} \\
 &= \frac{808}{\sqrt{404} \times \sqrt{1616}} = \frac{808}{\sqrt{652864}} = \frac{808}{808} = 1.
 \end{aligned}$$

Thus the coefficient of correlation between  $x$  and  $y$  is positive and perfect.

## 8.6 SPEARON'S RANK CORRELATION COEFFICIENT

The coefficient of rank correlation is based on the various values of the variates and is denoted by  $R$ . It is applied in the problems in which data cannot be measured quantitatively but qualitative assessment is possible such as beauty, honesty, etc. In this case, the best individual is given rank number 1, next rank 2 and so on. The coefficient of rank correlation is given by formula

$$R = 1 - \frac{6 \sum D^2}{n(n^2 - 1)},$$

where  $D^2$  is the square of the difference of corresponding ranks, and  $n$  is the number of pairs of observations.

### Working Rule

**Step I.** Assign ranks to each item of both the series.

- Step II.** Calculate the difference of ranks of  $x$  from the ranks of  $y$  and write it under the column headed by  $D$ .
- Step III.** Square the difference  $D$  and write it under the column  $D^2$ .
- Step IV.** Apply the formula

$$R = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

The above method is explain with the help of the following example.

**Example 4 :** Ten students got the following percentage of marks in mathematics and physics.

Mathematics ( $X$ )	8	36	98	25	75	82	92	62	65	35
Physics ( $Y$ )	84	51	91	60	68	62	86	58	35	49

Find the coefficient of rank correlation.

**Solution :**

$X$	$Y$	$x_i = \text{Rank in } X$	$y_i = \text{Rank in } Y$	$D = x_i - y_i$	$D^2$
8	84	10	3	7	49
36	51	7	8	-1	1
98	91	1	1	0	0
25	60	9	6	3	9
75	68	4	4	0	0
82	62	3	5	-2	4
92	86	2	2	0	0
62	58	6	7	-1	1
65	35	5	10	-5	25
35	49	8	9	-1	1
				$\Sigma D = 0$	$\Sigma D^2 = 90$

$$R = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = \frac{6 \times 90}{10(100 - 1)} = 1 - \frac{6}{11} = 1 - 0.545 = 0.455.$$

## 8.7 SCATTER OR DOT DIAGRAM — GRAPHICAL METHOD

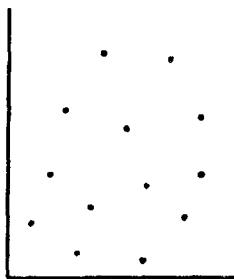
Scatter diagram is a graphical method of showing the correlation between the two variables. Let  $(x_i, y_i)$ ,  $i = 1, 2, 4, \dots, n$  be a bivariate distribution. Let the values of the variables  $x$  and  $y$  be plotted along the  $x$ -axis and  $y$ -axis in a coordinate plane by choosing a suitable scale, so that it measures the range of the data of both the variates (series) under consideration. Then corresponding to every ordered pair  $(x_i, y_i)$ , there corresponds a point or a dot in the coordinate plane.

The diagram of dots or points so obtained is called a **scatter diagram** or a **dot diagram**.

The scatter diagram may indicate both degree and the type of correlation.

In many problems, it is more important to know the exact mathematical relationship between two variables than to have a measure of it expressed in terms of the correlation coefficient. Also such measure can be determined only when the relationship between the two variables is linear. It is necessary, therefore, to know the pattern of the relationship.

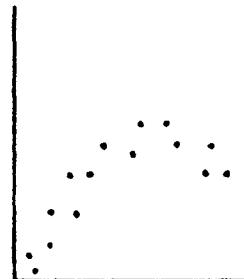
As a first step in determining the pattern of the relationship between two quantitatively measured variables, data are plotted on a scatter diagram. Examination of the arrangement of the dots representing the individual pairs of values gives certain information as to the possible relationship. The dots may or may not form a distinct pattern. If, as in Figure 8.5, they scatter all over the diagram, this is evidence that little relationship between the two variables is present. The dots may, however form a distinct pattern, which may be of a linear nature as in Figure 8.6, or they may form a pattern which is in the form of curve as in Figure 8.7. The relationship represented in Figure 8.6 is linear in nature; that of Figure 8.7 is curvilinear. If the red blood count and the white blood count are determined the same time for a group of normal individuals and plotted in a scatter diagram, the resulting graph will in all probability resemble that of Figure 8.5.



*Fig. 8.5 : No relationship*



*Fig. 8.6 : Linear relationship*



*Fig. 8.7 : Curvilinear relationship*

Nor is there reason to expect any relationship between blood sugar and the weight of an individual. However, if the body temperature and pulse rate of normal persons are compared, linear relationship is usually found. This is also true of surface area of the body and blood plasma volume. The urinary nitrogen of pregnant women goes down in a linear relationship as the stage of pregnancy increases. Growth curves by contrast represent a curvilinear relationship; for example, growth of the embryo and length of pregnancy and growth of the individual represented either by height or weight when compared with age or growth of bacteria with time. In the first examples growth precedes rapidly at first and then gradually shows up with duration of time. In the latter example bacteria go through a lag phase immediately after inoculation that is followed by a period of rapid growth, then a gradual slowing up, ending by a decrease in growth. All these patterns are curvilinear in nature.

## EXERCISES

1. The following results were obtained for calculating the coefficient of correlation between the two variables  $x$  and  $y$  from 25 pairs of observations:

$$\Sigma x = 125, \Sigma x^2 = 50, \Sigma y = 100, \Sigma y^2 = 460, \Sigma xy = 50.$$

Calculate the coefficient of correlation between  $x$  and  $y$ .

2. Calculate Karl Pearson's coefficient between the marks in Geology and Botany obtained by 10 students:

Marks in Geology:	20	35	15	40	10	35	30	25	45	30
Marks in botany :	25	30	20	35	20	25	25	35	35	30

3. Calculate the correlation coefficient between the price and consumption for the following data:

Price :	5	5.50	6	6.50	7	7.50	8
Consumption :	10	10	8	7	7	6	6

4. Calculate Karl Pearson's correlation coefficient between  $x$  and  $y$  for the following data:

$x :$	43	54	59	68	76
$y :$	105	98	84	63	50

5. Calculate the coefficient of correlation of ranks obtained by 10 students of a class in Hindi and English:

Hindi :	1	2	3	4	5	6	7	8	9	10
English :	3	8	1	7	10	2	9	4	6	5

[Hint :  $D = -2 \quad -6 \quad 2 \quad -3 \quad -5 \quad 4 \quad -2 \quad 4 \quad 3 \quad 5$

$$\Sigma D^2 = 4 + 36 + 4 + 9 + 25 + 16 + 4 + 16 + 9 + 25 = 148$$

$$R = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6 \times 148}{10 \times 99} = 1 - 0.897 = 0.103$$

6. Calculate the coefficient of correlation between  $x$  and  $y$  for the following data:

$x :$	1	2	3	4	5	6	7	8	9
$y :$	12	11	13	15	14	17	16	19	10

7. Calculate the rank correlation coefficient for the following data:

$x :$	92	89	87	86	83	77	71	63	53	50
$y :$	86	83	91	77	68	85	52	82	37	57

8. Calculate the correlation coefficient between the heights of fathers and sons from the given data:

Heights of fathers (in inches) :	64	65	66	67	68	69	70
Heights of sons (in inches) :	66	67	65	68	70	68	72

## ANSWERS

1.  $r = 0.2$       2.  $r = 0.76$       3.  $r = 0.95$       4.  $r = 0.41$   
 5.  $r = 0.103$       6.  $r = 0.933$       7.  $r = 0.73$       8.  $r = 0.81$



## 9

# Regression Analysis

## 9.1 REGRESSION ANALYSIS

**Regression :** We know that the correlation studies the relationship between two variables  $x$  and  $y$ . In this section we shall consider the related problem of **prediction or estimation** of the value of variable from a known value of other variable to which it is related. This would be discussed by means of **regression lines**.

*Regression* means to return or to go back. So it implies the act of returning or going back. Now the question arises what returns and where it returns. Natural phenomena generally have a tendency to return to normal. In statistics the term '**regression**' is used to denote backward tendency which means going back to average or normal.

The term regression was first used by **Sir Francis Galton** in study of heredity. He found that though "tall fathers have all sons", the average height of sons of tall fathers is  $x$  above the general height; the average height of their sons is  $(2/3) x$  above the general height. The recession in the average height was described by Galton as **regression to Mediocrity**.

Though Sir Galton used the term 'regression' in studying heredity the concept of regression is now extended to other spheres of phenomena which have a tendency to regress or set back to the normal or the general average. The concept of regression is very helpful in the study of correlation.

*Regression shows a relationship between the average values of two variables. Thus regression is very helpful in estimating and predicting the average value of one variable for a given value of the other variable. The estimate or prediction may be made with the help of regression line which shows the average value of one variable  $x$  for a given value of the other variable  $y$ .* The best average value of one variable associated with the given value of the other variable may also be estimated or predicted by means of an equation and the equation is known as a **regression equation**.

In linear regression the relationship between the two variables  $x$  and  $y$  is linear. In order to estimate the best average values of the two variables, two regression equations are required

and they are used separately. One equation is used for estimating the value of  $X$  variable for a given value of  $Y$  variable and the second equation is used for estimating the value of  $Y$  variable for a given value of  $X$  variable. In both cases the assumption is that one is an independent variable and the other is a dependent variable and *vice versa*.

The two lines of regression are

**1. Regression Equation of  $X$  on  $Y$  is**

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

[It estimates  $X$  for a given value of  $Y$ ]

**2. Regression Equation of  $Y$  on  $X$  is**

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

[It estimates  $Y$  for a given value of  $X$ ]

where  $X$  = Value of  $X$ ,  $Y$  = Value of  $Y$ ,

$\bar{X}$  = Arithmetic Mean of  $X$  series,

$\bar{Y}$  = Arithmetic Mean of  $Y$  series,

$\sigma_x$  = Standard Deviation of  $X$  series,

$\sigma_y$  = Standard Deviation of  $Y$  series,

$r$  = Correlation Coefficient between  $X$  and  $Y$ .

It is important to note that the regression equation of  $X$  on  $Y$  should be used for predicting or timing the value of  $X$  for a given value of  $Y$  and the regression equation of  $Y$  on  $X$  should be used for predicting or estimating the value of  $Y$  for a given value of  $X$ .

## 9.2 REGRESSION COEFFICIENTS

The regression coefficient of  $y$  on  $x$  is  $b_{yx} = \frac{r\sigma_y}{\sigma_x}$  and that of  $x$  on  $y$  is :  $b_{xy} = \frac{r\sigma_x}{\sigma_y}$ .

## 9.3 PROPERTIES OF REGRESSION COEFFICIENTS

(a) The coefficients of correlation is the geometric mean of the coefficients of regression i.e.,

$$r = \sqrt{b_{yx} \times b_{xy}}.$$

(b) If one of the regression coefficients is greater than unity, then the other is less than unity.

(c) Arithmetic mean of the regression coefficient is greater than the correlation coefficient.

(d) Regression coefficient are independent of change of origin but not of scale.

## 9.4 STANDARD ERROR OF ESTIMATE OR PREDICTION

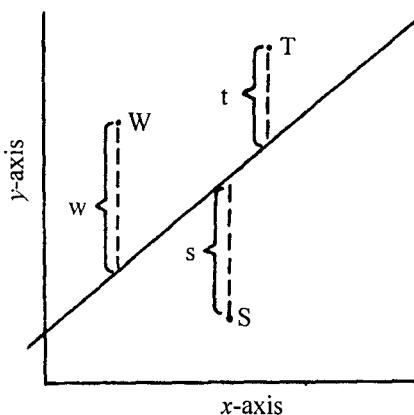
The standard error of estimate of  $Y$  in the line of regression of  $Y$  on  $X$  is given by

$$\Sigma x^2 = \sigma_x^2 (1 - r^2).$$

## 9.5 LINEAR REGRESSION LINE OR EQUATION

A method, which express in the form of a mathematical equation the relationship between the two variables  $x$  and  $y$ , is relatively simple when the **relationship is linear in character**. This equation is known as the **regression equation**. Methods are also available for determining the regression equation when the **relationship is not linear**, but are beyond the scope of this presentation. An understanding of the method used in determining the linear regression equation will aid the student in understanding curvilinear relationship, however, when they are encountered by the student.

Since in most instances of linear regression the dots representing the individual pairs of values for the two variables will not all fall in such a way that every dot will fit on a straight line (perfect correlation), the problem resolves itself into determining the equation of the straight line that best fits the dots; or in other words, the line around which the dots have the least scatter. This line is called a *regression line* and its equation can be found by a method known as **method of least squares**. This involves finding the line for which the sum of the squares of the distances of each observed point from the line shall be as small as possible. Consider Figure 9.1 where for simplicity only three points,  $W$ ,  $S$  and  $T$  have been plotted. If the distance from point  $W$  to the line be measured by  $w$ ; that from  $S$  by  $s$ , and that from  $T$  by  $t$ , the problem is to find the equation of the line which will make  $w^2 + s^2 + t^2$  a minimum. Let the equation of this line be, in general terms,  $y = a + bx$ , where  $y$  and  $x$  represent the two variables, ' $a$ ' is the  $y$ -intercept or the distance between the  $x$ -axis and the point where the line crosses the  $y$ -axis and ' $b$ ' is the slope or the increase in the  $y$  value per unit change in the  $x$  value (Fig. 9.2).



*Fig. 9.1 : Illustration of deviations from the line in the  $y$  direction.*

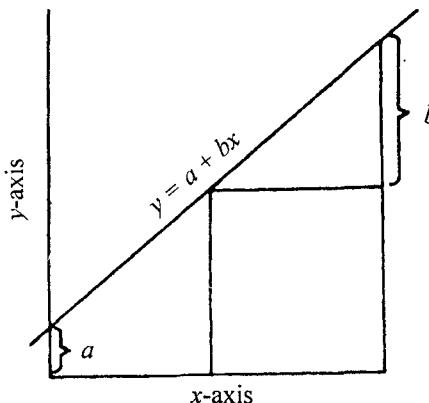
Using mathematical principles developed in calculus, the value of  $w^2 + s^2 + t^2$  will be a minimum if the value of  $a$  and  $b$  are such that they satisfy two simultaneous linear equations known as the **normal equations**. These are :

1.  $\Sigma y = na + b\Sigma x$
2.  $\Sigma xy = a\Sigma x + b\Sigma x^2$

Solving the two equations by the ordinary methods of algebra gives the value of  $b$ . In general terms :

$$b_{yx} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2}$$

After the value of  $b_{yx}$  is found, this value together with the values of  $\Sigma y$ ,  $n$ ,  $\Sigma x$  can be substituted in equation (1) and the value of  $a$  calculated. The accuracy of the computations can be checked by substituting the values of  $a$  and  $b$  in equation (2).



*Fig. 9.2 : Illustration of intercept and slope for the regression line.*

For the data previously used showing the correlation between surface area of the body (designated as the  $x$  variable) and circulating protein of the blood (designated as the  $y$  variable), the following sums were given as:

$$\Sigma x = 117.76$$

$$\Sigma x^2 = 225.4290$$

$$\Sigma y = 11016$$

$$\Sigma y^2 = 2000748$$

$$\Sigma xy = 21050.38$$

$$n = 62$$

To determine the equation of the regression line that minimizes the squares of the  $y$  deviations (that is, the difference between the observed  $y$  for any given value of  $x$  and the  $y$  value calculated from the equation for the same value of  $x$ ) these values may be substituted in the above formula for  $b$  to get:

$$b = \frac{62(21050.38) - (11016)(117.76)}{62(225.4290) - (117.76)^2} = \frac{7879.40}{109.1804} = 72.1686.$$

Substituting this value for  $b$  together with the other necessary values in equation (1) gives:

$$11016 = 62a + (72.1686)(117.76)$$

Solving,       $11016 = 62a + 8498.5743$

$$62a = 2517.4257$$

$$a = 40.6036.$$

Checking by substituting the necessary values in equation (2) gives :

$$21050.38 = (40.6036) + (72.1686)(225.4290)$$

$$21050.38 = 21050.38$$

These values of  $a$  and  $b$  can now be substituted in the equation of the line  $y = a + bx$ , giving for the regression line  $y$  on  $x$  as:

$$y = 40.604 + 72.169x.$$

To plot the regression line on the scatter diagram, it is necessary to find two paired values of  $x$  and  $y$ . First, one value of  $x$  such as 1.6 is substituted for  $x$  in the equation  $y = 40.604 + 72.169x$ , and the corresponding value of  $y$  is obtained. Then a second value of  $x$  such as 2.0 is substituted for  $x$  and the value of  $y$  obtained. These two paired values for  $x$  and  $y$ , i.e., (1.6, 156.1) and (2.0, 184.9) are now plotted on the diagram and the line connecting them is the regression line.

**What do the values of  $a$  and  $b$  indicate?** The value of  $b$  indicates for each unit change in surface area of the body, the circulating protein increases (because  $b$  is positive) by 72.169 gm. The  $a$  value tells simply what the value of  $y$  is when  $x$  is 0, or tells where the regression line crosses the  $y$ -axis.

A second alternate formula for calculation of  $b_{yx}$  is especially useful when dealing with grouped data. This is

$$b_{yx} = r \frac{\sigma_y}{\sigma_x}.$$

The subscript  $yx$  is used to show that the value of  $b_{yx}$  being calculated is that which is to be substituted in the line  $y = a + bx$ . The value of  $a$  can then be calculated, as before, by substituting the appropriate figure in the equation:

$$\Sigma y = na + b\Sigma x$$

This equation can also be written as:

$$\frac{\Sigma y}{n} = a + b \frac{\Sigma x}{n}$$

or

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

Therefore knowing  $b$ ,  $\bar{x}$ , and  $\bar{y}$ , it becomes easy to find the value of  $a$ . Use of this formula gives the same line as we got by the first method.

This last formula emphasizes that fact that the regression line will always pass through the mean of the  $x$  values and the mean of the  $y$  values. This is a further check on the accuracy of computations.

This line has been plotted on the scatter diagram (Fig 9.3). Inspection shows that there is considerable scatter of the dots about the line. This is to be expected because of the low correlation coefficient previously found for these data.

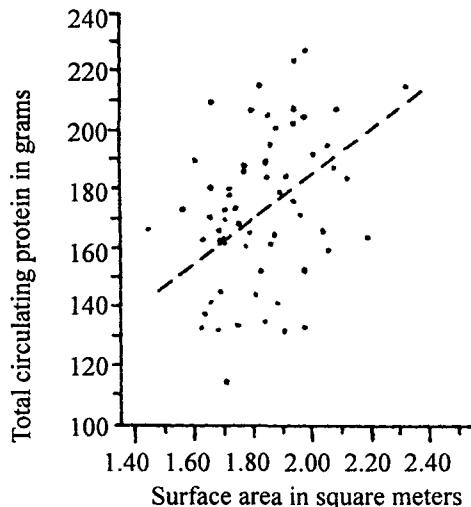
A measure of the variation of these points around the line can be determined. This measure is comparable to the standard deviation of a distribution; the only difference is that this measures the variation around the line rather than around the mean. This measure, known as  $\sigma_{yx}$  is equal

to :  $\sigma_{yx} = \sqrt{\frac{\sum (y - Y)^2}{n}}$ , where  $y$  represents the observed value for a given  $x$  and  $Y$  represent the

value of  $y$  calculated by substituting the same value of  $x$  in the equation of the regression line. In other words, it is the square root of the average of the squares of the  $y$  deviations from the regression line. It can be shown algebraically that :

$$\sqrt{\frac{\sum (y - \bar{Y})^2}{n}} = \sigma_y \sqrt{1 - r^2}$$

when  $n$ , or the number of observations is large, the value of  $\sigma_{yx}$  may be calculated more easily from the alternate formula. If the first formula is used the value of  $y$  must be calculated for every value of  $x$ ,  $n$  differences must be calculated, squared and summed. All the added calculations necessary for the second formula are the summation of the squares of the  $y$  values, and the determination of  $\sigma_y$  and  $r$  from the appropriate formulas.



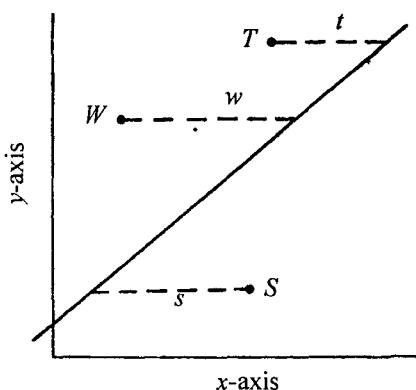
**Fig. 9.3 : Relationship between total circulating protein and surface area with regression line,**  
 $y = 40.604 + 72.169x$ .

For these particular data, the alternate formula is much simpler. Substituting in the alternate formula,

$$\sigma_{yx} = 26.47 \sqrt{1 - (.46)^2} = 23.48 \text{ gm.}$$

This **standard error of estimate**, as it is often called, is a **measure of the residual variation about the line**. It measures the agreement between the  $y$  values observed and those predicated from the observed value of  $x$ . If there is correlation between the two variables,  $\sigma_{yx}$ , will always be smaller than  $\sigma_y$ , the standard deviation of the  $y$  values from the mean of all the  $y$  observations. In this particular example the variation around the line is not much less than the variation of the  $y$  values from  $\bar{y}$  since the correlation coefficient is small.

This standard error of estimate is also useful in determining confidence intervals for the predication of the  $y$  value for any given value of  $x$ . For details of this method a good text book at the more advanced level should be consulted.



*Fig. 9.4 : Illustration of deviations from the line in the x direction*

It would be possible to minimize the  $x$  deviations as well as the  $y$  deviations. The distance of the points  $W$ ,  $S$  and  $T$  which were shown as distances in the  $y$  direction in Figure 9.1, have been measured now as distances from the line in the  $x$  directions (Figure 9.4). To do this involves the same method as before except that the  $x$  and  $y$  values are interchanged. The equation of the line is :  $x = A + By$ . The normal equations which must be solved to determine  $A$  and  $B$  are:

$$\begin{aligned}\Sigma x &= nA + B \Sigma y \\ \Sigma xy &= A \Sigma y + B \Sigma y^2.\end{aligned}$$

In this equation  $B$  is the increase in the average value of  $x$  per unit increase in  $y$  and  $A$  is the value of  $x$  when  $y$  is zero, or it tells where the regression line crosses the  $x$ -axis.

The variation around this line is measured as before except that the differences are those between the  $x$  values for a given  $y$  value and the calculated from the regression line for the same  $y$  value. The formula written in the two forms used before is :

$$\left( \sigma_{xy} = \sqrt{\frac{\sum (x - X)^2}{n}} \text{ or } \sigma_x \sqrt{1 - r^2} \right)$$

Fundamentally, what is the difference in these two regression lines and under what circumstances will there be only one line? ***The first line ( $y = a + bx$ ) is known as the regression line of  $y$  on  $x$ . That is, it is the line that would be used for the prediction of value of  $y$ , from given values of  $x$ .*** It was calculated in such a way that the sum of the squares of the differences between the observed  $y$  and the predicted  $Y$  (determined by substitution in the equation of the regression line) would be the smallest possible.

***The second line ( $x = A + By$ ) is known as the regression line of  $x$  on  $y$ . It is the line that would be used for the prediction of  $x$  values for given  $y$  values.*** It was calculated in such a way that the sum of the squares of the differences between the observed  $x$  value and the predicted  $x$  should be minimum.

When there is perfect correlation, the two lines will be identical. The higher the degree of relationship or the higher the value of  $r$ , the smaller will be the residual variation. **The two lines**

will always intersect at the point that represents the mean of the  $x$  values and the mean of the  $y$  values, since both must pass through the point  $\bar{x}$ ,  $\bar{y}$ .

When the correlation is not perfect, the problem often arises as to which of the regression line shall be calculated. In many problems, one of the variables may be relatively easy to determine, while the other may be more difficult. For example, in considering the relation of the specific gravity of the blood to the concentration of the blood, specific gravity is more easily calculated than in protein concentration. If it could be shown that the correlation between these two determinations was sufficiently high so that the amount of protein in the blood could be estimated by determining the specific gravity, considerable laboratory work could be saved. In such a case, if specific gravity were represented  $x$  and blood protein by  $y$ , the regression line to be determined would be  $y = a + bx$ ; what is the line so calculated that the  $y$  deviations are minimized. When only one line is to be fitted it is customary, although not necessary, to designate the variable to be estimated (sometimes called the dependent variable) as the  $y$  variable and the variable from which the other is to be estimated as the  $x$  variable (called the independent variable).

If estimations of both  $y$  from  $x$  and  $x$  from  $y$  are to be made, which is very uncommon, both regression lines *must* be calculated. To be useful in estimations, the correlation between the two variables must be high and the variability about the line must be low. It should also be stated that in most instances, prediction of values beyond the range of those used in determining the equation of the regression line should not be made. Regression lines are used also in certain problems on calibration in which only one variable is measured at fixed points on a scale of time, temperature, dilution etc. Such a problem would be a measurement of optical density for solutions of varying but predefined dilution. Here only the regression of optical density on dilution has any meaning.

**Example 1 :** From the data given below estimate the most likely weight of a father whose son's height is 65".

Father's : Mean height is 67" and a s.d. of 3.5".

Son's : Mean height is 65" with s.d. of 2.5".

The coefficient of correlation between the heights of fathers and sons is +0.8.

**Solution :** Let  $x$ ,  $y$  be the variables corresponding to the heights of sons and fathers.

∴ The  $\bar{x} = 65$ ,  $\bar{y} = 67$ ,  $\sigma_x = 2.5$ ,  $\sigma_y = 3.5$ ,  $r_{xy} = 0.8$ .

$$\text{Now } b_{yx} = \frac{r\sigma_y}{\sigma_x} = \frac{(0.8)(3.5)}{2.5} = 1.12.$$

∴ The equation of line of regression of  $y$  on  $x$  is :

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

$$\Rightarrow y - 67 = 1.12 (x - 65)$$

$$\Rightarrow y = 1.12x - 5.8.$$

.... (1)

Height of a father whose son's height is 70" is = Estimate of  $y$  for 70.

Putting  $x = 70$  in (1), we get

$$y = (1.12) \times 70 - 5.8 = 78.4 - 5.8 = 72.6".$$

**Example 2 :** Given the following results of the height and weight of 1,000 students:

$\bar{Y} = 68$  inches,  $\bar{X} = 150$  lbs.,  $r = 0.60$ ,  $\sigma_y = 2.50$  inches,  $\sigma_x = 10$  lbs. Amit weighs 100 lbs. Sumeet is 5 feet tall. Estimate the height of Amit from his weight and the weight of Sumeet from his height.

**Solution :** Here

$$\text{Height} = Y$$

$$\text{Weight} = X$$

$$\bar{Y} = 68 \text{ inches}$$

$$\bar{X} = 150 \text{ lbs}$$

$$r = 0.60$$

$$\sigma_x = 2.50 \text{ inches}$$

$$\sigma_y = 10 \text{ lbs}$$

- (a) The regression equation of  $Y$  on  $X$  is :  $Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$

$$\text{or } Y - 68 = 0.60 \times \frac{2.50}{10} (X - 150)$$

$$\Rightarrow Y - 68 = 0.15 \times (-15)$$

$$\Rightarrow Y = 0.15X - 15 + 68$$

$$\therefore Y = 0.15X + 53$$

When Amit's weight  $X = 100$  lbs, his height

$$Y = 0.15 \times 100 + 53 = 15 + 53 = 68 \text{ inches.}$$

$\therefore$  Required height of Amit = 68 inches.

- (b) The regression equation of  $X$  on  $Y$  is :  $X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$

$$\text{or } X - 150 = 0.60 \times \frac{10}{2.5} (Y - 68)$$

$$\text{or } X - 150 = 2.4 (Y - 68)$$

$$\text{or } X - 150 = 2.4Y - 163.2$$

$$\text{or } X = 2.4Y - 163.2 + 150$$

$$\therefore X = 2.4Y - 13.2.$$

When Sumeet's height  $Y = 5$  feet = 60 inches

his weight  $X = 2.4 \times 60 - 13.2 = 144 - 13.2 = 130.8$  lbs.

$\therefore$  Required weight of Sumeet = 130.8 lbs.

**Example 3 :** Find the regression of  $x$  on  $y$  from the following data:

$$\Sigma x = 24, \quad \Sigma y = 44, \quad \Sigma xy = 306, \quad \Sigma x^2 = 164, \quad \Sigma y^2 = 574, \quad N = 4$$

Find the value of  $x$ , when  $y = 6$ .

**Solution :** Here  $\bar{x} = \frac{\Sigma x}{N} = \frac{24}{4} = 6$  ;  $\bar{y} = \frac{\Sigma y}{N} = \frac{44}{4} = 11$ .

$$\text{Regression coefficient : } b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{N \Sigma xy - \Sigma x \Sigma y}{N \Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{4 \times 306 - 24 \times 44}{4 \times 574 - (44)^2} = \frac{1222 - 1056}{2296 - 1936} = \frac{168}{360} = 0.47.$$

The regression equation of  $x$  on  $y$  is :  $x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$

$$\text{or } x - 6 = 0.47(y - 11) \quad \text{or } x - 6 = 0.47y - 5.17$$

$$\text{or } x = 0.47y - 5.17 + 6 \quad \therefore \quad x = 0.47y + 0.83$$

$$\text{When } y = 6, \quad x = 0.47 \times 6 + 0.83 = 2.82 + 0.83 = 3.65$$

**∴ Required value of  $x = 3.65$ .**

## EXERCISES

1. Find the lines of regression  $X$  on  $Y$  and  $Y$  on  $X$  for the following data :

$X:$	3	5	6	6	9
$Y:$	2	3	4	6	5

2. For 10 observations on price ( $x$ ) and supply ( $y$ ), the following data were obtained (in appropriate units):

$$\Sigma x = 130, \quad \Sigma y = 220, \quad \Sigma x^2 = 2238, \quad \Sigma y^2 = 5506, \quad \Sigma xy = 3467.$$

Obtain the line of regression of  $y$  on  $x$  and estimate the supply when the price is 16 units.

[Hint : The lines  $y$  on  $x$  is  $y = a + bx$ , where  $a$  and  $b$  given by the normal equation  $220 = 10a + 130b, 2367 = 130a + 2288b$ .

$$\Rightarrow \quad a = 8.8 \quad \text{and} \quad b = 1.015$$

$$\text{The line is} \quad y = 8.8 + 1.015x$$

$$\text{Put} \quad x = 16 \text{ to get } y = 25.04].$$

3. The following data give the correlation coefficient, means and standard deviation of rainfall and yield of paddy in a certain tract:

*Yield per acre in lbs.*      *Annual rainfall*

Mean	973.5	18.3
S.d.	38.4	2.0

$$\text{Coefficient of correlation} = 0.58$$

Estimate the most likely yield of paddy when the annual rainfall is 22", other factors being assumed to remain the same.

[Hint : Regression of  $y$  on  $x$  is :  $y - 973.5 = 11.136(x - 18.3)$

$$\text{Put} \quad x = 22 \text{ to estimate } y = 1014.7]$$

4. Find the correlation coefficient and the equation of the regression line of the following values of  $x$  and  $y$ :

$$\{(x, y)\} = \{(1, 2), (2, 5), (3, 3), (5, 8), (5, 7)\}.$$

## ANSWERS

1.  $Y = 0.6X + 0.4$ ,  $X = 1.2Y + 1.20$ .      2.  $Y = 8.8 + 1.05X$ ; 25.04.  
3. 1014.7.      4.  $Y = 1.3X + 1.1$ ,  $X = 0.5Y + 0.5$ ,  $r = 0.8$ .



# 10

# Probability and Baye's Theorem

## 10.1 INTRODUCTION

Probability theory was originated from gambling theory. A large number of problems exist even today which are based on the game of chance, such as coin tossing, dice throwing and playing cards. The utility of probability in business and economics is most emphatically revealed in the field of predictions for future. We have to anticipate consequences of the each of these plans and finally we compare the results. *Uncertainty plays an important role in business and probability is a concept which measures the degree of uncertainty and that of certainty also as a corollary. The probability when defined in the simplest way is chance of occurrence of a certain event when expressed quantitatively.* The probability is defined in two different ways:

- (i) Mathematical (or *a priori*) definition.
- (ii) Statistical (or *empirical*) definition.

Before we study the probability theory in details, it is appropriate to give the definition of certain terms, which are essential for the study of *probability theory*.

## 10.2 SOME IMPORTANT TERMS AND CONCEPTS

### 10.2.1 Random Experiment or Trial

An experiment is characterised by the property, that its observations under a given set of circumstances do not always lead to the same observed outcome but rather to different outcomes which follows a sort of statistical regularity. It is also called a **Trial**.

For example, tossing a coin, or throwing a dice.

### 10.2.2 Sample Space

A set of all possible outcomes from an experiment is called a **sample space**. Let us toss a coin the result is either head or tail. Let 1 denote head and 0 denote tail. Mark the point 0, 1 on a straight line. These points are called **sample points or event points**. For a given experiment there are different possible outcomes and hence different sample points. The collection of all

such sample points is a **sample space**. Toss two coins simultaneously, then the possible results are four pairs  $(0, 0)$ ,  $(1, 0)$ ,  $(0, 1)$ ,  $(1, 1)$  and they can be represented as points in a coordinate plane. These points constitute a sample space of four sample points. Similarly, by tossing three coins simultaneously, we get eight points which can be denoted by the triplets  $(0, 0, 0)$ ,  $(0, 1, 0)$ ,  $(0, 0, 1)$ ,  $(1, 0, 0)$ ,  $(1, 1, 0)$ ,  $(0, 1, 1)$ ,  $(1, 0, 1)$ ,  $(1, 1, 1)$ . These 8 points form a sample space of 3-dimension. In general, *in tossing  $n$  coins simultaneously we can have an  $n$ -dimensional sample space consisting of  $2^n$  sample points.*

### 10.2.3 Discrete Sample Space

*A sample space whose elements are finite or infinite but countable is called a discrete sample space.*

**For example**, if we toss a coin as many times as we require for turning up one head, then the sequence of points  $S_1 = (1)$ ,  $S_2 = (0, 1)$ ,  $S_3 = (1, 0, 0)$ ,  $S_4 = (0, 0, 0, 1)$  etc., is a discrete sample space.

### 10.2.4 Continuous Sample Space

*A sample space whose elements are infinite and uncountable or assume all the values on a real line  $R$  or on an interval of  $R$  is called a continuous sample space.* In this case the sample points build up a continuum, and the **sample space** is said to be **continuous**.

**For example**, all the points on a line or all points on a plane is a sample space.

### 10.2.5 Event

*A sub-collection of a number of sample points under a definite rule or law is called an event.*

**For example**, let us take a dice. Let its faces 1, 2, 3, 4, 5, 6 be represented by  $E_1, E_2, E_3, E_4, E_5, E_6$  respectively. Then all the  $E_i$ 's are sample points. Let  $E$  be the event of getting an even number on the dice. Obviously,  $E = \{E_2, E_4, E_6\}$ , which is a **subset** of the set  $\{E_1, E_2, E_3, E_4, E_5, E_6\}$ .

### 10.2.6 Null Event

*An event having no sample point is called a null event and is denoted by  $\emptyset$ .*

### 10.2.7 Simple Event

*An event consisting of only one sample point of a sample space is called a simple event.*

**For example**, let a dice be rolled once and  $A$  be the event that face number 5 is turned up, then  $A$  is a **simple event**.

### 10.2.8 Compound Events

*When an event is decomposable into a number of simple events, then it is called a compound event.*

**For example**, the sum of the two numbers shown by the upper faces of the two dice is seven in the simultaneous throw of the two unbiased dice, is a **compound event** as it can be decomposable.

### 10.2.9 Exhaustive Cases or Events

*It is the total number of all the possible outcomes of an experiment.*

**For example,** when we throw a dice, then any one of the six faces (1, 2, 3, 4, 5, 6) may turn up and therefore, there are six possible outcomes. Hence there are six **exhaustive cases or events** in throwing a dice.

### 10.2.10 Mutually Exclusive Events

*If in an experiment the occurrence of an event precludes or prevents or rules out the happening of all other events in the same experiment, then these events are said to be **mutually exclusive events**.*

**For example,** in tossing a coin, the events head and tail are **mutually exclusive**, because if the outcome is head, then the possibility of getting a tail in the same trial is ruled out.

### 10.2.11 Equally Likely Events

*Events are said to be **equally likely** if there is no reason to expect any one in preference to other.*

**For example,** in throwing a dice, all the six faces (1, 2, 3, 4, 5, 6) are equally likely to occur.

### 10.2.12 Collectively Exhaustive Events

*If total number of events in a population exhausts the population. So they are known as **collectively exhaustive events**.*

### 10.2.13 Equally Probable Events

*If in an experience all possible outcomes have equal chances of occurrence, then such events are said to be **equally probable events**.*

**For example,** in throwing a coin, the events head and tail have equal chances of occurrence, therefore, they are **equally probable events**.

### 10.2.14 Favourable Cases

*The cases which ensure the occurrence of an event are said to be **favourable to the event**.*

### 10.2.15 Independent and Dependent Events

*When the experiments are conducted in such a way the occurrence of an event in one trial does not have any effect on the occurrence of this or other events at a subsequent experiment, then the events are said to be **independent**. In other words, two or more events are said to be **independent** if the happening of any one does not depend on the happening of the other. Events which are not independent are called **dependent events**.*

**Illustration 1.** If we draw a card in a pack of well shuffled cards and again draw a card from the rest of pack of cards (containing 51 cards), then the second draw is **dependent** on the first. But if on the other hand, we draw a second card from the pack by replacing the first card drawn, the second draw is known as **independent** of the first.

## 10.3 DEFINITIONS OF PROBABILITY

### 10.3.1 Classical Definition of Probability

If an experiment has  $n$  mutually exclusive, equally likely and exhaustive cases, out of which  $m$  are favourable to the happening of the event  $A$ , then the probability of the happening of  $A$  is denoted by  $P(A)$  and is defined as:

$$P(A) = \frac{m}{n} = \frac{\text{No. of cases favourable to } A}{\text{Total (Exhaustive) number of cases}}$$

**Notes :**

1. *Probability of an event which is certain to occur is 1 and the probability of an impossible event is zero.*
2. *The probability of occurrence of any event lies between 0 and 1, both inclusive.*

**Example 1 :** What is the probability of getting an even number in a single throw with a dice?

**Solution :** The possible cases in the throw of a dice are six, viz., 1, 2, 3, 4, 5, 6.

Favourable cases are those which are marked with 2, 4, 6 and these are three in number.

$$\therefore \text{Probability of getting an even number} = \frac{3}{6} = \frac{1}{2}.$$

**Example 2 :** What is the probability of getting tail in a throw of a coin?

**Solution :** When we toss a coin, there are two possible outcomes, viz., Head or Tail. In this case the number of possible cases  $n = 2$ .

$$\text{No. of favourable cases} = m = 1 \quad (\because \text{The outcome of tail is a favourable event})$$

$$\therefore \text{Probability of getting an even number} = \frac{m}{n} = \frac{1}{2}.$$

**Example 3 :** A bag contains 6 white balls, 9 black ball. What is the probability of drawing a black ball?

**Solution :** The total number of equally likely and exhaustive cases  $= n = 6 + 9 = 15$ .

$$\text{No. of favourable cases} = m = 9 \quad (\because \text{Number of black balls} = 9)$$

$$\therefore \text{Probability of drawing a black ball} = \frac{9}{15} = \frac{3}{5}.$$

**Example 4 :** What is the probability that if a card is drawn at random from an ordinary pack of cards, it is (i) a red card, (ii) a club, (iii) one of the court cards (Jack or Queen or King).

**Solution :** No. of exhaustive cases  $= 52$ .

(i) There are 26 red cards and 26 black cards in an ordinary pack.

$$\therefore \text{Favourable cases} = n = 26 \quad (\text{number of red cards})$$

$$\therefore \text{Probability of getting a red card} = \frac{26}{52} = \frac{1}{2}.$$

(ii) Number of clubs in a pack  $= 13$ .

$$\therefore \text{Favourable cases} = 13.$$

$$\therefore \text{Probability of getting a club} = \frac{13}{52} = \frac{1}{4}.$$

(iii) There are  $4 \times 3 = 12$  court cards in a pack of cards.

$$\therefore \text{Number of favourable cases} = m = 12.$$

$$\text{Number of exhaustive cases} = n = 52.$$

$$\therefore \text{Probability of getting a court card} = \frac{12}{52} = \frac{3}{13}.$$

**Example 5 :** What is the probability of throwing a number greater than 3 with an ordinary dice?

**Solution :** The six faces of a dice are marked with numbers 1, 2, 3, 4, 5, 6 and thus there are only 6 exhaustive cases, i.e.,  $n = 6$ .

There are three numbers 4, 5, 6 which are greater than 3 in a dice.

$$\therefore \text{Favourable case} = m = 3.$$

$$\therefore \text{Probability of getting a number greater than 3} = \frac{3}{6} = \frac{1}{2}.$$

**Example 6 :** What is the probability that a leap year, selected at random, will have 53 Sundays?

**Solution :** There are 366 days in a leap year and it has 52 weeks and 2 days over. These two extra days can occur in following possible ways.

- |                               |                               |
|-------------------------------|-------------------------------|
| (i) Sunday and Monday ;       | (ii) Monday and Tuesday ;     |
| (iii) Tuesday and Wednesday ; | (iv) Wednesday and Thursday ; |
| (v) Thursday and Friday ;     | (vi) Friday and Saturday ;    |
| (vii) Saturday and Sunday.    |                               |

$$\therefore \text{No. of exhaustive cases} = 7.$$

$$\text{No. of favourable cases} = 2$$

[ $\because$  There are two cases which have Sunday (i) and (vii)]

$$\therefore \text{Probability} = \frac{2}{7}.$$

**Example 7 :** What is the probability of getting a total of more than 10 in a single throw with two dice?

**Solution :** The number of exhaustive cases when two dice are thrown simultaneously =  $6^2 = 36$ .

Out of these 36 cases, there will be only three favourable cases. These are (5, 6), (6, 5), (6, 6).

$$\therefore \text{Probability} = \frac{3}{36} = \frac{1}{12}.$$

**Example 8 :** A card is drawn from an ordinary pack of playing cards and a person bets that it is a spade or an ace. What are the odds against his winning this bet?

**Solution :** In a pack of 52 cards, 1 card can be drawn in 52 ways. Since there are 13 spades and 3 aces (one ace is also present in spade), therefore, the favourable cases =  $m = 13 + 3 = 16$ .

No. of exhaustive cases =  $n = 52$ .

$$\text{Probability of getting a spade or an ace} = \frac{16}{52} = \frac{4}{13} = \frac{4}{9+4}.$$

$\therefore$  Odds against winning the bet are 9 to 4.

### 10.3.2 Statistical or Empirical Definition of Probability

Von Mises has given the following statistical or empirical definition of probability.

"If the experiment be repeated a large number of times under essentially identical conditions, the limiting value of the ratio of the number of times the event A happens to the total number of trials of the experiment as the number of trials increases indefinitely is called the probability of happening of the event A".

**Symbolically :** Let  $P(A)$  denote the probability of the occurrence of  $A$ . Let  $m$  be the number of times in which an event  $A$  occurs in a series of  $n$  trials, then

$$P(A) = \lim_{n \rightarrow \infty} \frac{m}{n}, \text{ provided the limit is finite and unique.}$$

## 10.4 THEOREMS ON PROBABILITY

There are two important theorems of probability, namely,

1. **Addition Theorem or Theorem on Total Probability.**
2. **Multiplication Theorem or Theorem on Compound Probability.**

### 10.4.1 Addition Theorem or Theorem on Total Probability

**Statement :** If  $n$  events are mutually exclusive, then the probability of happening of any one of them is equal to the sum of the probabilities of the happening of the separate events, i.e., in other words, if,  $E_1, E_2, E_3, \dots, E_n$  be  $n$  events and  $P(E_1), P(E_2), \dots, P(E_n)$ , be their respective probabilities, then,

$$P(E_1 + E_2 + E_3 + \dots + E_n) = P(E_1) + P(E_2) + \dots + P(E_n)$$

**Proof :** Let  $E_1, E_2, E_3, \dots, E_n$  be  $n$  mutually exclusive events with probabilities  $P(E_1), P(E_2), \dots, P(E_n)$  respectively. Let  $N$  be the total number of trials. Let  $m_1, m_2, m_3, \dots, m_n$  be the cases favourable to events  $E_1, E_2, E_3, \dots, E_n$  respectively, then,

$$P(E_i) = \frac{m_i}{N}, \quad i = 1, 2, \dots, n$$

Since the events  $E_1, E_2, \dots, E_n$  are mutually exclusive, therefore, the cases favourable to the happening of any one of the events  $E_1, E_2, \dots, E_n$  are  $m_1 + m_2 + \dots + m_n$ . Hence the probability of happening of any one of the events  $E_1, E_2, \dots, E_n$  is

$$\begin{aligned} P(E_1 + E_2 + \dots + E_n) &= \frac{m_1 + m_2 + \dots + m_n}{N} = \frac{m_1}{N} + \frac{m_2}{N} + \frac{m_3}{N} + \dots + \frac{m_n}{N} \\ &= P(E_1) + P(E_2) + P(E_3) + \dots + P(E_n). \end{aligned}$$

**Example 9 :** A dice is rolled. What is the probability that a number 1 or 6 may appear on the upper face?

**Solution :** The probability of appearing the number 1 on the upper face =  $\frac{1}{6}$ .

The probability of appearing 6 on the upper face =  $\frac{1}{6}$ .

$$\therefore \text{The probability of 1 or 6 appearing on the face} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

**Example 10 :** If the probability of the horse A winning the race is  $\frac{1}{5}$  and the probability of the horse B winning the same race is  $\frac{1}{6}$ , what is the probability that one of the horses will win the race?

**Solution :** Probability of winning of the horse A =  $\frac{1}{5}$ .

Probability of winning of the horse B =  $\frac{1}{6}$ .

$$\therefore P(A + B) = P(A) + P(B) = \frac{1}{5} + \frac{1}{6} = \frac{11}{30}.$$

#### 10.4.2 Multiplicative Theorem or Theorem on Compound Probability

Before we proceed further in stating and proving this theorem, we must learn the following definitions.

**Simple and compound events.** A single event is called a **simple event**. When two or more than simple events occur in connection with each other, then their simultaneous occurrence is called a **compound event**. If A and B are two **simple events**, the simultaneous occurrence of A and B is called a **compound event and is denoted by AB**.

**Conditional probability.** The probability of the happening of an event B, when it is known that A already happened, is called the **conditional probability of B** and is denoted by  $P(B/A)$ .

i.e.,  $P(B/A) \Rightarrow$  conditional probability of B given that A has already occurred.

Similarly,  $P(A/B) \Rightarrow$  conditional probability of A given that B has already happened.

**Mutually Independent Events :** An event A is said to be independent of the event B if  $P(A/B) = P(A)$ , i.e., the probability of the happening of A is independent of the happening of B.

#### THEOREM ON COMPOUND PROBABILITY

**Statement :** The probability of the simultaneous occurrence of the two events A and B is equal to the probability of one of the events multiplied by the conditional probability of other given the occurrence of the first, i.e.,

$$P(AB) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B).$$

**Proof :** Let an experiment have  $N$  mutually exclusive, exhaustive and equally likely cases out of which  $M$  are favourable to the happening of the event  $A$ . Let the cases which are favourable to the simultaneous occurrence of the events  $A$  and  $B$  be  $m$ . Obviously,  $m$  cases are included in  $M$  cases which are favourable to  $A$ . Now,

$$P(AB) = \frac{m}{N} = \frac{m}{M} \cdot \frac{M}{N}. \quad \dots (1)$$

We know that the cases favourable to the event  $B$  have to come out of  $M$  cases favourable to the event  $A$ , therefore, the exhaustive number of cases for  $B$  is  $M$ . Thus the ratio  $\frac{m}{M}$  represents the happening of  $B$  given that  $A$  has already occurred. Also  $\frac{M}{N} = P(A)$ , so from (1), we have

$$P(AB) = \frac{m}{N} \times \frac{m}{M} = P(A) \cdot P(B/A).$$

**Cor. 1.** If the events  $A$  and  $B$  are statistically independent then  $P(B/A) = P(B)$  and  $P(A/B) = P(A)$ .

$$\therefore P(A)B = P(A) \quad P(B/A) = P(A) = P(B) \quad [\because (P(B/A) = P(B))]$$

**Example 11 :** If five coins are tossed, what is the probability that all will show a head?

**Solution :** Let  $P(A)$  denote the probability of showing a head.

$$\therefore P(A) = \frac{1}{2}$$

Now probability of showing a head when 5 coins are tossed

$$= P(A) \times P(A) \times P(A) \times P(A) \times P(A) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{32}.$$

**Example 12 :** A card is drawn from a pack of 52 cards and then a second card is drawn. What is the probability that both the cards drawn are queen?

**Solution :** First draw : Probability of getting a queen =  $\frac{4}{52} = \frac{1}{13}$ .

Second draw : After drawing the first queen, we are left with 51 cards having 3 queens.

$$\therefore \text{Probability of getting a queen in second draw} = \frac{3}{51} = \frac{1}{17}.$$

$$\therefore \text{Probability that both the cards are queen} = \frac{1}{13} \times \frac{1}{17} = \frac{1}{221}.$$

**Example 13 :** A bag contains 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both the balls drawn are black.

**Solution :** Let  $P(A)$ ,  $P(B/A)$  denote the probability of drawing a black ball in the first and second attempt respectively.

$$\therefore \text{Probability of drawing a black ball in the first attempt is } P(A),$$

where

$$P(A) = \frac{\text{Favourable cases}}{\text{Total exhaustive cases}} = \frac{3}{5+3} = \frac{3}{8}.$$

Probability of drawing the second black ball given the first ball drawn is black is

$$P(B/A) = \frac{\text{Favourable cases}}{\text{Total number of exhaustive cases}} = \frac{2}{5+2} = \frac{2}{7}.$$

$$\begin{aligned}\text{Probability that the both balls drawn are black is } &= P(AB) = P(A) P(B/A) \\ &= \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}.\end{aligned}$$

## 10.5 COMPLIMENTARY EVENTS

The event '**A occurs**' and the event '**A does not occur**' are called **complementary events**.

The "event **A does not occur**" is denoted by  $A^C$  or  $\bar{A}$  and is read as **complementary of A**. It is important to note that

$$\begin{aligned}P(A) + P(A^C) &= 1 \\ \therefore P(A^C) &= 1 - P(A).\end{aligned}$$

**Proof :** Let there be  $n$  exhaustive, mutually exclusive outcomes of an experiment. Let  $m$  of these outcomes be favourable to the **happening of the event A**. Thus the event '**A does not occur**' in **remaining ( $n - m$ ) outcomes**. Thus these  $(n - m)$  outcomes are favourable to the event "**A does not occur**" i.e., to the event  $A^C$ .

$$\therefore P(A^C) = \frac{n-m}{n} = \frac{n}{m} - \frac{m}{n} = 1 - \frac{m}{n} = 1 - P(A) \quad \left[ \because P(A) = \frac{m}{n} \right]$$

$$\text{Thus } P(A^C) = 1 - P(A) \text{ or } P(A) + P(A^C) = 1.$$

**Example 14 :** Find the chance of throwing at least one up in single throw with three dices.

**Solution :** When we throw a dice, it results in six equally likely, mutually exclusive and exhaustive cases. Therefore the total number of outcomes of throwing three dices is  $6 \times 6 \times 6 = 216$  (By fundamental principle).

Let **A** be the event that **at least 1 appears on a dice**. Then the complementary event  $A^C$  is **1 does not appear on the dice**.

There are 5 outcomes on each dice when 1 does not appear, i.e., occurrence of 2 or 4 or 4 or 5 or 6. Hence, the number of cases favourable to  $A^C = 5 \times 5 \times 5 = 125$ .

$$\therefore P(A^C) = \frac{125}{216}$$

$$\text{Now } P(A) = 1 - P(A^C) = 1 - \frac{125}{216} = \frac{91}{216}.$$

## 10.6 HAPPENING OF AT LEAST ONE EVENT

Let **A** and **B** be the independent events and  $p_1$  and  $p_2$  be the probabilities of their happening, then the chance that **both A and B happen** is  $p_1 \times p_2$ . Also the probabilities of not happening of **A** and **B** are  $1 - p_1$ ,  $1 - p_2$  respectively.

**The probability that both do not happen is  $(1 - p_1) \times (1 - p_2)$ .**

Now the chance that at least one of them happens:

$$p_1 p_2 + p_1 (1 - p_2) + p_2 (1 - p_1) = 1 - (1 - p_1) (1 - p_2)$$

Thus the above result can be generalised for  $n$  events  $E_1, E_2, \dots, E_n$  with  $p_1, p_2, \dots, p_n$  respective probabilities.

**Then the probability that at least one of them happens is**

$$= 1 - (1 - p_1) (1 - p_2) \dots (1 - p_n).$$

**Example 15 :** 4 coins are tossed. Find the probability that at least one head turns up.

**Solution :** Probability of getting a head in a toss of a coin = 1/2.

Probability of getting a tail in each case = 1/2.

$$\text{Probability of getting a tail in all the four cases} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}.$$

$$\therefore \text{Probability of getting at least one head} = 1 - \frac{1}{16} = \frac{15}{16}.$$

**Example 16 :** A problem in mathematics is given to three students Dayanand, Ramesh and Naresh whose chances of solving it are  $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$  respectively. What is the probability that the problem will be solved?

**Solution :** The probabilities of Dayanand, Ramesh and Naresh solving the problem are  $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$  respectively.

$\therefore$  The probabilities of all the three, i.e., Dayanand, Ramesh and Naresh not solving the problem =  $\frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} = \frac{1}{4}$ .

$$\text{The probability that the problem will be solved by at least one of them} = 1 - \frac{1}{4} = \frac{3}{4}.$$

**Example 17 :** A card is drawn at random from a well shuffled pack of 52 cards. Find the probability of getting a two of heart or a diamond.

**Solution :** Since there is only one two of hearts and only one two of diamonds, therefore, the probability of getting a two of hearts =  $\frac{1}{52}$  and the probability of getting a two of diamonds =  $\frac{1}{52}$ .

Since these are mutually exclusive cases, therefore, the probability of getting a two of heart or a two of diamond =  $\frac{1}{52} + \frac{1}{52} = \frac{2}{52} = \frac{1}{26}$ .

**Example 18 :** A man and his wife appear for an interview for two posts. The probability of the husband's selection is  $\frac{1}{7}$  and that of the wife's selection is  $\frac{1}{5}$ . What is the probability that only one of them will be selected?

**Solution :** The probability that husband is not selected =  $1 - \frac{1}{7} = \frac{6}{7}$ .

The probability that wife is not selected =  $1 - \frac{1}{5} = \frac{4}{5}$ .

Probability that only husband is selected =  $\frac{1}{7} \times \frac{4}{5} = \frac{4}{35}$ .

Probability that only wife is selected =  $\frac{1}{5} \times \frac{6}{7} = \frac{6}{35}$ .

Probability that only one of them is selected =  $\frac{4}{35} + \frac{6}{35} = \frac{10}{35} = \frac{2}{7}$ .

**Example 19 :** A salesman has a 60% chance of making a sale to each customer. The behaviour of successive customers is independent. If two customers A and B enter, what is the probability that the salesman will make a sale to A or B.

**Solution :** Probability making of sale to a customer =  $\frac{60}{100} = 0.6$ .

Probability of making no sale to a customer =  $1 - 0.6 = 0.4$ .

Probability of making no sale to A =  $P(A^C) = 0.4$ .

Probability of making no sale to B =  $P(B^C) = 0.4$ .

Probability of making no sale to A and no sale to B =  $P(A^C) \times P(B^C) = 0.4 \times 0.4 = 0.16$ .

Probability of making a sale to A or B =  $1 - P(A^C) \times P(B^C)$   
 $= 1 - 0.16 = 0.84 = 84\%$ .

## 10.7 ADDITION THEOREM FOR COMPATIBLE EVENTS

**Theorem :** The probability of the occurrence of at least one of the events A and B (not mutually exclusive) is given by

$$P(A + B) = P(A) + P(B) - P(AB).$$

**Proof :** If A and B are two events, then the event A + B means the occurrence of at least one of the two events A and B. It is also written as (A or B). It is the union of the following mutually exclusive events.

(i)  $A\bar{B}$  which means that A happens but B does not happen.

(ii)  $\bar{A}B$  which means B happens but A does not happen.

(iii)  $AB$  which means both A and B occur. AB is also written as A or B.

$$\therefore A + B = A\bar{B} + \bar{A}B + AB.$$

Applying the theorem of addition of probabilities, we have

$$P(A + B) = P(A\bar{B}) + P(\bar{A}B) + (AB) \quad \dots (i)$$

But,

$$A = A\bar{B} + AB$$

∴

$$P(A) = P(A\bar{B}) + P(AB)$$

or

$$P(A\bar{B}) = P(A) - P(AB) \quad \dots (ii)$$

Similarly,

$$P(B) = P(\bar{A}B) + P(AB)$$

or

$$P(\bar{A}B) = P(B) - P(AB) \quad \dots (iii)$$

Now (i), (ii) and (iii) imply

$$\begin{aligned} P(A + B) &= P(A) - P(AB) + P(B) - P(AB) + P(AB) \\ &= P(A) + P(B) - P(AB) \end{aligned}$$

Hence,

$$P(A + B) = P(A) + P(B) - P(AB)$$

It is also written as  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ .

**Example 20 :** Two students X and Y work independently on a problem. The probability that X will solve it is  $(3/4)$  and the probability that Y will solve it is  $(2/3)$ . What is the probability that the problem will be solved.

**Solution :** Let  $P(X)$  and  $P(Y)$  denote the probability that X and Y respectively will solve the problem.

$$\therefore P(X) = \frac{3}{4}, \quad P(Y) = \frac{2}{3}.$$

The events X and Y are mutually exclusive as both of them may solve the problem.

$$\therefore P(X \text{ and } Y) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}.$$

∴ The probability of solving the problem is:

$$\begin{aligned} P(X \text{ or } Y) &= P(X) + P(Y) - P(X \text{ and } Y) \\ &= \frac{3}{4} + \frac{2}{3} - \frac{1}{2} = \frac{9 + 8 - 6}{12} = \frac{11}{12}. \end{aligned}$$

**Second Method :**

$$P(X^C) = 1 - \frac{3}{4} = \frac{1}{4};$$

$$P(Y^C) = 1 - \frac{2}{3} = \frac{1}{3};$$

$$P(X^C \text{ and } Y^C) = \frac{1}{4} \times \frac{1}{3} = \frac{1}{12}.$$

$$\therefore \text{Probability that problem will be solved} = 1 - P(X^C \text{ and } Y^C) = 1 - \frac{1}{12} = \frac{11}{12}.$$

**Example 21 :** The probability that a student passes a Physics test is  $(2/3)$  and the probability that he passes both physics and English test is  $(14/45)$ . The probability that he passes at least one test is  $(4/5)$ . What is the probability that the student passes the English test.

**Solution :** Let  $A$  denote that the student passes Physics test :

$$\therefore P(A) = \frac{2}{3}.$$

Let  $B$  denote that the student passes English test : then  $P(B) = ?$

It is given that the probability that the student passes both Physics and English test is  $\frac{4}{5}$ .

$$\therefore P(A \text{ and } B) = \frac{14}{45}$$

Also the probability that he passes at least one test is  $\frac{4}{5}$ :

$$\therefore P(A \text{ or } B) = \frac{4}{5}$$

Now using the addition theorem

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

We get

$$\frac{4}{5} = \frac{2}{3} + P(B) - \frac{14}{45}$$

or  $P(B) = \frac{4}{5} + \frac{14}{45} - \frac{2}{3} = \frac{36 + 14 - 30}{45} = \frac{20}{45} = \frac{4}{9}.$

Hence the probability that the passes English test is  $\frac{4}{9}$ .

## 10.8 DEFINITION OF PROBABILITY IN TERMS OF ODD IN FAVOUR OR ODDS AGAINST THE EVENT

**Odds in favour of an event.** If the odds in favour of an event  $A$  are in the ratio  $x : y$  (or  $x$  to  $y$ ), then

$$P(A) = \frac{x}{x+y} \quad \text{and} \quad P(A^C) = \frac{y}{x+y} \quad \dots (1)$$

It is important to note that  $P(A^C) = 1 - P(A) = 1 - \frac{x}{x+y} = \frac{y}{x+y}$ , which is same as given

by (1).

**Odds against an event.** If the odds against an event  $B$  are  $m : n$ , (or  $m$  to  $n$ ), then

$$P(B) = \frac{n}{m+n} \quad \text{and} \quad P(B^C) = \frac{m}{m+n} \quad \dots (2)$$

It is important to note that

$$P(B^C) = 1 - P(B) = 1 - \frac{n}{m+n} = \frac{m+n-n}{m+n} = \frac{m}{m+n},$$

which is the same as given by (2).

This concept is illustrated by the following examples.

**Example 22 :** Find the probability of the event A:

- (a) if the odds in favour are 3 : 2.      (b) if the odd against it are 1 : 4.

**Solution :**

$$(a) \text{ Required } P(A) = \frac{3}{3+2} = \frac{3}{5}$$

$$(b) \text{ Required } P(A) = \frac{4}{1+4} = \frac{4}{5}.$$

**Example 23 :** A problem in statistics is given to two students A ad B. The odds in favour of A solving the problem are 6 to 9 and against B solving the problem are 12 to 10. If A and B attempt, find the probability of the problem being solved.

**Solution :**

$$\text{Probability of } A \text{ solving the problem } P(A) = \frac{6}{6+9} = \frac{6}{15} \quad (\text{odds in favour})$$

$$\text{Probability of } B \text{ solving the problem } P(B) = \frac{10}{12+10} = \frac{5}{11} \quad (\text{odds against})$$

$$P(A^C) = 1 - P(A) = 1 - \frac{6}{15} = \frac{9}{15}$$

$$P(B^C) = 1 - P(B) = 1 - \frac{5}{11} = \frac{6}{11}$$

Probability that both of them will fail to solve the problem

$$P(A^C \text{ and } B^C) = P(A^C) \times P(B^C) = \frac{9}{15} \times \frac{6}{11} = \frac{18}{55}$$

$$\therefore \text{Probability that the problem will be solved} = 1 - P(A^C \text{ and } B^C) \\ = 1 - \frac{18}{55} = \frac{37}{55}.$$

**Second Method :**

$$P(A \text{ and } B) = \text{Probability that both will solve it} = \frac{6}{15} \times \frac{5}{11}.$$

$$\therefore P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \\ = \frac{6}{15} + \frac{5}{11} - \frac{6}{15} \times \frac{5}{11} = \frac{66 + 75 - 30}{165} = \frac{141 - 30}{165} = \frac{111}{165} = \frac{37}{55}.$$

**Example 24 :** The odds against a certain event are 5 to 3 and the odds in favour of another event, independent of the former are 7 to 5. Find the chance that at least one of the events will happen.

**Solution :** Let the two events be  $A$  and  $B$ .

$$\text{The probability of the first event } A \text{ not taking place } P(A^C) = \frac{5}{5+3} = \frac{5}{8}.$$

$$\text{The probability of second event not taking place } P(B^C) = \frac{5}{7+5} = \frac{5}{12}$$

The probability that neither of the two events takes place is :

$$P(A^C \text{ and } B^C) = P(A^C) \times P(B^C) = \frac{5}{8} \times \frac{5}{12} = \frac{25}{96}.$$

$$\text{Probability of happening of at least one of the events} = 1 - P(A^C \text{ and } B^C)$$

$$= 1 - \frac{25}{96} = \frac{71}{96}.$$

**Example 25 :** Suppose that it is 11 to 5 against to a person who is now 38 years of age living till he is 73 and 5 to 3 against  $B$  now 43 living till he is 78 years. Find the chance that at least one of these persons will be alive 35 years hence.

**Solution :** The probability that  $A$  will die within 35 year.

$$= \frac{11}{11+5} = \frac{11}{16}, \quad i.e., \quad P(A^C) = \frac{11}{16}$$

The probability that  $B$  will die within 35 years

$$= \frac{5}{5+3} = \frac{5}{8}, \quad i.e., \quad P(B^C) = \frac{5}{8}$$

$$\text{The probability that both will die within 35 years} = \frac{11}{6} \times \frac{5}{8} = \frac{55}{128}$$

$$i.e., \quad P(A^C \text{ and } B^C) = \frac{55}{128}$$

∴ Probability that both of them will not die, i.e., or at least one of them will be alive

$$= 1 - P(A^C \text{ and } B^C) = 1 - \frac{55}{128} = \frac{73}{128} \text{ or } \frac{73}{128} \times 100 = 57\%.$$

**Example 26 :** The chances of living a person who is now 35 years old till he is 75 are 8 : 6 and of living another person now 40 years old till he is 80 are 4 : 5. Find the probability that at least one of these persons would die before completing 40 years hence.

**Solution ;** Probability that the first person lives till he is 75 years =  $\frac{8}{14}$ .

Probability that the second person lives till he is 80 years =  $\frac{4}{9}$ .

Probability of the compound event that both the persons live 40 years

$$= \frac{8}{14} \times \frac{4}{9} = \frac{32}{126} = \frac{16}{63}$$

Probability that at least one of them would die without living 40 years  $= 1 - \frac{16}{63} = \frac{47}{63}$ .

## 10.9 APPLICATION OF PERMUTATION AND COMBINATION

**Example 27 :** From a pack of 52 cards, two are drawn at random. Find the chance that one is a king and the other is a queen.

**Solution :** The total number of ways in which two cards can be drawn out of 52 cards is  ${}^{52}C_2$ .

Since in a pack of cards there are 4 kings and 4 queens, therefore, the number of favourable cases are  ${}^4C_1 \times {}^4C_1$ .

$$\text{Thus required probability is } = \frac{{}^4C_1 \times {}^4C_1}{{}^{52}C_2} = \frac{4 \times 4}{\frac{52 \times 51}{2}} = \frac{4 \times 4 \times 2}{52 \times 51} = \frac{8}{663}.$$

**Example 28 :** There are two bags. One bag contains 4 white and 2 black balls. Second bag contains 5 white and 4 black balls. Two balls are transferred from first bag to second bag. Then one ball is taken from the second bag. Find the probability that it is a white ball.

**Solution :** There are 3 mutually exclusive and exhaustive ways in which 2 balls can be transferred from first bag to second bag.

**First case :** Two white balls have been transferred from first bag to the second bag so that

$$\text{probability for that is } \frac{{}^4C_2}{{}^6C_2} = \frac{2}{5}.$$

In the second bag we have 7 white and 4 black balls and the probability of getting a white ball is  $\frac{7}{11}$ .

$$\therefore \text{Required probability} = \frac{{}^2C_2}{{}^6C_2} \times \frac{7}{11} = \frac{2}{5} \times \frac{7}{11} = \frac{14}{55}.$$

**Second case :** Two black balls have been transferred from first bag to the second bag so that probability for that is  $\frac{{}^2C_2}{{}^4C_2} = \frac{1}{15}$ .

In the second bag we have five white and 6 black balls and probability of getting a white ball is  $\frac{5}{11}$ .

$$\therefore \text{Required probability} = \frac{1}{15} \times \frac{5}{11} = \frac{1}{33}.$$

**Third case :** One black and one white ball have been transferred from first bag to the second so that the probability for this is  $\frac{{}^4C_1 \times {}^2C_1}{{}^4C_1} = \frac{8}{15}$ .

In the second bag we have 6 white and 5 black and the probability of drawing a white ball is  $\frac{6}{11}$ .

$$\therefore \text{Required probability} = \frac{8}{15} \times \frac{6}{11} = \frac{16}{55}.$$

Since these three cases are mutually exclusive, therefore, the required probability of drawing a white ball

$$= \frac{14}{55} + \frac{1}{33} + \frac{16}{55} = \frac{42 + 5 + 48}{165} = \frac{95}{165} = \frac{19}{33}.$$

**Example 29 :** Two cards are drawn from a well shuffled pack of playing cards. Determine the probability that both are aces.

**Solution :** The total number of ways in which 2 cards can be drawn out of 52 is

$${}^{52}C_2 = \frac{52!}{2! \times 50!} = \frac{52 \times 51}{2} = 26 \times 51 = 1326.$$

Since there are 4 aces in a pack of cards, so the number of favourable cases is  ${}^4C_2$ ,

$$\begin{aligned} \therefore \text{Required probability} &= \frac{{}^4C_2}{{}^{52}C_2} = \frac{4!}{2! \times 2!} \times \frac{2! \times 50!}{52!} \\ &= \frac{4 \times 3 \times 2 \times 1}{2 \times 2} \times \frac{2}{52 \times 51} = \frac{12}{52 \times 51} = \frac{1}{221}. \end{aligned}$$

**Example 30 :** Four persons are chosen at random from a group containing 3 men, 2 women and 4 children. Show that the chance that exactly two of them will be children is  $\frac{10}{21}$ .

**Solution :** The total number of ways in which four persons can be selected out of 9 persons is  ${}^9C_4$ .

For favourable cases, we want that 2 out of the four selected should be children. Two children can be selected out of 4 in  ${}^4C_2$  ways. The other two are to be selected out of 5 persons (3 men and 2 women). Two persons can be selected out of 5 in  ${}^5C_2$  ways.

$$\therefore \text{The number of favourable cases} = {}^4C_2 \times {}^5C_2$$

$$\therefore \text{Required probability} = \frac{{}^4C_2 \times {}^5C_2}{{}^9C_4} = \frac{4!}{2! \times 2!} \times \frac{5!}{2! \times 3!} / \frac{9!}{4! \times 5!} = \frac{10}{21}.$$

**Example 31 :** In a factory, there are 6 skilled workers and 4 unskilled workers. What is the probability that (a) worker selected is skilled worker ; (b) the two workers selected are skilled.

**Solution :** (a) Total number of favourable cases =  ${}^6C_1 = 6$ .

Total number of equally likely cases =  ${}^{10}C_1 = 10$ .

$$\therefore \text{Required probability} = \frac{6}{10}.$$

(b) Number of favourable cases = Selection of two workers out of 6 skilled workers.

This is done in  ${}^6C_2 = \frac{6 \times 5}{1 \times 2}$  ways.

$\therefore$  Total number of equally likely cases = selection of 2 workers out of a total of 10  
ways =  ${}^{10}C_2$ .

$$\therefore \text{Required probability} = \frac{{}^6C_2}{{}^{10}C_2} = \frac{6 \times 5}{1 \times 2} \times \frac{1 \times 2}{10 \times 9} = \frac{1}{3}.$$

**Example 32 :** Four cards are drawn from a full pack of cards. Find the probability that two are spades and two are hearts.

**Solution :** Total number of possible selections =  ${}^{52}C_4$ .

Number of ways of selecting 2 spades cards =  ${}^{13}C_2$ .

Number of ways of selecting 2 hearts cards =  ${}^{13}C_2$ .

$$\begin{aligned} \text{Total favourable cases} &= {}^{13}C_2 \times {}^{13}C_2 \\ \therefore \text{Required probability} &= \frac{{}^{13}C_2 \times {}^{13}C_2}{{}^{52}C_4} \\ &= \frac{13 \times 12}{1 \times 2} \times \frac{13 \times 12}{1 \times 2} \times \frac{1 \times 2 \times 3 \times 4}{52 \times 51 \times 50 \times 49} = \frac{468}{20825}. \end{aligned}$$

**Example 33 :** Ten students are seated at random in a row. Find the probability that two particular students are not seated side-by-side.

**Solution :** Let us tie these two particular students into one. Then it amounts to selection of two students out of 9. This is done in  ${}^9C_2$  ways.

But the two students can be permuted in 2 ways.

$\therefore$  Total number of ways for selection of two students such that two particular students are not seated side-by-side =  $2 \times {}^9C_2$  ways.

$$\therefore \text{Favourable cases} = 2 \times {}^9C_2$$

$$\text{Total exhaustive cases} = {}^{10}C_2$$

$$\text{Required probability} = \frac{2 \times {}^9C_2}{{}^{10}C_2} = \frac{2 \times 9!}{7! \times 2!} \times \frac{8! \times 2!}{7! \times 2!} = \frac{4}{5}.$$

**Example 34 :** A committee of three is to be chosen from a group consisting of 4 men and 5 women. If the selection is made at random, find the probability that (a) all three are men; (b) two are men.

**Solution : Exhaustive cases.** It is the selection of 3 out of  $(4 + 5) = 9 = {}^9C_3$ ,

(a) **Favourable cases.** Selection of 3 men out of 4 =  ${}^4C_3$ .

$$\therefore \text{Required probability} = \frac{{}^4C_3}{{}^9C_3} = \frac{4!}{3!} \times \frac{3! \times 6!}{9!} = \frac{1}{21}.$$

(b) In this case two are men and 1 is woman. It amounts to selection of 2 men out of 4 men and 1 woman out of 5 women =  ${}^4C_2 \times {}^4C_1$ .

$$\therefore \text{Required probability} = \frac{{}^4C_2 \times {}^5C_1}{{}^9C_3} = \frac{4!}{2! \times 2!} \times \frac{5!}{4!} \times \frac{3! \times 6!}{9!} = \frac{5}{84}.$$

**Example 35 :** A bag contains 5 white and 8 red balls. Two drawings of 3 balls are made such that (a) balls are replaced before the second trial and (b) the balls are not replaced before the second trial. Find the probability that the first drawing will give 3 white and the second red balls in each case.

**Solution :** (a) When the balls are replaced

Possible cases : Selection of 3 balls out of 13 =  ${}^{13}C_3$  ways.

Favourable cases : Selection of 3 white balls out of 5 white ball in first draw =  ${}^5C_3$ .

$$\text{Probability of 3 white in first trial} = \frac{{}^5C_3}{{}^{13}C_3} = \frac{5}{143}.$$

$$\text{Probability of 3 red in second trial} = \frac{{}^8C_3}{{}^{13}C_3} = \frac{28}{143}.$$

Probability of 3 white in first trial  $\times$  Probability of 3 red in second trial

$$= \frac{5}{143} \times \frac{28}{143} = \frac{140}{20449}$$

(b) When the balls are not replaced

In this case the probability of drawing 3 white in first draw =  $\frac{5}{143}$ .

When the balls are not replaced, after the first draw, the total number of balls left

$$= 2 \text{ white} + 8 \text{ red} = 10 \text{ balls.}$$

$\therefore$  In this case probability of drawing 3 red balls in the second draw =  $\frac{{}^{10}C_3}{{}^{10}C_3} = \frac{7}{15}$ .

$$\text{Required compound probability} = \frac{5}{143} \times \frac{7}{15} = \frac{7}{429}.$$

**Example 36 :** An urn contains 25 balls numbered 1 through 25. Two balls are drawn from the urn with replacement. Find the probability of selecting:

(a) both odd numbers.

(b) one odd and one even number.

- (c) at least one odd number.
- (d) no odd number.
- (e) both even number.

**Solution :** Total number of cases = 25.

Let  $E$  and  $D$  denote the events of drawing 'even' and 'odd' numbered ball. There are 13 odd numbered balls and 12 even numbered balls.

Then

$$P(E) = \frac{12}{25}$$

$$P(D) = \frac{13}{25}.$$

- (a) When the balls are replaced back in the urn

Let  $P(DD)$  be the probability of getting both odd numbered balls, then

$$\therefore P(DD) = P(D) \cdot P(D) = \frac{13}{25} \times \frac{13}{25} = \frac{169}{625}.$$

- (b) The total number of such cases consists of (first odd and second even) + (first even and second odd)

$$\begin{aligned} \therefore P(\text{one odd and one even}) &= P(ED) = P(E) \cdot P(D) + P(D) \cdot P(E) \\ &= \frac{12}{25} \times \frac{13}{25} + \frac{13}{25} \times \frac{12}{25} = \frac{312}{625}. \end{aligned}$$

- (c) At least one odd

Probability of getting both the balls even numbered

$$P(EE) = P(E) \cdot P(E) = \frac{12}{25} \times \frac{12}{25} = \frac{144}{625}$$

$$P(\text{at least one odd}) = 1 - P(EE) = 1 - \frac{144}{625} = \frac{481}{625}.$$

- (d) No odd number

$$P(\text{no odd number}) = 1 - P(\text{at least one odd}) = 1 - \frac{481}{625} = \frac{144}{625}.$$

- (e) Probability of (both even numbered) =  $\frac{12}{25} \times \frac{12}{25} = \frac{144}{625}.$

**Example 37 :** *n* cadets have to stand in a row. If all possible permutations are equally likely, find the probability that two particular cadets stand side-by-side.

**Solution :** The number of ways in which  $n$  cadets can stand in row is  $= n!$ .

Combine the two particular cadets together and regard them as one cadet.

The number of ways new  $(n - 1)$  cadets stand in a row =  $(n - 1)!$

Now the two cadets tied together can be arranged in  $2! = 2$  ways.

$\therefore$  Number of ways in which  $n$  cadets can stand in a row where two of them stand together  
 $= 2 \times (n - 1)!$

$$\therefore \text{Required probability} = \frac{2 \times (n-1)!}{n!} = \frac{2 \times (n-1)}{n(n-1)} = \frac{2}{n}.$$

**Example 38 :** One bag contains five white and four black balls. Another bag contains seven white and nine black balls. A ball is transferred from the first bag to the second and then a ball is drawn from the second. Find the probability that the ball is white.

**Solution : Case I.** Let a white ball be transferred from the first bag to the second bag.

The probability of selecting white ball in the first bag is  $p_1 = \frac{5}{9}$ .

Now the second bag has 8 white and 9 black balls.

The probability of selecting a white ball from the second bag is

$$p_2 = \frac{8}{8+9} = \frac{8}{17}.$$

The probability that both these events take place simultaneously is

$$= p_1 \times p_2 = \frac{5}{9} \times \frac{8}{17} = \frac{40}{153}.$$

**Case II :** Let a black ball be transferred from the first bag to the second bag.

Its probability is  $p_3 = \frac{4}{9}$ .

Now the second bag contains 7 white and 10 black balls.

The probability of drawing a white ball from the second bag is

$$p_4 = \frac{7}{7+10} = \frac{7}{17}.$$

$\therefore$  The probability of both these events taking place simultaneously is

$$= p_3 \times p_4 = \frac{4}{9} \times \frac{7}{17} = \frac{28}{135}.$$

$$\therefore \text{The required probability} = p_1 p_2 + p_3 p_4 = \frac{40}{153} + \frac{28}{135} = \frac{68}{153}.$$

**Example 39 :** A speaks truth in 75% and B in 80% of the cases. In what percentage of cases are they likely to contradict each other narrating the same incident?

**Solution :** Let  $P(A)$ ,  $P(B)$  be the probability of A and B speaking the truth, then

$$P(A) = \frac{75}{100} = \frac{3}{4}, \quad P(B) = \frac{80}{100} = \frac{4}{5}.$$

$$P(\bar{A}) = P(A \text{ tell a lie}) = 1 - P(A) = 1 - \frac{3}{4} = \frac{1}{4}$$

$$P(\bar{B}) = P(B \text{ tell a lie}) = 1 - P(B) = 1 - \frac{4}{5} = \frac{1}{5}$$

$$\begin{aligned} \text{Now, } P(A \text{ and } B \text{ will contradict}) &= P(A) P(\bar{B}) + P(B) P(\bar{A}) \\ &= \frac{3}{4} \times \frac{1}{5} + \frac{4}{5} \times \frac{1}{4} = \frac{7}{20} \quad \text{or} \quad 35\%. \end{aligned}$$

**Example 40 :** Two dice are tossed. What is the probability that total is divisible by 3 or 4?

**Solution :** A dice has 6 six faces, viz., 1, 2, 3, 4, 5, 6.

Total number of possible cases of 2 dice for the event =  $6 \times 6 = 36$ .

Number of outcomes which are multiple of 3 = 12.

(viz., 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36)

Number of outcomes which are multiple of 4 = 9.

(viz., 4, 8, 12, 16, 20, 24, 28, 32, 36)

Number of outcomes which are multiple of both 3 and 4, i.e., 12 = 3.

(viz., 12, 24, 36)

Let,  $A$  = the event that the outcome is a multiple of 3. Then  $P(A) = \frac{12}{36} = \frac{1}{3}$ ;

$B$  = the event that the outcomes, is a multiple of 4. Then  $P(B) = \frac{9}{36} = \frac{1}{4}$ .

Also

$$P(A \cap B) = \frac{3}{36}$$

Probability that the total is divisible by 3 or 4.

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{12}{36} + \frac{9}{36} - \frac{3}{36} = \frac{12+9-3}{36} = \frac{18}{36} = \frac{1}{2}. \end{aligned}$$

**Example 41 :** The probability that a person A who is now 25 years old, lives for another 30 years is  $2/5$ ; and the probability that the person B who is now 45 years old lives for another 30 years is  $7/16$ . Find the probability that at least one of these persons will be alive 30 years hence.

**Solution :** Let  $X, Y$  be the events that the persons A and B respectively live for another 30 years.

$$P(X) = \frac{2}{5}; \quad P(Y) = \frac{7}{16}.$$

Probability of living at least one of them live for another 30 years.

$$\begin{aligned} P(X \cup Y) &= P(X) + P(Y) - P(X \cap Y) \\ &= P(X) + P(Y) - P(X) \cdot P(Y) \\ &= \frac{2}{5} + \frac{7}{16} - \frac{2}{5} \times \frac{7}{16} = \frac{67}{80} - \frac{14}{80} = \frac{53}{80}. \end{aligned}$$

Hence the probability that at least one of the two persons will be alive for another 30 years is  $\frac{53}{80}$ .

## 10.10 CONDITIONAL PROBABILITY

Let  $A$  be any event in the sample space  $S$ , and  $P(A) > 0$ . The probability that an event  $B$  occurs subject to the condition that  $A$  has already occurred is called the Conditional Probability of  $B$  given that  $A$  has already occurred. It is denoted by  $P(B/A)$ .

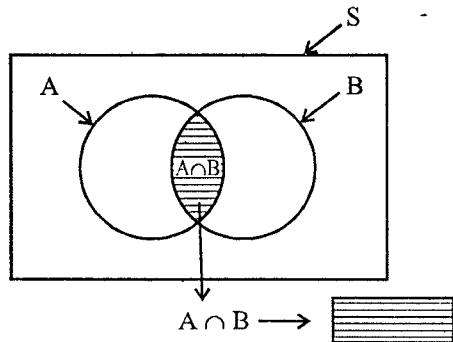
### Expression for $P(B/A)$

Let  $S$  be a finite equiprobable space. Let  $n(S)$  = total number of sample points  $S$ . Let  $A$  be any event with sample points  $n(A)$ . Let  $B$  be another event such that the sample points in  $A \cap B$  is  $n(A \cap B)$ . Then we know that

$$P(A) = \frac{\text{Sample points in } (A)}{\text{Sample points in } S} = \frac{n(A)}{n(S)}$$

Also,  $P(A \cap B) = \frac{n(A \cap B)}{n(S)}$

Under the assumption that  $A$  has occurred, the sample space is restricted to the points lying inside the region  $A$ . (see figure). It has  $n(A)$  sample points. This new sample space is called the **Reduced Sample Space**.



Under this assumption the occurrence of the event  $B$  is restricted to the shaded region in figure 0.9, which is  $A \cap B$  having  $n(A \cap B)$  sample points.

$\therefore P(B/A) = \text{conditional probability of } B \text{ given that } A \text{ has occurred} = \frac{n(A \cap B)}{n(S)} = \text{the}$

ratio of the sample points of the events  $(A \cap B)$  to the reduced sample space consisting of equiprobable sample points.

$$\text{Ratio} = \frac{n(A \cap B)/n(S)}{n(A)/n(S)} = \frac{P(A \cap B)}{P(A)}$$

Hence

$$P(B/A) = \frac{P(A \cap B)}{P(A)}.$$

or

$$P(A \cap B) = P(A) \times P(B/A).$$

It is obvious that  $P(B/A)$  is defined if and only if  $P(A) \neq 0$ . Hence the condition  $P(A) > 0$ .

Similarly, the conditional probability of  $A$  given that  $B$  has occurred is  $P(A/B)$ , where,

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

$\Rightarrow$

$$P(A \cap B) = P(B) \times P(A/B)$$

**Example 42 :** Let  $A$  and  $B$  be events with  $P(A) = \frac{1}{3}$ ,  $P(B) = \frac{1}{4}$ ,  $P(A \cap B) = \frac{1}{12}$ . Find

- (i)  $P(A/B)$ , (ii)  $P(B/A)$ , (iii)  $P(B/A^C)$ , (iv)  $P(A \cap B^C)$ .

**Solution :**

$$(i) \text{ Here } P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/12}{1/4} = \frac{1}{3}.$$

$$(ii) P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/12}{1/3} = \frac{1}{12} \times \frac{3}{1} = \frac{1}{4}.$$

$$(iii) \text{ Now, } P(B/A^C) = \frac{P(B \cap A^C)}{P(A^C)} = \frac{P(B) - P(B \cap A)}{1 - P(A)}$$

$$= \frac{(1/4) - (1/12)}{(1 - 1/3)} = \frac{1}{6} \times \frac{3}{2} = \frac{1}{4}.$$

$$(iv) P(A \cap B^C) = P(A) - P(A \cap B) = \frac{1}{3} - \frac{1}{12} = \frac{3}{12} = \frac{1}{4}.$$

## 10.11 INDEPENDENT EVENTS

The events  $A$  and  $B$  are said to be independent if

$$P(AB) = P(A) \cdot P(B)$$

Thus in case of independent events the above multiplication theorem becomes

$$\text{or } P(A \cap B) = P(A) \times P(B) \quad [\because P(A/B) = P(A) \text{ and } P(B/A) = P(B)]$$

**Example 43 :** Given that  $P(A) = \frac{3}{8}$ ,  $P(B) = \frac{5}{8}$  and  $P(A \cup B) = \frac{3}{4}$ , find  $P(A/B)$  and  $P(B/A)$ . Show whether  $A$  and  $B$  are independent.

**Solution :** We have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\Rightarrow \frac{3}{4} = \frac{3}{8} + \frac{5}{8} - P(A \cap B) \Rightarrow P(A \cap B) = \frac{3}{8} + \frac{5}{8} - \frac{3}{4} = \frac{1}{4}$$

$$\text{Now, } P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{5/8} = \frac{1}{4} \times \frac{8}{5} = \frac{2}{5}$$

$$\text{and } P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/4}{3/8} = \frac{1}{4} \times \frac{8}{3} = \frac{2}{3}.$$

$$\text{Again, } P(A \cap B) = \frac{1}{4} \text{ and } P(A) \cdot P(B) = \frac{3}{8} \times \frac{5}{8} = \frac{15}{64}.$$

$A$  and  $B$  will be independent events, if  $P(A \cap B) = P(A) \cdot P(B)$ .

Thus we notice that  $P(A \cap B) \neq P(A) \cdot P(B)$ , so the events  $A$  and  $B$  are not independent.

**Example 44 :** The odds against a student  $X$  solving a business statistics problem are 8 to 6 and odds in favour of the student  $Y$  solving the problem are 14 to 16.

- (a) What is the chance that the problem will be solved if they both try independent of each other.  
 (b) What is the probability that none of them is able to solve the problem.

**Solution :** Let  $A$  = the event that the student  $X$  solves the problem.

$$\therefore P(A) = \frac{6}{8+6} = \frac{6}{14}$$

Let  $B$  = the event that the student  $Y$  solves the problem.

$$\therefore P(B) = \frac{14}{14+16} = \frac{14}{30}.$$

(a) Probability that the problem will be solved  
 $= P(\text{at least one of them solve the problem})$   
 $= P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A) \cdot P(B)$   
 $= \frac{6}{14} + \frac{14}{30} - \frac{6}{14} \times \frac{14}{30} = \frac{73}{105}.$

(b) Probability that none will solve the problem

$$= P(A^C) \times P(B^C) = \{1 - P(A)\} \{1 - P(B)\} = \frac{8}{14} \times \frac{16}{30} = \frac{32}{105}.$$

**Example 45 :** Show that for the three events  $A$ ,  $B$ ,  $C$ , the probability that at least one of the events will occur is

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

**Solution :** Let us put  $B \cup C = D$ , so that

$$P(D) = P(B \cup C) = P(B) + P(C) - P(B \cap C) \quad \dots \quad (1)$$

$$\therefore P(A \cup B \cup C) = P(A \cup D) = P(A) + P(D) - P(B \cap C) \\ = P(A) + P(B) + P(C) - P(B \cap C) - (A \cap D) \quad \dots \quad (2)$$

But

$$P(A \cap D) = P[A \cap (B \cap C)] = P[(A \cap B) \cup (A \cap C)] \\ = P(A \cap B) + P(A \cap C) - P[(A \cap B) \cup (A \cap C)] \\ = P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \quad \dots \quad (3)$$

Using (3) in (2), we get

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) + P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

**Example 46 :** The probability that  $A$  can solve the problem is  $\frac{4}{5}$ ,  $B$  can solve it is  $\frac{2}{3}$ , and  $C$  can solve it is  $\frac{3}{7}$ . If all of them try independently, find the probability that the problem will be solved.

**Solution :** Here  $P(A) = \frac{4}{5}$ ,  $P(B) = \frac{2}{3}$  and  $P(C) = \frac{3}{7}$ .

$$\begin{aligned}
 P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) + P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \\
 &= P(A) + P(B) + P(C) - P(A) \cdot P(B) - P(A) \cdot P(C) - P(B) \cdot P(C) + P(A) \cdot P(B) \cdot P(C) \\
 &= \frac{4}{5} + \frac{2}{3} + \frac{3}{7} - \left( \frac{4}{5} \times \frac{2}{3} \right) - \left( \frac{4}{5} \times \frac{3}{7} \right) - \left( \frac{2}{3} \times \frac{3}{7} \right) + \frac{4}{5} \times \frac{2}{3} \times \frac{3}{7} \\
 &= \frac{4}{5} + \frac{2}{3} + \frac{7}{3} - \frac{8}{15} - \frac{12}{35} - \frac{2}{7} + \frac{8}{35} = \frac{101}{105}.
 \end{aligned}$$

**Example 47 :** A bag contains 6 red and 5 blue balls and another bag contains 5 red and 8 blue balls. A ball is drawn from the first bag and without noticing its colour is put in the second. A ball is then drawn from the second bag. Find the probability that the ball drawn is blue in colour.

**Solution :** There are two possibilities:

**Case I :** When the ball drawn from the first bag is red.

$$p_1 = P(\text{red ball from 1st bag}) = \frac{6}{11}$$

Now the second bag contains 6 red and 8 blue balls.

$$p_2 = P(\text{blue ball from second bag}) = \frac{8}{14}$$

Therefore, the probability of both these events taking place simultaneously is:

$$= p_1 \times p_2 = \frac{6}{11} \times \frac{8}{14} = \frac{48}{154}$$

**Case 2 :** When the ball drawn from first bag is blue.

$$p_3 = P(\text{blue ball from 1st bag}) = \frac{5}{11}$$

Now the second bag has 5 red and 9 blue balls.

$$p_4 = P(\text{blue ball from second bag}) = \frac{9}{14}$$

Probability of both these events taking place simultaneous is:

$$= p_3 \times p_4 = \frac{5}{11} \times \frac{9}{14} = \frac{45}{154}$$

But the Case 1 and Case 2 are mutually exclusive, therefore,

$$\text{the required probability} = p_1 p_2 + p_3 p_4 = \frac{48}{154} + \frac{45}{154} = \frac{93}{154}.$$

**Example 48 :** Two cards are drawn from a pack of 52 cards at random and kept out. Then one card is drawn from the remaining 50 cards. Find the probability that it is an ace.

**Solution :** Let A denote the event 'getting 2 aces in the first draw of 2 cards'. 'B' denote the event getting 1 ace and 1 non-ace in the first draw of 2 cards. Let 'C' denote the event getting 2 non-aces in the first draw of 2 cards.

Then

$$P(A) = \frac{^4C_2}{^{52}C_2} = \frac{4 \times 3}{52 \times 51} = \frac{1}{221}$$

$$P(B) = \frac{^4C_1 \times ^{48}C_1}{^{52}C_2} = \frac{32}{221}$$

$$P(C) = \frac{^{48}C_2}{^{52}C_2} = \frac{188}{221}.$$

Let E be the event of drawing an ace in the second draw of one card.

$$\begin{aligned}\text{Required probability} &= P(E) = P[(A \cap E) \cup (B \cap E) \cup (C \cap E)] \\&= P(A \cap E) + P(B \cap E) + P(C \cap E) \\&= P(A) P(E/A) + P(B) P(E/B) + P(C) P(E/C) \\&= \frac{1}{221} \times \frac{2}{50} + \frac{32}{221} \times \frac{3}{50} + \frac{188}{221} \times \frac{4}{50} \\&= \frac{2 + 96 + 752}{221 \times 50} = \frac{850}{11050} = \frac{1}{13}.\end{aligned}$$

### EXERCISE 10.1

1. List the reasons for classical definition of probability not being very satisfactory. Give the modern definition of probability.
2. State the addition and multiplication theorem on probability.
3. Give the mathematical and statistical definition of probability.
4. State the addition and multiplication rules of probability giving one example of each case.
5. The probability that machine A will be performing a usual function in 5 years time is  $(1/4)$ , while the probability that machine B will still be operating usefully at the end of same period is  $(1/3)$ . find the probability that in five years time.

(a) both machines will be performing a usual function.

(b) at least one of the machines will be operating.

[Hint : (a) Required probability =  $\frac{1}{4} \times \frac{1}{3} = \frac{1}{12}$ .

$$(b) \text{ Required probability} = \frac{1}{4} + \frac{1}{3} - \frac{1}{12} = \frac{1}{2}.$$

6. The probability that (i) A can solve a problem in statistics is  $4/5$ ; (ii) B can solve it is  $2/3$  and (iii) C can solve it is  $3/7$ . If all the them try independently, find the probability that the problem will be solved.

[Hint : Required probability =  $1 - \frac{4}{5} \times \frac{2}{3} \times \frac{3}{7} = \frac{81}{105}$ .]

7. The odds against X solving a problem are 8 to 6, and odds in favour of Y solving the same problem are 14 to 16.

(a) What is the chance that the problem will be solved if both try?

(b) What is the probability that neither solves the problem?

$$[\text{Hint : Here } P(X) = \frac{6}{8+6} = \frac{6}{14}; \therefore P(X^C) = 1 - \frac{6}{14} = \frac{8}{14}]$$

$$P(Y) = \frac{14}{14+16} = \frac{14}{30}; \quad P(Y^C) = \frac{16}{30}.$$

$$(a) \text{Probability that the problem is not solved} = \frac{8}{14} \times \frac{16}{30} = \frac{32}{105}.$$

$$\text{Probability that the problem is solved} = 1 - \frac{32}{105} = \frac{73}{105}.$$

$$(b) \text{Probability that none will solve the problem} = \frac{32}{105}.$$

8. The probability that a boy will get a scholarship is 0.9 and that a girl will get is 0.8. What is the probability that at least one of them will get the scholarship?

$$[\text{Hint : Required probability} = 1 - (1 - 0.9) \times (1 - 0.8) = 1 - 0.02 = 0.98.]$$

9. Which event has the greatest probability (i) getting 4 with one dice; (ii) getting 8 with two dice and (iii) getting 12 with three dice.

$$[\text{Hint : (i)} \quad P(\text{getting a four}) = \frac{1}{6}$$

$$(\text{ii}) \quad P(\text{getting 8 with two dice}) = \frac{\text{Possible pairs are } (6, 2), (4, 4), (5, 3), (3, 5), (2, 6)}{36} = \frac{5}{36}$$

$$(\text{iii}) \quad P(\text{getting 12 with 3 dice}) = \frac{25}{216}$$

Hence getting 4 with one dice has the greatest probability.]

10. A bag contains 10 white, 15 red and 8 green balls. A single draw of 3 balls is made.

(a) What is the probability of getting all the three white balls?

$$[\text{Hint : (a)} \quad \frac{\binom{10}{1} \times \binom{10}{1} \times \binom{10}{1}}{\binom{33}{3}} = \frac{75}{341}, \quad (\text{b}) \quad \frac{\binom{10}{3}}{\binom{33}{3}} = \frac{15}{642}.]$$

11. A bag contains 7 red balls and 5 white balls. 4 balls are drawn at random. What is the probability that (i) all of them are red; (ii) two of them are red and two white?

$$[\text{Hint : (i)} \quad \frac{C(7, 4)}{C(12, 4)}, \quad (\text{ii}) \quad \frac{C(7, 2) \times C(5, 2)}{C(12, 4)}.]$$

12. The odds against A solving a problem are 8 : 6 and odds in favour of B solving it are 9 : 12. What is the probability that if both of them try it, will be solved?

13. A piece of equipment will function only when all the three components A, B and C are working. The probability of A failing during one year is 0.15, that of B failing is 0.05 and that of C failing is 0.10. What is the probability that the equipment will fail before the end of the year?

[Hint : Probability that all components will work =  $(1 - 0.15)(1 - 0.05)(1 - 0.10)$   
 $= 0.85 \times 0.95 \times 0.90 = 0.727$

Probability that equipment fails =  $1 - 0.727 = 0.273.$ ]

14. There are 3 economists, 4 engineers, 2 statisticians and 1 doctor. A committee of 4 from among them is to be formed. Find the probability that the committee.

(i) consists of one each kind.

(ii) has at least one economist.

(iii) has the doctor as a member and three others.

15. A candidate is selected for interview for 3 posts. For the first there are 3 candidates, for the second there are 4 and for the third there are 2. What are the chances of his getting at least one?

[Hint : Let  $A, B, C$  be the posts, then  $P(A) = \frac{1}{3} \Rightarrow P(\bar{A}) = 1 - \frac{1}{3} = \frac{2}{3}.$

Similarly,  $P(\bar{B}) = 1 - \frac{1}{4} = \frac{3}{4}, P(\bar{C}) = 1 - \frac{1}{2} = \frac{1}{2};$

$$P(\bar{A} \bar{B} \bar{C}) = P(\bar{A}) P(\bar{B}) P(\bar{C}) = \frac{2}{3} \times \frac{3}{4} \times \frac{1}{2} = \frac{1}{4}.$$

16. There are four hotels in a certain city. If 3 men check into hotels in a day, what is the probability they each are into a different hotel?

[Hint : 3 men can check into a hotel in  $= 4 \times 4 \times 4 = 64$  ways.

They can check into a different hotel in only  $4 \times 3 \times 2 = 24$  ways.

Required probability =  $\frac{24}{64} = 0.375.$ ]

17. What is the probability that a leap year selected at random will contain either 53 Thursdays or 53 Fridays?

[Hint : Leap year has 366 days out of which 52 complete weeks and 2 days. These two days can occur in any of the following combinations:

(i) Sunday and Monday

(ii) Monday and Tuesday

(iii) Tuesday and Wednesday.

(iv) Wednesday and Thursday

(v) Thursday and Friday.

(vi) Friday and Saturday.

(vii) Saturday and Sunday.

Let  $P(A) =$  Probability that leap year will contain 53 Thursday =  $\frac{2}{7}.$

$P(B) =$  Probability that leap year will contain 53 Fridays =  $\frac{2}{7}.$

$$\therefore P(A \text{ and } B) = 1/7.$$

$$\therefore P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = \frac{2}{7} + \frac{2}{7} - \frac{1}{7} = \frac{3}{7}.$$

18. If a pair of dice is thrown, find the probability that the sum is neither 7 or 11.
19. Three horses  $A$ ,  $B$  and  $C$  are in a race;  $A$  is twice as likely to win as  $B$  and  $B$  is twice as likely to win as  $C$ . What are their respective probabilities of winning?
20. An investment consultant predicts that the odds against the price of certain stock going up during the next week are 2 : 1 and the odds in favour of price remaining the same are 1 : 3. What is the probability that the price of stock will go down during the next week?
21. A pack of 50 tickets numbered 1 to 50 is shuffled and then two tickets are drawn. Find the probability that
- both the tickets drawn have prime numbers.
  - none of the tickets drawn has prime numbers.

[Hint : (i) Required probability =  ${}^{15}C_2 / {}^{50}C_2 = \frac{3}{35}$ .

(ii) Required probability =  ${}^{35}C_2 / {}^{50}C_2 = \frac{17}{35}$ .]

22. There are three urns  $A$ ,  $B$  and  $C$ . Urn  $A$  contains 4 red balls and 3 black balls. Urn  $B$  contains 5 red balls and 4 black balls. Urn  $C$  contains 4 red balls and 4 black balls. One ball is drawn from each of these urns. What is the probability that the three balls drawn consist of a red ball and two black balls?

[Hint :  $P(1 \text{ red and } 2 \text{ black}) = P(\text{red from } A, \text{black from } B, \text{black from } C) + P(\text{black from } A, \text{red from } B, \text{black from } C) + P(\text{black from } A, \text{black from } B, \text{red from } C)$ .

$$= \frac{4}{7} \times \frac{4}{9} \times \frac{4}{8} + \frac{3}{7} \times \frac{5}{9} \times \frac{4}{8} + \frac{3}{7} \times \frac{4}{9} \times \frac{4}{8} = \frac{43}{126}.$$

23. A coin is tossed six times. What is the probability of getting at least two heads?

[Hint : Let  $X$  denotes the number of heads, then the required probability

$$= P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$$

$$\text{Required probability} = 1 - P(X = 0) - P(X = 1) = \frac{57}{64}.$$

24. Two persons throw a dice alternately, till one of them gets a multiple of three and wins the game. Find their respective probabilities of winning.

[Hint :  $P(\text{multiple of 3}) = P(\text{getting 3 or 6}) = \frac{2}{6} = \frac{1}{3}$ . Here  $p = \frac{1}{3}$ ,  $q = 1 - \frac{1}{3} = \frac{2}{3}$ .

Let the two persons be  $A$  and  $B$ . Let  $A$  start throwing the dice. He wins in the first throw, 3rd, 5th, 7th and so on. Let  $P(A)$  = probability of  $A$ 's winning,  $P(\bar{A})$  = Probability of  $A$ 's not winning. Similarly,  $P(B)$  and  $P(\bar{B})$  denote the probability of  $B$ 's winning and not winning respectively.

$$\begin{aligned} \text{Probability of } A\text{'s winning} &= P(A) + P(\bar{A})P(\bar{B})P(A) + P(\bar{A})P(\bar{B})P(\bar{A})P(\bar{B})P(A) + \dots \\ &= \frac{1}{3} + \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} + \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} + \dots \end{aligned}$$

$$= \frac{1}{3} \left[ 1 + \left(\frac{2}{3}\right)^2 + \left(\frac{2}{3}\right)^4 + \dots \right] = \frac{1}{3} \cdot \frac{1}{1 - (2/3)^2} = \frac{1}{3} \times \frac{9}{5} = \frac{3}{5}$$

$$\text{Probability of } B\text{'s winning the first game} = 1 - \frac{3}{2} = \frac{2}{5}.$$

25. A problem in statistics is given to three students  $A$ ,  $B$  and  $C$  whose chances of solving it are  $1/2$ ,  $3/4$  and  $1/4$  respectively. What is the probability that the problem is solved?

$$[\text{Hint : } P(\bar{A}) = \frac{1}{2}, \quad P(\bar{B}) = \frac{1}{4}, \quad P(\bar{C}) = \left(\frac{3}{4}\right). \quad \therefore \quad P(\overline{ABC}) = \frac{1}{2} \times \frac{1}{4} \times \frac{3}{4} = \frac{3}{32}]$$

$$\text{Probability that the problem is solved} = 1 - \frac{3}{32} = \frac{29}{32}.$$

26. Find the chance of drawing an ace, a queen, a king, and a knave, in this order from an ordinary pack of four consecutive draws, the card drawn not being replaced.
27. A dice has two of its sides painted red, two black and two yellow. If the dice is rolled twice, what is the probability that the same colour appears both the times.
28. A bag contains  $n$  distinct white and  $n$  distinct red balls. Pairs of balls are drawn without replacement until the bag is empty. Show that the probability that each pair consists of one white and one red ball is  $2^n/C(2n, n)$ .
29. A pack of 50 tickets numbered 1 to 50 is shuffled and two tickets are drawn at random without replacement. Find the probability that at least one of the tickets drawn has a prime number.

[Hint : There 15 prime numbers and 35 non-prime numbers between 1 and 50.]

$$\text{Let } P(A) = P(\text{None is prime number}) = \frac{C(35, 2)}{C(50, 2)} = \frac{17}{35}.$$

$$P(\text{at least one prime}) = 1 - P(A) = 18/25.$$

30. The face cards are removed from a full pack of 52 playing cards. Out of the remaining 40 cards are drawn at random. What is the probability that they belong to different suits?
31. In a group there are 2 men and 3 women. Three persons are selected at random from the group. Find the probability that 1 man and 2 women or 2 men and 1 woman are selected.

[Hint : Required probability =  $P(1M \text{ and } 2W) + P(2M + 1W)$ .]

$$= \frac{C(2, 1) \times C(3, 2)}{C(5, 3)} + \frac{C(2, 2) \times C(3, 1)}{C(5, 3)} = \frac{9}{10}.$$

32. There are 3 red and 5 black balls in bag  $A$ ; and 2 red and 3 black balls in bag  $B$ . One ball is drawn from bag  $A$  and two from bag  $B$ . Find the probability that out of the 3 balls drawn, 1 is red and 2 are black.
33. A man alternatively tosses a coin and throws a die beginning with the coin. What is the probability that he will get a head before he gets '5 or 6' on the die?

[Hint : Let  $S$  = the event of getting a '5 or 6'.]

$$\therefore P(S) = \frac{1}{3} \text{ and } P(\bar{S}) = \frac{2}{3}. \quad \text{Also } P(H) = \frac{1}{2}, \quad P(T) = \frac{1}{2}.$$

$\therefore P(\text{head before '5 or 6'}) = P(H \text{ or } T\bar{S}H, T\bar{S}T\bar{S}H \text{ and so on})$

$$= \frac{1}{2} P(H) + P(T) P(\bar{S}) P(H) + P(T) + P(\bar{S}) P(T) P(\bar{S}) P(H) + \dots$$

$$= \frac{1}{2} + \frac{1}{2} \times \frac{2}{3} \times \frac{1}{2} + \frac{1}{2} \times \frac{2}{3} \times \frac{1}{2} \times \frac{2}{3} \times \frac{1}{2} + \dots \text{ to } \infty = \frac{(1/2)}{1 - (1/3)} = \frac{3}{4}.$$

34. A bag contains 4 red and 5 black balls. Another bag contains 3 red and 6 black balls. One ball is drawn at random from the first bag and two balls are drawn from the second bag. Find the probability that out of the three, two are black and one is red.

[Hint : Case I.]  $p_1 = P(\text{one Red from bag I and 2 blacks from bag II})$

$$= \frac{C(4, 1)}{C(9, 1)} \times \frac{C(6, 2)}{C(9, 2)} = \frac{60}{324}.$$

Case II.

$p_2 = P(1 \text{ Black from bag I, 1 Black and one Red from bag II})$

$$= \frac{C(5, 1)}{C(9, 1)} \times \frac{C(3, 1) \times C(6, 1)}{C(9, 2)} = \frac{90}{324}$$

$$\text{Required probability} = p_1 + p_2 = \frac{60}{324} + \frac{90}{324} = \frac{25}{54}.$$

35. A bag contains 8 red, 3 white and 9 blue balls. Three balls are drawn at random from the bag. Determine the probability that none of the balls drawn is white.

[Hint : Required probability =  $\frac{C(17, 3)}{C(20, 3)} = \frac{34}{57}$ .]

36. A, B and C play a game and chances of their winning it in an attempt are  $2/3$ ,  $1/2$  and  $1/4$  respectively. A has the first chance, followed by B and then by C. This cycle is repeated till one of them wins the game. Find their respective chances of winning the game.

37. In a class there are 3 boys and 2 girls. 3 students are selected at random from the class. Find the probability that 2 boys and 1 girl or 1 boy and 2 girls are selected.

[Hint : Required probability =  $\frac{C(3, 1) \times C(2, 2)}{C(5, 3)} + \frac{C(3, 2) \times C(2, 1)}{C(5, 3)} = \frac{6}{10} + \frac{3}{10} = \frac{9}{10}$ .]

38. A card is drawn from a well shuffled pack of 52 cards. The events A and B are:

A : getting a card of spade; B : getting an ace. Determine whether A and B are independent or not.

[Hint :  $P(A) = \frac{13}{52}$ ,  $P(B) = \frac{4}{52}$ ,  $P(A \cap B) = \frac{1}{52}$ ;  $P(A \cap B) = P(A) \cdot P(B)$

$\Rightarrow A$  and  $B$  are independent.]

39. Out of the numbers 1 to 120, one is selected at random. What is the probability that it is divisible by 8 or 10?

[Hint : Let  $A$  : the event that the number is divisible by 8 and  $B$  : the event that the number is divisible by 10.

$$\therefore A = \{8, 16, 24, 32, 40, 48, 56, 64, 72, 80, 88, 96, 104, 112, 120\}$$

$$B = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 120\}$$

$$\therefore A \cap B = \{30, 40, 80, 120\}$$

$$\therefore n(A) = 15, n(B) = 12, n(A \cap B) = 4.$$

$$\text{Now, } P(A \text{ or } B) = P(A) + P(B) - P(A \cap B) = \frac{15}{120} + \frac{12}{120} - \frac{4}{120} = \frac{23}{120}.$$

40. Probability that a trainee will remain with company is 0.8. The probability that an employee earns more than Rs. 20,000 per year is 0.4. The probability that an employee, who was a trainee and remained with the company or who earns more than 20,000 per year is 0.9. What is the probability that an employee earns more than Rs. 20,000 per year given that he is a trainee, who stayed with the company?

[Hint :  $A$  = a trainee who remain with company  $\Rightarrow P(A) = 0.8$

$B$  = A employee who earn more than Rs. 20,000  $\Rightarrow P(B) = 0.4$

$$\therefore P(A \text{ or } B) = 0.9$$

$$\therefore P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

$$\Rightarrow 0.9 = 0.8 + 0.04 - P(A \cap B) \Rightarrow P(A \cap B) = 0.3$$

$$\therefore \text{Required probability} = P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.3}{0.8} = 0.375.$$

41. Three machines  $A$ ,  $B$  and  $C$  produce respectively 50%, 30% and 20% of the total number of items of a factory. The percentage of defective output of these machines are 3%, 4% and 5%. If an item is selected at random, find the probability that the item is defective.

[Hint :  $P$  = Prob. (defective item from  $A$ ) + Prob. (defective item from  $B$ )

+ Prob. (defective item from  $C$ ).

$$= 50\% \times 3\% + 30\% \times 4\% + 20\% \times 5\% = 0.5 \times 0.03 + 0.3 \times 0.04 + 0.2 \times 0.05$$

$$= 0.15 + 0.012 + 0.010 = 0.037.$$

42. A box contains three coins; one coin is fair, one coin is two-headed, and one is weighted so that the probability of heads appearing is  $1/3$ . A coin is selected at random and tossed. Find the probability  $P$  that head appears.

$$[\text{Hint : Required probability} = \left(\frac{1}{3} \times \frac{1}{2}\right) + \left(\frac{1}{3} \times 1\right) + \frac{1}{3} \times \frac{1}{3} = \frac{11}{18}].$$

43. The odds that a book will be favourable received by 3 independent critics are 5 to 2, 4 to 3 and 3 to 4 respectively. What is the probability that of the three reviews, a majority will be favourable?

[Hint : Let  $A, B, C$  denote the events of favouring the book by the first, the second and the third critic respectively, then  $P(A) = \frac{5}{7}$ ,  $P(B) = \frac{4}{7}$  and  $P(C) = \frac{3}{7}$ .

$$\begin{aligned}\therefore \text{Required probability} &= P(\text{two favour the book or three favour the book}) \\ &= P(\text{two favour the book}) + P(\text{three favour the book}) \\ &= P(AB\bar{C} \text{ or } A\bar{B}C \text{ or } \bar{A}BC) + P(ABC) \\ &= P[(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) (\bar{A} \cap B \cap C)] \\ &\quad + P(A \cap B \cap C).\end{aligned}$$

44. A student takes his examination in four subjects  $\alpha, \beta, \gamma, \delta$ . He estimates his chance of passing in  $\alpha$  as  $\frac{4}{5}$ , in  $\beta$  as  $\frac{3}{4}$ , in  $\gamma$  as  $\frac{5}{6}$  and  $\delta$  as  $\frac{2}{3}$ . To qualify he must pass in  $\alpha$  and at least two other subjects. What is the possibility that he qualifies?

[Hint : Required probability =  $P(\alpha \beta \gamma \bar{\delta}) + P(\alpha \beta \delta \bar{\gamma}) + P(\alpha \gamma \delta \bar{\beta}) + P(\alpha \beta \gamma \delta)$ .]

45. Two cards are drawn at random from a pack of 52 cards. What is the probability that it will be  
(a) a club and a heart ; (b) a king and a queen.

[Hint : (a)  $\frac{C(13, 1) C(13, 1)}{C(52, 2)} = \frac{13}{102}$  ; (b)  $\frac{C(4, 1) \times C(4, 1)}{C(52, 2)} = \frac{8}{663}$ .]

46. A salesman is known to sell a product in 3 out of 5 attempts while another salesman in 2 out of 5 attempts. Find the probability that

(i) no sales will be effected when both try to sell the

$$P(A) = \frac{3}{5}, P(\bar{A}) = \frac{2}{5}, P(B) = \frac{2}{5} \text{ and } P(\bar{B}) = \frac{3}{5}$$

(ii) Probability that both will not be able to sell is :

$$P(\bar{A} \cap \bar{B}) = P(\bar{A}) P(\bar{B}) = \frac{2}{5} \times \frac{3}{5} = \frac{6}{25} = 0.24.$$

(iii) Probability that either  $A$  or  $B$  will succeed is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{5} + \frac{2}{5} - \frac{6}{25} = 0.76.$$

47. Three balls are drawn successively from an urn containing 6 red balls, 4 white balls and 5 blue balls. Find the probability that these are drawn in order as red, white and blue if each ball is not replaced.

[Hints : Let  $A$  = the event first ball is red ;  $B$  = the event the second ball is white and  $C$  = the event the third ball is blue.

$$\text{Now, } P(ABC) = P(A) P(B/A) P(C/AB) = \frac{6}{15} \times \frac{4}{14} \times \frac{5}{13} = \frac{4}{91}.$$

48. A bag contains 7 green, 4 white and 5 red balls. If four balls are drawn one by one without replacement, what is the probability that none is red?
49. A bag contains 4 white and 2 black balls. Another contains 3 white and 5 black balls. If one ball is drawn from each bag, find the probability that (i) both are white; (ii) both are black; (iii) one is white and one is black.
50. There are two bags containing 5 red and 7 white; 3 red and 12 white balls respectively. A ball is drawn from one of the two bags. Find the probability of drawing a red balls.
51. Bag  $A$  contains 5 red and 3 green balls, and bag  $B$  contains 3 red and 5 green balls. One ball is drawn from bag  $A$  and two from bag  $B$ . Find the probability that the balls are green.
52. A bag contains 5 white and 3 black balls. Four balls are drawn one at a time without replacement. Find the probability that the balls are alternatively of different colours.
53. An article manufactured by a company consists of two parts  $A$  and  $B$ . In the process of manufacturing of the parts  $A$ , 9 out of 100 parts may be defective. Similarly, 5 out of 100 are likely to be defective in the manufacture of part  $B$ . Find the probability that the product will not be defective.
54. Arun and Tarun appeared for an interview for two vacancies. The probability of Arun's selection is  $1/4$  and that of Tarun's rejection of  $2/3$ . Find the probability that at least one of them will be selected.

[Hint : Required probability =  $1 - \left( \frac{3}{4} \times \frac{2}{3} \right) = \frac{1}{2}$ .]

55.  $A$  speaks truth in 60% of the cases and  $B$  in 90% of the cases. In what percentage of cases are likely to contradict each other in stating the same fact?
56. Find the probability of drawing a one rupee coin from a purse with two compartments. One of which contain 3 fifty-paise coins and 2 one rupee coins; and the other contains 2 fifty-paise coins and 3 one rupee coins.
57. Events  $E$  and  $F$  are known to be independent. Examine if the following  
 (i)  $\bar{E}$  and  $\bar{F}$     (ii)  $\bar{E}$  and  $F$ ;    (iii)  $E$  and  $\bar{F}$  are independent.
58. A lot contains 50 defective and 50 non-defective bulbs. Two bulbs are drawn at random, one at a time with replacement. The events  $A$ ,  $B$  and  $C$  are defined as:  
 $A = \{\text{the first bulb is defective}\}$ ;  $B = \{\text{the second bulb is defective}\}$ ; and  
 $C = \{\text{the two bulbs are both defective or non-defective}\}$ .  
 Show that (a)  $A$ ,  $B$  and  $C$  are pairwise independent;  
 (b)  $A$ ,  $B$  and  $C$  are not independent.
59. Tickets are numbered 1 to 10. Two tickets are drawn one after the other with replacement. Find the probability that the number on one of the tickets is a multiple of 5 and on the other is a multiple of 4.

[Hint : Let  $A = \{(5, 4), (5, 8), (10, 4), (10, 8), (4, 5), (4, 10), (8, 5), (8, 10)\}$

$$\therefore P(A) = \frac{8}{10 \times 10} = \frac{2}{25}.$$

60. Out of 9 outstanding students in a school, there are 4 boys and 5 girls. A team of 4 students is to be selected for a quiz programme. Find the probability that 2 are girls and 2 are boys.

### ANSWERS

- |  |  |  |  |
|--|--|--|--|
| <b>5.</b> (a) $\frac{1}{12}$ ; (b) $\frac{1}{2}$   | <b>6.</b> $\frac{101}{105}$                            | <b>7.</b> (a) $\frac{73}{105}$ ; (b) $= \frac{32}{105}$                    | <b>8.</b> 0.98                                       |
| <b>9.</b> Throwing a four with one dice.   |  | <b>10.</b> (a) $\frac{75}{341}$ ; (b) $= \frac{15}{682}$ .                 |  |
| <b>11.</b> (i) $\frac{C(7, 4)}{C(12, 4)}$ ; (ii) $\frac{C(7, 2) \times C(5, 2)}{C(12, 4)}$ |  | <b>12.</b> $\frac{37}{49}$   | <b>13.</b> 0.273                                     |
| <b>14.</b> (i) $\frac{4}{35}$ ; (ii) $\frac{5}{6}$ ; (iii) $\frac{2}{5}$                   | <b>15.</b> $\frac{1}{4}$                               | <b>16.</b> 0.375   | <b>17.</b> $\frac{3}{7}$                             |
| <b>18.</b> $\frac{7}{9}$   | <b>19.</b> $\frac{1}{7}$                               | <b>20.</b> $\frac{5}{12}$  | <b>21.</b> (i) $\frac{3}{35}$ , (ii) $\frac{17}{35}$ |
| <b>22.</b> $\frac{43}{126}$  | <b>23.</b> $\frac{57}{64}$                             | <b>24.</b> $\frac{2}{5}$   | <b>25.</b> $\frac{29}{32}$                           |
| <b>26.</b> $\frac{32}{812175}$   | <b>27.</b> $\frac{1}{3}$                               | <b>29.</b> $\frac{18}{35}$   | <b>30.</b> $\frac{1000}{9139}$                       |
| <b>31.</b> $\frac{9}{10}$  | <b>32.</b> $\frac{39}{80}$                             | <b>33.</b> $\frac{3}{4}$   | <b>34.</b> $\frac{25}{54}$                           |
| <b>35.</b> $\frac{34}{57}$   | <b>36.</b> $\frac{16}{21}, \frac{4}{21}, \frac{1}{21}$ | <b>37.</b> $\frac{9}{10}$  | <b>38.</b> Independent                               |
| <b>39.</b> $\frac{23}{120}$  | <b>40.</b> 0.375                                       | <b>41.</b> 0.037   | <b>42.</b> $\frac{11}{18}$                           |
| <b>43.</b> $\frac{209}{343}$   | <b>44.</b> $\frac{61}{90}$                             | <b>45.</b> $\frac{8}{663}$   | <b>46.</b> (i) 0.24; (ii) 0.76                       |
| <b>47.</b> $\frac{4}{91}$  | <b>48.</b> $\frac{33}{182}$                            | <b>49.</b> (i) $\frac{1}{4}$ ; (ii) $\frac{5}{24}$ ; (iii) $\frac{13}{24}$ |  |
| <b>50.</b> $\frac{37}{120}$  | <b>51.</b> $\frac{15}{112}$                            | <b>52.</b> $\frac{1}{7}$   | <b>53.</b> 0.8645                                    |
| <b>54.</b> $\frac{1}{2}$   | <b>55.</b> 42%   | <b>56.</b> $\frac{1}{2}$   |  |
| <b>57.</b> (i) Independent; (ii) Independent; (iii) Independent                            |  | <b>59.</b> $\frac{2}{25}$  | <b>60.</b> $\frac{10}{21}$                           |

## 10.12 BAYES' THEOREM

**Bayes' Theorem is based on the concept that probabilities should be revised when some new information is available.** The idea of revising the probabilities is used by all of us in daily life even though we may not be knowing anything about probability.

**Illustration 7 :** A student while going to a college may start without a rain coat, but as soon as he comes out of his home and sees a thick mass of clouds in the sky he may decide to take a rain coat with him. Thus, he has revised his earlier decision of going to a college without a rain coat. In the same way **the probabilities are revised as soon as some new information is available about the problem concerned.**

*The necessity of revising probabilities arises from a need to make better use of available information and, thereby reduce the element of risk involved in decision-making.* The idea of revising probabilities on the bases of new information was given by British Mathematician Thomas Bayes (1702-61). His theorem known as **Bayes' Theorem or Rule**, was actually published in 1763 after his death in the form of a small paper.

*Bayes' Theorem offers a powerful statistical method of evaluating new information and revising our prior estimates (based on the limited information) of the probability. Modern decision theory is often called Bayesian Decision Theory.*

## 10.13 PRIOR PROBABILITIES (OR PRIORI) AND POSTERIOR PROBABILITIES

### Prior Probabilities or Priori

*Probabilities before revision, by Bayes' Rule, are called Prior probabilities or simply Priori, because they are determined before the sample (or new) information is taken into account.*

### Posterior Probabilities

*The posterior probability is the revision of probability with added information. A probability which has been revised in the light of sample information (by Bayes' Rule) is called a Posterior probability since it represents a probability calculated after the additional information is taken into account.* Posterior probabilities are also called **Revised Probabilities** because they are obtained by revising the prior probabilities in the light of the additional information gained. *Posterior probabilities are always conditional probabilities, the conditional event being the sample information.* Thus, *a prior probability which is unconditional probability becomes a posterior probability, which is a conditional probability, by using Bayes' Rule.*

## 10.14 INVERSE PROBABILITY

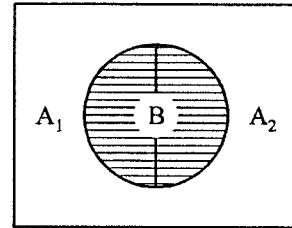
The concept of conditional probability discussed in earlier section takes into account the information about the occurrence of one event to predict the probability of another event. In other words, in conditional probability we studied only those problems on probability in which our knowledge of *factors affecting the event was sufficient to enable us to determine the chances of happening of the event. But now we shall discuss the problems of reverse nature.* By this we mean that if an event has happened as a result of one of the several causes, we may be interested to find out the probability of a particular cause which really affected the event to happen. The problem of such a type are called the problems of *Inverse probability.*

**Illustration 1 :** Suppose a bag  $A$  contains 5 white and 3 green balls. Another bag  $B$  contains 4 white and 6 green balls. A white ball has been drawn from one of the two bags. A white ball must have been drawn either from the bag  $A$  or from bag  $B$ . To find out the probability that it came from the bag  $B$ , (or the white ball has come from bag  $A$ ) is an example of **inverse probability**.

In order to study these types of problems, British Mathematician **Thomas Bayes** (1702-61), gave a theorem known as **Bayes' theorem**. It is also known as **inverse probability theorem**.

**Illustration 2 :** Consider a sample output of 5 defectives in 100 trials (event  $A$ ) might be used to estimate the probability that a machine is not working correctly (event  $B$ ).

Let  $A_1$  and  $A_2$  be the set of events which are mutually exclusive (the two events cannot occur together) and exhaustive (the combination of the two events is the entire experiment), i.e.,  $A = A_1 \cup A_2$  and  $A_1 \cap A_2 = \emptyset$ .



Let  $B$  = a simple event which intersects each of the events  $A_1$  and  $A_2$  as shown in the figure. We observe the following:

- (i) The part of  $B$  which is in  $A_1$  represents “the area ( $A_1$  and  $B$ )” or  $A_1 \cap B$ .
  - (ii) The part of  $B$  which is in  $A_2$ , represents the “the area ( $A_2$  and  $B$ )” or  $A_2 \cap B$ .
  - (iii) The event  $B = (A_1 \cap B) \cup (A_2 \cap B)$
- $$\therefore P(B) = P[(A_1 \cap B) \cup (A_2 \cap B)] \\ = P(A_1 \cap B) + P(A_2 \cap B) = P(A_1 \text{ and } B) + P(A_2 \text{ and } B)$$
- or  $P(B) = P(A_1) P(B/A_1) + P(A_2) P(B/A_2)$ . Also  $P(B) \neq 0$ .

Then the probability of event  $A_1$  given the event  $B$ , is

$$P(A_1/B) = \frac{P(A_1 \text{ and } B)}{P(B)} = \frac{P(A_1 \cap B)}{P(B)}$$

Similarly, the probability of event  $A_2$ , given the  $B$ , is

$$P(A_2/B) = \frac{P(A_2 \text{ and } B)}{P(B)} = \frac{P(A_2 \cap B)}{P(B)}$$

The conditional probabilities  $P(A_1/B)$  and  $P(A_2/B)$  (which are computed after getting the sample information  $B$ ) are called posterior (or inverse) probabilities of event  $A_1$  and  $A_2$ .

## 10.15 STATEMENT OF BAYES' THEOREM

**Theorem :** Let  $E_1, E_2, \dots, E_n$  be the set of  $n$  mutually exclusive and exhaustive events whose union is the random sample space  $S$ , of an experiment. If  $A$  be any arbitrary event of the sample space of the above experiment with  $P(A) \neq 0$ , then the probability of the event  $E_i$ , when the event  $A$  has actually occurred is given by  $P(E_i/A)$ , where

$$P(E_i/A) = \frac{P(A \cap E_i)}{P(A \cap E_1) + P(A \cap E_2) + \dots + P(A \cap E_n)} \quad (\text{First Form})$$

We know that

$$P(A \cap E_i) = P(E_i) P(A/E_i)$$

$$\therefore P(E_i/A) = \frac{P(E_i) P(A/E_i)}{\sum P(E_i) P(A/E_i)} \quad (\text{Second Form})$$

In this theorem, we come across three types of probabilities  $P(E_i)$ ,  $P(E_i/A)$  and  $P(A/E_i)$ . We shall explain each one of them in the following manner:

$P(E_i)$  = It is the probability associated with the happening of the event  $E_i$ ,  $i = 1, 2, \dots, n$ ; such that  $P(E_1) + P(E_2) + \dots + P_n(E_n) = 1$ .  $P(E_i)$  are known as Prior or a Prior probabilities.

$P(A/E_i)$  = It is the conditional probability of the event  $A$  given that  $E_i$  has already occurred.

$P(E_i/A)$  = The probabilities  $P(E_i/A)$  are called the Posterior probabilities. The information that  $A$  has occurred allow us to re-assess the probability  $P(E_i)$  assigned to the event  $E_i$ .

**Example 49 :** A new pregnancy test was given to 100 pregnant women and 100 non-pregnant women. The test indicated pregnancy of 92 of 100 pregnant and to 12 of the 100 non-pregnant women. If a randomly selected woman takes this test and the test indicates that she is pregnant. What is the probability that she was not pregnant?

**Solution :** Let  $E_1$  : the event that a pregnant women is selected.

$E_2$  : the event that a non-pregnant women is selected.

$A$  : the event that test shows that the selected woman is pregnant.

$$\therefore P(E_1) = \frac{100}{200} = \frac{1}{2}, \quad P(E_2) = \frac{100}{200} = \frac{1}{2}; \quad P(A/E_1) = \frac{92}{100}, \quad P(A/E_2) = \frac{12}{100}.$$

We are interested in finding  $P(E_2/A)$

$$\begin{aligned} \therefore P(E_2/A) &= \frac{P(E_2) P(A/E_2)}{P(E_1) P(A/E_1) + P(E_2) P(A/E_2)} \\ &= \frac{(1/2)(12/100)}{(1/2)(92/100) + (1/2)(12/100)} = \frac{12}{104} = \frac{3}{26}. \end{aligned}$$

**Example 50 :** There are 4 boys and 2 girls in room A and 5 boys and 3 girls in room B. A girl from one of the two rooms laughed loudly. What is the probability that the girl who laughed was from room B.

**Solution :** Let

$E_1$  : the event that the girl laughed from room A.

$E_2$  : the event that the girl laughed from room B.

$$P(E_1) = \frac{1}{2}; \quad P(E_2) = \frac{1}{2}$$

Let  $G$  denote the event that the girl laughed

$P(G/E_1)$  = Probability that the girl selected belongs to room  $A$

$$\therefore P(G/E_1) = \frac{2}{6} = \frac{1}{3}.$$

Similarly,  $P(G/E_2) = \frac{8}{3}$ .

We are interested in finding the probability

$$\begin{aligned} P(E_2/G) &= \frac{P(E_2) P(G/E_2)}{P(E_1) P(G/E_1) + P(E_2) P(G/E_2)} \\ &= \frac{(1/2) \times (3/8)}{(1/2) \times (1/3) + (1/2) \times (3/8)} = \frac{(3/16)}{(17/48)} = \frac{9}{17}. \end{aligned}$$

### TREE DIAGRAM

We can also solve this problem by a Tree diagram of the following type:

**Event                  Probability**

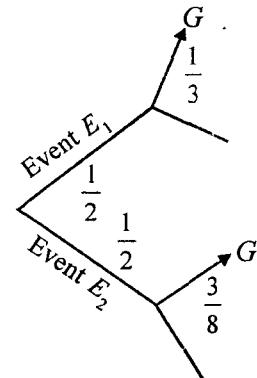
$$E_1 \cap G \quad P(E_1 \cap G) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}.$$

$$E_2 \cap G \quad P(E_2 \cap G) = \frac{1}{2} \times \frac{3}{8} = \frac{3}{16}.$$

$$\text{Total} = P(E_1 \cap G) + P(E_2 \cap G)$$

$$\begin{aligned} \text{Also } P(G) &= P(E_1 \cap G) + P(E_2 \cap G) \\ &= \frac{1}{6} \times \frac{3}{16} = \frac{17}{48}. \end{aligned}$$

$$P(G/E_2) = \frac{P(E_2 \cap G)}{P(G)} = \frac{(3/16)}{(17/48)} = \frac{9}{17}.$$



Hence, the probability that the girl who laughed was from room  $B$  is  $\frac{9}{17}$ .

**Example 51 :** The probability of  $X$ ,  $Y$  and  $Z$  becoming managers are  $4/9$ ;  $2/9$  and  $1/3$  respectively. The probabilities that the Bonus scheme will be introduced if  $X$ ,  $Y$  and  $Z$  becomes managers are  $3/10$ ,  $1/2$  and  $4/5$  respectively.

(a) What is the probability that the bonus scheme will be introduced?

(b) If the bonus scheme has been introduced, what is the probability that the manager appointed was  $X$ ?

**Solution :** Let,

$P(X)$  = Probability that  $X$  becomes manager

$P(Y)$  = Probability that  $Y$  becomes manager

$P(Z)$  = Probability of  $Z$  becomes manager.

Let  $P(B/X)$  = Probability that Bonus scheme is introduced when  $X$  becomes manager. Similarly, we can define  $P(B/Y)$  and  $P(B/Z)$ .

We are given       $P(X) = \frac{4}{9}$ ;     $P(Y) = \frac{2}{9}$ ;     $P(Z) = \frac{1}{3}$

Also,                 $P(B/X) = \frac{3}{10}$ ;     $P(B/Y) = \frac{1}{2}$ ;     $P(B/Z) = \frac{4}{5}$ .

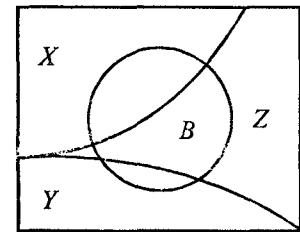
Let  $B$  = denote the event that the bonus scheme is introduced.

(a) Required probability  $P(B) = P(B \cap X) + P(B \cap Y) + P(B \cap Z)$

$$\begin{aligned} &= P(X) P(B/X) + P(Y) P(B/Y) + P(Z) P(B/Z) \\ &= \frac{4}{9} \times \frac{3}{10} + \frac{2}{9} \times \frac{1}{2} + \frac{1}{3} \times \frac{4}{5} = \frac{12 + 10 + 24}{90} = \frac{46}{90} = \frac{23}{45}. \end{aligned}$$

(b) Using Bayes' theorem, the required probability is:

$$\begin{aligned} P(X/B) &= \frac{P(X \cap B)}{P(B)} = \frac{P(X) \cdot P(B/X)}{P(B)} \\ &= \frac{4/9 \times 3/10}{23/45} = \frac{12}{90} \times \frac{45}{23} = \frac{6}{23}. \end{aligned}$$



[Note :

$$B = X \cap B + Y \cap B + Z \cap B$$

∴

$$P(B) = P(X \cap B) + P(Y \cap B) + P(Z \cap B)$$

$$= P(X) P(B/X) + P(Y) P(B/Y) + P(Z) P(B/Z)]$$

**Example 52 :** A factory has two machines  $A$  and  $B$ . Past records show that machine  $A$  produces 30% of the total output and machine  $B$  the remaining 70%. Machine  $A$  produces 5% defective articles and machine  $B$  produces 1% defective items. An item is drawn at random and found to be defective. What is the probability that it was produced (a) by machine  $A$ , (b) by machine  $B$ ?

**Solution :** Let  $D$  denote the item drawn be defective.

(i) Probability that the item came from machine  $A$  =  $P(A_1)$  :  $P_1 = \frac{30}{100}$ .

(ii) Probability that machine  $A$  produces a defective item :  $P(D/A) = P_1 p_1 = \frac{5}{100}$ .

(iii) Joint probability of the two events:

$$P(A \cap D) = P(A) P(D/A) = P_1 p_1 = \frac{30}{100} \times \frac{5}{100} = \frac{150}{10,000}.$$

(iv) Probability that the item came from machine  $B$  =  $P(B)$  :  $P_2 = \frac{70}{100}$ .

(v) Probability that machine  $B$  produces a defective item :  $P(D/B) = p_2 = \frac{1}{100}$ .

(vi) Joint probability of the two events :

$$P(B \cap D) = P(B) P(D/B) = P_2 p_2 = \frac{70}{100} \times \frac{1}{100} = \frac{70}{10000}.$$

$$\therefore P(D) = P(A \cap D) + P(B \cap D) = \frac{150}{10000} + \frac{70}{10000} = \frac{220}{10000}.$$

Now, the probability that the defective item came from machine A:

$$\begin{aligned} P(A/D) &= \frac{P_1 p_1}{P_1 p_1 + P_2 p_2} \\ &= \frac{(150 / 10,000)}{(150 / 10,000) + (70 / 10,000)} = \frac{150}{220} = 0.682. \end{aligned}$$

The probability that the defective machine B:

$$\begin{aligned} P(B/D) &= \frac{P(B \cap D)}{P(D)} = \frac{P_2 p_2}{P_1 p_1 + P_2 p_2} \\ &= \frac{(70 / 10,000)}{(150 / 10,000) + (70 / 10,000)} = \frac{70}{220} = 0.318. \end{aligned}$$

### ALTERNATE METHOD

#### Tree Diagram Event Probability

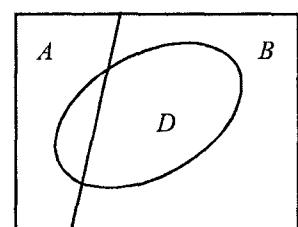
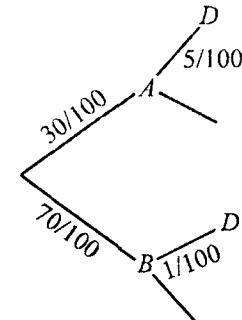
$$A \cap D \quad P(A \cap D) = \frac{30}{100} \times \frac{5}{100} = \frac{150}{1000}.$$

$$B \cap D \quad P(B \cap D) = \frac{70}{100} \times \frac{1}{100} = \frac{70}{10000}$$

$$\therefore P(D) = P(A \cap D) + P(B \cap D) \\ = \frac{70}{10000} + \frac{150}{10000} = \frac{220}{10000} = \frac{22}{1000}.$$

$$\begin{aligned} P(A/D) &= \frac{P(A \cap D)}{P(D)} \\ &= \frac{(150 / 10000)}{(220 / 10000)} = \frac{150}{220} = 0.682. \end{aligned}$$

$$\begin{aligned} P(B/D) &= \frac{P(B \cap D)}{P(D)} \\ &= \frac{(70 / 10000)}{(220 / 10000)} = \frac{70}{220} = 0.318. \end{aligned}$$



Hence, (i) The probability that the defective item came from machine A = 0.682.

(ii) The probability that the defective item came from machine B = 0.318.

**Example 53 :** A company has two plants to manufacture scooters. Plant I manufactures 80% of the scooters and Plant II manufacture, 20%. At plant I; 85 out of 100 scooters are rated standard quality or better. At Plant II only 65 out of 100 scooters are rated standard quality or better.

- What is the probability that a scooter selected at random came from plant I if it is known that the scooter is of standard quality?
- What is the probability the scooter came from plant II if it is known that the scooter is of standard quality.

**Solution :** Let  $P_1$  = Probability that the scooter is manufactured by plant I = 0.80.

$P_2$  = Probability that the scooter is manufactured by Plant II = 0.20.

Let the probabilities of standard quality scooter coming from plant I and II respectively be  $p_1$  and  $p_2$ .

$$\therefore p_1 = 0.85, \quad p_2 = 0.65.$$

Let  $E$  = denote the event that the scooter is of standard quality.

The joint probabilities of standard quality scooter coming from Plant I and Plant II are

$$\text{Plant I : } P_1 p_1 = 0.80 \times 0.85 = 0.68.$$

$$\text{Plant II : } P_2 p_2 = 0.20 \times 0.65 = 0.13.$$

$$\text{Probability of the event } E : P(E) = P_1 p_1 + P_2 p_2 = 0.68 + 0.13 = 0.81.$$

- The probability that the standard quality scooter came from plant I would be:

$$P(\text{Plant I}/E) = \frac{P_1 p_1}{(P_1 p_1) + (P_2 p_2)} = \frac{P_1 p_1}{P(E)} = \frac{0.68}{0.81} = 0.84.$$

- The probability that the standard quality scooter came from Plant II would be:

$$P(\text{Plant II}/E) = \frac{P_2 p_2}{(P_1 p_1) + (P_2 p_2)} = \frac{P_2 p_2}{P(E)} = \frac{0.13}{0.81} = 0.16.$$

**Example 54 :** A factory produces a certain type of output by three types of machines. The respective daily production figures are Machine I : 3,000 units, Machine II : 2,500 units and Machine III : 4,500 units.

Post experience shows that 1% of the output produced by Machine I is defective. The corresponding fraction of defectives for the other two machines are 1.2% and 2% respectively. An item is drawn at random from the day's production run and is found to be defective. What is the probability that it comes from the output of (a) Machine I, (b) Machine II and (c) Machine III?

**Solution :** Let  $A$  = the event that an item is produced by Machine I

$B$  = the event an item is produced by Machine II

$C$  = the event that an item is produced by Machine III.

Then  $P(A) = \frac{3000}{3000 + 2500 + 4500} = \frac{3000}{10000} = \frac{3}{10} = 0.3;$

$$P(B) = \frac{2500}{10000} = 0.25 ; \quad P(C) = \frac{4500}{10000} = 0.45.$$

Now, let  $D$  = the event that an item drawn is defective.

Then, the conditional probabilities are

$$P(D/A) = \frac{1}{100} = 0.01 ; \quad P(D/B) = \frac{1.2}{100} = 0.012 ; \quad P(D/C) = \frac{2}{100} = 0.02.$$

Again, the joint probability that an item is produced by three machines and it is defective:

$$P(A \cap D) = P(A) \cdot P(D/A) = 0.3 \times 0.01 = 0.003$$

$$P(B \cap D) = P(B) \cdot P(D/B) = 0.25 \times 0.012 = 0.003$$

$$P(C \cap D) = P(C) \cdot P(D/C) = 0.45 \times 0.02 = 0.009$$

Also  $P(D) = P(A \cap D) + P(B \cap D) + P(C \cap D)$   
 $= 0.003 + 0.003 + 0.009 = 0.015.$

Now, the revised or posterior probability is obtained by Bayes' Theorem.

$$P(A/D) = \frac{P(A \cap D)}{P(D)} = \frac{P(A \cap D)}{P(A \cap D) + P(B \cap D) + P(C \cap D)} = \frac{0.003}{0.015} = \frac{3}{15} = 0.2.$$

$$P(B/D) = \frac{P(B \cap D)}{P(D)} = \frac{P(B \cap D)}{P(A \cap D) + P(B \cap D) + P(C \cap D)} = \frac{0.003}{0.015} = \frac{3}{15} = 0.2.$$

$$P(C/D) = \frac{P(C \cap D)}{P(D)} = \frac{P(C \cap D)}{P(A \cap D) + P(B \cap D) + P(C \cap D)} = \frac{0.009}{0.015} = \frac{9}{15} = 0.6.$$

Hence, the probability that the defective item comes from (a) Machine I = 0.2 ; from Machine II = 0.2 ; and from Machine III = 0.6.

### ALTERNATE METHOD

#### Tree Diagram Event Probability

$$A \cap D \quad P(A \cap D) = 0.3 \times 0.01 = 0.003.$$

$$B \cap D \quad P(B \cap D) = 0.25 \times 0.012 = 0.003.$$

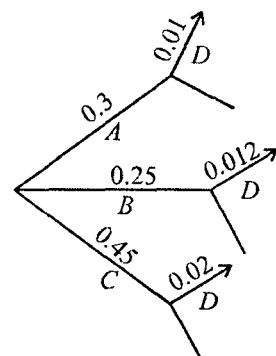
$$C \cap D \quad P(C \cap D) = 0.45 \times 0.02 = 0.009.$$

$$P(A/D) = P(A \cap D) + P(B \cap D) + P(C \cap D) \\ = 0.003 + 0.003 + 0.009 = 0.015.$$

$$P(A/D) = \frac{P(A \cap D)}{P(D)} = \frac{0.003}{0.015} = 0.2,$$

$$P(B/D) = \frac{P(B \cap D)}{P(D)} = \frac{0.003}{0.015} = 0.2,$$

$$P(C/D) = \frac{P(C \cap D)}{P(D)} = \frac{0.009}{0.015} = 0.6.$$



**Example 55 :** Two sets of candidates are competing for the positions on the Board of Directors of a company. The probabilities that the first and the second sets will win are 0.6 and 0.4 respectively. If the first set wins, the probability of introducing a new product is 0.8 and corresponding probability if the second set wins is 0.3. What is the probability that the product will be introduced by the second set?

**Solution :** Let the probabilities of the winning of the first and second sets be  $P_1$  and  $P_2$ .

$$\therefore P_1 = 0.6; \quad P_2 = 0.4$$

Let the probabilities of introducing of new product if set A wins be  $p_1$ ; and if B wins be  $p_2$ .

$$\therefore p_1 = 0.8; \quad p_2 = 0.3$$

**Joint probabilities  $P_1 p_1$  and  $P_2 p_2$  are:**

$$P_1 p_1 = 0.6 \times 0.8 = 0.48$$

$$P_2 p_2 = 0.4 \times 0.3 = 0.12$$

Let  $E$  = denote the event that the product will be introduced.

$$\therefore P(E) = P_1 p_1 + P_2 p_2 = 0.48 + 0.12 = 0.60.$$

$$\therefore P(\text{Second set}/E) = \frac{P_2 p_2}{P(E)} = \frac{P_2 p_2}{P_1 p_1 + P_2 p_2} = \frac{0.12}{0.60} = \frac{1}{5} = 0.2.$$

Hence the probability that the product will be introduced by the second set is 0.2.

**Example 56 :** You note that your officer is happy on 60% of your calls, so you assign a probability of his being happy on your visit as 0.6. You have noticed also that if he is happy, he accedes to your request with a probability of 0.4 whereas if he is not happy, he accedes to the request with a probability of 0.1. You call one day, and he accedes to your request. What is the probability of his being happy?

**Solution :** Let  $H$  be the hypothesis that the officer is happy and  $\bar{H}$  be the hypothesis that the officer is not happy.

$$P(H) = \frac{6}{10}, \quad P(\bar{H}) = \frac{4}{10}$$

Let  $A$  be the event that he accedes to request.

$$(A/H) = 0.4 = \frac{4}{10},$$

$$\text{Also } P(A/\bar{H}) = 0.1 = \frac{1}{10}.$$

$$\therefore P(H/A) = \frac{P(H) \times (A/H)}{P(H) \times P(A/H) + P(\bar{H}) \times P(A/\bar{H})} \quad (\text{Bayes' Theorem})$$

$$= \frac{\frac{6}{10} \times \frac{4}{10}}{\frac{6}{10} \times \frac{4}{10} + \frac{4}{10} \times \frac{1}{10}} = \frac{24 \times 100}{100 \times 28} = \frac{24}{28} = \frac{6}{7} = 0.857.$$

**Example 57 :** A certain production process produces items that are 10 per cent defective. Each item is inspected before being supplied to customers but the inspector incorrectly classifies an item 10 per cent of the time. Only items classified as good are supplied. If 820 items in all have been supplied, how many of them are expected to be defective?

**Solution :** Let  $P(D)$  = Probability of defective item =  $\frac{10}{100} = 0.1$ .

$$P(G) = \text{Probability of good item} = 1 - P(D) = 1 - 0.2 = 0.9.$$

$$P(G/D) = P(\text{classified as good when it is defective}) = \frac{10}{100} = 0.1.$$

$$P(G/G) = P(\text{classified as good when it is good}) = 1 - P(G/D) = 1 - 0.1 = 0.9.$$

**P (Defective/classified as good when actually it is defective)**

$$\begin{aligned} &= \frac{P(D) \times P(G/D)}{P(D) \times P(G/D) + P(G) \times P(G/G)} \\ &= \frac{0.1 \times 0.1}{0.1 \times 0.1 + 0.9 \times 0.9} = \frac{0.01}{0.82} = \frac{1}{82}. \end{aligned}$$

$$\therefore \text{Expected number of defective items out of 820 items} = 820 \times \frac{1}{82} = 10.$$

## EXERCISE 10.2

1. A coin is tossed. It turns up heads, two balls will be drawn from urn  $A$ ; otherwise two balls will be drawn from urn  $B$ . Urn  $A$  contains three black and five white balls. Urn  $B$  contains seven black and one white ball. In both cases, selections are to be made with replacement. What is the probability that urn  $A$  is used given that both the balls drawn are black?
2. Assume that a factory has two machines. Past records show that machine  $A$  produces 30% of the items of output and machine  $B$  produces 70% of the items. Further 5% of the items produced by machine  $A$  were defective and only 10% produced by machine  $B$  were defective. If a defective item is drawn at random, what is the probability that it was produced by machine  $B$ ?
3. A factory has three units,  $A$ ,  $B$  and  $C$ .  $A$  produces 25% of its product, unit  $B$  produces 25% and unit  $C$  produces 50%. If the percentages of defective items produced by three units  $A$ ,  $B$  and  $C$  are respectively 10%, 20% and 3% and an item selected randomly from the total products of the factory is found to be defective. What is the probability that it is produced by the unit  $C$ ?
4. A box  $A$  contains 1 red and 2 white marbles and another box  $B$  contains 3 red and 2 white marbles. One marble is drawn at random from one of the boxes and it is found to be red. Find the probability that it was drawn from box  $B$ .
5. In a bulb factory, machines  $A$ ,  $B$  and  $C$  manufacture 60%, 25% and 15% respectively. Of the total of their output 1%, 2% and 1% are defective bulbs. A bulb is drawn at random from the

total production and found to be defective. From which machine, the defective bulb is expected to have been manufactured?

6. A company manufactures scooters at two plants  $A$  and  $B$ . Plant  $A$  produces 80% and plant  $B$  produces 20% of the total product, 85% of the scooters produced at plant  $A$  are standard in quality, while 65% of the scooters produced at plant  $B$  are standard in quality. A scooter produced by the company is selected at random and it is found to be standard in quality. What is the probability that it was manufactured at plant  $A$ ?

[Hint : Let  $E_1$  and  $E_2$  be the event of the scooter being manufactured at plants  $A$  and  $B$  respectively. Let  $S$  be the event of getting a scooter of standard quality.

$$\text{Then, } P(E_1) = \frac{4}{5}; \quad P(E_2) = \frac{1}{5}; \quad P(S/E_1) = \frac{17}{20} \quad \text{and} \quad P(S/E_2) = \frac{13}{20}$$

$$\begin{aligned}\therefore P(E_1/S) &= \frac{P(S/E_1) \cdot P(E_1)}{P(S/E_1) \cdot P(E_1) + P(S/E_2) \cdot P(E_2)} \\ &= \frac{(4/5) \times (17/20)}{(4/5) \times (17/20) + (1/5) \times (13/20)} = \frac{68}{81}.\end{aligned}$$

7. There are 2000 scooter drivers, 4000 car drivers and 6000 truck drivers all insured. The probabilities of an accident involving a scooter, a car, a truck are 0.01, 0.03 and 0.15 respectively. One of the insured drives met with an accident. What is the probability he is a scooter driver?

$$[\text{Hint : } P(S) = \frac{2000}{2000 + 4000 + 6000} = \frac{1}{6}; \quad P(C) = \frac{4000}{12000} = \frac{1}{3}; \quad P(T) = \frac{6000}{12000} = \frac{1}{2}$$

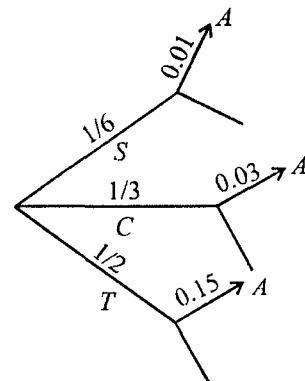
$$P(A \cap S) = P(A) P(S/A) = \frac{1}{6} \times 0.01 = \frac{1}{600}$$

$$P(A \cap C) = P(A) P(C/A) = \frac{1}{3} \times 0.03 = \frac{1}{100}$$

$$P(A \cap T) = P(A) P(T/A) = \frac{1}{2} \times 0.15 = \frac{15}{200}$$

$$\begin{aligned}P(A) &= P(A \cap S) + P(A \cap C) + P(A \cap T) \\ &= \frac{1}{600} + \frac{1}{100} + \frac{15}{200} = \frac{52}{600}\end{aligned}$$

$$P(S/A) = \frac{P(A \cap S)}{P(A)} = \frac{(1/600)}{(52/600)} = \frac{1}{52}.$$



8. A letter is known to have come either from TATA NAGAR or CALCUTTA. On the envelop just two consecutive letters TA are visible. Show that the probability that the letter have come from CALCUTTA is  $4/11$ .

[Hint : Let  $E_1$  : the event that letter came from TATA NAGAR  
 $E_2$  : the event that the letter came from CALCUTTA

$\therefore A$  : the event that two consecutive visible letters on the envelop are TA

$$\therefore P(E_1) = \frac{1}{2} = P(E_2); \quad P(A/E_1) = \frac{2}{8} = \frac{1}{4}; \quad P(A/E_2) = \frac{1}{7}$$

$$P(E_2/A) = \frac{(1/2) \times (1/7)}{(1/2) \times (1/4) + (1/2) \times (1/7)} = \frac{4}{11}.$$

9. By examining the chest X-ray probability that T.B. is detected when a person is actually suffering from T.B. is 0.99. The probability that the doctor diagnose incorrectly that a person has a T.B. A person selected at random is diagnosed to have T.B. What is the chance that he actual has T.B.?
10. In a competition an examine either guesses or copies or knows the answer to a multiple choice question with four choices. The probability that he makes a guess is  $1/3$  and the probability that he copies the answer is  $1/6$ . The probability that his answer is correct, given that he copied it, is  $1/8$ . Find the probability that he knew the answer to the question, given that he correctly answered it.

[Hint : Let  $G$ ,  $C$ ,  $K$  respectively denote the event that he guesses, copies and knows the answer to question. Let  $H$  denote the event that his answer is correct.

$$\therefore P(G) = 1/3, \quad P(C) = 1/6, \quad P(K) = 1 - \frac{1}{3} - \frac{1}{6} = \frac{1}{2}.$$

$$P(H/G) = 1/4 \quad (\because \text{only one choice is correct of four choices})$$

$$P(H/C) = 1/8, \quad P(G/K) = 1.$$

$$\therefore P(K/H) = \frac{(1/2) \times 1}{(1/2) \times 1 + (1/3)(1/4) + (1/6)(1/8)} = \frac{24}{29}.$$

11. T.V. components are produced by two machines  $A$  and  $B$ . 50% of the components are produced by machine  $A$  with an essential of 10% of them being defective. On machine  $B$ , 20% of the components produced are defective. If a component taken at random is found to be defective, what is the probability that the component was produced by machine  $A$ ?
12. Suppose that a product is produced in three factories  $A$ ,  $B$  and  $C$ . It is known that factory  $A$  produces twice as many items as factory  $B$ ; and that factories  $B$  and  $C$  produce the same number of products. Assume that it is known that 2 per cent of the items produced by each of the factories  $A$  and  $B$  are defective while 4 per cent of those manufactured by factory  $C$  are defective. All the items produced in the three factories are stocked and an item of product is selected at random. What is the probability that this product is defective?

[Hint : Let the number of items produced by each of the factories  $B$  and  $C$  be  $n$ . Then the number of items produced by the factory  $A$  is  $2n$ . Let  $A_1$ ,  $A_2$  and  $A_3$  denote the event that the item is produced by factory  $A$ ,  $B$  and  $C$  respectively and let  $E$  be the event of the item being defective. Then we have

$$\therefore P(A_1) = \frac{2n}{2n+n+n} = \frac{2n}{4n} = 0.5, \quad P(A_2) = \frac{n}{4n} = 0.25, \quad P(A_3) = \frac{n}{4n} = 0.25.$$

$$P(E/A_1) = P(E/A_2) = 0.02 \text{ and } P(E/A_3) = 0.04 \text{ (Given)}$$

The probability that an item selected at random from the stock is defective is given by

$$\begin{aligned} P(E) &= P(A_1 \cap E) + P(A_2 \cap E) + P(A_3 \cap E) \\ &= P(A_1) P(E/A_1) + P(A_2) P(E/A_2) + P(A_3) P(E/A_3) = 0.07. \end{aligned}$$

13.  $A, B$  and  $C$  are bidding on a contract for the construction of a bridge. The probabilities that  $A, B, C$  will get the contract are 0.5, 0.3 and 0.2 respectively. If  $A$  gets it, he will select  $E$  as the sub-contractor with probability 0.8. If  $B$  or  $C$  gets it,  $E$  will be chosen with probabilities 0.4 and 0.1 respectively.  $E$  gets the sub-contract. What is the probability that  $E$  gets the sub-contract when  $A$  is selected?
14. An insurance company insured 1500 scooter drivers, 3,500 car drivers and 5,000 truck drivers. The probability of an accident is 0.05, 0.02 and 0.10 respectively in case of scooter, car and truck drivers. One of the insured persons meets an accident. What is the probability that he is a car driver?
15. A factory has a machine shop in which three machines  $A, B$  and  $C$  produce 100 cm. aluminium tubes. An inspector is equally likely to sample tubes from  $A$  and  $B$ , and three times as likely to select tubes from  $C$  as he is from  $B$ . The defective rates from the three machines are:  $A = 10\%$ ,  $B = 10\%$  and  $C = 20\%$ . What is the probability that a tube selected by the inspector.
  - (i) is from machine  $A$ ?
  - (ii) is defective?
  - (iii) comes from machine  $A$ , given that it is defective?
16. Three machines  $A, B, C$  produce respectively 50%, 30% and 20% of the total number of items of a factory. The percentages of defective outputs of these machines are respectively 3%, 4% and 5%. If an item is selected at random and is found to be defective. What is the probability that this item is from machine  $A$ ?
17. Suppose that one of three men, a politician, a businessman, and an educator will be appointed as the vice-chancellor of a university. The respective probabilities of their appointments are 0.50, 0.30, 0.20. The probabilities that research activities will be promoted by these people if they appointed are 0.30, 0.70 and 0.80 respectively. A research activity is promoted. What is the probability that research activity is promoted by the politician vice-chancellor?
18. There are three urns having the following composition and white balls:

Urn I :              7 white and 3 black balls.

Urn II :              4 white and 6 black balls

Urn III :              2 white and 8 black balls

One of the Urns is chosen at random with probabilities 0.2, 0.6 and 0.2 respectively. From the chosen Urn, two balls are drawn at random without replacement. Both the balls happen to be white. Calculate probability that the balls drawn were from Urn III.

19. A factory produces a certain type of output by three types of machines. The respective daily production figures are Machine I: 3000 units, Machine II: 2500 units, Machine III: 4500 units. Past experience shows that 3 per cent of the output produced by Machine I is defective. The corresponding fraction of defectives, for the other two machines are

respectively, 1.2 per cent and 2 per cent respectively. An item is drawn at random from the day's production and is found to be defective. What is the probability that it comes from the output of (a) Machine I, (b) Machine II and (c) Machine III?

20. The chance that a female worker in a chemical factory will contract an occupational disease is 0.04 and the chance for a male workers is 0.06. Out of 1,000 workers in a factory 200 are females. One worker is selected at random and the worker is found to have contracted the disease. What is the probability that the worker is a female?
21. Suppose that there are three Urns containing 2 white and 3 black balls; 2 white and 2 black balls and 4 white and one black balls respectively. There is equal probability of each Urn being taken. One ball is drawn from an urn at random. We are told that a white ball has been drawn. Find the probability that it was drawn from the first urn.
22. A factory has two machines  $A$  and  $B$ . Past record shows that machine  $A$  produced 60% of the items of output and machine  $B$  produce 40% of the items of output. Further 20% of items produced by machine  $A$  were defective and 1% produced by machine  $B$  were defective. If a defective item is drawn at random. What is the probability that it was produced by machine  $A$ .

[Hint : Let  $D$  be defective events,

$$\text{Here, } P(A) = 0.6 ; P(B) = 0.4 ; P(D/A) = 0.02, P(D/B) = 0.01$$

$$\therefore P(A/D) = \frac{P(A) P(D/A)}{P(A) P(D/A) + P(B) P(D/B)} = \frac{0.6 \times 0.02}{0.6 \times 0.02 + 0.4 \times 0.01} = \frac{3}{4}.$$

23. A car manufacturing factory has two plants. Plant  $A$  manufacture 70% of cars and Plant  $B$  manufacture 30%. At plant  $A$ , 80% of cars are rated of standard quality and at plant  $B$ , 90% of cars are rated of standard quality. A car is pieced up at random and is found to be of standard quality. What is the probability that it has come from plant  $A$ ?

[Hint : Let  $E$  be the event that cars is of standard quality.

$$P(A) = 0.7 ; P(B) = 0.3 ; P(E/A) = 0.8, P(E/B) = 0.9.$$

$$\therefore P(A/D) = \frac{P(A) P(D/A)}{P(A) P(D/A) + P(B) P(D/B)} = \frac{0.7 \times 0.8}{0.7 \times 0.8 + 0.3 \times 0.9} = \frac{56}{83}.$$

24. Suppose 5 men out of 100 and 25 women out of 1000 are good orators. An orator is chosen at random. Find the probability that a male person is selected. Assume that there are equal number of men and women.

[Hint : Let  $E$  be the event of being an orator.

$$P(M) = \frac{1}{2} ; P(W) = \frac{1}{2} ; P(E/M) = \frac{5}{100}, P(E/W) = \frac{25}{1000}$$

$$\therefore P(M/E) = \frac{P(M) P(E/M)}{P(M) P(E/M) + P(W) P(E/W)} = \frac{\frac{1}{2} \times \frac{5}{100}}{\frac{1}{2} \times \frac{5}{100} + \frac{1}{2} \times \frac{25}{1000}} = \frac{2}{3}.$$

25. A bag contains 7 white balls and 5 red balls, of which 3 white balls and 3 red balls marked  $A$ , and 4 white balls and 2 red balls marked  $B$ . A white ball is drawn randomly from the bag. Find the probability that the white ball drawn is (i) marked  $A$  and (ii) also marked  $B$ .

26. Suppose that  $A$  is known to tell the truth in five cases out of six and he states that a white ball was drawn from a bag containing 9 black and one white ball. What is the probability that the white ball was really drawn?

[Hint : The probability that a white ball is drawn in any case is  $1/10$ . Also the probability that the white ball was drawn and that  $A$  told the truth is  $1/10 \times 5/6$ .

Furthermore, the probability that a black ball was drawn and  $A$  told a lie about it is  $9/10 \times 1/6$ . Hence the probability that a white ball was drawn is

$$= \frac{(1/10) \times (5/6)}{(1/10) \times (5/6) + (9/10) \times (1/6)} = \frac{5}{14}.$$

27. A bag contains 4 white and 2 red balls. Another bag contains 1 white and 3 red balls. One ball is transferred from the first bag to the second bag and then one ball is drawn from the second bag. If it is found to be a red ball, what is the probability that the ball transferred is white?

28. A company has two plants to manufacture scooters. Plant I manufactures 70% of the scooters and Plant II manufactures 30%. At Plant I, 80% of the scooters produced are of standard quality and at Plant II, 90% of the scooters produced are of standard quality. A scooter is picked at random and found to be of standard quality. What is the chance that it has come from Plant II.

29. A firm produces pipes in two Plants I and II with daily production of 500 and 1000 pipes respectively. Plant I produces 5% defective pipes and Plant II produces 8% defective pipes. A defective pipe is selected at random from the total production, what is the chance that the pipe is produced by (i) Plant I or (ii) Plant II?

30. A pack of playing cards was found to contain only 51 cards. If the first 13 cards which are examined are all red, what is the probability that the missing card is black?

[Hint : Let  $A_1, A_2$  be the events that black and red card is lost respectively. Let  $B$  denote the occurrence of first 13 cards.

$$P(A_1) = \frac{1}{2}, \quad P(A_2) = \frac{1}{2}, \quad P(B/A_1) = \frac{^{26}C_{13}}{^{51}C_{13}}; \quad P(B/A_2) = \frac{^{25}C_{13}}{^{51}C_{13}}$$

$$P(A_1/B) = \frac{P(A_1) P(B/A_1)}{P(A_1) P(B/A_1) + P(A_2) P(B/A_2)} = \frac{(1/2) \cdot ^{26}C_{13}}{(1/2) [^{26}C_{13} + ^{25}C_{13}]} = \frac{2}{3}.$$

31. State and illustrate Bayes' theorem.

32. Distinguish between a priori and posterior probability.

33. Explain the concept of posterior probability.

34. Explain the significance of Bayes' theorem.

35. A company has three plants to manufacture 8000 scooters in a month. Out of 8000 scooters, Plant I manufactures 4000 scooters, Plant II manufactures 3000 scooters and Plant III manufactures 1000 out of 100 scooters are rated of standard quality or better and at Plant III, 60 out of 100 scooters are rated of standard quality or better. What is the probability that the scooter selected at random came from (i) Plant I, (ii) Plant II and (iii) Plant III, if it is known that the scooter is of standard quality.

[Hint : Let  $S$  be the event that the scooter selected of standard quality.

$$\therefore P(A_1) = \frac{1}{2}, \quad P(A_2) = \frac{3}{8}; \quad P(A_3) = \frac{1}{8}$$

$$P(S/A_1) = \frac{85}{100} = 0.85, \quad P(S/A_2) = \frac{65}{100} = 0.65, \quad P(S/A_3) = 0.60$$

$$\begin{aligned} P(S) &= P(S \cap A_1) + P(S \cap A_2) + P(S \cap A_3) \\ &= \frac{1}{2} \times \frac{85}{100} + \frac{3}{8} \times \frac{65}{100} + \frac{1}{8} \times \frac{60}{100} = \frac{595}{800}. \end{aligned}$$

$$(i) \quad P(A_1/S) = \frac{P(S \cap A_1)}{P(S)} = \frac{(85/100)}{(595/800)} = \frac{340}{595} = \frac{69}{119}.$$

$$(ii) \quad P(A_2/S) = \frac{P(S \cap A_2)}{P(S)} = \frac{(195/100)}{(595/800)} = \frac{39}{119}.$$

$$(iii) \quad P(A_3/S) = \frac{P(S \cap A_3)}{P(S)} = \frac{(60/100)}{(595/800)} = \frac{12}{119}.$$

36. Police plan to enforce speed limits by using radar traps at 4 different locations within the city limits. The radar traps at each of the locations  $L_1, L_2, L_3$  and  $L_4$  are operated 40%, 30%, 20% and 10% of the time and if a person is speeding on his way to work has probabilities 0.2, 0.1, 0.5 and 0.2 respectively, of passing through these locations. What is the probability that he will receive speeding ticket?

[Hint : Let  $A_1, A_2, A_3$  and  $A_4$  be the event that the radar traps will be active at locations  $L_1, L_2, L_3$  and  $L_4$ . Let  $S$  denote the event of speeding ticket.

$$P(A_1) = 0.4, \quad P(A_2) = 0.3, \quad P(A_3) = 0.2, \quad P(A_4) = 0.1$$

$$P(S/A_1) = 0.2, \quad P(S/A_2) = 0.1, \quad P(S/A_3) = 0.5, \quad P(S/A_4) = 0.2$$

$$\begin{aligned} P(S) &= P(S \cap A_1) + P(S \cap A_2) + P(S \cap A_3) + P(S \cap A_4) \\ &= P(A_1) P(S/A_1) + P(A_2) P(S/A_2) + P(A_3) P(S/A_3) + P(A_4) P(S/A_4) \\ &= 0.4 \times 0.2 \times 0.1 + 0.2 \times 0.5 + 0.1 \times 0.2 = 0.23. \end{aligned}$$

**Probability that a person will receive a speeding ticket = 0.23.]**

37. In a railway reservation office, two clerks are engaged in checking reservation forms. On an average, the first clerk checks 55% of the forms, while the second does the remaining. The first clerk has an error rate of 0.03 and second has an error rate of 0.02. A reservation form is selected at random from the total number of forms checked during a day, and is found to have an error. Find the probabilities that it was checked by first and second clerk respectively.

$$P(A_1) = 0.55, \quad P(A_2) = 0.45; \quad P(E/A_1) = 0.03; \quad P(E/A_2) = 0.02.$$

$$P(E) = P(A_1 \cap E) + P(A_2 \cap E) = 0.55 \times 0.03 + 0.45 \times 0.02 = 0.255.$$

$$(i) \quad P(A_1/E) = \frac{P(A_1 \cap E)}{P(E)} = \frac{0.55 \times 0.03}{0.0255} = \frac{11}{17}.$$

$$(ii) \quad P(A_2/E) = \frac{P(A_2 \cap E)}{P(E)} = \frac{0.45 \times 0.02}{0.0255} = \frac{6}{17}.$$

38. A factory has three units  $A$ ,  $B$  and  $C$ .  $A$  produces 25% of its product, unit  $B$  produces 25% and the unit  $C$  produces 50%. If the percentages of defective items produced by three units  $A$ ,  $B$  and  $C$  are respectively 1%, 2% and 3% and an item selected randomly from the total production of the factory is found to be defective, what is the probability that it is produced by the unit  $C$ ?
39. Suppose that a product is produced in three factories  $A$ ,  $B$  and  $C$ . It is known that factory  $A$  produces thrice as many items as factory  $B$ , and that factories  $B$  and  $C$  produce the same number of product. Assume that it is known that 4 per cent of the items produced by each of the factories  $A$  and  $B$  are defective while 5 per cent of those manufactured by factory  $C$  are defective. All the items produced in three factories are stocked, and an item of product is selected at random. What is the probability that this item is defective and is produced by the factory  $A$ ?
40. In a population of workers, suppose 40% are school graduates, 50% are high school graduates and 10% are college graduates. Among the school graduates, 10% are unemployed; among the high school graduates, 5% are unemployed, and among the college graduates 2% are unemployed. If a worker is chosen at random and found to be unemployed, what is the probability that he is a college graduate?
41. A company uses a ‘selling aptitude test’ in the selection of salesman. Past experience has shown that only 70% of all persons applying for a sales position achieved a classification “dissatisfactory” in actual selling, whereas the remainder were classified as “satisfactory”. 80% has scored a passing grade on the aptitude test. Only 25% of those classified unsatisfactory has passed the test on the basis of this information. What is the probability that a candidate would be a satisfactory salesman given that he passed the aptitude test?
- [Hint : Let  $S$  stand for a ‘satisfactory’ classification as a salesman and  $P$  stand for ‘passing the test’. The probability that a candidate would be “satisfactory” salesman given that he passed the aptitude test is:

$$P(S/P) = \frac{(0.70)(0.85)}{(0.70)(0.85) + (0.30)(0.25)} = \frac{0.595}{0.595 + 0.075} = 0.888.$$

The result indicates that the tests are of value in screening candidates. Assuming no change in the type of candidates applying for the selling positions, the probability that a random applicant would be satisfactory is 70%. On the other hand, if the company only accepts an applicant if he passed the test, the probability increases to 0.888.]

42. A manufacturing firm produces units of a product in four plants. Define event  $A_i$ : a unit is produced in plant  $i$ ,  $i = 1, 2, 3, 4$  and event  $B$ : a unit is defective. From the past records of the proportions of defectives produced at each plant the following conditional probabilities are set:

$$P(B/A_1) = 0.05, \quad P(B/A_2) = 0.10, \quad P(B/A_3) = 0.15, \quad P(B/A_4) = 0.02$$

The first plant produces 30 per cent of the units of product, the second plant 25 per cent, third plant 40 per cent and the fourth plant 5 per cent. A unit of the product made at one of these plants is tested and is found to be defective. What is the probability that the unit was produced in plant 3?

43. If a machine is correctly set up it will produce 90% acceptable items. If it is incorrectly set up it will produce 30% acceptable items. Past experience shows that 80% of set ups are correctly done. If after a certain set up, first items produced is acceptable, what is the probability that the machine is correctly set up?

[Hint :  $P(A_1) = 0.80, \quad P(A_2) = 0.20; \quad P(E/A_1) = 0.90, \quad P(E/A_2) = 0.30$ .

$$E = (A_1 \cap E) \cup (A_2 \cap E)$$

$$P(E) = (A_1 \cap E) + (A_2 \cap E) = 0.90 \times 0.80 + 0.30 \times 0.20$$

$$P(A \cap E) = P(A_1) P(E/A_1); \quad P(A_2 \cap E) = P(A_2) P(E/A_2)$$

$$\therefore P(A_1/E) = \frac{P(A_1 \cap E)}{P(E)} = \frac{0.90 \times 0.80}{0.90 \times 0.80 + 0.30 \times 0.20} = \frac{12}{13}.$$

## ANSWERS

1. 0.125

2.  $\frac{7}{22}$

3.  $\frac{2}{3}$

4.  $\frac{9}{14}$

5. Machine A

6.  $\frac{68}{81}$

7.  $\frac{1}{52}$

8.  $\frac{4}{11}$

9.  $\frac{110}{221}$

10.  $\frac{24}{29}$

11.  $\frac{1}{3}$

12. 0.07

13. 0.54

15. (i) 0.02 ; (ii) 0.16 ; (iii) 0.125

16.  $\frac{3}{74}$

17.  $\frac{15}{52}$

18.  $\frac{1}{40}$

19. (a) 0.43 ; (b) 0.14 ; (c) 0.43

20.  $\frac{1}{7}$

21.  $\frac{2}{9}$

22.  $\frac{3}{4}$

23.  $\frac{56}{83}$

24.  $\frac{2}{3}$

25.  $\frac{3}{7}, \frac{4}{7}$

26.  $\frac{5}{14}$

27.  $\frac{3}{5}$

28. 0.325

29.  $P(A_1/E) = \frac{5}{21}; \quad P(A_2/E) = \frac{16}{21}.$

30. 0.656

31. (i)  $\frac{69}{119}$ , (ii)  $\frac{39}{119}$ , (iii)  $\frac{12}{119}$

36. 0.23

37. (i)  $\frac{11}{17}$ , (ii)  $\frac{6}{17}$

38.  $\frac{2}{3}$

39. (i) 0.042, (ii) 0.571

40. 0.03

41. 0.888

42.  $\frac{12}{13}$

44.  $\frac{1}{10}$

45.  $\frac{25}{52}$

# 11

# *Binomial, Poisson and Normal Distributions*

## 11.1 THEORETICAL DISTRIBUTIONS

The distributions which are based on actual data of experiments are called **Observed frequency distribution**. It is sometimes possible, by assuming a certain hypothesis, to derive mathematically the frequency distribution of a certain population. Such distributions are known as **theoretical distributions**. In other words, a **theoretical distribution is the frequency distribution of certain events in which frequencies are obtained by mathematical computations**. Some of the important theoretical distributions are:

1. Binomial distribution
2. Poisson distribution
3. Normal distribution

There are two types of theoretical distribution.

- (i) Discrete distribution. Binomial and Poisson are discrete distribution
- (ii) Continuous distributions. Normal distribution is an example of a continuous distribution.

## 11.2 THE BINOMIAL DISTRIBUTION

We will attempt to set the stage for the **Binomial distribution** by presenting an example. Suppose that a treatment for a particular disease has 0.6 probability of alleviating all signs and symptoms in an individual patient. For brevity, we will say that such a patient is cured. Find the probability that in a series of four treated patients exactly two will be cured, assuming that the outcomes cure or non-cure in the individual patients are mutually independent.

In attempting to solve this problem we can dissect it into several fundamental parts. If the patients are denoted by  $A, B, C, D$  then the outcome "exactly two cures" can occur in several different ways:  $A$  and  $B$  cured,  $C$  and  $D$  not cured;  $A$  and  $C$  cured,  $B$  and  $D$  not cured; and so on. In fact, using  $E_1, E_2, \dots, E_6$  to denote the specific ways of obtaining exactly two cures, the situation is as follows:

	Patient			
	A	B	C	D
$E_1$ :	cured	cured	not cured	not cured
$E_2$ :	cured	not cured	cured	not cured
$E_3$ :	cured	not cured	not cured	cured
$E_4$ :	not cured	cured	cured	not cured
$E_5$ :	not cured	cured	not cured	cured
$E_6$ :	not cured	not cured	cured	cured

That there are six results with the attribute "exactly two cures" is no accident. Infact,  $6 = {}^4C_2$ , and arises from the fact that from the four patients we must pick a set of two patients to be cured. Order is important, because, for example, "*A* and *B* cured" is identical to "*B* and *A* cured".

The  $E_i$  are mutually exclusive, and thus

$$P(E_1 \text{ or } E_2 \text{ or } \dots \text{ or } E_6) = P(E_1) + P(E_2) + \dots + P(E_6)$$

But  $E_1$  is the event: *A* cured and *B* cured and *C* not cured and *D* not cured, and since the outcomes for the individual patients are mutually independent we have,

$$\begin{aligned} P(E_1) &= P(\text{*A* cured and *B* and *C* not cured and *D* not cured}) \\ &= P(\text{*A* cured}) \cdot P(\text{*B* cured}) \cdot P(\text{*C* not cured}) \cdot P(\text{*D* not cured}) \\ &= (0.6)(0.6)(0.4)(0.4) = (0.6)^2(0.4)^2 \end{aligned}$$

The numerical result follows from the fact that any individual patient is cured with probability is 0.6, and thus any individual patient is not cured with probability  $1 - 0.6 = 0.4$ .

Similarly,  $P(E_2) = (0.6)(0.4)(0.6)(0.4) = (0.2)^2(0.4)^2$ . In fact,  $P(E_i) = (0.6)^2(0.4)^2$  for any  $i = 1, 2, \dots, 6$ . Therefore,

$$P(\text{exactly 2 cures}) = \sum_{i=1}^6 P(E_i) = \sum_{i=1}^6 (0.6)^2(0.4)^2 = 6 \cdot (0.6)^2(0.4)^2 = 0.3456$$

Similarly,  $P(\text{exactly 1 cure}) = 4(0.6)^2(0.4)^2 = 0.1536$ , since the result "exactly one cure" in four equiprobable mutually exclusive ways corresponding to the cure of any of the four patients. A complete listing of the probabilities of any outcomes of treating the four patients is

$$P(\text{exactly 0 cures}) = 1 \cdot (0.4)^4 = 0.0256$$

$$P(\text{exactly 1 cure}) = 4 \cdot (0.6)(0.4)^3 = 0.1536$$

$$P(\text{exactly 2 cures}) = 6 \cdot (0.6)(0.4)^2 = 0.3456$$

$$P(\text{exactly 3 cures}) = 4 \cdot (0.6)^3(0.4) = 0.3456$$

$$P(\text{exactly 4 cures}) = 1 \cdot (0.6)^4 = 0.1296$$

It is instructive to examine several properties of these probabilities. First,  $P(\text{exactly } r \text{ cures}) = {}^4C_r (0.6)^r (0.4)^{4-r}$  for  $r = 0, 1, 2, 3, 4$ ; recall that any non-zero quantity raised to the zero power equals 1; for example,  $(0.6)^0 = 1$ . Second, since the event (exactly 0 cures) or (exactly 1 cure), or (exactly 4 cures) includes all possible results of treatment, and

since the subevents connected by “or” are mutually exclusive, the sum of the probabilities of the subevents should be 1. That is,

$$\begin{aligned} P(\text{exactly 0 cures or exactly 1 cure or .... or exactly 4 cures}) \\ = P(\text{exactly 0 cures}) + \dots + P(\text{exactly 4 cures}) = 1 \end{aligned}$$

Addition verifies that this property holds numerically in the present example.

### 11.3 BINOMIAL DISTRIBUTION LAW

*Binomial distribution is a discrete probability distribution which is obtained when the probability  $p$  of the happening of an event is same in all the trials, and there are only two events in each trial.*

For example, the probability of getting a head, when a coin is tossed a number of times,

must remain same in each toss, i.e.,  $p = \frac{1}{2}$ .

Let an experiment consisting of  $n$  trials be performed and let the occurrence of an event in any trial be called a success and its non-occurrence a failure. Let  $p$  be the probability of success and  $q$  be the probability of its failure in a single trial, where  $q = 1 - p$ , so that  $p + q = 1$ .

*Let us assume that trials are independent and the probability of success is same in each trial. Let us claim that we have  $n$  trials, then the probability of happening of an event  $r$  times and failing  $(n - r)$  times in any specified order is  $p^r q^{n-r}$  (by the theorem on multiplication of probability) But the total number of ways in which the event can happen  $r$  times exactly in  $n$  trials is  $C(n, r)$ . These  $C(n, r)$  ways are equally likely, mutually exclusive and exhaustive.*

Therefore, the probability of  $r$  successes and  $(n - r)$  failures in  $n$  trials in any order, whatsoever is  $= {}^n C_r p^r q^{n-r}$ .  $\therefore C(n, r) = {}^n C_r$

It can also be expressed in the form

$$P(X = r) = P(r) = {}^n C_r p^r q^{n-r}; \quad r = 0, 1, 2, 3, \dots, n$$

where  $P(x = r)$  or  $P(r)$  is the probability distribution a random variable  $X$  of the number of  $r$  successes. Giving different values to  $r$ , i.e., putting  $r = 0, 1, 2, 3, \dots, n$ , we get the corresponding probabilities  ${}^n C_0 q^n, {}^n C_1 q^{n-1} p, {}^n C_2 q^{n-2} p^2, {}^n C_3 q^{n-3} p^3, \dots, p^n$ , which are the different terms in the Binomial expansion of  $(q + p)^n$ .

As a result of it, the distribution  $P(r) = {}^n C_r p^r q^{n-r}$  is called Binomial Probability distribution. The two independent constants, viz.,  $n$  and  $p$  in the distribution are called the parameters of the distribution

Again if the experiment (each consisting of  $n$  trials) be repeated  $N$  times, the frequency function of the Binomial distribution is given by

$$f(r) = N \times P(r) = N \times C(n, r) p^r q^{(n-r)}$$

The expected frequencies of 0, 1, 2, ...,  $n$  successes in the above set of experiment are the successive terms in the binomial expansion of  $N(q + p)^n$ ; where  $p + q = 1$ , which is also called the binomial frequency distribution.

## 11.4 MEAN AND VARIANCE OF BINOMIAL DISTRIBUTION

The probability distribution of binomial distribution for  $r$  successes in  $n$  trials is given by

$$P(r) = {}^n C_r q^{n-r} p^r$$

It has Mean =  $np$  and Variance =  $npq$ .

## 11.5 CONDITIONS FOR APPLICATION OF BINOMIAL DISTRIBUTION

1. The variable should be discrete, i.e., the values of  $x$  could be 1, 2, 3, 4 or 5 etc., and never 1.5, 2.1 or 3.41 etc.
2. A dichotomy exists. In other words, the happening of an event must have two alternatives. It must be either a success or a failure.
3. The number of trials 'n' should be finite and small.
4. The trials or events must be independent. The happening of one event must not affect the happening of other events. In other words, statistical independence must exist.
5. The trials or events must be repeated under identical conditions.

## 11.6 PASCAL'S TRIANGLE

**Method.** In the 'Pascal's Triangle' table each term is derived by adding together the two terms of the line lying immediately above it on either side of it. Thus when  $n=9$ , the fourth term in this line is 84, which is obtained by adding together the 3rd term and fourth term of the line for  $n=8$  (i.e.,  $84 = 28 + 56$ ).

Number is sample $n$	Binomial coefficients in the expansion $(q + p)^n$	Sum
1	1 1	2
2	1 2 1	4
3	1 3 3 1	8
4	1 4 6 4 1	16
5	1 5 10 10 5 1	32
6	1 6 15 20 15 6 1	64
7	1 7 21 35 35 21 7 1	128
8	1 8 28 56 70 56 28 8 1	256
9	1 9 36 84 126 126 84 36 9 1	512
10	1 10 45 120 210 252 210 120 45 10 1	1,024

## 11.7 CHARACTERISTICS OF BINOMIAL DISTRIBUTION

1. It is a discrete distribution which gives the theoretical probabilities.
2. It depends on the parameters  $p$  or  $q$ , the probability of success or failure and  $n$  (the number of trials). The parameter  $n$  is always a positive integer.
3. The distribution will be symmetrical if  $p = q$ . It is skew-symmetric or asymmetric if  $p \neq q$  although with  $n$  tending to be large it is approximately so.

4. The parameters of the binomial distribution are mean =  $np$ ; Variance =  $npq$ ; and standard deviation =  $\sqrt{npq}$ .
5. The mode of the binomial distribution is equal to that value of  $X$  which has the largest frequency.
6. It can be graphically, taking the X-axis to represent the number of successes and Y-axis to represent the probabilities or frequencies. Its graph will be vertical lines, with spaces in between them. Drawing a smooth curve by free hand is inadmissible as the variable  $X$  is a discrete one.
7. The shape and location of a binomial distribution changes as  $p$  changes for a given  $n$  or  $n$  changes for a given  $p$ .
8. The binomial coefficients are given by the Pascal's triangle.

## 11.8 RECURSION FORMULA OR RECURRENCE RELATION FOR BINOMIAL DISTRIBUTION

We know that for the binomial distribution :

$$\begin{aligned}
 P(r) &= {}^nC_r p^r q^{n-r} \quad \text{and} \quad P(r+1) = {}^nC_{r+1} p^{r+1} q^{n-r-1}. \\
 \frac{P(r+1)}{P(r)} &= \frac{{}^nC_{r+1} p^{r+1} q^{n-r-1}}{{}^nC_r p^r q^{n-r}} \\
 &= \frac{n!}{(r+1)! (n-r-1)!} \times \frac{r! (n-r)!}{n!} \\
 &= \frac{(n-r) \times (n-r-1)! \times r!}{(r+1) \times r! \times (n-r-1)!} \times \frac{p}{q} \\
 \Rightarrow P(r+1) &= \frac{(n-r)}{(r+1)} \cdot \frac{p}{q} \cdot P(r).
 \end{aligned}$$

Hence,  $P(r+1) = \frac{(n-r)}{(r+1)} \times \frac{p}{q} \times P(r)$  is the required recurrence relation for the binomial distribution.

**Example 1 :** In a binomial distribution, the mean and standard deviations are 12 and 2 respectively. Find  $n$  and  $P$ .

**Solution :** We are given that Mean =  $np = 12$ , S.D. =  $\sqrt{npq} \Rightarrow npq = 4$ .

$$\text{Now, } \frac{npq}{np} = \frac{4}{12} = \frac{1}{3} \Rightarrow q = \frac{1}{3}, \quad p = 1 - q = 1 - \frac{1}{3} = \frac{2}{3}$$

$$\text{Also, } np = 12 \Rightarrow n \times \frac{2}{3} = 12 \Rightarrow n = 18$$

$$\text{Hence, } n = 18, \quad p = \frac{2}{3}$$

**Example 2 :** Using the formula for binomial distribution, find the probability of rolling at most 2 sixes in 5 rolls of a dice.

**Solution :** Here  $n = 5$ .

$$\text{Probability of rolling a six on a dice} = p = \frac{1}{6} \quad \therefore q = \frac{5}{6}$$

$$\begin{aligned} P(x \leq 2) &= P(x = 0) + P(x = 1) + P(x = 2) \\ &= {}^5C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^5 + {}^5C_1 \left(\frac{1}{6}\right) \left(\frac{5}{6}\right)^4 + {}^5C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 \\ &= \left(\frac{5}{6}\right)^5 + 5 \left(\frac{1}{6}\right) \left(\frac{5}{6}\right)^4 + \frac{5 \times 4}{2 \times 1} \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^3 \\ &= \frac{5^4}{6^5} [5 + 5 + 2] = 12 \times \frac{5^4}{6^5} = \frac{625}{648}. \end{aligned}$$

**Example 3 :** If 10 coins are tossed 100 times, how many times would you expect 7 coins to fall head upward?

**Solution :** The probability of getting a head in a single toss of a coin is  $p = \frac{1}{2}$ .

$$\therefore q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

Also we are given  $n = 10$ ,  $N = 100$  and  $r = 7$ .

**∴ The required frequency :**  $P(r) = N \times {}^nC_r p^r q^{n-r}$

$$\begin{aligned} &= 100 \times {}^{10}C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3 = 100 \times \frac{10}{7! \times 3!} \left(\frac{1}{2}\right)^{10} \\ &= 100 \times \frac{10 \times 9 \times 8}{3 \times 2 \times 1} \left(\frac{1}{2}\right)^{10} = \frac{375}{32} = 11.7 = 12. \end{aligned}$$

**Example 4 :** There are 64 beds in a garden and 3 seeds of particular type of flower are sown in each bed. The probability of a flower being white is  $1/4$ . Find the number of beds with 3, 2, 1 and 0 white flowers.

**Solution :** The probability  $p$  of a white flower  $p = \frac{1}{4}$ .

$$\therefore q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}. \quad \text{Here } n = 3, N = 64.$$

$$f(r) = N \times C(3, r) \left(\frac{1}{4}\right)^r \left(\frac{3}{4}\right)^{3-r}$$

**∴ Number of beds with zero white flower** =  $N f(0)$

$$= 64 \times C(3, 0) \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^3 = 64 \times \frac{27}{64} = 27.$$

$$\text{Beds with 1 white flower} = 64 f(1) = 64 \times C(3, 1) \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^2 = \\ = 64 \times 3 \times \frac{1}{4} \times \frac{9}{16} = \frac{27}{64} \times 64 = 27.$$

$$\text{Beds with 2 white flowers} = 64 \times C(3, 2) \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^0 = 64 \times \frac{9}{64} = 9.$$

$$\text{Beds with 3 white flowers} = 64 \times C(3, 3) \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^0 = 64 \times \frac{1}{64} = 1.$$

## 11.9 FITTING OF A BINOMIAL DISTRIBUTION

The probability of 0, 1, 2, 3, ...n successes would be obtained from the binomial expansion of  $(q + p)^n$ .

The probability of  $r$  successes in a single throw is given by

$$P(r) = {}^nC_r p^r q^{n-r}$$

Suppose this experiment is repeated for  $N$  times, then

**the frequency of  $r$  success is**  $= N \times P(r) = N \times {}^nC_r p^r q^{n-r}$

Putting  $r = 0, 1, 2, \dots, n$ , we get the expected or theoretical frequencies of the binomial distribution as follows :

Number of successes ( $r$ )	Expected or theoretical frequency $N P(r)$
0	$N q^n$
1	$N \times {}^nC_1 p q^{n-1}$
2	$N \times {}^nC_2 p^2 q^{n-2}$
$r$	$N \times {}^nC_r p^r q^{n-r}$
$n$	$N p^n$

**Example 5 :** As a result of a certain experiment, the data obtained was :

$$\begin{array}{ccccc} x : & 0 & 1 & 2 & 3 & 4 \\ f : & 8 & 32 & 34 & 24 & 5 \end{array}$$

Fit a binomial distribution to the above data.

$$\text{Solution : Mean : } np = \frac{0 + 32 + 68 + 72 + 20}{103} = \frac{192}{103} = 1.864$$

$$\therefore p = \frac{1.864}{4} = 0.466; \text{ and } q = 1 - p = 0.534.$$

The expected frequencies are given by

$$N (q + p)^n = 103 (0.534 + 0.466)^4 \quad [\because n = 4]$$

Putting  $r = 0, 1, 2, 3, 4$ , we get:

$$103 [(0.534)^4 ; 4 (0.534)^3 (0.466) ; 6 (0.534)^2 (0.466)^2 ; 4 (0.534) (0.466)^3 ; (0.466)^4]$$

**Example 6 :** State the conditions under which binomial distribution is used. Find the distribution if the mean is 48 and standard deviation is 4.

**Solution :** Binomial distribution is used under the following conditions:

1. A random experiment is performed repeatedly  $n$  number of times, where  $n$  is a finite positive integer.
2. The trials are independent.
3. There are only two outcomes, viz., success or failure with constant probability  $p$  of success at each trial, i.e., a dichotomy exists).

Given,  $np = 48$ ;  $npq = 16$

$$\therefore q = \frac{npq}{np} = \frac{16}{48} = \frac{1}{3}, \quad p = 1 - q = 1 - \frac{1}{3} = \frac{2}{3}$$

$$np = 48 \Rightarrow n \times \frac{2}{3} = 48 \text{ or } n = 72$$

The required binomial distribution is given by  $(q + p)^{72} = \left(\frac{1}{3} + \frac{2}{3}\right)^{72}$ .

**Example 7 :** Find the probability of success  $p$  for a binomial distribution, if  $n = 6$  and  $4 \times P(X = 4) = P(X = 2)$ .

**Solution :** We know that in binomial distribution the probability of  $r$  successes in  $n$  trials is given by

$$P(X = r) = {}^nC_r p^r q^{n-r}$$

According to the given condition

$$\begin{aligned} 4 P(X = 4) &= P(X = 2) \Rightarrow 4 ({}^6C_4 p^4 q^2) = {}^6C_2 p^2 q^4 \\ \Rightarrow 4p^2 &= q^2 \Rightarrow 4p^2 = (1 - p)^2 \quad [\because {}^6C_4 = {}^6C_2 \text{ and } q = 1 - p] \\ \Rightarrow 3p^2 + 2p - 1 &= 0 \quad \text{or} \quad p = \frac{-2 \pm \sqrt{4 + 12}}{6} = -1 \text{ or } \frac{1}{3}. \end{aligned}$$

Since  $p$  cannot be negative so  $p = \frac{1}{3}$ .

**Example 8 :** A die is thrown three times. Getting a '3' or a '6' is considered a success. Find the probability of at least two successes.

**Solution :** Let  $P(x)$  be the probability of getting the number  $x$  in a toss of a dice. Then

$$P(3) = \frac{1}{6} \quad P(6) = \frac{1}{6}$$

$$\therefore P(3 \text{ or } 6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} = p. \quad \therefore q = 1 - \frac{1}{3} = \frac{2}{3}$$

$$\text{Now, } P(r) = {}^nC_r p^r q^{n-r} \quad \text{Hence } n = 3.$$

$$P(2) = {}^3C_2 \left(\frac{2}{3}\right)^{3-2} \left(\frac{1}{3}\right)^2 = 3 \times \frac{2}{3} \cdot \frac{1}{9} = \frac{2}{9}$$

$$P(3) = {}^3C_3 \left(\frac{2}{3}\right)^{3-3} \left(\frac{1}{3}\right)^3 = 1 \cdot 1 \cdot \frac{1}{27} = \frac{1}{27}$$

$$P(\text{at least 2 successes}) = P(2) + P(3) = \frac{2}{9} + \frac{1}{27} = \frac{7}{27}.$$

**Example 9 :** 12% of the items produced by a machine are defective. What is the probability that out of a random sample of 20 items produced by the machine, 5 are defective? (Simplification is not necessary).

**Solution :** Let  $p$  be the probability that an item produced by the machine is defective. We are given :

$$n = 20; \quad p = 0.12, \quad q = 1 - p = 0.88$$

$$\begin{aligned} P(x) &= \text{Probability of } x \text{ defective items} \\ &= {}^nC_x p^x q^{n-x} = {}^{20}C_x (0.12)^x (0.88)^{20-x} \end{aligned}$$

Required probability is given by :  $P(5) = {}^{20}C_5 (0.12)^5 (0.88)^{15}$ .

**Example 10 :** On an average 2% of the population in an area suffers from T.B. What is the probability that out of 5 persons chosen at random from this area atleast two suffer form T.B. (simplification is not necessary).

**Solution :** Here,  $n = 5; \quad p = 2\% = \frac{2}{100} = 0.02; \quad q = 1 - p = 0.98$ .

The probability that in a random sample of  $n$  persons,  $r$  persons are suffering from T.B. is given by the binomial probability distribution.

$$P(X = r) = {}^nC_r p^r q^{n-r} = {}^5C_r (0.02)^r (0.98)^{5-r}.$$

The required probability  $P$  that at least two of the 5 persons are suffering from T.B. is given by

$$\begin{aligned} P &= P(X > 2) = 1 - P(X < 2) = 1 - [P(0) + P(1)] \\ &= 1 - [{}^5C_0 (0.02)^0 \cdot (0.98)^5 + {}^5C_1 (0.02) (0.98)^4] \\ &= 1 - [(0.98)^5 + 5 \times 0.02 \times (0.98)^4] \\ &= 1 - (0.98)^4 [0.98 + 0.10] = 1 - (0.98)^4 \times 1.08 = 1 - 0.84 = 0.16. \end{aligned}$$

**Example 11 :** The incidence of occupational disease in an industry is such that the workmen have a 20% chance of suffering from it. What is the probability that out of six workmen, 4 or more contract the disease?

**Solution :** Here,  $n = 6; \quad p = 20\% = \frac{20}{100} = \frac{1}{5}; \quad \therefore q = 1 - p = 1 - \frac{1}{5} = \frac{4}{5}$

By binomial probability distribution, the probability that out of 6 workmen,  $r$  will contract the disease, is:

$$P(r) = {}^6C_r p^r q^{6-r} = {}^6C_r \left(\frac{1}{5}\right)^r \left(\frac{4}{5}\right)^{6-r} = \frac{1}{5^6} {}^6C_r \times 4^{6-r}$$

Thus the probability that in a sample of 6 workmen, 4 or more will contract the disease is:

$$\text{Required probability} = P(4) + P(5) + P(6)$$

$$\begin{aligned} &= \frac{1}{5^6} [{}^6C_4 \times 4^2 + {}^6C_5 \times 4 + {}^6C_6] \\ &= \frac{1}{5^6} \left[ \frac{6 \times 5}{2} \times 16 + 16 \times 4 + 1 \right] = \frac{265}{5^6} = \frac{53}{3125} = 0.01696. \end{aligned}$$

**Example 12 :** An oil exploration firm finds that 55 of the test cells it drills yield a deposit of natural gas. If it drills 6 wells, find the probability that at least one well will yield gas. (Simplification is not necessary.)

**Solution :** Here,  $n = 6$ ; gas = 0.05.

$$\therefore p = \text{Probability that a test well has a deposit of natural gas} = 0.05$$

$$\therefore q = 1 - p = 0.95.$$

$$P(xr) = \text{Probability that in a drilling of 6 wells, } r \text{ wells have deposit of natural gas} = {}^6C_r (0.05)^r (0.95)^{6-r}$$

$$\therefore P(0) = {}^6C_0 (0.05)^0 (0.95)^6 = (0.95)^6$$

**Probability that atleast one well will yield a deposit of natural gas**

$$= 1 - (\text{Probability that the well does not yield gas})$$

$$= 1 - P(0) = 1 - (0.95)^6.$$

**Example 13 :** A manufacturing process turns out articles that on the average are 10% defective. Compute the probability of 0, 1, 2 and 3 defective articles that might occur in a sample of 3 articles.

**Solution :** Let  $p$  = Probability of a defective article = 10% = 0.10. Then  $q = 1 - p = 0.10 = 0.9$ ;  $n = 3$ .

The probability of  $x$  defectives in a sample of 3 articles is given by the binomial probability distribution :

$$P(x) = {}^3C_x p^x q^{3-x} = {}^3C_x (0.1)^x (0.9)^{3-x}; x = 0, 1, 2, 3.$$

Putting  $x = 0, 1, 2$  and  $3$ , we get probability of 0, 1, 2 and 3 defective articles respectively:

$$P(0) = (0.9)^3 = 0.729.$$

$$P(1) = {}^3C_1 (0.1) (0.9)^2 = 3 \times 0.1 \times 0.81 = 0.243.$$

$$P(2) = {}^3C_2 (0.1)^2 (0.9) = 3 \times 0.01 \times 0.9 = 0.027.$$

$$P(3) = (0.1)^3 = 0.001.$$

**Example 14 :** If the probability of a defective bolt is 0.1, find the mean and standard deviation for the distribution of defective bolt in a total of 500.

**Solution :** Let  $(q + p)^n$ ,  $q + p = 1$ , be the binomial distribution. Here,  $p = 0.1$ ,  $n = 500$

$$\therefore \text{Mean} = np = 0.1 \times 500 = 50$$

$$\text{Now, } p = 0.1 \Rightarrow q = 1 - p = 1 - 0.1 = 0.9$$

$$\therefore \text{Variance} = npq = 500 \times 0.1 \times 0.9 = 45.$$

$$\therefore \text{Standard deviation} = \sqrt{45} = 6.7$$

**Example 15 :** Six dice are thrown 729 times. How many times do you expect at least three dice to show a five or a six?

**Solution :** Let  $N(q + p)^n$ ,  $q + p = 1$ , be the binomial distribution.

$$\text{Now, } p = \text{the probability of getting 5 or 6 with one die} = \frac{2}{6} = \frac{1}{3}$$

$$q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}$$

Here,  $N = 729$  and  $n = 6$

The binomial distribution is  $N(q + p)^n = 728 \left(\frac{2}{3} + \frac{1}{3}\right)^6$ .

The expected number of times at least three dice showing 5 or 6.

$$\begin{aligned} &= 729 \left[ {}^6C_3 \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^3 + {}^6C_4 \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^4 + {}^6C_5 \left(\frac{2}{3}\right) \left(\frac{1}{3}\right)^5 + {}^6C_6 \left(\frac{1}{3}\right)^6 \right] \\ &= 729 [160 + 60 + 12 + 1] = 233. \end{aligned}$$

**Example 16 :** Eight coins are thrown simultaneously. Find the probability of getting at least six heads.

**Solution :** Let  $p$  denote the probability of getting a head and  $q$  be the probability of not getting a head, then  $p = \frac{1}{2}$ ,  $q = \frac{1}{2}$ .

∴ Probability of getting at least six heads when 8 coins are thrown simultaneously

$$\begin{aligned} &= P(6) + P(7) + P(8) \\ &= {}^8C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^2 + {}^8C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right) + {}^8C_8 \left(\frac{1}{2}\right)^8 \\ &= \left(\frac{1}{2}\right)^8 [{}^8C_6 + {}^8C_7 + {}^8C_8] = \frac{1}{256} [28 + 8 + 1] = \frac{37}{256}. \end{aligned}$$

**Example 17 :** Assuming that half the population are consumers of rice so that the chance of an individual being a rice consumer is  $1/2$  and assuming that 100 investigations each take 10 individuals to see whether they are rice consumers. How many investigations would you expect to report that three people or less were consumers?

**Solution :** Here,  $p = \frac{1}{2}$ ,  $q = \frac{1}{2}$ ,  $n = 10$ ,  $N = 100$ .

∴ The probability that  $r$  persons out of 10 persons are consumers of rice is given by

$$P(r) = {}^{10}C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{10-r}$$

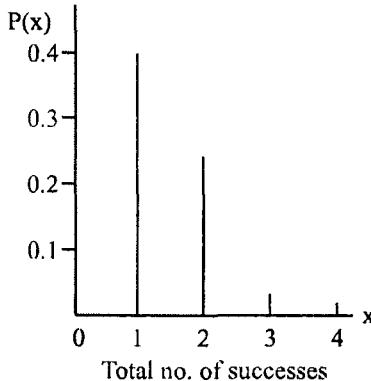
∴ The expected number of investigations (i.e., expected frequencies) who would report that three or less people were consumers of rice

$$\begin{aligned}
 &= 100 [P(0) + P(1) + P(2) + P(3)] \\
 &= 100 \left[ {}^{10}C_0 \left(\frac{1}{2}\right)^{10} + {}^{10}C_1 \left(\frac{1}{2}\right)^{10} + {}^{10}C_2 \left(\frac{1}{2}\right)^{10} + {}^{10}C_3 \left(\frac{1}{2}\right)^{10} \right] \\
 &= \frac{100}{2^{10}} [1 + 10 + 45 + 120] = \frac{17600}{1024} = 17 \text{ (approx.)}
 \end{aligned}$$

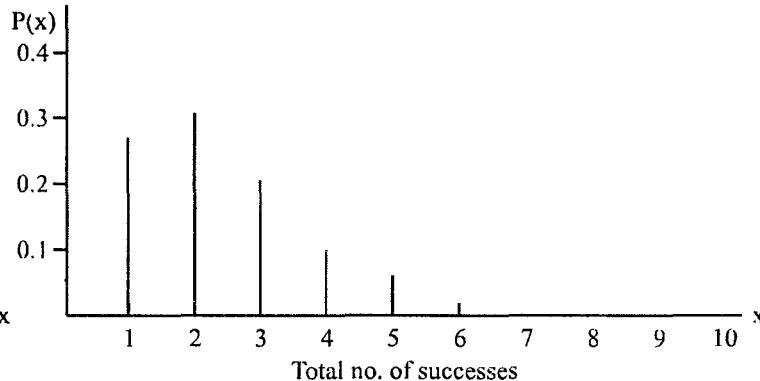
## 11.10 SOME REMARKS

The mean and variance of the binomial distribution of  $x$ , the total number of successes in  $n$  independent trials, with probability  $p$  of success on an individual trial are  $np$  and  $np(1 - p)$  =  $npq$ , respectively.

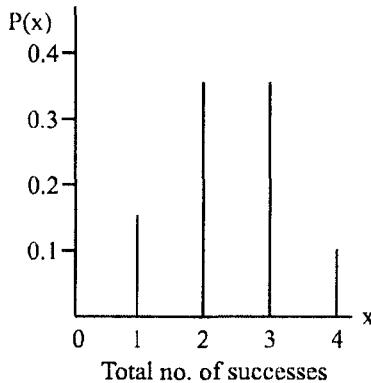
$$\begin{array}{l}
 \text{A} \\
 n = 4, \quad B = 0.2 \\
 \mu_x = 0.8, \quad \sigma_x^2 = 0.64
 \end{array}$$



$$\begin{array}{l}
 \text{B} \\
 n = 10, \quad B = 0.2 \\
 \mu_x = 2.0, \quad \sigma_x^2 = 1.6
 \end{array}$$



$$\begin{array}{l}
 \text{C} \\
 n = 4, \quad B = 0.6 \\
 \mu_x = 2.4, \quad \sigma_x^2 = 0.96
 \end{array}$$



$$\begin{array}{l}
 \text{D} \\
 n = 10, \quad B = 0.6 \\
 \mu_x = 6.0, \quad \sigma_x^2 = 2.4
 \end{array}$$

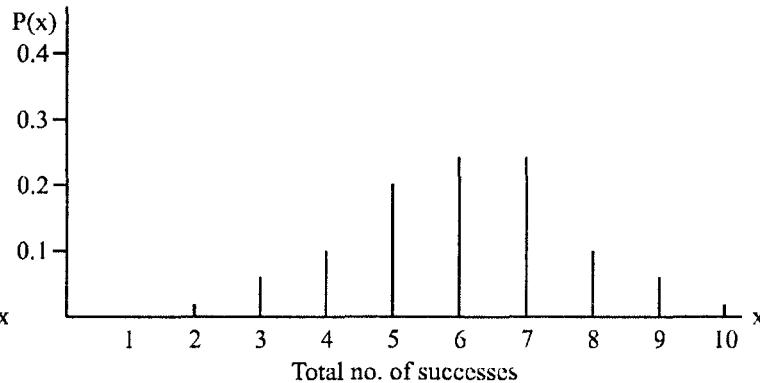


Fig. 11.1 : Graphs of binomial distributions.

A little reflection may convince you that the mean should depend directly on both  $n$  and  $p$ . Certainly, for a fixed number of trials, a large value of  $p$ , the single-trial success probability,

should be associated with a large mean number of successes. Also, for fixed  $p$  an increase in  $n$ , the total number of trials, should result in an increased mean. The quantity  $np$  exhibits both these properties.

Figure 11.1 shows four specimen binomial distributions for  $n = 4$  and  $10$  and  $p = 0.2$  and  $0.6$ . The vertical scale of each graph gives  $P(x)$ , the probability of  $x$  successes. Several observations concerning the relative shapes of these four distribution are instructive. The distributions for  $p = 0.6$  are better balanced, that is, more symmetric than are the distribution for  $p = 0.2$ . In general, the nearer  $p$  is to  $0.5$ , the more symmetric is the corresponding distribution. In fact, when  $p = 0.5$  the binomial distribution is perfectly symmetric for any value of  $n$ . We also notice that the distribution for  $p = 0.6$  are more spread than those for  $p = 0.2$ . This conforms with the fact that the corresponding variances are larger. There is also more spread in the distributions for  $n = 10$  than for  $n = 4$ , again in conforming with larger variances.

A useful notational device for the binomial probability uses the symbol  $P(x; n, p)$  for the probability that the total number of successes is  $x$  in  $n$  independent trials with success probability  $p$  on a single trial. For example, with four patients and single patient cure probability  $0.6$ , the probability of exactly two cures is  $B(2:4, 0.6) = 0.3456$ . Formally,

$$P(x, n, p) = {}^n C_x p^x (1 - p)^{n-x}$$

The construction of tables giving binomial probability can be simplified by taking advantage of the property that

$$P(x, n, p) = .B(n - x, n, 1 - p)$$

It is as easy to establish this relation semantically as it is analytically. The statement says that the probability of  $x$  successes in  $n$  independent trials when the probability of success on a single trial is  $p$  is the same as the probability of  $n - x$  failures in  $n$  independent trials when the probability of failure on a single trial is  $1 - p$ . The designation of one kind of outcome as success and the other as failure is, as we have noted, purely arbitrary. The labels can be interchanged, provided we remember that if the probability of one outcome is  $p$  then the probability of the other is  $1 - p$ . This verbal or semantic argument establishes the relation. This relation makes it unnecessary to tabulate binomial probabilities for values of  $p$  larger than  $0.5$ . If, for example, we want  $P(6; 10, 0.8)$ , we notice that it is numerically equal to  $P(4; 10, 0.2)$ .

**Example 18 :** Suppose that the probability of a single human birth producing a male infant is  $0.5$ , ignoring twin and other multiple births. (In fact, this probability is slightly greater than  $0.5$ ). Assuming that different deliveries are independent with respect to sex, find the probability that a family of five children will include: (a) exactly three boys, (b) at least one boy, (c) at most one boy.

**Solution :**

$$(a) P(\text{exactly 3 boys}) = B(3; 5, 0.5) = 0.3125$$

$$(b) P(\text{at least 1 boy}) = 1 - P(\text{no boys}) = 1 - P(0; 5, 0.5) = 1 - 0.0312 = 0.9688$$

$$(c) P(\text{at least 1 boy}) = P(\text{exactly 0 boys or exactly 1 boy})$$

$$= P(0; 5, 0.5) + P(1; 5, 0.5) = 0.0312 + 0.1562 = 0.1874$$

### EXERCISE – 11.1

1. If two parents, both heterozygous carriers of autosomal recessive gene causing cystic fibrosis have five children. What is the probability that three will be normal?

[Hint :  $p$  = probability of having a normal child during pregnancy

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

$$q = \text{probability of affected child (cystic fibrosis)} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$\therefore P(x=r) = {}^nC_r p^r q^{n-r} = {}^4C_3 \left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right)^2 = 0.26]$$

2. Out of 1000 families of 3 children each, how many families would you expect to have two boys and one girl assuming that boys and girls are equally likely?

[Hint : Here  $N = 1000$ ;  $n = 3$ ,  $p$  = probability of a boy =  $\frac{1}{2}$ ;  $q = \frac{1}{2}$ .

The expected number of families who have  $x$  boys and  $n - x$  girls is:

$$\begin{aligned} N \cdot {}^nC_r p^r \cdot q^{n-r} &= 1000 \times {}^3C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{3-2} \\ &= 1000 \times {}^3C_2 \left(\frac{1}{2}\right)^3 = 1000 \times 3 \times \frac{1}{8} = 375] \end{aligned}$$

3. The probability that an evening college student will graduate is 0.4. Determine the probability that out of 5 students (a) none, (b) one and (c) at least one will be graduate.

[Hint : Here  $n = 5$ ;  $p = 0.4$ ,  $q = 0.6$ .  $P(r) = {}^5C_r p^r q^{5-r}$ ;  $r = 0, 1, 2, 3, 4, 5$

$$(a) P(0) = {}^5C_0 (0.6)^5 = 0.80 ; \quad (b) P(1) = {}^5C_1 (0.4) (0.6)^4 = 0.26 ;$$

$$(c) \text{Probability atleast one will graduate} = 1 - P(0) = 1 - 0.80 = 2.20]$$

4. Eight coins are thrown simultaneously. Show that the probability of obtaining of atleast 6 heads is  $37/256$ .

[Hint :  $P(X \geq 6) = P(6) + P(7) + P(8)]$ .

5. Assuming that it is true that 2 out of 10 industrial accidents are due to fatigue, find that exactly 2 of 8 industrial accidents will be due to fatigue.

[Hint : Here  $p = (2/10) = 0.2$ ;  $q = 1 - p = 0.8$ ;  $n = 8$ .

$$P(X=r) = {}^nC_r p^r q^{n-r} \Rightarrow P(X=2) = {}^8C_2 (0.2)^2 (0.8)^6 = 0.29].$$

6. From the past weather record, it has been found that, on an average, rain falls on 12 days in June. Find the probability that in a given week of June,

(a) first four days will be dry and the remaining three days wet;

(b) Exactly three days will be wet.

[Hint : Let  $D$  and  $W$  respectively denote dry and wet day.

$$\therefore p = P(W) = \frac{12}{30} = \frac{2}{5} = 0.4; q = P(D) = 1 - p = 0.6$$

$$(a) P(DD DD WWW) = P(D)P(D)P(D)P(D)P(W)P(W)P(W) = (0.6)^4(0.4)^3 = 0.008$$

$$(b) \text{ Required probability} = {}^7C_3 p^3 q^{7-3} = {}^7C_3 (0.4)^3 (0.6)^4 = 0.291]$$

7. An insurance salesmen sells policies to 5 men, all of identical age and good health. According to the actuarial tables the probability that a man of this particular age will be alive 30 years hence is  $2/3$ . Find the probability that 30 years hence (a) at least 1 man will be alive; (b) at least 3 men will be alive.

[Hint :  $p$  = Probability that the man will be alive 30 years hence =  $\frac{2}{3}$ ;  $q = \frac{1}{3}$ .

$$P(X=r) = {}^nC_r p^r q^{n-r}; P(0) = {}^5C_0 \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^5 = \left(\frac{1}{3}\right)^5 = \frac{1}{243}.$$

$$(a) \text{ Required probability} = 1 - P(0) = 1 - \frac{1}{243} = \frac{242}{243}.$$

$$(b) P(3) + P(4) + P(5) = \frac{64}{81}.$$

8. If hens of a certain breed lay eggs on 5 days a week on an average, find how many days during a season of 100 days a poultry keeper with 5 hens of this breed, will expect to receive atleast 4 eggs?

[Hint : Here  $n = 5$ ,  $N = 100$ ;  $p = \frac{5}{7}$ ,  $q = \frac{2}{7}$ ;

$$P(X=r) = {}^nC_r p^r q^{n-r}; \text{ Required probability } P(X \geq 4) = P(X=4) + P(X=5) \\ = {}^5C_4 (5/7)^4 (2/7) + {}^5C_5 (5/7)^5 = 0.56$$

$$\text{Required number of days} = N \times P(X \geq 4) = 100 \times 0.56 = 56]$$

9. Twelve coins are tossed. What are the probabilities in a single tossing getting (a) 9 or more heads, (b) less than 3 heads, (c) at least 8 heads?

[Hint : (a) Required probability =  $P(0) + P(1) + P(2) + P(3)$

(b) Required probability =  $P(0) + P(1) + P(2)$

(c) Required probability = ]

10. The average percentage of failure in a certain examination is 40. What is the probability that out of a group of 6 candidates, at least 4 pass in the examination.

[Hint :  $p = 0.4$ ,  $q = 0.6$

$$P(x \geq 4) = P(x=4) + P(x=5) + P(x=6).$$

$$= {}^6C_4 (0.6)^4 (0.4)^2 + {}^6C_5 (0.6)^5 (0.4) + {}^6C_6 (0.6)^6 = 0.5443]$$

11. In a particular market 40% of the consumers prefer readymade clothing. A sample of  $n = 5$  consumers is to be drawn. Give the probabilities of having 0, 1, 2, 3, 4 and 5 preferring readymade clothes.

[Hint : Probability of  $r$  consumers preferring readymade clothes out of a sample of 5 =  ${}^5C_r (0.4)^r (0.6)^{5-r}$ . Put  $r = 0, 1, 2, 3, 4$  and 5.]

12. A sample of 10 pieces was examined out of large consignment which has 5% defective pieces. Give the probability of 1 defective in the sample of 10.

[Hint : Here  $p = 0.05$ ,  $q = 0.95$ ,  $n = 10$ . Required probability =  ${}^{10}C_1 (0.05) (0.95)^9$ ]

13. In 100 sets of ten tosses of an unbiased coin, in how many cases should we expect seven heads and three tails?

[Hint : Here probability of Head =  $p = \frac{1}{2}$ ,  $q = \frac{1}{2}$ ,  $n = 10$ ,  $N = 100$ .

Required number of cases =  $100 \times {}^{10}C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3$ .

14. Assuming that one in 80 births is a case of twins, calculate the probability of 2 or more sets of twins on a day when 30 births occur.

[Hint :  $p$  (twin birth) =  $\frac{1}{80} = 0.0125$ ;  $q = 1 - p = 0.9875$ ;  $n = 30$

$$P(x = r) = {}^nC_r p^r q^{n-r}]$$

## ANSWERS

- |             |  |                                    |
|-------------|--|------------------------------------|
| 1. 0.26.    | 2. 375.  | 3. (a) 0.80 ; (b) 0.26 ; (c) 0.20. |
| 5. 0.29.    | 6. (a) 0.008 ; (b) 0.291.                                    | 7. (a) (242/243) ; (b) (64/81).    |
| 8. 56.      | 9. (a) (299/4096) ; (b) (79/4096) ; (c) 0.1938               |                                    |
| 10. 0.5443. | 11. ${}^5C_r (0.4)^r (0.6)^{5-r}$ , $r = 0, 1, 2, 3, 4, 5$ . |                                    |

12.  ${}^{10}C_1 (0.05) (0.95)^9$ . 13.  $100 \times {}^{10}C_7 \left(\frac{1}{2}\right)^{70}$ .

## 11.11 POISSON DISTRIBUTION

**Poisson Distribution.** The Poisson distribution was discovered by a French Mathematician Simen Denie Poisson in 1837. It is a discrete distribution and is very widely used. Poisson distribution is a limiting form of the Binomial distribution in which  $n$ , the number of trials, becomes very large and  $p$ , the probability of the success of the event is very small such that  $np$  is a finite quantity and not necessarily large. In Binomial distribution, the probability of  $x$  successes is  ${}^nC_x p^x q^{n-x}$ . We require the limiting value of the expression

$f(x) = {}^nC_x p^x q^{n-x}$ , where  $p = \frac{m}{n}$  where  $n$  ultimately tends to  $\infty$  and  $m$  is a constant.

**Definition.** The probability distribution of a random variable  $X$  is said to have a Poisson Distribution if it takes only non-negative values and if its distribution is given by

$$P(X = x) = \frac{m^x e^{-m}}{x!}, \quad x = 0, 1, 2, \dots$$

where  $m$  is the parameter and  $e = 2.7183 = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \dots$

Hence the probability of 0, 1, 2, 3, ....  $x$  successes are given by

$$e^{-m}, \frac{e^{-m} m}{1!}, \frac{e^{-m} m^2}{2!}, \dots, \frac{e^{-m} m^x}{x!}, \text{ respectively.}$$

Also the formula of Poisson Distribution for computing the theoretical or relative frequencies of a random variable  $x$  is given by

$$\text{Frequency} = N \left( \frac{e^{-m} m^x}{x!} \right), \text{ where } N \text{ is number of trials, } m = \text{mean.}$$

The following are the **statistical measures of the Poisson distribution**.

1. Mean =  $np = m$ .
2. Variance =  $np = m$
3. Standard deviation =  $\sqrt{np} = \sqrt{m}$ .

#### 11.11.1 Some Examples of Poisson Distribution

1. The number of cars passing through a certain street (say Janpath) in a time  $t$ .
2. The number of deaths in a city in one year by a rare disease.
3. The number of defective screws per box of 100 screws.
4. The number of suicides or deaths by heartattack in time  $t$ .
5. The number of printing mistakes in each page of the first proof of a book.
6. The emission of radio-active (alpha) particles.
7. The number of air accidents in India in one year.
8. The number of pieces of a certain merchandise sold by a Super Bazar in time  $t$ .
9. The number of telephone calls received at a particular telephone exchange in some unit of time.
10. The number of defective materials in a packing material manufactured by a goods concern.
11. The number of fragments received by a surface area  $A$  from a fragment Atom bomb.
12. Occurrence of a number of scratches on a sheet of glass.
13. Flashing of a number of lightings per second.
14. Scoring of number of goals in a game etc.

## 11.12 CONDITIONS UNDER WHICH POISSON DISTRIBUTION IS USED

1. *The random variable  $X$  should be discrete.*
2. *A dichotomy exists, i.e., the happening of the events must be of two alternatives such as success and failure; occurrence and non-occurrence etc.*
3. *It is applicable in those cases where the number of trials  $n$  is very large and the probability of success  $p$  is very small but the mean  $np = m$  is finite.*
4.  *$p$  should be very small (close to zero). If  $p \rightarrow 0$ , then the distribution is J-shaped and unimodal.*
5. *Statistical independence is assumed. In other words, it is applicable to those cases where the happening of one event does not affect the happening of the other events.*

## 11.13 CHARACTERISTICS OF POISSON DISTRIBUTION

1. *Poisson distribution is a discrete distribution. It gives theoretical probabilities and theoretical frequencies of a discrete variable.*
2. *It depends mainly on the value of the mean  $m$ .*
3. *This distribution is positively skewed to the left. With the increase in the value of the mean  $m$ , the distribution shifts to the right and the skewness diminishes.*
4. *Its arithmetic mean in relative distribution is  $p$  and in absolute distribution is  $np$ .*
5. *If  $n$  is large and  $p$  is small, this distribution gives a close approximation to Binomial distribution. Since the arithmetic mean of Poisson is same as that of Binomial, so the Poisson distribution can be used instead of binomial if  $n$  or  $p$  is not known.*
6. *Poisson distribution has only one parametric, viz.,  $m$ , the arithmetic mean. Thus the entire distribution can be determined once the arithmetic mean is known.*
7. *Poisson distribution is based on the following assumptions :*
  - (i) Statistical independence is assumed, i.e., the occurrence or non-occurrence of an event does not influence the other events.
  - (ii) The probability of happening of more than one even in a very small interval is negligible.
  - (iii) The probability of success for a small space or a short interval of time is proportional to the space or length of time interval as the case may be.

## 11.14 BINOMIAL APPROXIMATION TO POISSON DISTRIBUTION

. Poisson distribution can be derived from Binomial distribution under the following conditions :

- (i)  *$p$ , the probability of the occurrence of the event is very small*
- (ii)  *$n$  is very very large, where  $n$  is number of trials, i.e.,  $n \rightarrow \infty$ .*
- (iii)  *$np$  is a finite quantity, say  $np = m$ , then  $m$  is called the parameter of the Poisson distribution. In Poisson distribution, the probability of  $r$  success is given by*

$$P(r) \text{ or } P(X = r) = \frac{e^{-m} m^r}{r!}.$$

The probability of 0, 1, 2, ...,  $x$ , ... successes are given by

$$e^{-m}, \frac{e^{-m}m}{1!}, \frac{e^{-m}m^2}{2!}, \dots, \frac{e^{-m}m^x}{x!}, \text{ respectively.}$$

where  $m$  is called the parameter of the Poisson distribution and

$$e = 2.7183 = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots$$

### 11.15 MEAN AND VARIANCE OF POISSON DISTRIBUTION

Poisson distribution is :  $P(r) = \frac{e^{-m}m^r}{r!}, r = 0, 1, 2, 3\dots$

$$\text{Mean} = m; \quad \text{Variance} = m.$$

$$\text{Standard deviation} = \sqrt{\text{Variance}} = \sqrt{m}.$$

### 11.16 RECURRENCE RELATION

$$\text{We have, } P(r) = \frac{e^{-m}m^r}{r!} \quad \therefore \quad P(r+1) = \frac{e^{-m}m^{r+1}}{(r+1)!}.$$

$$\text{Now, } \frac{P(r+1)}{P(r)} = \frac{m^{r+1}e^{-m}}{(r+1)!} \times \frac{r!}{m^r e^{-m}} = \frac{m}{r+1}.$$

$$P(r+1) = \frac{m}{r+1} P(r).$$

which is a recurrence relation.

With this formula we can find  $P(1), P(2), P(3), P(4), \dots$  if  $P(0)$  is given.

**Example 19 :** If the variance of the Poisson distribution is 2, find the distribution for  $r = 1, 2, 3, 4$  and 5 from the recurrence relation of the distribution. (Use  $e^2 = 0.1353$ ).

**Solution :** Here variance =  $m = 2$ .  $P(0) = e^{-2} = 0.1353$  (given). We know that

$$P(r+1) = \frac{m}{r+1} P(r) = \frac{2}{r+1} P(r).$$

$$P(1) = \frac{2}{0+1} \times P(0) = 2e^{-2} = 2 \times 0.1353 = 0.2706 \quad [\because P(0) = 0.1353]$$

$$P(2) = \frac{2}{2} \times P(1) = P(1) = 0.2706$$

$$P(3) = \frac{2}{3} \times P(2) = \frac{2}{3} \times 0.2706 = 0.1804$$

$$P(4) = \frac{2}{4} \times P(3) = \frac{1}{2} \times P(3) = \frac{1}{2} \times 0.1804 = 0.0902$$

$$P(5) = \frac{2}{5} \times P(4) = \frac{2}{5} \times 0.0902 = 0.0361.$$

**Example 20 :** Suppose a book of 285 pages contains 43 typographical errors. If these errors are randomly distributed throughout the book, what is the probability that 10 pages, selected at random, will be free from errors? (Use  $e^{-0.735} = 0.4795$ ).

**Solution :** Here  $n = 10$ ,  $p = \frac{43}{585} = 0.0735$ , mean  $= np = 0.735$

$$\text{Then Poisson distribution is given by : } P(r) = \frac{e^{-m} m^r}{r!} = \frac{e^{-0.735} \times (0.735)^r}{r!}$$

$$\therefore \text{Probability of zero error} = P(0) = \frac{e^{-0.735} \times (0.735)^0}{0!} = e^{-0.735} = 0.4795 \quad (\text{given})$$

**Example 21 :** The mortality rate for a certain disease is 7 in 1000. What is the probability for just 2 details on account of this disease in a group of 400? Given  $e^{-2.8} = 0.06$ .

**Solution :** Here  $p = \frac{7}{1000}$ , and  $n = 400$ ,  $m = np = 400 \times \frac{7}{1000} = 2.8$

Now required probability is given by :

$$P(2) = \frac{m^2 e^{-m}}{2!} = \frac{(2.8)^2}{2!} e^{-2.8} = \frac{2.8 \times 2.8}{2!} \times 0.06 = 0.2352 = 23.52\%.$$

**Example 22 :** It is given that 3% of the electric bulbs manufactured by a company are defective. Using Poisson distribution, find the probability that a sample of 100 bulbs will contain no defective bulb. Given that  $e^{-3} = 0.05$ .

**Solution :** Let  $p$  be the probability of defective bulb. Then

$$p = \frac{3}{100}, n = 100. \text{ Also } \lambda = np = 100 \times \frac{3}{100} = 3$$

$$P(r) = \frac{e^{-\lambda} \lambda^r}{r!}.$$

Now probability that the sample will contain no defective bulb is

$$P(0) = \frac{e^{-3}(3)^0}{0!} = e^{-3} = 0.05.$$

**Example 23 :** In a certain manufacturing process, 5% of the tools produced turn out to be defective. Find the probability that in a sample of 40 tools, utmost 2 will be defective. (Given that  $e^{-2} = 0.135$ ).

**Solution :** The parameter of the Poisson distribution is given by  $m = np = 40 \times 0.05 = 2$ .

The probability of  $r$  defects, i.e.,  $P(X = r)$ , i.e.,  $P(X = r) = \frac{e^{-m} m^r}{r!}$ .

$$\begin{aligned} P(\text{utmost 2 defects}) &= P(0 \text{ defect}) + P(1 \text{ defect}) + P(2 \text{ defects}) \\ &= e^{-2} (1 + 2 + 2) = 0.135 \times 5 = 0.675. \end{aligned}$$

## 11.17 FITTING OF A POISSON DISTRIBUTION

The probability of ' $r$ ' successes in Poisson distribution is given by

$$P(r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, 3, \dots, n$$

and ' $m$ ' is mean of the distribution.

Suppose this experiment is repeated  $N$  times.

$$\text{Then the frequency of } 'r' \text{ successes is : } N \times p(r) = N \times \frac{e^{-m} m^r}{r!}.$$

Putting  $r = 0, 1, 2, \dots, n$ , we can get the expected or theoretical frequencies as given below.

Number of successes ( $r$ ) :	0	1	2	3	....	$e^n$
Expected frequency	$N e^{-m}$	$N \times e^{-m} \times \frac{m}{1!}$	$N \times \frac{e^{-m} m^2}{2!}$	$N \times \frac{e^{-m} m^3}{3!}$	....	$N \times \frac{e^{-m} m^n}{n!}$

The distribution given above is the **Poisson distribution with mean  $m$** . Thus, we have

$$N P(0) = N e^{-m},$$

$$N P(1) = N \times e^{-m} \times \frac{m}{1!} = N P(0) \times \frac{m}{1};$$

$$N P(2) = N \times e^{-m} \times \frac{m^2}{2!} = N P(1) \times \frac{m}{2};$$

$$N P(3) = N \times e^{-m} \times \frac{m^3}{3!} = N \times \left( e^{-m} \times \frac{m^2}{2!} \right) \times \frac{m}{3} = N P(2) \times \frac{m}{3};$$

$$N P(4) = N P(3) \times \frac{m}{4} \quad \text{and so on and}$$

$$\text{In general, } N P(r) = N \times P(r - 1) \times \frac{m}{r}.$$

**Example 24 :** Assuming that the typing mistakes per page committed by a typist follows a Poisson distribution, find the expected frequencies for the following distribution of typing mistakes:

No. of mistakes per page :	0	1	2	3	4	5
----------------------------	---	---	---	---	---	---

No. of pages :	40	30	20	15	10	5
----------------	----	----	----	----	----	---

(Value of  $e^{-1.5} = 0.22313$ ).

**Solution :** Here  $N = 120$ .

$$\text{Mean} = m = \frac{40 \times 0 + 30 \times 1 + 20 \times 2 + 3 \times 15 + 4 \times 10 + 5 \times 5}{120} = \frac{180}{120} = 1.5$$

Frequencies are  $P(0), P(1), P(2), \dots, P(5)$ , where

$$P(0) = e^{-1.5} = 0.22313 ;$$

$$P(1) = e^{-1.5} \times 1.5 = 0.334695; [\because e^{-1.5} = 0.22313]$$

$$P(2) = e^{-1.5} \times \frac{(1.5)^2}{2!} = 0.25 ;$$

$$P(3) = e^{-1.5} \times \frac{(1.5)^3}{3!} = 0.13;$$

$$P(4) = e^{-1.5} \times \frac{(1.5)^4}{4!} = 0.05 ;$$

$$P(5) = e^{-1.5} \times \frac{(1.5)^5}{5!} = 0.01.$$

The expected frequencies are given by  $N \times \frac{e^{-\lambda} \lambda^r}{r!}$ ;  $r = 0, 1, 2, 3, 4, 5$ .

No. of Mistakes	No. of Pages	Expected Frequency $Ne^{-\lambda} \lambda^x/x!$
0	40	$120 \times 0.22313 = 27$
1	30	$120 \times 0.334695 = 40$
2	20	$120 \times 0.25 = 30$
3	15	$120 \times 0.13 = 16$
4	10	$120 \times 0.05 = 6$
5	5	$120 \times 0.01 = 1$
	$N = 120$	120

**Example 25 :** The average number of customers, who appear at a counter of a red cross blood bank per minute is two. Find the probability that during a given minute.

(i) No customer appears, (b) Three or more customers appear. (Given  $e^{-2} = 0.1353$ ).

**Solution :** Here  $m = 2$ , so that the Poisson distribution is

$$P(x) = \frac{e^{-2} 2^x}{x!}; r = 0, 1, 2, \dots;$$

$$(i) P(0) = e^{-2} = 0.1353$$

$$\begin{aligned} (ii) P(x \geq 3) &= 1 - P(x \leq 2) = 1 - [P(x = 0) + P(x = 1) + P(x = 2)] \\ &= 1 - (e^{-2} + 2e^{-2} + 2e^{-2}) = 1 - 5e^{-2} = 1 - 5(0.1353) \\ &= 1 - 0.6765 = 0.3235. \end{aligned}$$

**Example 26 :** Which probability distribution is appropriate to describe the situation where 100 misprints are distributed randomly throughout the 100 pages of a book? For this distribution, find the probability that a page selected at random will contain atleast three misprints.

(Value of  $e^{-1} = 1/2.718$ ).

**Solution :** Since the number of trials is large and the probability of occurrence of printing mistake is very small, so Poisson distribution is appropriate.

$$m = np = 100 \times \frac{1}{100} = 1, P(x = r) = \frac{e^{-m} m^r}{r!}$$

$$\begin{aligned}
 P(x \leq 3) &= 1 - P(0) - P(1) - P(2) \\
 &= 1 - \left[ \frac{e^{-1}}{0!} + \frac{e^{-1}}{1!} + \frac{e^{-1}}{2!} \right] = 1 - \left[ \frac{1}{2.718} + \frac{1}{2.718} + \frac{1}{5.436} \right] \\
 &= 1 - \frac{5}{5.436} = 1 - 0.92.
 \end{aligned}$$

**Example 27 :** Suppose that the chance of an individual cool-miner being killed in a mine accident during a year is  $(1/1400)$ . Use the Poisson distribution to calculate the probability that in the mine employing 350 miners, there will be atleast one fatal accident in a year. (Use  $e^{-0.25} = 0.78$ ).

**Solution :** Here,  $p = \frac{1}{1400}$ ,  $n = 350$ ;  $\therefore$  Mean =  $m = np = \frac{350}{1400} = 0.25$

The poisson distribution is :  $P(r) = \frac{e^{-0.25}(0.25)^r}{r!}$ ;  $r = 0, 1, 2, 3, \dots$

Probability of zero fatal accident :  $P(0) = \frac{e^{-0.25}(0.25)^0}{0!} = e^{-0.25} = 0.78$

**P (atleast one fatal accident) =  $1 - P(0) = 1 - 0.78 = 0.22$ .**

**Example 28 :** In a certain factory turning out optical lenses there is a small chance of  $1/500$  for any blade to be defective. The lenses are supplied in a packet of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective, one defective, two defective and three defective lenses in a consignment of 10,000 packets. (Given  $e^{-0.02} = 0.9802$ ).

**Solution :** Here,  $N = 10,000$ ;  $p = \frac{1}{500}$ ,  $n = 10$ .  $\therefore$  Mean =  $m = np = 10 \times \frac{1}{500} = \frac{1}{50} = 0.02$

Let  $x$  be the number of defective blades in a packet.

Let  $P(x)$  be the number of packets containing  $x$  defective blades, then

$$P(x) = N \times \frac{e^{-m} m^x}{x!}; \quad x = 0, 1, 2, 3, \dots$$

$P(0)$  = Number of packets with no defective blades

$$= 10,000 \times \left[ 0.9802 \times \frac{(0.02)^0}{0!} \right] = 10,000 \times 0.9802 = 9802.$$

$\therefore$  Number of packets with no defective blade = 9802

$P(1)$  = Number of packets with one defective blade

$$\begin{aligned}
 &= 10,000 \times \left[ 0.9802 \times \frac{0.02}{1!} \right] = 10,000 [0.9802 \times 0.02] \\
 &= 10,000 [0.019604] = 196.04
 \end{aligned}$$

$\therefore$  Number of packets with one defective blade = 196.04 or 196.

$P(2)$  = Number of packets with 2 defective blades

$$\begin{aligned}
 &= 10,000 \times \left[ 0.9802 \times \frac{(0.02)^2}{2!} \right] = 10,000 \left[ 0.9802 \times \frac{0.0004}{2} \right] \\
 &= 10,000 [0.00019604]^2 = 1.96 = 2 \text{ Packets}
 \end{aligned}$$

**P(3) = Number of packets with 3 defective blades**

$$= 10,000 \times \left[ 0.9802 \times \frac{(0.02)^3}{3!} \right] = 0.0134 = \text{no Packets.}$$

**Example 29 : In a Poisson distribution 3 P(x = 2) = P(x = 4). Find P (x = 3).**

**Solution :** Since,

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!} ; x = 0, 1, 2, 3, \dots$$

Now,

$$3P(x = 2) = P(x = 4) \Rightarrow 3 \left( \frac{\lambda^2 e^{-\lambda}}{2!} \right) = \frac{\lambda^4 e^{-\lambda}}{4!}$$

$$\Rightarrow \frac{3\lambda^2 e^{-\lambda}}{2 \times 1} = \frac{\lambda^4 e^{-\lambda}}{4 \times 3 \times 2 \times 1} \Rightarrow \frac{3}{2} = \frac{1}{24} \lambda^2$$

$$\Rightarrow \lambda^2 = 36 \Rightarrow \lambda = \pm 6.$$

Since mean is always positive, therefore,  $\lambda = 6$

$$\text{Now, } P(x = 3) = \frac{\lambda^3 e^{-\lambda}}{3!} = \frac{(6)^2 e^{-6}}{6} = 36 e^{-6}.$$

**Example 30 : A car hire firm has two 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 mean 1.5. Calculate the proportion of days on which no car is used and the proportion of days on which some demand is refused ( $e^{-1.5} = 0.2231$ ).**

**Solution :** Since the number of demands for a car on each day is distributed as Poisson distribution with mean 1.5, therefore, the probability of having  $r$  demands on a day is given by

$$P(r) = \frac{(1.5)^r e^{-1.5}}{r!} ; r = 0, 1, 2, 3, \dots$$

And the proportion of days on which no car is used

$$= \text{Probability of no demand} = P(0) = e^{-1.5} = 0.2231$$

And the proportion of days on which some demand is refused = Probability of more

$$\begin{aligned}
 \text{than two demands in a day} &= 1 - [P(0) + P(1) + P(2)] = 1 - \left[ e^{-1.5} + 1.5 e^{-1.5} + \frac{(1.5)^2}{2} e^{-1.5} \right] \\
 &= 1 - 0.2231 \times 3.625 = 0.1913.
 \end{aligned}$$

**Example 31 : It is known from the past experience that in a certain plant they are on the average 4 industrial accidents per month. Find the probability that in a given year, there will be less than 4 accidents. Assume Poisson distribution. (Given  $e^{-4} = 0.0183$ ).**

**Solution :** Let the random variable  $X$  denote the number of accidents in the plant per month.

$$\therefore P(X = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, 3, \dots, n \quad \text{Here } m = 4$$

$$\therefore P(X = r) = \frac{4^r e^{-4}}{r!}, \quad r = 0, 1, 2, 3, \dots, n$$

$$\begin{aligned} \text{Again, } P(X < 4) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= e^{-4} \left[ 1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!} \right] = e^{-4} [1 + 4 + 8 + 10.67] \\ &= e^{-4} \times 23.67 = 0.0183 \times 23.67 = 0.4332. \end{aligned}$$

**Example 32 :** Six coins are tossed 6,400 times. Using the Poisson distribution, what is approximate probability of getting six heads  $x$  times.

**Solution :** Probability of getting a head in a throw of a coin =  $\frac{1}{2}$ .

$$\text{Probability of getting six heads in a throw of six coins} = \frac{1}{2^6} = \frac{1}{64}.$$

$$\therefore \text{Mean} = m = np = 6400 \times \frac{1}{64} = 100.$$

Hence, the probability of getting six heads  $x$  times according to Poisson distribution

$$\text{is } P(x) = \frac{e^{-100} (100)^x}{x!}.$$

**Example 33 :** A Hospital receives patients at the rate of 3 patients per minute on average. What is the probability of receiving no patient in one minute interval.

**Solution :** Let the random variable  $x$  denote the number of patients per minute, then  $x$  follows Poisson distribution with parameter  $m = 3$  and probability density function.

$$P(x = r) = \frac{m^r e^{-m}}{r!} = \frac{3^r e^{-3}}{r!}, \quad r = 0, 1, 2, 3, \dots \quad [\because m = 3]$$

Probability of no patients in one minute =  $P(0)$ , where

$$P(0) = \frac{3^0 e^{-3}}{0!} = 0.04979 \quad [e^{-3} = 0.4979 \text{ from table}]$$

## EXERCISE 11.2

- Between the hours 2 p.m and 4 p.m the average number of phone calls per minute coming into the switchboard of a company is 2.35. Find the probability that during one particular minute there will be at most 2 phone calls. [Given  $3^{-2.35} = 0.095374$ ].

[Hint :  $P(r) = \frac{e^{-m} m^r}{r!} = \frac{e^{-2.35} \times (2.35)^r}{r!}, \quad r = 0, 1, 2, 3, \dots$

$$P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) = e^{-2.35} \left[ 1 + \frac{2.35}{1!} + \frac{(2.35)^2}{2!} \right] = 0.5828543$$

2. If 5% of the electric bulbs manufactured by a company are defective. Use poisson distribution to find the probability that in a sample of 100 bulbs will be defective. [Given  $e^5 = 0.007$ ].

[Hint : Here  $m = np = 100 \times 0.05 = 5$

$$P(X=r) = \frac{e^{-m} m^r}{r!} = \frac{e^{-5} 5^r}{r!}, r = 0, 1, 2, \dots$$

$$\therefore P(X=5) = \frac{e^{-5} 5^5}{5!} = \frac{0.007 \times 625}{5!} = \frac{4.375}{24} = 0.1823.$$

3. A manufacturer of lenses knows that 5% of his product is defective. If he sells lenses in boxes of 100 and guarantees that not more than 10 lenses will be defective, what is the probability (approximately) that a box will fail to meet the guaranteed quality?

[Hint : Here  $m = np = 100 \times 0.05 = 5$ .  $P(X=r) = \frac{e^{-5} 5^r}{r!} = r = 0, 1, 2, 3, \dots$

$$\therefore P(X > 10) = 1 - P(X \leq 10) = 1 - \sum_{r=0}^{10} \frac{e^{-5} 5^r}{r!}$$

4. If the random variable  $X$  follows poisson distribution such that  $P(x=1) = P(x=2)$ , find (a) the mean of the distribution, (b)  $P(x=0)$ .

[Hint : (a)  $P(X=1) = P(X=2) \Rightarrow \frac{e^{-m} m}{1!} = \frac{e^{-m} m}{2!} \Rightarrow m = 2$ .

$$(b) P(X=0) = \frac{e^{-m} m^0}{0!} = e^{-m} = e^{-2} = 0.1353.$$

5. In a certain factory turning out lenses, there is a 0.2% probability for any lenses to be defective. Lenses are supplied in packets of 10. Using poisson distribution, calculate the approximate number of packets containing no defective, one defective, two defective, three defective lenses respectively in a consignment of 20,000 packets. [Given  $e^{-0.02} = 0.9802$ ].

6. The number of accidents in a year attributed to taxi drivers in a city follows poisson distribution with mean 3. Out of 1,000 taxi drivers, find approximately the number of drivers with (i) no accidents in a year, and (ii) more than 3 accidents in a year. [Given  $e^{-1} = 0.3679$ ,  $e^{-2} = 0.1353$ ,  $e^{-3} = 0.498$ ].

7. A manufacturer, who produces medicine bottles, finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes from the producer of bottles. Using poisson distribution, find how many boxes will contain, (i) no defective, (ii) at least two defectives.

[Hint : (i)  $P(r) = \frac{e^{-0.5}(0.5)^r}{r!} \times 100$ . (ii)  $P(0) = 100 e^{-0.5} = 60.65 \approx 61$ .

(ii) Number of boxes containing at least 2 defectives

$$= 100 [1 - e^{-0.5} - e^{-0.5} \times 0.5] = 9.025 \approx 9.$$

8. Suppose that a local appliances shop has found from experience that the demand for tube lights is roughly distributed as poisson with a mean of 4 tube lights per week. If the shop keep 6 tube lights during a particular week, what is the probability that the demand will exceed the supply during the week?
9. A certain firm uses a large fleet of delivery vehicles. Their records over a long period of time (during which their fleet size utilization may be assumed to have remained suitably constant) show that the average number of vehicles per day is 3. Estimate the probability on a given day when,
- (i) all their vehicles will be serviceable,
  - (ii) more than 2 vehicles will be unserviceable, and
  - (iii) exactly 4 vehicles will be unserviceable.
10. Suppose that a manufacture product has 2 defects per unit of product inspected. Using Poisson distribution, calculate the probabilities of finding a product without any defect, with 3 defects and with 4 defects. (Given  $e^{-2} = 0.135$ ).
11. Fit a poisson distribution to the following data and calculate theoretical frequencies.
- |               |     |     |     |    |    |   |   |   |
|---------------|-----|-----|-----|----|----|---|---|---|
| Deaths :      | 0   | 1   | 2   | 3  | 4  | 5 | 6 | 7 |
| Frequencies : | 305 | 365 | 210 | 80 | 28 | 9 | 2 | 1 |
12. The probability that a man aged 50 years will die within a year is 0.01125. What is the probability that out of 12 such men at least 11 will reach their fifty-first birthday.
13. A machine produces large quantities of items and past experiment shows that on an average it produces 1% defective items. A sample of 20 items is drawn from a large number of items. Use Poisson distribution to determine the probability of 2 defectives.
14. In a town 10 accidents took place in a span of 50 days. Assuming that the number of accidents per day follows the poisson distribution, find the probability that there will be three or more accidents in a day.
- [Hint :  $(x \geq 3) = 1 - [P(0) + P(1) + P(2)]$ .
15. Accidents occur on a particular stretch of highway at an average rate of 3 per week. What is the probability that there will be exactly two accidents in a given week? (Given  $e^3 = 20.08$ ).

[Hint :  $P(X=r) = \frac{e^{-m} m^r}{r!}$ ,  $r = 0, 1, 2, \dots$ . Find  $P(X=2) = 0.221$ ].

## ANSWERS

1. 0.5828543.

2. 0.1823.

3.  $1 - e^{-5} \sum_{r=0}^{10} \frac{5^r}{r!}$ .

4. (a)  $m = 2$ ,  $P(2) = 4$ , (b) 0.1353.      5.  $P(0) = 19604$ ,  $P(4) = 0.2161$ ,  $P(1) = 392$ .  
 6. (a)  $100 \times e^{-3} = 49.8$ , (ii) 353 nearly.
7. (i) 61 ; (ii)  $100 [1 - 0.6065 - 0.6065 \times 0.5] \approx 9$ .    8.  $1 - e^{-4} \sum_{r=0}^6 \frac{4^r}{r!} \approx 0.1107$ .
9. (i)  $P(0) = 0.0498$  or 4.98% of the vehicles will be serviceable.  
 (ii)  $P(\text{more than 2 vehicles will be unserviceable}) = 1 - P(X \leq 2) = 1 - 0.4232 = 0.5768$ .
10.  $P(1) = 0.27$ ;  $P(2) = 0.27$ ,  $P(3) = 0.18$ ;  $P(4) = 0.09$ .
11. 301.2, 361.4, 216.8, 86.7.      12. 0.9916.
13. 0.0164.      14. 0.002.      15. 0.224.

### 11.18 NORMAL DISTRIBUTION

Normal distribution is a *continuous probability distribution in which the relative frequencies of a continuous variable are distributed according to the normal probability law*. In other words, it is a symmetrical distribution in which the frequencies are distributed evenly about the mean of the distribution.

The normal distribution of a variable, when represented graphically, takes the shape of a symmetrical curve, known as the *Normal curve*. This curve is asymptotic to x-axis (base line) on its either sides. It is also known as *Normal curve or Error* as the normal curve is extensively used to describe errors made in repeated measurements. It helps us to find the proportion of measurement that falls within a certain range above, below or between selected values.

Normal distribution is a limiting form of Binomial distribution under the following conditions.

- (i)  $n$ , the number of trials is infinitely large i.e.,  $n \rightarrow \infty$
- (ii) neither  $p$  (or  $q$ ) is very small, i.e.,  $p$  and  $q$  are fairly near equally.

A random variable  $x$  is said to have a normal distribution with mean  $\mu$  and standard deviation  $\sigma$  if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

where  $e = 2.7183$ ,  $\sqrt{2\pi} = 2.5060$

The probability density function with mean zero and standard deviation  $\sigma$  is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \quad -\infty < x < \infty.$$

Normal distribution was first discovered by British mathematician De-Movire in 1733. Normal distribution is also known as Gaussian distribution named after Karl Friedrich Gauss who used this distribution to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies. Now-a-days normal probability model is one of the most important probability models in statistical analysis.

## 11.19 STANDARD NORMAL DISTRIBUTION

A random variable  $Z$  which has a normal distribution with mean  $\mu = 0$  and a standard deviation  $\sigma = 1$  is said to be a standard normal distribution. Its probability density function is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, -\infty < z < \infty$$

It is denoted by  $N(0, 1)$ . In short, standard normal variate is written as S.N.V. or s.n.v.

The area under any normal curve is found from the table of a standard normal probability distribution showing the area between the mean and any value of the normally distributed random variable. For a given value of  $\mu$  and  $\sigma$ , and a specific value,  $X$ , of the random variable the standardized normal variate  $Z$  is derived from the following formula:

$$Z = \frac{x - \mu}{\sigma}, \text{ where } \mu \text{ is the mean}$$

The purpose of standardization of normal distribution is to enable us to make use of the tables of the area of the standard curve  $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$  for various points along the  $x$ -axis.

The standard normal distribution is also known as Unit Normal Distribution or Z-distribution.

The standard normal curve helps us to find the areas within two assigned limits under the curve. The areas between the standard normal curve drawn at two assigned limits  $a$  and  $b$  will give the proportion of cases for which the values of  $Z$  lie between  $a$  and  $b$ . Thus the area between two assigned limits  $a$  and  $b$  under the standard normal curve will represent the probability that  $Z$  will be between  $a$  and  $b$ . It is denoted by  $P(a \leq Z \leq b)$ .

## 11.20 PROPERTIES OF NORMAL CURVE

The normal probability curve with mean  $\mu$  and standard deviation  $\sigma$  has the following properties:

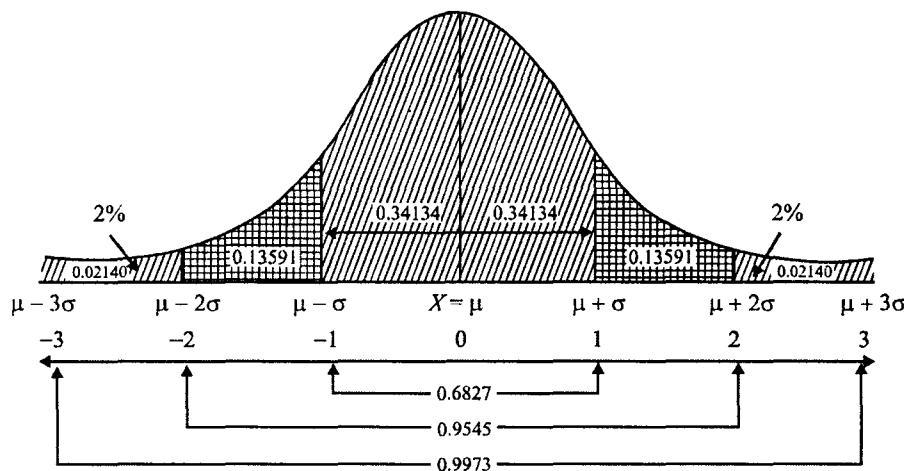
1. The equation of the curve is

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty$$

and it is bell-shaped. The top of the bell is directly above the mean  $\mu$ .

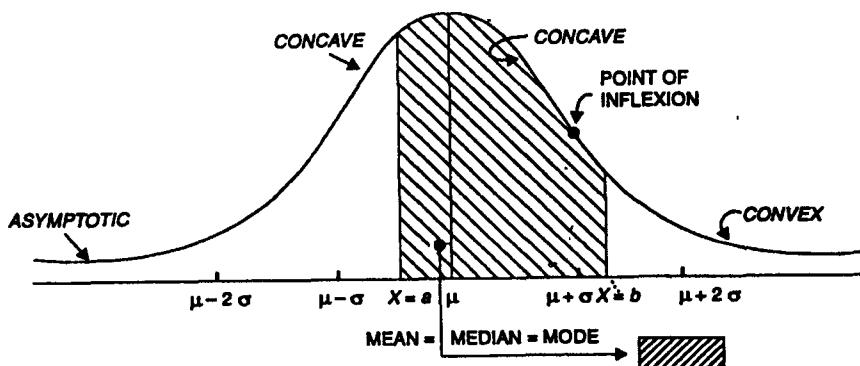
2. The curve is symmetrical about the line  $x = \mu$  and  $x$  ranges from  $-\infty$  to  $+\infty$ .
3. Mean, mode and median coincide at  $x = \mu$  as the distribution is symmetrical.
4. X-axis is asymptote to the curve.
5. The points of inflection of the curve are at  $x = \mu + \sigma$ ,  $x = \mu - \sigma$  and the curve changes from concave to convex at  $x = \mu + \sigma$  to  $x = \mu - \sigma$ .
6. The total area under the normal curve is equal to unity and the percentage distribution of area under the normal curve is given below and is shown also in the figure.

- (i) About 68% of the area falls between  $\mu - \sigma$  and  $\mu + \sigma$ .
- (ii) About 95.5% of the area falls between  $\mu - 2\sigma$  and  $\mu + 2\sigma$ .
- (iii) About 99.7% the area falls between  $\mu - 3\sigma$  and  $\mu + 3\sigma$ .



*Fig. 11.2 : Percentage distribution of area under the normal curve.*

8. In a normal distribution : Q.D. : M.D. : S.D. : 10 : 12 : 15;  
where Q.D. = Quartile Deviation, M.D. = Mean Deviation and S.D. = Standard Deviation.
9. The mean deviation from the mean in normal distribution is equal to  $(4/5)$  of its standard deviation.
10. All the odd moments about the mean are zero.
11. The maximum ordinate lies at the mean, i.e., at  $x = \mu$ .
12. The curve of normal distribution has a single peak, i.e., it is a unimodal.
13. The two tails of the curve extend indefinitely and never touch the horizontal line.
14. The mathematical equation is completely determined if  $\mu$  and  $\sigma$  are known.



*Fig. 11.3*

15. Since mean = median =  $\mu$ , the coordinate at  $x = \mu$  (or  $Z = 0$ ) divides the whole area into two equal parts. The area to the right of the ordinate as well as to the left of the ordinate at  $x = \mu$  (or  $Z = 0$ ) is 0.5.
16. No portion of the curve lies below the  $x$ -axis as  $f(x)$ , being the probability function can never be negative.

## 11.21 APPLICATIONS OR USES OF NORMAL DISTRIBUTION

1. The normal distribution can be used to approximate the binomial and poisson distributions.
2. It has extensive use in sampling theory. It helps us to estimate parameter from statistic and to find confidence limits of parameter.
3. It has a wide use in testing Statistical Hypothesis and Tests of significance in which it is always assumed that the population from which the samples have been drawn should have normal distribution.
4. It has significant applications in statistical quality control as the control chart in statistical quality control is closely related to normal distribution.
5. It can be used for smoothing and graduating a distribution which is not normal, simply by contracting a normal case.
6. It serves as a guiding instrument in the analysis and interpretation of statistical data.

## 11.22 METHOD TO FIND THE PROBABILITY WHEN THE LIMITS OF STANDARD NORMAL VARIATES ARE GIVEN

Let  $Z$  be a standard normal variate  $N(0, 1)$ . The area under the standard normal curve is given by the table at the end of the book. Let us find area under the curve between 0 and 1.53, i.e.,  $0 \leq z \leq 1.53$ .

- I. Find 1.5 at the left of the table and single out that row.
- II. Find 0.3 at the top of the table and single out that column.
- III. In the body of the table, where the row 1.5 and the column for 0.3 meet, is the entry 4370. This is the area under the standard normal curve between 0 and 1.53.

[Note : We use the row for 1.5 and column for .03 because  $1.53 = 1.5 + 0.3$ ].

- I. Similarly to find the area between 0 and 0.96, use row 0.9 and column 0.06 ( $\because 0.96 = 0.9 + 0.06$ ), the answer is 0.3315.
- II. Since the standard normal variate  $Z$  is symmetrical about zero, so the area between  $-Z$  and 0 = (Area between  $Z$  and 0).
- III. Total area under the normal curve is 1 as the area is proportional to the probability.
- IV. If  $X$  is a random variable with mean  $\mu$  and variance  $\sigma^2$ , then  $Z = (X - \mu)/\sigma$  has mean 0 and variance 1. If  $X$  is normally distributed then  $Z$  is a standard normal variate  $N(0, 1)$ . The variance  $Z$  can be looked on as the number of standard deviation from the mean.

**For example,** If  $Z = 1.7$ , then  $Z = \frac{x - \mu}{\sigma} = 1.7 \Rightarrow x - \mu = 1.7 \sigma$ . Thus saying  $Z$  is 1.7

is the same thing as saying that the distance from  $x$  to mean  $\mu$  is 1.7, when the standard deviation  $\sigma$  is taken as the unit of measurement.

**Example 34 :** What is the probability that a standard normal variate  $Z$  will be (a) greater than 1.09 ; (b) less than -1.65, (c) lying between -1.00 and 1.96 ; (d) lying between 1.25 and 2.75?

**Solution :** (a) The shaded area to the right of  $Z = 1.09$  is the probability that  $Z$  will be greater than 1.09. In the table the area between 0 and 1.09 is 0.3621. The total area under the curve is equal to 1, so that the area to the right of zero must be 0.5000 ( $\because$  curve is symmetrical at  $x = 0$ ).

$$P(Z > 1.09) = 0.5000 - 0.3621 = 0.1375$$

(b) The shaded area to the left of  $Z = -1.65$  is the probability that  $Z$  will be less than -1.65. Also the area between -1.65 and 0 is the same as area between 0 and

1.65. In the table the area between zero and 1.65 is 0.4505. But the area to the left of zero is 0.05.

$$\therefore P(Z \leq 1.65) = 0.5000 - 0.4505 = 0.0495.$$

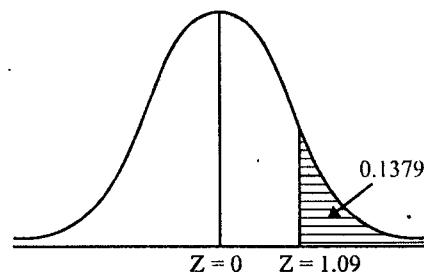


Fig. 11.4

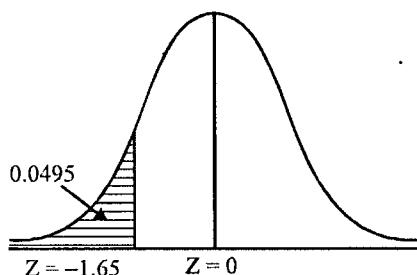


Fig. 11.5

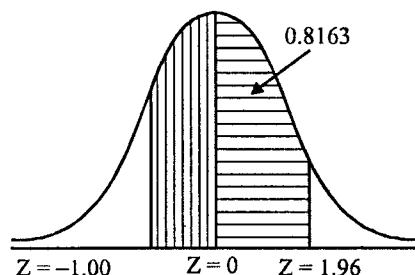


Fig. 11.6

(c) The probability that the random variable  $Z$  is between -1.00 and 1.96 is the shaded area in the figure 11.5 it is found by adding the corresponding areas shown in the figure.

Also area between -1.00 and 1.96 = Area between  
(-1.00 and 0) + area between (0 and 1.96).

$$P(-1.00 < Z < 1.96)$$

$$= P(-1.00 < Z < 0) + P(0 < Z < 1.96) \\ = 0.3413 + 0.4750 = 0.8163$$

(d) The shaded area between  $Z = 1.25$  and  $2.75$  is the probability that  $Z$  will be between 1.25 and 2.75.

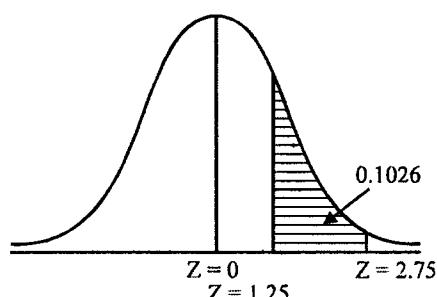


Fig. 11.7

But area between  $z = 1.25$  and  $z = 2.75$  = (Area between  $Z = 0$  and  $Z = 2.75$ ) – (Area between  $Z = 0$  and  $Z = 1.25$ ).

$$\therefore P(1.25 < Z < 2.75) = P(0 < Z < 2.75) - P(0 < Z < 1.25)$$

$$= 0.4970 - 0.3944 = 0.1026 \quad (\text{From the table})$$

**Example 35 :** Find the value of  $Z$  such that the probability of a larger value is 0.7881.

**Solution :** The area to the right of  $Z$  is greater than 0.5000. We know that  $Z$  must be negative and that the area between  $Z$  and 0 must be  $0.7881 - 0.5000 = 0.2881$ . In the table we search out the entry 0.2881 and find the corresponding  $Z$  is  $-0.80$ .

$$\text{Hence } Z > -0.80$$

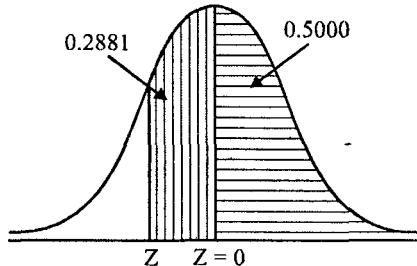


Fig. 11.8

### 11.23 METHOD TO FIND THE PROBABILITY WHEN THE VARIATE IS NORMALLY DISTRIBUTED

Let  $X$  be a normal variate with mean  $\mu$  and standard deviation  $\sigma$ . Suppose we want to find the probability that a randomly selected value of  $X$  will lie between  $a$  and  $b$  (i.e.,  $P(a < X < b)$ ).

**Step 1.** Convert  $X$  into a standard normal variate by the formula.

$$Z = \frac{X - \mu}{\sigma} \quad \dots (1)$$

**Step 2.** Find the limits of  $Z$  corresponding to the limits of  $X$ .

$$\text{When } X = a, \text{ then } Z = \frac{a - \mu}{\sigma} \quad [\text{Put } X = a \text{ in (1)}]$$

$$\text{When } X = b, \text{ then } Z = \frac{b - \mu}{\sigma} \quad [\text{Put } X = b \text{ in (1)}]$$

Thus the limits of  $Z$  are  $\frac{a - \mu}{\sigma}$  to  $\frac{b - \mu}{\sigma}$ , when  $X = a$  to  $X = b$ .

**Step 3.** Thus  $P(a < X < b) = P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right)$ .

**Step 4.** From the table find the probability that  $Z$  lies between  $((a - \mu)/\sigma)$  and  $((b - \mu)/\sigma)$ . The method is illustrated by the following examples.

**Example 36 :** A certain type of wooden beam has a mean breaking strength of 1500 kgs and a standard deviation of 100 kgs. Find the relative frequency of all such beams whose breaking strengths lie between 1450 and 1600 kgs.

**Solution :** Let  $X$  be the breaking strength. Then we are to find  $P(1450 < X < 1600)$ .

Here  $X$  is a normal variate with mean 1500 and standard deviation 100.

$$\therefore Z = \frac{X - 1500}{100} \text{ is a standard normal variate } N(0, 1).$$

Also when  $X = 1450$ , then  $Z = \frac{1450 - 1500}{100} = -0.5$

When  $X = 1600$ , then  $Z = \frac{1600 - 1500}{100} = 1$

Thus  $P(1450 < X < 1600) = P(-0.5 < Z < 1) = 0.1951 + .3413 = .5328$  [See tables]

$\Rightarrow$  53% of the beams have the breaking strength between 1450 kgs and 1600 kgs.

**Example 37 :** Assume the mean height of children to be 68.22 cm with a variance of 10.8 cm. How many children in a school of 1,000 would you expect to be over 72 cm tall?

**Solution :** Let the distribution of height be normal. Let  $X$  be a normal variate with mean 68.22 and standard deviation  $\sqrt{10.8}$ .

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 68.22}{\sqrt{10.8}}$$

When  $X = 72$  cm, then  $Z = \frac{72 - 68.22}{\sqrt{10.8}} = \frac{3.78}{3.286} = 1.15$ .

Now  $P(X > 72) = P(Z > 1.15) = \text{Area to the right of the coordinate at } Z = 1.15 = (\text{Total area on the right of } Z = 0) - (\text{Area from } Z = 0 \text{ to } Z = 1.15)$   
 $= (0.5000 - 0.3749) = 0.1251$ .

Expected number of children to be above 72 cm out of 1000  $= 0.1251 \times 1000 = 125.1$  or 125 children.

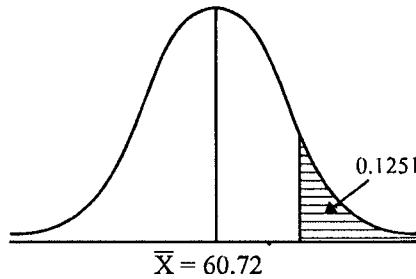


Fig. 11.9

**Example 38 :** The life time of a certain kind of pace maker has a mean of 300 days and a standard deviation of 35 days. Assuming that the distribution of life times, which are measured to the nearest day is normal, find the percentage of pace makers which have life time of more than 370 days.

**Solution :** Let  $X$  be a random normal variate measuring the lift time of pace maker.

Here mean  $\mu = 300$ ,  $\sigma = 35$ .

$$\therefore Z = \frac{X - \mu}{\sigma} = \frac{X - 300}{35} \text{ is a standard normal variate.}$$

When,  $X = 370$ , then  $Z = \frac{370 - 300}{35} = 2$ .

$$\begin{aligned}
 P(X > 370) &= P(Z > 2) \\
 &= \text{Area between } Z = 0 \text{ and } Z = 2 \\
 &= 0.4772 \text{ (from the table).}
 \end{aligned}$$

$$\begin{aligned}
 \text{Area of } Z > 2 &= 0.5000 - 0.4772 \\
 &= 0.0228 = (\text{Shaded Region})
 \end{aligned}$$

$\therefore$  The percentage of batteries having life time more than 370 days  $= 0.0228 \times 100 = 2.28\%$ .

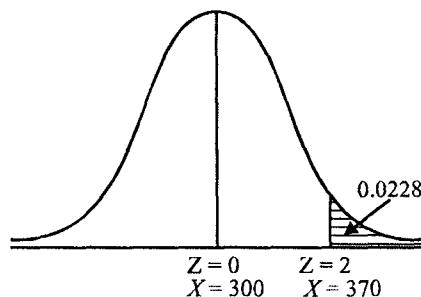


Fig. 11.10

**Example 39 :** (a) Find the area to the left of  $Z = 1.90$ .

(b) Find the area to the right of  $Z = 0.25$ .

**Solution :** (a) Total area to the left of  $Z = 0$  is 0.5000

.... (I)

Area between  $Z = 0$  and  $Z = 1.90 = 0.4713$  (from the table)

.... (II)

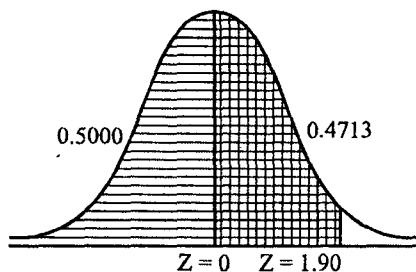


Fig. 11.11

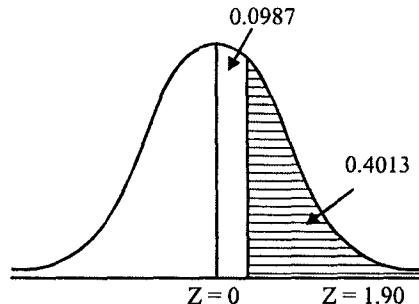


Fig. 11.12

$\therefore$  Total area to the left of  $Z = 1.90 = 1 + II = 0.5000 + 0.4713 = 0.9713$  (Shaded region).

(b) Total area to the right of  $Z = 0$  is 0.5000

.... (III)

Area from  $Z = 0$  to  $Z = 0.25$  is 0.0987 (from the table)

.... (IV)

**Area to the right of  $Z = 0.25$**

$$\begin{aligned}
 &= (\text{Total area to the right of } Z = 0) - (\text{Area from } Z = 0 \text{ to } Z = 0.25) \\
 &= III - IV = 0.5000 - 0.0987 = 0.4013 \text{ (Shaded region).}
 \end{aligned}$$

**Example 40 :** 1,000 light bulbs with a mean life of 120 days are installed in a new factory. Their length of life is normally distributed with a standard deviation of 20 days.

(a) How many bulbs will expire in less than 90 days?

(b) If it is decided to replace all bulbs together, what interval should be allowed between replacement if not more than 10% should expire before replacement.

**Solution :** Let  $X$  be the normal variate of life of light bulbs. Then its mean  $\mu = 120$  and standard deviation  $\sigma = 20$ .

Now  $Z = \frac{X - \mu}{\sigma} = \frac{X - 120}{20}$  is a S.N.V

(a) When  $X = 90$ , then  $Z = \frac{90 - 120}{20} = -1.5$ .

$\therefore$  Number of bulbs expected to expire less than 90 days out of 1,000 bulbs  $= 1000 \times 0.0668 \approx 66.8$  or 67 bulbs.

(b) Since not more than 10% or 0.1 expire before replacement so the value  $Z$  to an area  $0.5 - 0.1 = 0.4$  is 1.28.

Thus the value of  $Z$  is less than -1.28.

$$\therefore Z = \frac{x - 120}{20} = -1.28$$

$$\Rightarrow X = 120 - 25.6 = 94.4 \text{ or } 94 \text{ days.}$$

Then the bulbs may be replaced after 94 days.

**Example 41 :** The scores made by candidate in a certain test are normally distributed with mean 500 and standard deviation 100. What percentage of candidate receives the scores between 400 and 600?

**Solution.:** Let  $X$  the normal variate showing scores of candidates. Its mean  $\mu = 500$ ,  $\sigma = 100$ .

Now  $Z = \frac{X - \mu}{\sigma} = \frac{X - 500}{100}$  is

a standard normal variate  $N = (0, 1)$ .

$$\text{When } X = 400, \text{ then } Z = \frac{400 - 500}{100} = -1.$$

$$\text{When } X = 600, \text{ then } Z = \frac{600 - 500}{100} = 1.$$

When  $X$  lies between 400 and 600, then  $Z$  lies between -1 and 1.

**Area under the curve from ( $Z = -1$  to  $Z = 1$ )** = Area under the curve from ( $Z = -1$  to  $Z = 0$ ) + Area under the curve from ( $Z = 0$  to  $Z = 1$ ).

$$= 2 \times (\text{area under the curve from } Z = 0 \text{ to } Z = 1) \quad (\text{Symmetric property}) \\ = 2 \times .34134 = 0.6827.$$

$\therefore$  Percentage of candidates securing marks between 400 and 600  
 $= 0.6827 \times 100 = 68.27\%$ .

**Example 42 :** 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.

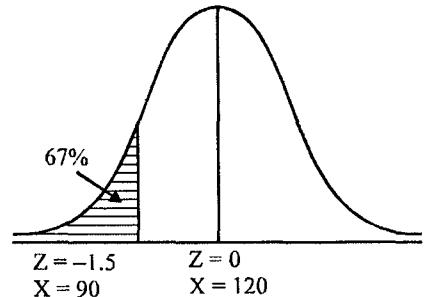


Fig. 11.13

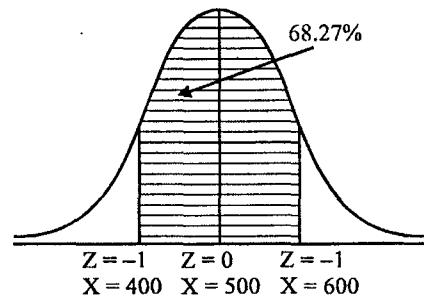


Fig. 11.14

**Solution :** 31% of items are under 45  $\Rightarrow$  Area to the left is 0.31. But area right from this point (0.31) to the point 0.5 is  $(.50 - .31) = 0.19$ .

Let  $X$  be the random variate which is normally distributed with mean  $\mu$  and standard deviation  $\sigma$ . Then,

$$Z = \frac{X - \mu}{\sigma} \text{ is a S.N.V.}$$

The S.N.V. corresponding to  $X = 45$  and  $X = 64$  are as below:

$$\text{When } X = 45, \text{ then } Z = \frac{45 - \mu}{\sigma} = Z_1, \text{ (say) ... (I)}$$

$$\text{When } X = 64, \text{ then } Z = \frac{64 - \mu}{\sigma} = Z_2, \text{ (say) .... (II)}$$

From the figure it is obvious that

$$P(0 < Z < Z_2) = 0.42 \Rightarrow Z_2 = 1.405.$$

Fig. 11.15

$$P(-Z_1 < Z < 0) = 0.19 \Rightarrow P(0 < Z < Z_1) = 0.19$$

$$\Rightarrow Z_1 = 0.496.$$

[From Normal Table]

Substituting the values of  $Z_1$  and  $Z_2$  in (I) and (II), we get

$$\frac{45 - \mu}{\sigma} = -0.496 \Rightarrow 45 - \mu = -0.496 \sigma \quad \dots \text{(III)}$$

$$\frac{64 - \mu}{\sigma} = 1.405 \Rightarrow 64 - \mu = 1.405 \sigma \quad \dots \text{(IV)}$$

Solving (III) and (IV), we get  $\sigma = 10$ ,  $\mu = 49.96 \approx 50$  (approx.)

Hence, mean is 50 and standard deviation is 10.

**Example 43 :** Of a large number of group of children 5% are under 60 cm in height and 40% are between 60 and 65 cm. Assuming a normal distribution, find the mean height and standard deviation.

**Solution :** Let  $X$  be the normal variate of heights of children with mean  $\mu$  and standard deviation  $\sigma$ .

$$\text{Then } Z = \frac{X - \mu}{\sigma} \text{ is a S.N.V.} \approx N(0, 1)$$

The area lying to the left of  $X = 60$  is 0.05.

$\therefore$  The area between  $Z = 0$  and

$$\begin{aligned} Z &= \frac{60 - \mu}{\sigma} \\ &= 0.5 - 0.05 = 0.45 \end{aligned}$$

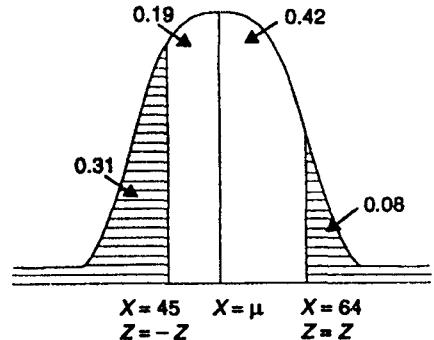


Fig. 11.15

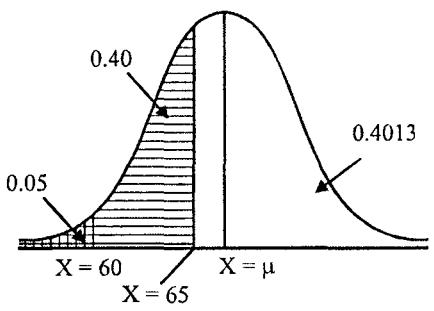


Fig. 11.16

From table, for area 0.45  $\Rightarrow Z = -1.64$  .... (I)

From (I),  $Z = -1.64 = \frac{60 - \mu}{\sigma}$  or  $60 - \mu = -1.64 \sigma$ . .... (II)

Height of children between 60 cm and 65 cm = 40%.

Thus the area between  $Z = 0$  and  $Z = \frac{65 - \mu}{\sigma}$  is 0.5. .... (III)

From table, area 0.5 corresponds to  $Z = 0.13$ . Substituting it in (III), we get

$$Z = -0.13 = \frac{65 - \mu}{\sigma} \Rightarrow 65 - \mu = -0.13 \sigma. \quad \dots \text{(IV)}$$

Solving (II) and (IV), we get

$$\sigma = 3.3 \text{ and } \mu = 65.412.$$

Hence, mean is 65.412 and standard deviation is 3.3.

**Example 44 :** An auditor has found that the credit rewards of a large mail order have approximately normally distributed and show an average account billing error of Rs. 10 and a standard deviation of Re. 1 (billing error may be positive or negative according to whether purchases were overcharged or undercharged). Suppose credit account is randomly selected from the files of the mail order house. Find the probability of a billing error between Rs. 10 and Rs. 11.50.

**Solution :** Let  $X$  be the random normal variate of billing error. Then its mean  $\mu = 10$  and S.D.  $\sigma = 1$ .

Let  $Z = \frac{X - \mu}{\sigma} = \frac{X - 10}{1}$  be a S.N.V.  $\sim N(0, 1)$

when  $X = 10$ , then  $Z = 10 - 10 = 0$ .

when  $X = 11.50$  then  $Z = 11.50 - 10 = 1.5$ .

Billing error between Rs. 10 and Rs. 11.50

= Area under the curve from  $Z = 0$  to  $Z = 1.5 = 0.4332$  [From Normal Table]

**Example 45 :** The mean yield for one-acre plot is 662 kgs with a standard deviation of 32 kgs. Assuming normal distribution, how many one acre plots in a batch of 10,000 plots would expect yield (a) over 700 kgs, (b) below 650 kgs, (c) What is the lowest yield of 1000 plots?

**Solution :** Here  $\mu = 662$ ,  $\sigma = 32$  of the normal variate  $X$ .

$$\therefore Z = \frac{X - \mu}{\sigma} = \frac{X - 662}{32}$$

When  $X = 700$ , then  $Z = \frac{700 - 662}{32} = 1.1875 \approx 1.19$ .

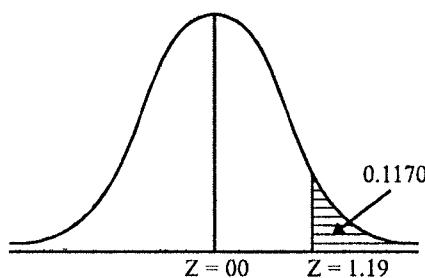


Fig. II.17

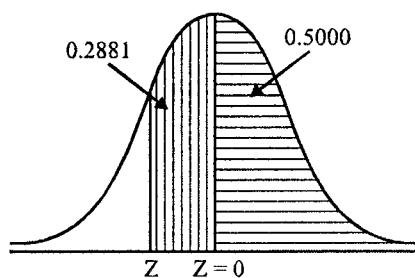


Fig. II.18

$$P(X > 700) = P(Z > 1.19) = 0.5000 - 0.3830 = 0.1170.$$

Expected number of plots out of 10,000 giving a yield over 700 kgs  
 $= 0.1170 \times 10,000 = 1170.$

(b) When  $X = 650$ , then  $Z = \frac{650 - 662}{32} = \frac{-12}{32} = -0.375 \approx -0.38$ .

$$\begin{aligned} P(X < 650) &= P(Z < -0.38) = P(0 < Z < -0.38) \\ &= 0.5000 - 0.1480 = 0.3520. \end{aligned}$$

[From table]

Expected number of plots giving a yield below 650 kgs.  $= 10,000 \times 0.3520 = 3520$

(c) Probability of 1,000 lowest yields plots  $= \frac{1,000}{10,000} = 0.1$

$\therefore$  Standard normal variates having area 0.1 to the left  $= 1.28$  [From Normal Table]

$$\therefore Z = \frac{X - \mu}{\sigma} = 1.28 \Rightarrow \frac{X - 662}{32} = 1.28$$

or

$$X = 662 + 32 \times 1.28 = 662 + 40.96 = 702.96.$$

**Example 46 :** The average weekly food expenditure of families in a certain area has a normal distribution with mean Rs. 125 and standard deviation Rs. 25. What is the probability that a family selected at random from this area will have an average weekly expenditure on food in excess of Rs. 175? What is the probability that out of eight such families selected atleast one family will have their weekly food expenditure in excess of Rs. 175?

**Solution :** Here mean

$$\mu = 125, \sigma = 25;$$

$\therefore$  Normal variable

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 125}{25}$$

$$\text{When } X = 175, \text{ then } Z = \frac{175 - 125}{25} = 2.$$

$$\begin{aligned} \therefore P(X > 175) &= P(Z > 2) = 1 - P(Z \leq 2) \\ &= 1 - [0.5 + P(0 \leq Z \leq 2)] = 1 - 0.9772 = 0.228. \end{aligned}$$

$$\begin{aligned} P(\text{atleast one family has food expenditure} > 175) &= 1 - (1 - 0.0228)^8 \\ &= 1 - (0.9772)^8 = 0.1685. \end{aligned}$$

**Example 47 :** The income of a group of 10,000 persons was found to be normally distributed with mean equal to Rs. 750 and standard deviation equal to Rs. 50. What was the lowest income among the richest 250?

**Solution :** We have  $\mu = 750$  and  $\sigma = 50$ .

Let  $x_1$  be the lowest income among the richest 250 people.

$$\text{The } P(X \geq x_1) = \frac{250}{10,000} \Rightarrow P\left(\frac{X - \mu}{\sigma} \geq \frac{x_1 - 750}{50}\right) = 0.025$$

$$\Rightarrow P\left(Z \geq \frac{x_1 - 750}{50}\right) = 0.025$$

$$\therefore \frac{x_1 - 750}{50} = 1.96 \Rightarrow x_1 = 750 + 50 \times 1.96 = 848.$$

Hence Rs. 848 is the lowest income among the richest 250 people.

**Example 48 :** The marks obtained in a certain examination follow normal distribution with mean 45 and standard deviation 10. If 1000 students appeared at the examination, calculate the number of students scoring (i) less than 40 marks and (ii) more than 60 marks.

**Solution :** Given  $\mu = 45$   $\sigma = 10$ ; Then  $Z = \frac{X - \mu}{\sigma} = \frac{X - 45}{10}$ .

$$(i) \text{ For } X = 40 \Rightarrow Z = \frac{40 - 45}{10} = -0.5; \quad P(X < 40) = P(Z < -0.5)$$

$$\text{or} \quad P(Z > 0.5) = 0.5 - P(0 < Z < 0.5) = 0.5 - 0.1915 = 0.3085.$$

[From Tables]

$$\text{Number of students} = 1000 \times 0.3085 = 309 \text{ (approximately)}$$

$$(ii) \text{ For } X = 60 \Rightarrow Z = \frac{X - \bar{X}}{\sigma} = \frac{60 - 45}{10} = 1.5.$$

$$\begin{aligned} P(X > 60) &= P(Z > 1.5) = 0.5 - P(0 < Z < 1.5) \\ &= 0.5 - 0.4332 = 0.0668. \end{aligned}$$

$$\text{Number of students} = 1000 \times 0.0668 = 67 \text{ (approximately).}$$

**Example 49 :** Let  $X$  be a continuous variable and follows a normal distribution with mean 12 and standard deviation 2. What is the probability that the value of  $X$  selected at random lies in the interval [11, 14]?

**Solution :** Here  $\bar{X} = 12$ ,  $\sigma = 2$ .

Converting the values from  $X$  scale to  $Z$  scale, we have

$$\text{For } X = 11, \quad Z = \frac{X - \bar{X}}{\sigma} = \frac{11 - 12}{2} = \frac{-1}{2} = -0.50$$

$$\text{For } X = 14, \quad Z = \frac{X - \bar{X}}{\sigma} = \frac{14 - 12}{2} = 1$$

We are, therefore, interested in finding the probability that Z lies between, -0.50 and 1.

$$\begin{aligned} \text{Thus } P(11 \leq X \leq 14) &= P(-0.50 \leq Z \leq 1) = P(0 \leq Z \leq 0.50) + P(0 \leq Z \leq 1) \\ &= 0.1915 + 0.3413 = 0.5328. \end{aligned}$$

**Example 49 :** Suppose that in a certain pediatric population, casual, sitting systolic blood pressure is normally distributed with  $\mu = 115 \text{ mm Hg}$ ,  $\sigma^2 = 225 (\text{mm Hg})^2$ . Find the probability that a child randomly selected from this population will have : (a) a systolic pressure less than 140 mm Hg, (b) a pressure greater than 100 mm Hg, (c) a pressure between 110 and 120. Also find (d) the systolic pressure below which pressures of 99 percent of the population lie, (e) the two systolic pressures between which the central 50 percent of the pressures in the population lie. Find (f) the quintiles of the distribution, (g) the percentile rank of 100 mm Hg.

**Solution :** In order to use Table A-4 to solve this set of problems we must first standardize the distribution. That is, if  $x$  denotes systolic blood pressure then we must convert numerical values  $x$  to values of  $z$  by the relation  $z = (x - \mu)/\sigma$ . If we require as a solution a value of  $x$ , we shall obtain directly from the table a value of  $z$  from which  $x$  is determined by the relation  $x = \mu + z\sigma$ .

(a) We must convert the  $x$  value 140 to a  $z$  value

$$z = \frac{140 - \mu}{\sigma} = \frac{140 - 115}{15} = 1.67$$

We find  $P(x < 140) = P(z < 1.67) = 0.9525$ .

(b) Here  $x = 100$  and  $z = \frac{100 - 115}{15} = -1$ .

$$P(x > 100) = P(z > -1) = P(z < 1) = 0.8413$$

$$x_1 = 110; \quad z_1 = \frac{100 - 115}{15} = -0.33$$

$$x_2 = 120; \quad z_2 = \frac{100 - 115}{15} = 0.33$$

$$P(110 < x < 120) = P(-0.33 < z < 0.33) = 0.2586$$

We have  $P(x < x_0) = P(z < z_0) = 0.99$ , where  $z_0 = 2.3263$  and

$$x_0 = 115 + (2.3263)(15) = 149.8945 \text{ or approximately } 150 \text{ mm Hg.}$$

We know that  $P(-z_0 < z < z_0) = 0.50$  and therefore  $z_0 = 0.6745$ .

Thus, between the values  $z_1 = -0.6745$  and  $z_2 = 0.6745$  lie the central 50 percent of pressures.

The  $x$ 's corresponding to these  $z$ 's are

$$x_1 = 115 + (-0.6745)(15) = 104.8825$$

$$\text{and } x_2 = 115 + (0.6745)(15) = 125.1175$$

The central 50 percent of pressures are found between the approximate values 105 and 125 mm Hg.

(f) The quintiles of the distribution divide the total area or probability into five equal parts. Again we must approach the problem by finding first the quintiles of the standard normal

distribution. These are quantities  $z_1, z_2, z_3, z_4$  such that  $P(z < z_1) = 0.20$ ,  $P(z < z_2) = 0.40$ ,  $P(z < z_3) = 0.60$  and  $P(z < z_4) = 0.80$ . From table we find  $z_1 = -0.8416$ ,  $z_2 = -0.2533$ ,  $z_3 = 0.2533$  and  $z_4 = 0.8416$ .

These  $z$  values are converted to  $x$ 's with the following results:  $x_1 = 1023760$ ,  $x_2 = 1112005$ ,  $x_3 = 1187995$  and  $x_4 = 1276240$  or in rounded form,  $x_1 = 102$ ,  $x_2 = 111$ ,  $x_3 = 119$ ,  $x_4 = 128$ , the required quintiles of the distribution of systolic blood pressure.

(g) The given value, 100 mm Hg, is a blood pressure and therefore an  $x$  value. Its corresponding  $z$  value is  $-1$  and  $P(z < -1) = P(z > 1) = 0.1587$ . The percentile rank of 100 is 15.87 percent.

### EXERCISE 11.3

- Find the probability that the standard normal variate lies between 0 to 1.5.
- Find the area under the normal curve for (a)  $Z = 1.64$ ; (b)  $Z = 1.32$ ; (c)  $Z = 0.56$ ; (d)  $Z = -1.54$ .
- Find the area to the (a) right of  $Z = 0.25$ ; (b) left of  $z = 1.96$ , (c) left of  $Z = -1.96$ ; (d) between  $Z = 0.4$  and  $Z = 0.6$ .
- A large number of measurements is normally distributed with mean of 65.5 cm and a standard deviation of 6.2 cm. Find the percentage of measurements that fall between 54.8 cm and 68.8 cm.
- In an intelligence test administered to 1,000 students the average score was 42 and standard deviation 24. Find (a) the number of students exceeding a score of 50; (b) the number of students trying between 30 and 54, (c) the value of score exceeded by top 100 students.

[Hint : (a) When  $X = 50$ , then  $Z = \frac{50 - 42}{24}$ .  $P(Z > 0.333) = 0.5 - 0.1304 = 0.3696$ .

Expected number of students exceeding a score of 50 =  $0.3696 \times 1000 = 369.6$  or 370

(b) When  $X$  lies between 30 and 54, then  $Z$  is between  $-0.5$  and  $0.5$ . Probability of having students with score between 30 and 54 =  $0.1915 + 0.1915 = 0.3830$ .

Expected number of students to score between 30 and 54 =  $1,000 \times 0.3830 = 383$ .

(c) Probability of getting top 100 students  $\frac{100}{1,000} = 0.1$ . Standard normal variate having 0.1 area to the right = 1.28.

$$\text{Also, } Z = \frac{X - \mu}{\sigma} \Rightarrow 1.28 = \frac{X - 42}{24} \Rightarrow X = 72.72 \approx 73.$$

- In a distribution exactly 7% of the items are 35 and 89% are under 63. What are the mean and standard deviations of the distribution?

[Hint :  $Z_1 = \frac{X - \mu}{\sigma} = \frac{35 - \mu}{\sigma} = -1.48 \Rightarrow 35 - \mu = -1.48 \sigma \quad \dots (1)$

$$Z_2 = \frac{63 - \mu}{\sigma} = 1.23 \Rightarrow 63 - \mu = 1.23 \sigma \quad \dots (2)$$

Solve (1) and (2) to get  $\mu = 50.3$ ,  $\sigma = 10.33$ ]

7. A sample of 100 dry battery cells tested to find the length produced the following results:  
 $\bar{X} = 12$  hours,  $\sigma = 3$  hours. Assuming the data to be normally distributed, what percentage of battery cells are expected to have life (a) more than 15 hours, (b) less than 6 hours (c) between 10 and 14 hours.

[Hint : (a) When  $X = 15$ ,  $Z = \frac{15 - 12}{3} = 1$ . Area to the right of  $Z = 1$  is  $0.5 - 0.3413 = 0.1587$ .

Percentage of battery cells having life more than 15 hours  $= 0.1587 \times 100 = 15.87\%$ .

(b) When  $X = 6$ ,  $Z = \frac{6 - 12}{3} = -2$  Area to the left of  $Z$  is  $-2 = 0.5 - 0.4772 = 0.228$ .

% of battery cells having life less than 6 hours  $= 0.0228 \times 100 = 2.28\%$ .

(c) When  $X = 10$ ,  $Z = \frac{10 - 12}{3} = -0.67$ ; when  $X = 14$ ,  $Z = \frac{14 - 12}{3} = 0.67$

Area between  $X = 10$  to  $X = 14$  is twice the area to the left of  $Z = .67 = 2 \times 0.2487 = 0.4974$

∴ % of battery cells having life span between 10 hours and 14 hours = **49.74%**.

8. The mean life time of 100 Watt light bulbs produced by Larsen and Tubro is 200 hours. It is known that the standard deviation is 20 hours. Assuming that the life time of light bulbs are normally distributed, what are the probabilities that a single 100 watt light bulb extracted from the production lot will (a) burn out between 180 hours and 210 hours? (b) burn out for a time greater than 250 hours.

[Hint : (a) when  $X = 180$ , then  $Z = \frac{180 - 200}{20} = -1$  and when  $X = 210$  then

$$Z = \frac{210 - 200}{20} = 0.5$$

Area between  $Z = -1$  and  $Z = 0.5$  is  $3413 + .1915 = 0.5328$ . Required probability = **0.5328**

(b) When  $X = 250$ ,  $Z = \frac{250 - 200}{20} = 2.5$ .

Required probability = Area to the right of  $Z$  equals  $2.5 = (0.5 - 0.4938) = .0062$ ]

9. In a sample of 1,000, the mean weight is 45 kgs with standard deviation 15 kgs. Assuming the normality of the distribution, find the number of items weighting between 40 and 60 kgs.

[Hint :  $P(40 \leq X \leq 60) = P\left(\frac{40 - 45}{15} < \frac{X - \mu}{\sigma} \leq \frac{60 - 45}{15}\right)$ .

$$= P(-0.333 \leq Z \leq 1) = P(-0.333 < Z < 0) + P(0 < Z < 1) = 0.1293 + 0.3413 = 0.4706$$

Required number of items  $= 1,000 \times 0.4706 = 470.6 \approx 471$ ]

10. In a sample of 120 workers in a factory the mean and standard deviation of wages were Rs. 11.35 and Rs. 3.03 respectively. Find the percentage of workers getting wages between Rs. 9 and Rs. 17 in the whole factory assuming that the wages are normally distributed.

11. The weekly wages of 1,000 workmen are normally distributed around a mean of Rs. 70 and with a standard deviation of Rs. 5. Estimate the number of workers whose weekly will be:
  - (a) more than 80, (b) less than Rs. 63, (c) between Rs. 69 and 72.
12. The customer accounts of a certain department at store have an average balance of Rs. 120 and a standard deviation of Rs. 40. Assuming that the account balance are normally distributed, find
  - (a) the proportion of account is more than Rs. 150.
  - (b) the proportion of account is between Rs. 100 and Rs. 150.
  - (c) the proportion of account is between Rs. 60 and Rs. 90.
13. The daily sales of a certain item are normally distributed with a mean of Rs. 8,000 and variance of Rs. 10,000.
  - (a) What is the probability that on a given day the sales will be less than Rs. 8,210?
  - (b) What percent of the days will the sales be between Rs. 8,100 and Rs. 8,210?
14. If the heights of 1000 soldiers in a regiment are distributed normally with a mean of 172 cms and a standard deviation of 5 cms, how many soldiers have heights greater than 183 cms?
15. The height distribution of a group of 2,989 individual is known to be normal distribution with mean 65" and standard deviation 2'1". Find also the number of individuals whose heights lie between 60.8" and 67.1". Find also the number of individuals whose heights are above 67.1".
16. Suppose that a doorway being constructed is to be used by a class of people whose heights are normally distributed with mean 70" and standard deviation 3". How much height the doorway should be, without causing more than 25% of the people to bump their heads? If the height of the door may be fixed at 76", how many persons out of 5,000 are expected to bump their heads?
17. Of a large group of men, 5 percent are under 60 inches in heights and 40 percent are between 60 and 65 inches. Assuming a normal distribution find the mean height and standard deviation.

### ANSWERS

1. 0.4332.
2. (a) 0.4484 ; (b) 0.4049 ; (c) 0.2123 ; (d) 0.4382.
3. (a) 0.4013 ; (b) 0.9750 ; (c) 0.0250 ; (d) 0.03811.
4. 0.6601.
5. (a) 370 ; (b) 383 ; (c) 73.
6. Mean  $\mu = 50.3$ , Standard deviation  $\sigma = 10.33$ .
7. (a) 15.87% ; (b) 2.28% ; 49.74%.
8. (a) 0.5328 ; (b) 0.0062.
9. 471.
10. 75.1.
11. (a) 23 ; (b) 81 ; (c) 235.
12. (a) 22.66% ; (b) 46.5% ; (c) 15.98%.
13. (a) 0.9821 ; (b) 46.5% ; (c) 15.98%.
14. 55.2.
15. 2446 ; 474.
16. 72.025 and 114.
17.  $\mu = 65.42$ ,  $\sigma = 3.29$ .

**MISCELLANEOUS EXERCISE**

1. Suppose that 60 per cent of voting population in a city, about to have referendum on adding fluoride to the drinking water, favour fluoridation.
  - (a) A sample of 10 persons are interviewed. What is the probability that five, six or seven persons favour fluoridation?
  - (b) A sample of 100 persons are interviewed. What is the probability that between 55 and 65 inclusive favour fluoridation?
2. The quoted figure for the 5 year mortality rate for a particular form of leukaemia is 80 per cent. In the hospital when you are a resident interested in malignant neoplasm research, of the last five cases with this form of leukaemia, four are cured and one died. Do you feel you should check to see if some new procedure was used on the patients or that they were special in some other way, or pass off the cures to chance.
3. Under microscopic investigation, on the average five particular microorganisms are found on a  $1 \text{ cm}^2$  untreated specimen. One such specimen was chemically treated. If it is assumed that the treated. If it is assumed that the treatment was in effective and if the Poisson distribution is used, what is the probability of finding:
  - (a) Exactly five?      (b) Fewer than three organisms?
  - (c) Two or three?      (d) More than six?
4. A dentist engaged in research states at a dental convention that in a survey on what makes any article acceptable to a journal, it was concluded that "Two out of every three published articles summarized the major findings in the first lines of the paper". Assuming that journals do not dictate the format of articles, if you were to write an article, would you structure it with the findings in the first lines? Discuss.
5. A student in quantitative analysis has been told that the probability of obtaining a successful endpoint in a particular titration is 0.7. This student carries out five such titrations and obtains only one successful end point. Should he think of a career that does not involve chemistry?
6. Suppose that a particular strain of staphlococcus produces a certain symptom in 2 per cent of persons infected. At a church picnic 200 persons are contaminated food and were infected with the organism. What is the probability of the following numbers having this symptom?
  - (a) None?      (b) 10 or fewer?      (c) Exactly 4?      (d) More than 4?
7. Suppose that weight  $x$  of 6-year-old boys is normally distributed with a mean  $\mu = 481\text{b}$  and a standard deviation  $\sigma = 5\text{ lb}$ .
  - (a) Find  $P(40 < x < 48)$       (b) Find  $P(x > 45)$       (c) Find  $P(x < 42)$ .
8. Suppose that a population of monkeys is given injections of a virus to cause them to develop antibodies preparatory to developing a vaccine for humans. The virus is in two lots, only one of which is effective. Eight per cent of the monkeys receive the potent virus. A lab technician selects 25 monkeys at random. What is the probability that between 15 and 20 inclusive have been effectively injected?

9. Suppose that diastolic blood pressure  $x$  in hypertensive women centers about a mean of 100 mm Hg with a standard deviation of 14 mm Hg and is normally distributed. Find :
- (a)  $P(x < 88)$
  - (b)  $x_0$  such that  $P(x < x_0) = 0.95$
  - (c)  $P(96 < x < 104)$
  - (d)  $P(x > 115)$
  - (e) Two symmetric values  $a$  and  $b$  that  $P(a < x < b) = 0.95$ .
10. Suppose that in a dental study on eighth graders it is assumed that the probability is that a student's teeth will have at least one cavity. What is the probability that:
- (a) Exactly six students will be examined to find the first decayed tooth?
  - (b) Fifteen students will have to be examined to find five students with cavities?
11. The probability an individual with a rare syndrome will be cured is  $p = \frac{1}{100}$ .
- (a) A random sample of 10 persons with this syndrome is selected; find  $p$  (1 person is cured), using the binomial distribution.
  - (b) A random sample of 600 persons with this syndrome is selected; find  $p$  (between 4 and 12 persons, inclusive, are cured) using the normal approximation to the binomial.
  - (c) A random sample of 300 persons with this syndrome is selected; find  $p$  (less than 4 are cured). Note that this problem requires values beyond the tabled values of the binomial distribution to solve such problems you may use the Poisson distribution to approximate the binomial distribution. Use  $\mu = np$ .



# 12

# Hypothesis Testing and Large Samples Tests

## 12.1 INTRODUCTION

The science of statistics can be broadly studied under the following two heading:

- (i) Descriptive statistics
- (ii) Inductive statistics

### Descriptive Statistics

*It consists in describing some characteristics of the numerical data* such as mean, median, mode, variance, coefficient of standard deviation, coefficient of variations, moments, skewness, kurtosis, index numbers, time series analysis etc.

### Inductive Statistics

The inductive statistics is also known as **statistical inference**. It may be termed as the *logic of drawing statistically valid conclusions about the population, in any statistical investigation on the basis of examining a part of the population, formed as sample and which is drawn from the population in a scientific manner.*

The statistical inference is classified into two heads:

- I. Estimation theory
- II. Testing of hypothesis

## 12.2 POPULATION AND SAMPLE

### 12.2.1 Population

Population in statistics means the whole of the information which comes under the purview of statistical investigation. It is the totality of all the observations of a statistical enquiry. In other words, an aggregate of objects animate or inanimate under study is the population. It is also known as the **universe**. For example, the population of the heights of students in a school, the population of the sum of points obtained in throwing three dice.

**Finite or Infinite Population :** A population may be finite or infinite according as the number of individuals in it are finite or infinite. The population of weights of students of class XII in a government school is an example of a finite population. The population of pressures at different points in the atmosphere is an example of an infinite population.

### 12.2.2 Sample

A part of the population selected for study is called a sample. In other words, *the selection of a group of individuals or items from a population in such a way that this group represents the population, is called a sample*. For example, a housewife tests a small quantity of rice to see whether it has been cooked or not. This small quantity of rice is a sample and represents the entire quantity of rice cooked. A sample is a selected portion of the population. A sample drawn from a population provides a valuable information about its parent population. It gives a fairly accurate result and a reliable picture of the total observations under investigations. It is always used to measure and estimate the corresponding characteristics of its parent population. *When the sample drawn is perfectly representative, it is identical with its parent population almost in every respect except that it is smaller than the population.*

**Size of sample.** The number of individuals included in the finite sample is called the size of the sample.

### 12.3 PARAMETER AND STATISTIC

There are various statistical measures in statistics such as mean, median, mode, standard deviation, coefficient of variation etc. These statistical measures can be computed both from population (or universe) data and Sample data.

**I. Parameter :** Any statistical measure computed from population data is known as parameter.

**II. Statistic :** Any statistical measure computed from sample data is known as statistic.

Thus a **parameter** is a statistical measure which relates to the population and is based on **population**, whereas a **statistic** is a statistical measure which relates to the **sample** and is based on sample data. Thus a **population mean, population median, population variance, population coefficient of variation etc., are all parameters**. Statistic computed from a **sample** drawn from the parent population plays an important role in (i) **The theory of estimation and, (ii) Testing of hypothesis.**

The usual notation used for **parameter** (in the case of population data) and **statistic** (in the case of sample data) are given below:

Statistical Measure	Notations	
	Population	Sample
Mean	$\mu$	$\bar{x}$
Standard deviation	$\sigma$	$s$
Population	$P$	$p$
Size	$N$	$n$

## 12.4 SAMPLING

A finite subset of the population, selected from it with the objective of investigating its properties is called as *sample* and the number of units in the sample is known as the *sample size*. *Sampling is a tool which enables us to draw conclusions about the characteristics of the population after studying only those items that are included in the sample.*

*In other words sampling is the procedure or process of selecting a sample from the population.* A sampling can also be defined as the process of drawing a sample from a population and of compiling a suitable **statistic** from such a sample in order to estimate the **parameter** of the parent population and to test the significance of the statistic computed from such sample.

## 12.5 SAMPLING THEORY

*The study of relationship existing between a population and the samples drawn from the population is called sampling.* **Sampling theory is based on sampling.** It deals with statistical inferences drawn from sampling results. Statistical inference made on the basis of sampling results are of the following three types:

- (i) **Statistical Estimation :** *It helps in estimating an unknown population parameter (such as population mean, median, mode, standard deviation, kurtosis etc.) on the basis of suitable statistic (such as sample mean, mode, median, variance, etc.) computed from the sample drawn from such parent population.*
- (ii) **Tests of Significance :** *Sampling theory helps in testing of significance about the population characteristics on the basis of suitable statistic computed from a sample drawn from such parent population.* In other words, it helps in determining whether observed differences between two samples are actually due to chance or whether they are really significant. Such questions help us in deciding whether one production is better than the other. The test of significance plays an important role in the theory of decisions.
- (iii) **Statistical Inference :** *Statistical inference means drawing conclusions about some matters on the basis of certain statistical results.* These statistical results are obtained by drawing samples from the population and then by computing suitable statistic from these samples so as to make statistical inferences. These statistical inferences enable us to draw statistical conclusions about some measures of a population on the basis of such statistic.

## 12.6 SAMPLING AND NON-SAMPLING ERRORS

A sample is a part of the whole population. A sample drawn from the population depends on chance and as such all the characteristics of the population may not be present in the sample drawn from the same population. **Any statistical measures say, mean of the sample, may not be equal to the corresponding statistical measure (mean) of the population from which the sample has been drawn.** Thus there can be discrepancies in the statistical measure of population, i.e., parameters and the statistical measures of sample drawn from the same population, i.e., statistic. These discrepancies are known as errors in sampling. Errors in sampling are of two types.

- (i) Sampling Errors
- (ii) Non-sampling errors or Bias.

## I. Sampling Errors

*Sampling errors is inherent in the method of sampling. Sampling depends on chance and due to the existence of chance in sampling, the sampling errors occur. Errors in sampling arise primarily due to the following reasons:*

1. **Faulty selection of the sample :** This may be due to selection of defective sampling techniques which may introduce the element of bias, e.g., purposive or judgement sampling, in which investigator deliberately selects a non-representative sample.
2. **Substitution :** Sometimes an investigator while collecting the information from a particular sampling unit, included in the random selection substitutes a convenient member of the population and this may lead to some bias as the characteristic possessed by substituted unit may be different from those possessed by the original unit included in sampling.
3. **Faulty demarcation of sampling units.**
4. **Variability of the population :** Sampling error may also depend on the variability or heterogeneity of the population from which the samples are to be drawn.

## II. Non-sampling Errors or Bias

Non-sampling errors or Bias automatically creep in due to human factors which always varies from one investigator to another. Bias may arise in the following different ways.

- (i) due to negligence and carelessness on the part of investigator;
- (ii) due to faulty planning of sampling;
- (iii) due to faulty selection of sample units;
- (iv) due to incomplete investigation and sample survey;
- (v) due to framing of a wrong questionnaires;
- (vi) due to negligence and non-response on the part of the respondents;
- (vii) due to substitution of a selected unit by another;
- (viii) due to error in compilation;
- (ix) due to applying wrong statistical measure.

## 12.7 SAMPLING FLUCTUATIONS

Any statistical measure or a parameter computed from a population is always fixed and constant. *Thus there cannot be any fluctuation in the value of the parameter computed from the population. But this is not true in the case of statistical measure. The statistic computed from different samples of the same size drawn from the same parent population may be different. Any statistical measure, i.e., statistic computed from each of such samples will also vary from sample to sample.*

The variation in the value of the statistic from one sample to another is called sampling fluctuation or sampling variation. Thus a statistic has sampling fluctuations whereas a parameter has no sampling fluctuations.

## 12.8 SAMPLING DISTRIBUTION OF A STATISTIC

Sampling distribution of a statistic is *the frequency distribution which is formed with various values of a statistic computed from different samples of the same size drawn from the same population*. We can draw a large number of samples of same size from a population of fixed size, each sample containing different population members. Any statistic (statistical measure of sample) like mean, median, variance, standard deviation etc., may be computed for each of the samples. As a result a series of various values of that statistic may be obtained. These various values can be arranged into a frequency distribution table, which is known as the sampling distribution of the statistic.

Sampling may be done with replacement or without replacement. Sampling with replacement means that the same unit of the population may be included in each sample more than once. Sampling without replacement means that the same unit of population may not be included in each sample more than once. In the case of sampling with replacement the total number of possible samples each of size  $n$  drawn out of population of size  $N$  is  $N^n$ . But if the sampling is without replacement the total number of possible samples will be  $C(N, n) = m$  (say): For each of these samples a value of statistic, say sample mean  $\bar{x}$ , is computed. As the samples are formed with different sample units so the value of each one of the sample means will be different. *The sample mean can be regarded as a random variable  $\bar{x}$  and each sample mean then constitute as the observed value of this new random variable  $\bar{x}$ . Let these values be  $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_m$ . These mean values  $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_m$  can be used to form a frequency distribution. Then this frequency distribution of the statistic  $\bar{x}$  is known as the Sampling distribution of Sample mean.* Similarly, the sampling distribution of standard deviation or coefficient of variation or variance may be constructed with the various values of standard deviation or coefficient of variation or variance respectively.

If the population size is infinitely large or sampling is done with replacement, then the total number of possible samples of the same size which may be drawn from the population cannot be determined. In such a case, *a large number of repeated random samples from the population of fixed size can be drawn and the values of statistic for these samples may be computed. These values of the statistic can be used to form a frequency distribution. This frequency distribution of the statistic is known as the Sampling Distribution of Statistic.* The main characteristic of the Sampling distribution of a statistic is that it approaches normal distribution even when the population distribution is not normal provided the sample size is sufficiently large (greater than 30). Another important feature of the sampling distribution of statistics is that the mean and the standard deviation of the sampling distribution of sample mean bear a definite relation to the corresponding parameters, i.e., mean and standard deviation of parent population. These characteristics of the sampling distribution help us

- (i) To estimate the unknown population parameter from the known statistic.
- (ii) To set up the confidence limits of the parameter within which the parameter values are expected to lie.
- (iii) To test a hypothesis and to draw a statistical inference from it.

Let  $t_1, t_2, t_3, \dots, t_k$  be the values of a statistic  $t$  (mean or standard deviation of sampling distribution) for  $k$  possible samples. Thus the statistic  $t$  may be regarded as a random variable which can take any one of the value  $t_1, t_2, t_3, \dots, t_k$ . The set of the values of  $t$  constitutes, what is known as the sampling distribution of the statistic  $t$ . We can compute the mean, variance and other statistical constants of the sampling distributions of the statistic  $t$ . For example,

$$\text{Mean} = E(t) = \frac{1}{k} \sum t = \bar{t} \quad (\text{say})$$

$$\text{Variance of } t = \text{Var}(t) = \frac{1}{k} \sum [t - E(t)]^2 = \frac{1}{k} \sum (t - \bar{t})^2.$$

## 12.9 STANDARD ERROR OF A STATISTIC

The statistical measure of **standard deviation** may be computed both from the observations of the **population** and also from the observations of the **Sampling distribution**. Also we know that the standard deviation is a measure of the average amount of the variability of all the observations of variable from their mean. When the average amount of the variability of the observations of a population is computed, it is called the **standard deviation**. But when the average amount of the variability of the observations of a sampling distribution is computed, it is known as **Standard Error**. Thus the standard deviation computed from the observations of a sampling distribution of a statistic is called the **standard error of the statistic**. In other words, the standard deviation used to measure the variability of the values of a statistic from sample to sample is called the **standard error of the statistic**. Thus the standard deviation and standard error have the same meaning and same connection but are used in different contexts and different circumstances. Both of them are used to measure the variability of observations.

**Standard error is used to measure the variability of the values of a statistic computed from the samples of the same size drawn from the population, whereas standard deviation is used to measure the variability of the observations of the population itself.**

Thus the standard deviation of the sampling distribution of a statistic ' $t$ ' is known as standard error of ' $t$ '. It is generally denoted by  $S.E.(t)$ . Thus

$$S.E.(t) = \sqrt{\text{Var}(t)} = \sqrt{\frac{1}{k} \sum (t - \bar{t})^2}$$

The following two standard errors are frequently used in statistics.

**1. Standard Error of Sample Mean :** It is the standard deviation of the sampling distribution of sample means. It is denoted by  $\sigma_{\bar{x}}$  and is given by  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ , where  $\sigma$  is the standard deviation of population and  $n$  is the sample size.

i.e.,  $S.E(\bar{x}) = \frac{\sigma}{\sqrt{n}}$ , where  $\sigma$  is known

or  $S.E(\bar{x}) = \frac{s}{\sqrt{n}}$ , when  $\sigma$  is not known

and  $s$  = standard deviation of sample is given

- 2. Standard error of difference of two sample means  $\bar{x}_1$  and  $\bar{x}_2$  drawn from the same population.**

$$\text{S. E. } (\bar{x}_1 - \bar{x}_2) = \sqrt{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

when  $\sigma$  = population standard deviation is known;

and  $n_1$  and  $n_2$  are the sizes of two samples, which are drawn from the same population.

- 3. Standard Error of difference of two samples means  $\bar{x}_1$  and  $\bar{x}_2$  drawn from two different populations with standard deviations  $\sigma_1$  and  $\sigma_2$  respectively**

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

where  $n_1$  and  $n_2$  are the sizes of the two samples.

## 12.10 UTILITY OF STANDARD ERROR OF A STATISTIC

The standard error of sample mean or the standard error of proportion is used in sampling in order to obtain the following facts:

1. Standard error is used to set up the confidence limits within which the population parameter may lie.
2. Standard error is used to test the hypothesis and to draw a statistical conclusion from it.
3. Standard error is used to measure the variability of the values of a statistic from its mean.
4. It is, generally, used for large samples and gives us the idea about the average amount of error which actually occurs in estimating the values of a parameter on the basis of a statistic.

## 12.11 ESTIMATION THEORY

Estimation of population parameters like mean, variance, proportion, correlation coefficient, etc., from the corresponding sample statistic is one of the very important problems of statistical inference. The theory of estimation was founded by Prof. R.A. Fisher round about 1930 and is divided into two groups.

- (i) Point Estimation and (ii) Interval estimation.

## 12.12 POINT ESTIMATION

In point estimation a single statistic (numerical value) is used to provide an estimate of the population parameter. A particular value of a statistic which is used to estimate a given parameter is known as a *point estimate* or *estimator* of the parameter.

A good estimator is one which is as close to the true value of the population parameter as possible. The following are some of the criteria which should be satisfied by a good estimator.

- (i) Unbiasedness; (ii) Consistency ; (iii) Efficiency ; (iv) Sufficiency

- 1. Unbiasedness :** A statistic  $t = t(x_1, x_2, \dots, x_n)$ , a function of the sample observations  $x_1, x_2, \dots, x_n$ , is said to be an unbiased estimate of the corresponding population parameter  $\theta$ , if

$$E(t) = \theta$$

i.e., if the mean value of the sampling distribution of the statistic is equal to parameter.

- 2. Efficiency :** If  $t_1$  and  $t_2$  are estimators of a parameter  $\theta$  such that

$$\text{Var}(t_1) < \text{Var}(t_2) \text{ for all } n,$$

then  $t_1$  is said to be more efficient than  $t_2$ . In other words, an estimator with lesser variability is said to be more efficient and consequently more reliable than the other.

Efficiency ( $E$ ) of  $t_1$  relative to  $t_2$  is defined as:

$$E = \frac{\text{Var}(t_1)}{\text{Var}(t_2)}$$

Obviously,  $E \leq 1$ .

- 3. Efficiency and sufficiency :** The concepts of *consistency* and *sufficiency* are beyond the scope of this book.

## 12.13 INTERVAL ESTIMATION

There are situations where the point estimation is not desirable and we are interested in finding such limits within which with a known probability or to a known degree of reliance, the value of the population parameter is expected to lie. Such a process of estimation is called the **interval estimation**. In other words, an interval estimation is the range of values used in making estimation of a population parameter. Thus the interval estimation of the population parameter is the estimation of the population parameter by an interval around the point. The interval estimation of a population parameter  $\theta$  is the estimation of the parameter  $\theta$  with the help of the interval  $[t - s, t + s]$ , where  $t$  is the sample statistic, i.e.,  $t - s \leq \theta \leq t + s$ .

Both point and interval estimation are ways of drawing inductive inferences and therefore involve an element of uncertainty. But the interval estimation has advantage over the point estimation as it provides a measure of degree of uncertainty in terms of probability attached to the interval. In the interval estimation, the estimate for the parameter lies between two limits. The two limits within which the estimate for the parameter lies are known as **Confidence Limits** or **Fiducial Limits** and the interval bounded by these two limits is called **Confidence Interval** or **Fiducial Interval**. The confidence interval depends upon the confidence that is required to set up. The probability that we associate with an interval is called the **confidence level**. It shows how confidently we can say that the interval estimate will include the population parameter. The higher the probability, the more is the confidence. Although any confidence level can be considered, but the most commonly used confidence levels are 90%, 95%, 98%, 98% and 99%. In the light of above discussion the interval estimation is defined as:

$P[C_1 < \theta < C_2, \text{ for given value of statistic } t] = 1 - \alpha$ , where  $\alpha$  is the level of significance, the interval is within which the unknown parameter  $\theta$  is expected to lie. It is known as

**Confidence Interval or Fiducial Interval** and  $1 - \alpha$  is called the confidence coefficient, depending upon the desired precision of the estimate [For example  $\alpha = 0.01$  gives 99% confidence limits],  $C_1, C_2$  are respectively known as lower limit and upper limit of the Confidence Interval.

If 't' is the sample statistic used to estimate the corresponding population parameter  $\theta$ , then  $(1 - \alpha) \%$  confidence limits for  $\theta = t \pm S.E. (t) \times t_\alpha$ , where S.E. (t) is the standard error of t, and  $t_\alpha$  is the significant or critical value of t at significance level  $\alpha$ . The quantity  $t + S.E. (t) \times t_\alpha$  is the upper limit and  $t - S.E. (t) \times t_\alpha$  is the lower limit of confidence interval

$$[t - S.E. (t) \times t_\alpha, t + S.E. (t) \times t_\alpha].$$

## 12.14 METHOD TO COMPUTE CONFIDENCE LIMITS FOR A POPULATION PARAMETER $\theta$

The following steps enable us to compute the confidence limits for the population parameter  $\theta$  in terms of the sample statistic t.

1. Compute or select the appropriate sample statistic t.
2. Obtain S.E. (t), the standard error of the sampling distribution for the sample statistic t.
3. Select the confidence level  $\alpha$  and corresponding to that "specific level of confidence" and note that critical value  $t_\alpha$  from the tables.
4. If 't' is the statistic used to estimate parameter  $\theta$ , then

$$(1 - \alpha) \% \text{ confidence limits for } \theta = t \pm S.E. (t) \times t_\alpha,$$

where  $t_\alpha$  is the significant or critical value of t at level of significance  $\alpha$ .

## 12.15 INTERVAL ESTIMATION FOR LARGE SAMPLES

As per central limit theorem, "For large samples, the underlying distribution of the standardized variate corresponding to the sampling distribution of the sample statistic t will be asymptotically normal, i.e.,

$$Z = \frac{t - E(t)}{S.E. (t)} \sim N(0, 1)$$

is a standardized normal variate with mean zero and variance one as  $n \rightarrow \infty$ .

## 12.16 CONFIDENCE INTERVAL OF THE MEAN

Let  $\mu$  be the population mean and  $\bar{x}$  be the sample mean of the sampling distributions of means. It is also assumed that the sample mean is normal or the sample is large. Then the interval estimate of population mean  $\mu$  by the sample mean  $\bar{x}$  of the sampling distribution of means is given by the following rule:

### Working Rule

**Step I.** Computer  $\bar{x}$  or take  $\bar{x}$ .

**Step II.** Select the confidence level and the corresponding to that specific level of confidence, note that the value of confidence Z or  $t_\alpha$  from the tables.

**Step III.** Compute S.E. ( $\bar{x}$ ) with the help of following results.

**Case I.** When  $\sigma$ , the standard deviation of the normal population  $\sigma$  is known

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

where  $n$  is the sample size.

**Case II.** When  $\sigma$ , the standard deviation of population is not known.

$$\text{S.E.} (\bar{x}) = \frac{s}{\sqrt{n-1}},$$

where  $n$  = sample size and  $n$  is large

$s$  = standard deviation of the sampling distribution of the sample.

**Step IV.** Construct a confidence level as follows:

**Case I.** When  $\alpha$  is known and population is normal or any population with large  $n$ .

$$\text{Confidence interval} = [\bar{x} - \text{S.E.} (\bar{x}) \times Z_\alpha, \quad \bar{x} + \text{S.E.} (\bar{x}) \times Z_\alpha]$$

$$\text{or } \left[ (\bar{x}) - \frac{\sigma}{\sqrt{n}} \times Z_\alpha \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} \times Z_\alpha \right] \quad \left( \because \text{S.E.} (\bar{x}) = \frac{\sigma}{\sqrt{n}} \right)$$

**Case II.** When  $\sigma$  is unknown and  $n$  is large.

In this case,

$$\text{Confidence interval} = [\bar{x} - \text{S.E.} (\bar{x}) \times Z_\alpha, \quad \bar{x} + \text{S.E.} (\bar{x}) \times Z_\alpha]$$

$$= \left[ \bar{x} - \frac{s}{\sqrt{n-1}} \times Z_\alpha, \bar{x} + \frac{s}{\sqrt{n-1}} \times Z_\alpha \right]$$

$$\text{or } \bar{x} - \frac{s}{\sqrt{n-1}} \times Z_\alpha \leq \mu \leq \bar{x} + \frac{s}{\sqrt{n-1}} \times Z_\alpha \quad \left( \because \text{S.E.} (\bar{x}) = \frac{s}{\sqrt{n-1}} \right)$$

**Step V.** Select the value  $Z_\alpha$  from the following table.

Table Value of  $Z_\alpha$

Confidence level ( $1 - \alpha$ )	Value of confidence coefficient $Z_\alpha$ or $t_\alpha$
90% or $\alpha = 0.10$	1.64
95% or $\alpha = 0.05$	1.96
98% or $\alpha = 0.02$	2.33
99% or $\alpha = 0.01$	2.58
Without any reference to the confidence level $\alpha$	3.00

Note : Where no reference to the confidence level is given, then always take  $t_\alpha$  or  $Z_\alpha = 3$ .

**Example 1 :** The quality control manager of a tyre company has sample of 100 tyres and has found the life time to be 30.214 km. The population s.d. is 860. Construct a 95% confidence interval for the mean life time for this particular brand of tyres.

**Solution :** Here  $n = 100$ ,  $\bar{x} = 30214$ ,  $\sigma = 860$ .

Also  $(1 - \alpha)\% = 95\%$  so that  $Z_{\alpha} = 1.96$ .

$$\therefore \text{S.E. } (\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{860}{\sqrt{100}} = \frac{860}{10} = 86. \quad [\text{From table}]$$

$$\begin{aligned} \therefore \text{Confidence interval} &= \left[ \bar{x} - \frac{\sigma}{\sqrt{n}} \times Z_{\alpha}, \bar{x} + \frac{\sigma}{\sqrt{n}} \times Z_{\alpha} \right] \\ &= [30214 - 86 \times 1.96, 30214 + 86 \times 1.96] \\ &= [30214 - 168.56, 30214 + 168.56] = [30045.44, 30382.56]. \end{aligned}$$

**Example 2 :** In a random selection of 64 of 600 road crossing in a town, the mean number of automobile accidents per year was found to 4.2 and the sample s.d. was 0.8. Construct a 95% confidence interval for the mean number of automobile accidents per crossing per year.

**Solution :** Here  $n = 64$ , s.d. of sample  $s = 0.8$ ,  $\bar{x} = 4.2$ .

$(1 - \alpha)\% = 95\%$ , so the value of  $Z_{\alpha} = 1.96$ .

$$\therefore \text{Confidential interval} = [\bar{x} - \text{S.E. } (\bar{x}) \times Z_{\alpha}, \bar{x} + \text{S.E. } (\bar{x}) \times Z_{\alpha}]$$

Since we have a finite population of size 600 and  $\frac{n}{N} = \frac{64}{600} = 0.107$ , which is more than 0.05, so we use the following result to calculate the standard error of  $\bar{x}$ .

$$\begin{aligned} \text{S.E. } (\bar{x}) &= \frac{s}{\sqrt{n-1}} \times \sqrt{\frac{N-n}{N-1}} = \frac{0.8}{\sqrt{63}} \times \sqrt{\frac{600-64}{600-1}} \\ &= \frac{0.8}{\sqrt{7.93}} \times \sqrt{\frac{536}{599}} = \frac{0.8}{7.93} \times 0.894 = 0.091. \end{aligned}$$

$$\begin{aligned} \text{Confidence interval} &= [4.2 - 0.091 \times 1.96, 4.2 + 0.091 \times 1.96] \\ &= [4.2 - 0.177, 4.2 + 0.177] = [4.023, 4.377]. \end{aligned}$$

**Example 3 :** A random sample of size 100 has mean 15, the population variance being 25. Find the interval estimate of the population mean with confidence level of (i) 99% and (ii) 95%.

**Solution :** Here  $n = 100$ ,  $\sigma^2 = 25$ ,  $\bar{x} = 15$ .

$$\text{S.E. } (\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{100}} = \frac{5}{10} = 0.5.$$

(i) 99% Confidence level  $\Rightarrow Z_{\alpha} = 2.58$

$$\begin{aligned} \therefore \text{Confidence interval} &= \bar{x} \pm \text{S.E. } (\bar{x}) \times Z_{\alpha} \\ &= 15 \pm 0.5 \times 2.58 = 15 \pm 1.29, \text{ i.e., } 13.71 \text{ to } 16.29. \end{aligned}$$

(ii) 95% Confidence level  $\Rightarrow Z_{\alpha} = 1.96$

$$\begin{aligned}\text{Confidence interval} &= \bar{x} \pm \text{S.E.} (\bar{x}) \times Z_{\alpha} = 15 \pm (0.5) \times 1.96 \\ &= 15 \pm 0.98, \text{ i.e., } 14.02 \text{ to } 15.98.\end{aligned}$$

## 12.17 TESTING OF HYPOTHESIS

Sampling theory deals with two types of problems, viz., estimation and testing of hypothesis. In this section we shall deal with the problem of Testing of Hypothesis. Modern theory of probability plays an important role in decision-making and the branch of statistics which helps us in arriving at the criterion for such decisions is known as **testing of hypothesis**, which was initiated by J. Neyman and E.S. Pearson. It employs statistical techniques to arrive at decisions in certain situation where there is an element of uncertainty on the basis of sample whose size is fixed in advance.

**Hypothesis :** A hypothesis is a statement about the population parameter. In other words, a hypothesis is a conclusion which is tentatively drawn on logical basis.

**Statistical Hypothesis :** It is tentative conclusion that specifies the properties of a distribution of a random variable. These properties generally refer to parameters of the population and the hypothetical values with which the values of statistic derived from a sample are compared in order to find the difference between statistic and corresponding parameter. In other words, **statistical hypothesis is some assumption or statement, which may or may not be true, about a population or about the probability distribution characterising the given population, which we want to test on the basis of the evidence from a random sample.**

**Test of Hypothesis :** Hypothesis testing can be regarded as an example of a decision process, in which data are assembled in a particular way to produce a quantity that leads to a choice between two decisions. Each decision then leads to an action. Because data arise from sampling process, there is some risk that an incorrect decision will be made with some loss attached to the resulting incorrect action. In this context, hypothesis testing is an example of a more general study known as decision theory. Many situations in life require a choice between two decisions, whether or not the choice is actually made on the basis of data from a sample. The decisions to purchase or not to purchase a particular piece of laboratory equipment or, to market or not to market a particular medicine. **The testing of hypothesis is a procedure that helps us to ascertain the likelihood of hypothesised population parameter being correct by making use of the sample statistic.** In other words, it is a process of testing of significance which concerns with the testing of some hypothesis regarding a parameter of the population on the basis of statistic from the sample. **In testing of hypothesis a statistic is computed from a sample drawn from the parent population and on the basis of this statistic it is observed whether the sample so drawn has come from the population with certain specified characteristic.** The value of sample statistic may differ from the corresponding population parameter due to sampling fluctuations. The test of hypothesis discloses the fact whether the difference between sample statistic and the corresponding hypothetical population parameters significant or not significant. Thus the test of hypothesis is also known as the **test of significance**.

## 12.18 PROCEDURE OF TESTING A HYPOTHESIS

The following are the steps involved in hypothesis testing problems.

**1. Setting up of Hypothesis :** There are two types of hypothesis:

- 1. Null hypothesis.
- 2. Alternative hypothesis

**Null Hypothesis :** The statistical hypothesis that is set up for testing a hypothesis is known as **null hypothesis**. The null hypothesis is set up for testing a statistical hypothesis only to decide whether to accept or reject the null hypothesis. It asserts that **there is no difference between the sample statistic and population parameter** and whatever difference is there, is attributable to sampling errors. **Null hypothesis is usually denoted by  $H_0$** .

### *Setting up null hypothesis*

The following steps must be taken into consideration while setting up a **null hypothesis**.

- (i) In order to test the significance of the difference between a sample statistic and the population parameter or between the two different sample statistic, we set up the null hypothesis  $H_0$  that the **difference is not significant**. There may be some difference but that is solely due to sampling fluctuations.
- (ii) To test any statement about the population, we hypothesise that it is true. For example, if we want to find the population mean has a specified value  $\mu_0$ , then the hypothesis  $H_0$  is set as follows.

$$H_0 : \mu = \mu_0.$$

Prof. R.A Fisher remarked, “**Null hypothesis is the hypothesis which is to be tested for possible rejection under the assumption it is true**”.

**Alternative Hypothesis :** The negation of null hypothesis is called the **Alternative hypothesis**. In other words, any hypothesis **which is not a null hypothesis is called an alternative hypothesis**. It is always denoted by  $H_1$  or  $H_\alpha$ . It is set in such a way that the rejection of null hypothesis implies the acceptance of alternative hypothesis. For example, if we want to find the null hypothesis that the average height of the student of a college is 165 cm, i.e.,  $\mu_0 = 165$  cms (say). Then the **Null hypothesis** is

$$H_0 = \mu = 165 (= \mu_0)$$

and the **Alternative hypothesis** could be

- |                            |   |                   |
|----------------------------|---|-------------------|
| (i) $H_1 : \mu \neq \mu_0$ | (i.e., $\mu > \mu_0$ or $\mu < \mu_0$ ) | [Two Tailed Test] |
| (ii) $H_1 : \mu > \mu_0$   |   | [One Tailed Test] |
| (iii) $H_1 : \mu < \mu_0$  |   | [One Tailed Test] |

**2. Computation of Test Statistic :** After setting up the null hypothesis and alternative hypothesis, we compute the **test statistic**. The **test statistic is a statistic based on appropriate probability distribution**. It is used to test whether the null hypothesis set up should be accepted or rejected. It is the main yard stick which help us to decide whether to accept or reject the null hypothesis. Different probability distribution values are used in appropriate cases while testing the null hypothesis. The following are the most commonly used test statistic.

**Z-Distribution :** We use Z-distribution under the normal curve for large sample (i.e., if the sample size  $n > 30$ ). The Z-statistic is defined as

$$Z = \frac{t - E(t)}{\text{S.E.}(t)} \sim N(0, 1) \text{ as } n \rightarrow \infty$$

is a standard normal variable with mean zero and variance one.

Also, S.E. ( $t$ ) = Standard error of the statistic  $t$

**t-test :** t-test is used for small sample, i.e., if the sample size  $n \leq 30$ . The student's t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{\text{S.E.}(\bar{x})} = \frac{\text{Difference of sample and population means}}{\text{Standard error of mean}}$$

with  $n - 1$  degrees of freedom, where  $n$  is the size of sample.

The following table shows the conditions for using Z-test and t-test in testing null hypothesis about the means.

TABLE

Size of sample	Population s.d. $\sigma$ is known	Population s.d. $\sigma$ is not known
$n > 30$	Z-test	Z-test
$n \leq 30$	t-test	t-test

**3. Types of Errors in Hypothesis Testing :** There is every chance that a decision regarding a null hypothesis may be correct or may not be correct. There are two types of errors.

- (a) **Type I Error :** *It is the error of rejecting null hypothesis  $H_0$  when it is true.* When a null hypothesis is true, but the difference (of mean) is significant and the hypothesis is rejected, then a **Type I Error** is made. **The probability of making a type I error is denoted by  $\alpha$ , the level of significance.** In order to control the type I error, the probability of type I error is fixed at a certain level of significance  $\alpha$ . The probability of making a correct decision is then  $(1 - \alpha)$ .
- (b) **Type II Error :** *It is the error of accepting the null hypothesis  $H_0$  when it is false.* In other words when a null hypothesis is false, but the difference of means is insignificant and the hypothesis is accepted, a type II error is made. **The probability of making a type II error is denoted by  $\beta$ .**

The following summary table in which  $H$  denotes the tested hypothesis may help fix the concepts of the two kinds of error:

	Truth	
	$H_0$ is True	$H_0$ is False
Decision Based on Data	Accept $H_0$	Correct Decision
	Reject $H_0$	Type I error ( $\alpha$ )
		Correct decision

*It may be helpful to think of hypothesis testing an analogous to a jury trial of H: the defendant is innocent. The type I error corresponds to convicting an innocent defendant, while the type II error corresponds to acquitting a guilty defendant.*

**Power :** It is probability of rejecting the tested hypothesis when it is false, i.e., when an alternative hypothesis is true; denoted by  $1 - \beta$ , where  $\beta$  is the probability of a type II error.

**4. Level of Significance :** The next step is the fixation of level of significance. The level of significance is the maximum probability of making a type I error and it is denoted by  $\alpha$ , i.e., **P (Rejecting  $H_0$  when  $H_0$  is true) =  $\alpha$ .** The probability of making a correct decision is then  $(1 - \alpha)$ . The best value for fixing the level of significance depends on the seriousness of the results of the types of error. The commonly used level of significance in practice are 5% ( $\alpha = 0.05$ ) and 1% ( $\alpha = 0.01$ ). If we use 5% level of significance ( $\alpha = 0.05$ ) we shall mean that the probability of making type I error is 0.05 or 5%, i.e.,

$$P [\text{Rejecting } H_0 \text{ when } H_0 \text{ is true}] = 0.05$$

This mean that there is a probability of making 5 out of 100 (or making 1 out of 20) type I error. This also mean that we are 95% confident that a correct decision has been made. Similarly 1% level of significance ( $\alpha = 0.01$ ) that there is a probability of making 1 error out of 100. It is important to note that if no level of significance is given, then we always take  $\alpha = 0.05$ .

**5. Critical Region or Rejection Region :** The rejection region or critical region is the region of the standard normal curve corresponding to predetermined level of significance  $\alpha$  (which is fixed for knowing the probability of making a type I error of rejecting the null hypothesis  $H_0$  when it is true). The region under the normal curve which is not covered by the rejection region is known as **Acceptance Region**. Thus the statistic which leads to rejection of null hypothesis  $H_0$  gives us a region known as **Rejection Region or Critical Region**. While those which lead to the acceptance of  $H_0$  give us a region called a **Acceptance region**. The value of the test statistic computed to test the null hypothesis  $H_0$  is known as the critical value. The critical value separates the **rejection region** from the **acceptance region**.

**6. Two Tailed Test and One Tailed Test :** The probability curve of the sampling distribution of the test statistic is a normal curve. In any test, the critical region is represented by a portion of the area under this normal curve. This curve has two sides (or ends) known as **two tails**. The rejection region may be represented by a portion of area on each of the two sides or by only one side of the normal curve and correspondingly the test is known as **Two Tailed Test (or Two sided Test) or one Tailed Test (or one sided test)**.

**Two Tailed Test or Two sided Test :** When the test of hypothesis is made on the basis of rejection region represented by both sides of the standard normal curve, it is called a *two tailed test or two sided test* (see Fig. 12.1). In other words, a test of statistical hypothesis where the alternative hypothesis  $H_1$  is **two sided or two tailed** such has

**Null Hypothesis :**  $H_0 : \mu = \mu_0$

**Alternative Hypothesis :**  $H_1 : \mu \neq \mu_0 \quad (\mu > \mu_0 \text{ or } \mu < \mu_0)$

It is called as **two tailed test or two sided test**.

**One Tailed or One sided Test :** A test of statistical hypothesis where either alternative hypothesis is one sided is called as one tailed test or one sided test. There are two types of one tailed or one sided tests.

1. **Right Tailed Test :** In the right tailed test the rejected region or critical region lies entirely on the right tail of the normal curve. (see Fig. 12.2).
2. **Left Tailed Test :** In the left tailed test the critical region or rejection region lies entirely on the left tail of the normal curve. (see Fig. 12.3).

For example, in the test for testing the mean ( $\mu$ ) of the population.

Null hypothesis	$H_0 : \mu = \mu_0$
Alternative hypothesis	$H_1 : \mu > \mu_0$ (Right Tailed Test)
and	$H_1 : \mu < \mu_0$ (Left Tailed Test)

Now the question arises which test is to be applied, i.e., when a two tailed test will be applied and where to apply one tailed test?

**Application of Two Tailed Test :** A two tailed test is applied in such cases when difference between the sample mean and population mean is tending to reject the null hypothesis  $H_0$ , the difference may be positive or negative.

**Application of One Tailed Test :** On the other hand, a one tailed test applied in such cases when it is considered to see whether the population mean is **at least as large** as some specified value of the mean or **atleast as small as some** specified value of the mean. In the former case the **right tailed tests** and in the latter case a **left tailed tests** are applied.

**7. Critical Value :** The value of the sample statistic that defines the regions of acceptance and rejection, is called the critical value. It is important to note that “the critical value of  $Z$  for a single tailed test (left or right) at a level of significant  $\alpha$  is the same as the critical value of  $Z$  for a two tailed test at a level of significance  $2\alpha$ ”. Further for Right Tailed Test :  $P(Z > Z_\alpha) = \alpha$ .

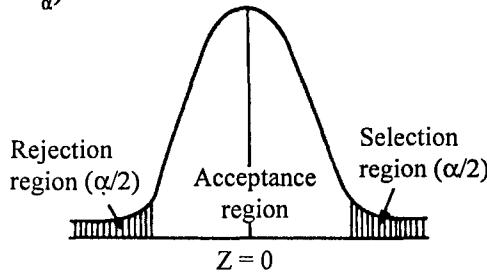


Fig. 12.1

$$\text{For Left Tailed Test} \quad P(Z < -Z_\alpha) = \alpha$$

$$\text{Also for a two tailed test: } P(|Z| > Z_\alpha) = \alpha$$

$$\Rightarrow P(Z > Z_\alpha) + P(Z < -Z_\alpha) = \alpha.$$

$$\Leftrightarrow P(Z > Z_\alpha) + P(Z < Z_\alpha) = \alpha. \quad (\text{By symmetry})$$

$$\Rightarrow P(Z > Z_\alpha) = \frac{\alpha}{2} \Rightarrow \text{the area of each tail is } \frac{\alpha}{2}.$$

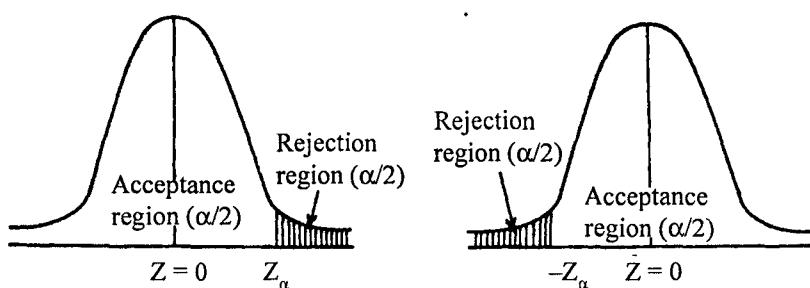


Fig. 12.2 : Right Tailed Test at significance level

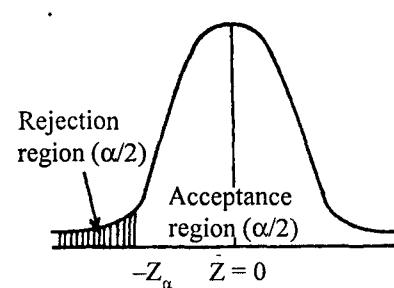


Fig 12.3 : Left Tailed Test at significance level

The following table will give the critical value of  $Z$  viz.,  $Z$  at 1%, 5%, 10% level of significance.

Table for Critical Values of  $Z_\alpha$  of  $Z$ 

Critical value ( $Z_\alpha$ )	Level of Significance $\alpha$		
	1%	5%	10%
Two tailed Test	$ Z_\alpha  = 2.58$	$ Z_\alpha  = 1.96$	$ Z_\alpha  = 1.645$
Right Tailed Test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left Tailed Test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = 1.28$

**8. Decision :** The last step is the decision about the null hypothesis, i.e., whether to accept it or to reject it. In this regard we compare the computed value of  $Z$  (obtained in step 2) with the critical value or significant value or tabled value  $Z_\alpha$  (given by table for critical value in step 7) at a given level of significance  $\alpha$  and decide as under:

- (i) If  $|Z| < |Z_\alpha|$ , then we accept the null hypothesis  $H_0 \Rightarrow$  If the calculated value of  $Z$  is less than the tabled  $Z_\alpha$  of  $Z$  at a level of significance  $\alpha$ , then the difference between  $[t - E(t)]$  is not significant (and this difference may be due to fluctuation of sampling), so we accept the null hypothesis. In this case the test statistic falls in the region of acceptance.
- (ii) If  $|Z| > |Z_\alpha|$ , then we reject the null hypothesis  $H_0$  and accept the alternative hypothesis  $H_1$ . In this case the computed value of  $Z$  is numerically greater than the critical value  $Z_\alpha$  at a level of significance  $\alpha$ , and therefore, the computed value of test statistic falls in the rejection region. So we reject the null hypothesis and accept the alternative hypothesis at a level of significance or confidence level  $(1 - \alpha)$ .

## 12.19 THE RELATIONSHIP BETWEEN HYPOTHESIS TESTING AND CONFIDENCE INTERVAL ESTIMATION

There is ordinarily a close relationship between a test of hypothesis concerning a parameter or parameters and the corresponding confidence interval. To illustrate this relationship we will examine the  $(1 - \alpha)$  level confidence interval for  $\mu$ , the mean of a normal distribution with known variance  $\sigma^2$  and sample size  $n$ .

The interval limits are:  $\bar{x} - z_{1-(\alpha/2)} (\sigma/\sqrt{n})$  and  $\bar{x} + z_{1-(\alpha/2)} (\sigma/\sqrt{n})$ .

Now consider a particular numerical value, say  $\mu_0$ , between the limit that is

$$\bar{x} - z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{x} + z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}}$$

By the opposite sequence algebraic steps to derive the interval, that is, reversing the order of terms in this inequality and using  $>$  signs instead of  $<$  signs, subtracting  $\bar{x}$ , multiplying by  $-1$ , and dividing by  $\sigma/\sqrt{n}$ , we have equivalent chain inequality

$$-z_{1-(\alpha/2)} < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{1-(\alpha/2)}$$

or recalling that  $Z_{\alpha/2} = -Z_{1-(\alpha/2)}$

$$Z_{\alpha/2} < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{1-(\alpha/2)}$$

This is the form of the acceptance region for testing  $H : \mu = \mu_0$  in a two sided test at  $\alpha$  level of significance, for known  $\sigma^2$  and sample size  $n$ . This says that a value  $\mu_0$  lying within the  $(1 - \alpha)$  confidence interval corresponds to a value of the test statistic  $(\bar{x} - \mu_0)/(\sigma/\sqrt{n})$  that would lead to acceptance of the hypothesis. If we start with a value of  $\mu_0$  lying outside the confidence interval, we find that the corresponding value of the test statistic will fall into the rejection region.

This is an important finding. We have shown that a  $(1 - \alpha)$  level confidence interval for  $\mu$  contains between its limits all values  $\mu_0$  of  $\mu$  for which  $H_0 : \mu = \mu_0$  would have been accepted in a two sided  $\alpha$  level test. Furthermore, all values  $\mu_0$  lying outside the interval correspond to hypothesis  $\mu = \mu_0$  which would have been rejected. Thus, the confidence interval simultaneously provides a decision for all *hypothesis* of the form  $\mu = \mu_0$ , that is, for any value of  $\mu_0$ . Similar results hold for other tests and intervals and in particular for all the hypothesis to be discussed in this chapter.

## LARGE SAMPLES TESTS

### 12.20 TEST OF SIGNIFICANCE OF MEAN — LARGE SAMPLE

#### Working Rule

We have the following steps:

**Step 1. Setting up of a Null hypothesis :** The null hypothesis is set up in the following form: "There is no significant difference between the sample mean and population mean" or the sample has been drawn from the parent population. Mathematically,

<b>Null hypothesis</b>	$H_0 : \bar{x} = \mu$	
<b>Alternative hypothesis</b>	$H_1 : \bar{x} \neq \mu$	[Two Tailed Test]
or	$H_1 : \bar{x} > \mu$	[One Tailed Test]
or	$H_1 : \bar{x} < \mu$	[One Tailed Test]

**Step 2. Computation of test statistic  $Z$  :** There are two ways of computing a test statistic  $Z$ .

(a) **When the standard deviation  $\sigma$  of population is known:**

In this case Standard Error of Mean = S.E. ( $\bar{x}$ ) =  $\frac{\sigma}{\sqrt{n}}$ ,

where  $n$  = sample size,  $\sigma$  = s.d. of population

$$\text{Test Statistic : } Z = \frac{\bar{x} - \mu}{S.E. (\bar{x})} = \frac{(\bar{x} - \mu) \sigma}{\sqrt{n}}.$$

where  $\mu$  is the sample mean.

(b) **Where the standard deviation  $\sigma$  of population is not known :** In this case we take  $s$ , the standard deviation of sample to calculate the standard error of mean

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

$$\text{Test Statistic } Z = \frac{\bar{x} - \mu}{\text{S.E.} (\bar{x})} = \frac{(\bar{x} - \mu) \sqrt{n}}{s}.$$

**Step 3. Level of significance.** Set the level of significance  $\alpha$ .

**Step 4. Critical value :** Find the critical value  $Z_\alpha$  of  $Z$  at the level of significance  $\alpha$  (of step 3), from the table "Areas under the normal curve:  $Z_\alpha$  – values". Suppose that we fix the level of significance  $\alpha$ . Then the critical region should consist of all those values of the test statistic  $Z$  so that they would only occur with probability  $\alpha$  when the hypothesis is true. Since  $Z$  is a standard normal variable when  $H_0$  is true, we can choose as critical region value of  $Z$  above  $Z_{1-(\alpha/2)}$  or below  $Z_{\alpha/2}$ ; these quantities  $Z_{1-(\alpha/2)}$  and  $Z_{\alpha/2}$  chop off area  $\alpha/2$  from the upper and lower tails of the distribution. Because the values  $Z_{1-(\alpha/2)}$  and  $Z_{\alpha/2}$  separate the critical region from the acceptance region, they are sometimes called *critical values*. Figure 12.4 shows the situation graphically. Notice in the figure that the critical region here is two-sided and consists of values of  $Z$  in both the upper and lower tails of the distribution. From the Table, we find that if  $\alpha = 0.05$ , the critical values are  $\pm 1.96$ , and if  $\alpha = 0.01$ , they are  $\pm 2.58$ .

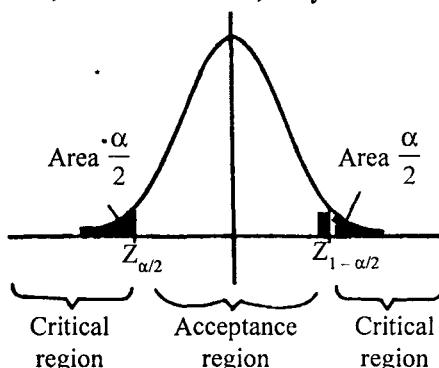


Fig. 12.4 : Critical and acceptance regions for testing  $m = m_0$  in a normal distribution when  $s^2$  is now.

- Step 5. Decision :** (a) If the computed value of  $|Z| < \text{critical value } |Z_\alpha|$  at a level of significance  $\alpha$ , then **accept the null hypothesis**.  
 (b) If computed value of  $|Z| > \text{critical value } |Z_\alpha|$  at a level of significance  $\alpha$ , then **we reject the null hypothesis  $H_0$  and accept the alternative hypothesis  $H_1$** .

**Example 4 :** A sample of size 400 was drawn and the sample mean was found to be 99. Test whether this sample could have come from a normal population with mean 100 and standard deviation 8 at 5% level of significance.

**Solution :** Here  $\bar{x} = 99$ ,  $n = 400$ ,  $\mu = 100$ ,  $\sigma = 8$ .

- 1. Null hypothesis :** The sample have come from a normal population with mean 100 and s.d. 8, i.e.,  $H_0 : \mu = 100$ .  
 ∴ Alternative Hypothesis.  $H_1 : \mu \neq 100 \Rightarrow$  It is a case of two tailed test.  
**2. Test Statistic.**

Here,

$$\text{S.E. } (\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{400}} = \frac{8}{20} = \frac{2}{5}$$

$$\therefore \text{Test Statistic } Z = \frac{\bar{x} - \mu}{\text{S.E. } (\bar{x})} = \frac{99 - 100}{(2/5)} = -\frac{5}{2} = -2.5$$

$$\therefore |Z| = 2.5$$

**3. Level of Significance :**  $\alpha = 0.05$ .

**4. Critical Value :** The value of  $Z_\alpha$  at 5% level of significance = 1.96. (from the table).

**5. Decision :** Since the calculated value of  $Z$  is greater than the critical value as  $2.5 > 1.96 \Rightarrow |Z| > |Z_\alpha| \Rightarrow$  Null hypothesis  $H_0$  is rejected  $\Rightarrow$  the sample has not been drawn from a normal population with mean 100 and s.d. 8.

**Example 5 :** A stenographer claims that she can type at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trials in which she demonstrates a mean of 116 words with a standard deviation of 15 words. Use 5% level of significance.

**Solution :** **1. Null Hypothesis :**  $H_0$  : Stenographer's claim is true,

$$\text{i.e., } H_0 : \mu = 120$$

Alternative hypothesis.  $H_1 : \mu \neq 120 \Rightarrow$  It is a case of two tailed test.

**2. Calculation of Test Statistic :** Here  $n = 100$ ,  $\bar{x} = 116$ ,  $\mu = 120$ ,  $s = 15$ .

$$\text{Standard Error of Mean} = \text{S.E. } (\bar{x}) = \frac{s}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5.$$

$$\therefore \text{Test Statistic} = Z = \frac{\bar{x} - \mu}{(s/\sqrt{n})} = \frac{116 - 120}{1/5} = -2.67.$$

**3. Level of Significance :**  $\alpha = 0.05$ .

**4. Critical value :**  $Z_\alpha = 1.96$  from the Normal table.

**5. Decision :** Since  $|Z| = 2.67$  is greater than  $|Z_\alpha| = 1.96$  at 5% level of significance, so the null hypothesis  $H_0$  is rejected. Thus the stenographer's claim of typing at the rate of 120 words per minute is not true.

**Example 6 :** Suppose that chest circumference of presumably normal newborn baby girls is normally distributed with  $\mu = 13.0$  in. (33 cm) and  $\sigma = 0.7$  in. (1.8 cm). A group of 49 newborn baby girls from a population group living in a remote region and thought perhaps to constitute a genetic isolate are studied and found to have an average chest circumference of 12.6 in. In this evidence that the group of 49 come from a population with parameter values different from the values  $\mu = 13.0$  and  $\sigma = 0.7$ ?

**Solution :** Recasting the problem in our terminology, we shall assume that the 25 observation are a random sample from a normally distributed population. If the population has the same standard deviation as the population of presumably normal infants, then  $\sigma = 0.7$ . We may test the hypothesis  $H : \mu = \mu_0 = 13.0$ . Suppose that we agree to use a level of significance  $\alpha = 0.05$ .

Then the critical values of the test statistic  $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$  are  $z_{0.025} = -1.96$  and  $z_{0.975} = 1.96$ . But for the sample we are given that  $\bar{x} = 12.6$  and thus

$$Z = \frac{12.6 - 13.0}{0.7/\sqrt{49}} = -\frac{0.4}{0.10} = 4.$$

Because this value is less than  $-1.96$ , it falls into the critical region and we reject the hypothesis. Our conclusion is that the infants cannot (at the 0.05 level of significance) be as the population which is normal.

Notice that a 95 per cent confidence interval for the mean of the population from which the sample of 49 is drawn would be

$$\bar{x} \pm 1.96 \left( \sigma/\sqrt{n} \right) = 12.6 \pm (1.96)(0.10) = 12.6 \pm 0.196 = 12.404$$

and 12.796. The hypothetical value 13.0 does not fall within this interval.

**Example 7 :** The mean life time of sample of 400 fluorescent light tube produced by a company is found to be 1570 hours with a standard deviation of 150 hours. Test the hypothesis that the mean life time of the bulbs produced by the company is 1600 hours against the alternative hypothesis that it is greater than 1600 hours at 1% level of significance.

**Solution :** Here  $n = 400 \Rightarrow$  it is a large sample.

Also  $\bar{x} = 1570$  standard deviation of sample mean  $s = 150$ ,  $\sigma = 1600$ .

**1. Null hypothesis :** The mean life time of the bulbs is 1600 hours, i.e.,  $H_0 : \mu = 1600$ .

**Alternative Hypothesis**  $H_1 : \mu > 1600 \Rightarrow$  It is a case of right tailed test.

**2. Computation of Test Statistic Z :**

$$\text{Standard error} = \text{S.E.}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{150}{\sqrt{400}} = \frac{150}{20} = 7.5.$$

$$\text{Test Statistic } Z = \frac{\bar{x} - \mu}{\text{S.E.}(\bar{x})} = \frac{1570 - 1600}{7.5} = -\frac{30}{7.5} = -4.$$

**3. Level of significance :** Here  $\alpha = 0.01$ .

**4. Critical value :** The critical value of  $Z$  at 1% level of significance for the Right tailed test (from the table) is  $Z_\alpha = 2.33$ .

**5. Decision :** Since the calculated value of  $|Z|$  is 4 which is greater than the critical value  $|Z_\alpha|$  (at  $\alpha = 0.01$ ) = 2.33, so the **null hypothesis is rejected and the alternative hypothesis is accepted.**

Hence the mean life time of bulbs is greater than 1600 hours.

**Example 8 :** A sample of 900 members has a mean 3.4 cms and s.d. 2.61 cms. Can the sample be regarded as one drawn from a population with mean 3.25 cms. Using the level of significance as 0.05, is the claim acceptable?

**Solution :** Here  $n = 900 \Rightarrow$  it is a large sample.

Also,  $\bar{x} = 3.4$  cm, s.d. of sample  $s = 2.61$  cm,  $\mu = 3.25$ .

**1. Null hypothesis :** The sample has been drawn from a population with mean 3.25 cms, i.e.,  $H_0 : \mu = 3.25$

**Alternative hypothesis.**  $H_1 : \mu \neq 3.25$ .

It is a case of two tailed test.

**2. Calculation of test statistic:**

$$\text{Standard error of mean} = \text{S.E.}(\bar{x}) = \frac{s}{\sqrt{n}}$$

$$= \frac{2.61}{\sqrt{900}} = \frac{2.61}{30} \\ = 0.087.$$

$$\therefore \text{Test statistics : } Z = \frac{\bar{x} - \mu}{\text{S.E.}(\bar{x})} \\ = \frac{3.4 - 3.25}{0.087} \\ = 1.72$$

**3. Level of Significance :** Here  $\alpha = 0.05$ .

**4. Critical value :** From the table  $|Z|$  the value of  $Z$  at  $\alpha = 0.05$  is  $|Z_\alpha| = 1.96$ .

**5. Decision :** Here calculated value of  $|Z| = 1.72 <$  Critical value  $|Z_\alpha| = 1.96$   
 $\Rightarrow$  Null hypothesis is accepted.

Hence, the sample has been drawn from a population with mean 3.25 cms.

### EXERCISE 12.1

1. 'Sampling is a necessity under certain conditions'. Explain this with illustrative examples.
2. Distinguish between a population and a sample. What is a random sample? Describe some methods of drawing random sample from a finite population.
3. What is a sampling? Give its objects and name the laws which form the basis of sampling.
4. Explain the terms 'Parameter and Statistic' as used in sampling.
5. What are the sampling and non-sampling errors?
6. What is meant by statistic and its standard error?
7. Give the expression for standard error of the sample mean.
8. Explain the concept of standard error. Distinguish between standard error and sampling error.
9. Explain the concept of sampling distribution of a statistic.
10. Explain the utility of standard error of a statistic.
11. Explain the following concepts:
 

(i) Point estimation ;	(ii) Interval estimation ;
(iii) A good estimator ;	(iv) Confidence interval ;
(v) Level of significance ;	(vi) Critical values
(vii) Type I error ;	(viii) Type II error.
12. Explain the procedure for 'testing of hypothesis'.
13. For a given sample of 200 items drawn from a large population, the mean is 65 and the standard deviation is 8. Find the 95% confidence limits for the population mean.  
**[Hint:** 95% confidence limits for population mean =  $\bar{x} \pm \frac{s}{\sqrt{n}} Z_{\alpha} = 65 \pm \frac{1.96 \times 8}{\sqrt{200}} = 65 \pm 1.109$ ]
14. In a random sample of 400 oranges from a large consignment, 40 were considered bad. Estimate the percentage of defective oranges in the whole consignment and assign limits within which the percentage probability lies.
15. In a sample survey of 1000 house-wives in a city 23% prefer a particular brand of pressure cooker. Find 99% confidence limits for the percentage of all housewives in the city preferring that brand of cooker.
16. The foreman of a company has estimated the average quantity of extracted ore to be 36.8 tonnes per shift and the sample standard deviation to be 2.8 tonnes per shift, based on a random selection of 4 shifts. Construct 90% confidence interval around this estimate.
17. A random sample of 50 items drawn from a particular population has a mean 30 with a standard deviation 28. Construct a 98% confidence interval estimate of the population mean.

### ANSWERS

- |                        |                     |                       |
|------------------------|---------------------|-----------------------|
| 13. 63.89 to 66.19.    | 14. 5.5% to 14.5%.  | 15. 19.57% to 26.43%. |
| 16. $36.8 \pm 3.802$ . | 17. 20.68 to 39.32. |                       |

## 12.21 TEST OF SIGNIFICANCE OF DIFFERENCE BETWEEN TWO MEANS - LARGE SAMPLES

Let  $A$  and  $B$  be two populations with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$  respectively. Let us take two independent samples of size  $n_1$  and  $n_2$  from these two populations. Let  $\bar{x}_1$  and  $\bar{x}_2$  be the corresponding sample means. Then our problem is

- (i) To test the equality of two population means, i.e., to test whether  $\mu_1 = \mu_2$ . Or
- (ii) To test the significance of the difference between two independent sample means viz.,  $\bar{x}_1 - \bar{x}_2$ .

The following steps should be taken for testing the **significance difference between two means**.

**1. Null Hypothesis :** The null hypothesis is set in any one of following forms.

$H_0 : \mu_1 = \mu_2$  i.e., the two samples have been drawn from different populations having the same means and equal standard deviation or,

$H_0 : \mu_1 = \mu_2$  i.e., the two samples have been drawn from the same parent population.

**Alternative hypothesis :**  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test)

OR  $H_1 : \mu_1 < \mu_2$  or  $H_1 : \mu_1 > \mu_2$  (One tailed test)

**2. Computation of Test Statistic :** We have the following two cases:

**Case I :** When the population standard deviations  $\sigma_1$  and  $\sigma_2$  are known. In this case we have:

$$\text{Standard Error of Difference of Means: } S.E. (\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\text{Test Statistic : } Z = \frac{\bar{x}_1 - \bar{x}_2}{S.E. (\bar{x}_1 - \bar{x}_2)}$$

**Case II. When the population standard deviations  $\sigma_1$  and  $\sigma_2$  are not known.** In this case we are given sample standard deviations  $s_1$  and  $s_2$  and we calculate the S.E. of difference of means by the formula.

$$\text{Standard Error of Difference of Means : } S.E. (\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

$$\therefore \text{Test Statistic : } Z = \frac{\bar{x}_1 - \bar{x}_2}{S.E. (\bar{x}_1 - \bar{x}_2)}$$

The other two steps such as (i) Level of significance; (ii) Critical value  $Z_\alpha$ ; (iii) Decision making for testing the significance of the difference between two means are the same as those given in testing the significance of a mean.

**Example 9 :** A college conducts both day and night classes intended to be identical. A sample of 100 day students yields examination results as:  $\bar{x} = 72.4$  and  $\sigma_1 = 14.8$ .

A sample of 200 night students yields examination results as:  $\bar{x}_2 = 73.9$  and  $\sigma_2 = 17.9$ . Are the two means statistically equal at 10% level?

**Solution :** Here  $n_1 = 100$ ;  $n_2 = 2000$ ,  $\bar{x}_1 = 72.4$ ;  $\bar{x}_2 = 73.9$ ,  $\sigma_1 = 14.8$ ;  $\sigma_2 = 17.9$

1. Null hypothesis :  $H_0 : \mu_1 = \mu_2$ .

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2 \Rightarrow$  It is a case of **Two tailed test**.

2. Calculation of Test Statistic :

$$\text{S.E. of difference means : S.E. } (\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(14.8)^2}{100} + \frac{(17.9)^2}{200}} = 1.95$$

$$\therefore \text{Test Statistic : } Z = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} = \frac{72.4 - 73.9}{1.95} = -0.769.$$

$$\therefore |Z| = 0.769$$

3. Level of significance : Here  $\alpha = 0.10$ .

4. Critical value : At  $\alpha = 0.10$  is  $Z_\alpha = 1.645$ .

5. Decision : Since calculated value of  $Z = 0.769 <$  critical value  $Z_\alpha$  (at  $\alpha = 0.10$ ) = 1.645  
 $\Rightarrow H_0$  is accepted. Hence we conclude that the **two means are statistically equal**.

**Example 10 :** A random sample of 1000 workers from south India show that their mean wages are Rs. 47 per week with a standard deviation of Rs. 28. A random sample of 1500 workers from north India gives a mean wage of Rs. 49 per week with a standard deviation of Rs. 40. Is there any significant difference between their mean level of wages?

**Solution :** We are given the following information.

	South India	North India
Size of Sample :	$n_1 = 1000$	$n_2 = 1500$
Mean of Sample :	$\bar{x}_1 = 47$	$\bar{x}_2 = 49$
S.D. of Sample :	$s_1 = 28$	$s_2 = 40$
Population Mean :	$\mu_1 = \text{Population Mean}$	$\mu_2 = \text{Population mean}$

1. Null hypothesis : There is no significant difference between two mean level of wages, i.e.,  $H_0 : \mu_1 = \mu_2$ .

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2 \Rightarrow$  It represents a **two tailed test**.

2. Calculation of Test Statistic :

$$\text{Standard Error of : S.E. } (\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(28)^2}{1000} + \frac{(40)^2}{1500}} = 1.36.$$

$$\therefore \text{Test Statistic : } Z = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} = \frac{47 - 49}{1.36} = \frac{-2}{1.36} = -1.47.$$

$$\therefore |Z| = 1.47$$

**3. Level of significance :** Since the level of significance is not given so we take  $\alpha = 0.05$ .

**4. Critical value :** Value of  $Z_\alpha$  at  $\alpha = 0.05$  (from the table) is 1.96.

**5. Decision :** Since the calculated value of  $|Z| = 1.47 <$  Critical value of  $|Z_\alpha| = 1.96 \Rightarrow$  the null hypothesis is accepted  $\Rightarrow$  there is no significant difference between the two means level of wages.

**Example 11 :** The research unit in an organisation wishes to determine whether scores on the scholastic aptitude test are different for male and female applicants. Random samples of applicant's file are taken and summarised below:

*Applicants*

	Female	Male
$\bar{x}$	502.1	510.5
$s$	86.2	90.4
$n$	399	204

Using the above sample data, test the null hypothesis that the average score is same for the population male and female applicants. Use 5% significance level and assume that the scores are normally distributed in each case.

**Solution :** Let  $x_1$  and  $x_2$  be the random variables of female and male applicants. Then we have

$$\bar{x}_1 = 502.1$$

$$\bar{x}_2 = 510.5$$

$$s_1 = 86.2$$

$$s_2 = 90.4$$

$$n_1 = 399$$

$$n_2 = 204$$

$$\mu_1 = \text{population mean}$$

$$\mu_2 = \text{population mean.}$$

**1. Null hypothesis :** The average score is same for the population of male and female, i.e.,  $H_0 : \mu_1 = \mu_2$ .

**Alternative hypothesis :**  $H_1 : \mu_1 \neq \mu_2 \Rightarrow$  It represents a Two tailed test.

**2. Calculation of Test Statistic :**

$$\begin{aligned}
 \text{S.E. of difference of Means} &= \text{S.E.} (\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
 &= \sqrt{\frac{(86.2)^2}{399} + \frac{(90.4)^2}{204}} = \sqrt{\frac{7430}{399} + \frac{8172.16}{204}} \\
 &= \sqrt{18.62 + 40.06} = \sqrt{58.68} = 7.66
 \end{aligned}$$

$$\begin{aligned}
 \therefore \text{Test Statistic : } Z &= \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E.} (\bar{x}_1 - \bar{x}_2)} = \frac{502.1 - 510.5}{7.66} = \frac{-8.4}{7.66} = -1.097 \\
 \therefore |Z| &= 1.097
 \end{aligned}$$

**3. Level of Significance :** Here  $\alpha = 0.05$ .

**4. Critical Value :** The value of  $Z_\alpha$  (at  $\alpha = 0.05$  from the table) = 1.96.

**5. Decision :** Calculated value of  $|Z| = 7.05 >$  Critical value  $|Z_\alpha| = 1.96 \Rightarrow H_0$  is rejected  $\Rightarrow$  Null hypothesis is accepted  $\Rightarrow$  Average score is same for the population of male and female.

**Example 12 :** The mean yield of wheat from a district A was 210 kgs, with standard deviation 10 kgs per acre from a sample of 100 plots. In another district B, the mean yield was 220 kgs, with standard deviation 12 kgs from a sample 150 plots. Assuming that the standard deviation of the yield in the entire state was 11 kgs., test whether there is any significant difference between the mean yield of crops in the two districts. Take  $d = 0.05$ .

**Solution :** We have the following data.

District A	District B
$n_1 = 100$	$n_2 = 150$
$\bar{x}_1 = 210$	$\bar{x}_2 = 220$
$s_1 = 10$	$s_2 = 12$
$\sigma_1 = 11$	$\sigma_2 = 11$
Population Mean $\mu_1$	$\mu_2$

**1. Null Hypothesis :** There is no significant difference between the mean yield of crops in the two districts i.e.,  $H_0 : \mu_1 = \mu_2$ .

**Alternative hypothesis :**  $H_1 : \mu_1 \neq \mu_2$ . It represents a Two Tailed Test.

**2. Calculation of Test Statistic :**

$$\begin{aligned} \text{Standard Error : S.E. } (\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(11)^2}{100} + \frac{(11)^2}{150}} \\ &= \sqrt{\frac{121}{100} + \frac{121}{150}} = \sqrt{1.21 + 0.807} = \sqrt{2.017} = 1.42 \end{aligned}$$

$$\therefore \text{Test Statistic : } Z = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} = \frac{210 - 220}{1.42} = \frac{-10}{1.42} = -7.05.$$

$$\therefore |Z| = 7.05$$

**3. Level of significance :** We take  $\alpha = 0.05$  as level of significance is mentioned in the question.

**4. Critical value at  $\alpha = 0.05$ .** The critical value of  $Z$ , i.e.,

$$Z_\alpha \text{ (at } \alpha = 0.05) = 1.96.$$

**5. Decision :** Since calculated  $|Z| = 7.05 >$  Critical value  $|Z_\alpha| = 1.96 \Rightarrow H_0$  is rejected  $\Rightarrow$  There is a significant difference between the mean yield of crops in the two districts.

## 12.22 TEST OF SIGNIFICANCE FOR DIFFERENCE OF TWO STANDARD DEVIATIONS – LARGE SAMPLES

Let  $A$  and  $B$  be two populations with standard deviations  $\sigma_1$  and  $\sigma_2$ . Let  $n_1$  and  $n_2$  be the size of two samples with s.d.  $s_1$  and  $s_2$  respectively drawn from the population  $A$  and  $B$ . Then we want to test the null hypothesis  $H_0 : \sigma_1 = \sigma_2$ . The sampling distribution of standard deviation  $s$  for large samples is normal with

$$\text{Var}(s) = \frac{\sigma^2}{2n}$$

$$\therefore \text{Standard Error of Standard Deviation} = \frac{\sigma}{\sqrt{2n}}$$

$$\text{Also } \text{Var}(s_1 - s_2) = \text{Var}(s_1) + \text{Var}(s_2) = \frac{\sigma_1^2}{2n} + \frac{\sigma_2^2}{2n_2}$$

**Null Hypothesis :**  $H_0$  : that the sample standard deviation does not differ significantly  
i.e.,  $H_0 : \sigma_1 = \sigma_2$ .

**Alternative Hypothesis :**  $H_1 : \sigma_1 \neq \sigma_2 \Rightarrow$  Two Tailed Test.

$$\text{Test Statistic : } Z = \frac{s_1 - s_2}{\text{S.E.}(s_1 - s_2)} = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

Also,  $Z \sim N(0.1)$ .

The other steps are the same as given in the test of significance of difference of two sample means.

When the population standard deviations are not known,

$$\text{Standard Error : S.E.}(s_1 - s_2) = \sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}$$

$$\therefore \text{Test Statistic : } Z = \frac{s_1 - s_2}{\text{S.E.}(s_1 - s_2)}$$

**Example 13 :** The standard deviation of the height of (Honours) students of a college is 4.0 cm. Two samples are taken. The standard deviation of 100 B.Com (Hons.) student is 3.5 cm and 50 B.A. Eco (Hons.) students is 4.5 cm. Test the significance of the difference of standard deviations of the samples.

**Solution :** We have the following data

B.Com (Hons.)

$$n_1 = 100$$

$$s_1 = 3.5 \text{ cm}$$

$$\sigma = 4.00 \text{ cm}$$

B.A. (Hons.) Eco.

$$n_2 = 50$$

$$s_2 = 4.5 \text{ cm}$$

**1. Null hypothesis :** There is no significant difference between the standard deviations of two samples, i.e.,

$$H_0 = \sigma_1 = \sigma_2$$

**Alternative hypothesis :**  $H_1 : \sigma_1 \neq \sigma_2 \Rightarrow$  It represents a Two tailed test.

**2. Calculation of Test Statistic :**

$$\begin{aligned} \text{Standard Error (S.E.)} &:= (s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}} \\ &= \sqrt{\frac{(4)^2}{2 \times 100} + \frac{(4)^2}{2 \times 50}} = \sqrt{\frac{16}{200} + \frac{16}{100}} = \sqrt{0.24} = 0.49. \\ \text{Test Statistic} : Z &= \frac{s_1 - s_2}{\text{S.E.}(s_1 - s_2)} = \frac{3.5 - 4.5}{0.49} = -2.04. \\ \therefore |Z| &= 2.04. \end{aligned}$$

**3. Level of Significance :** Since no level of significance is given so we take  $\alpha = 0.05$ .

**4. Critical Value :** Critical value of  $Z$  at  $\alpha = 0.05$  is  $Z_\alpha = 1.96$ .

**5. Decision :** Since  $|Z| = 2.04 > Z_\alpha = 1.96 \Rightarrow$  the null hypothesis  $H_0$  is rejected  $\Rightarrow$  there is a significant difference between the two standard deviations.

## EXERCISE 12.2

1. A sample of 400 items is taken from a normal population whose mean is 4 and whose variance is also 4. If the sample mean is 4.45 can the sample be regarded as truly random sample?

[Hint : Here the level of significance is not given so we take  $\alpha = 0.05$ . Also,  $\bar{x} = 4.45$ ,  $\mu = 4$ ,  $\sigma^2 = 4$ . Let  $H_0 : \mu = 4$ .

$$\therefore \text{S.E.}(\bar{x}) = \frac{2}{\sqrt{400}} = \frac{2}{20} = 0.1, \quad Z = \frac{\bar{x} - 4}{\text{S.E.}(\bar{x})} = \frac{4.45 - 4}{0.1} = 4.5$$

Also  $Z_\alpha$  at  $\alpha = 0.05$  is = 1.96.

Since  $|Z| = 4.5 > Z_\alpha = 1.96 \Rightarrow H_0$  is rejected at 5% level of significance  $\Rightarrow$  Sample cannot be regarded as having been drawn from the population with mean 4.]

2. A random sample of 100 students gave a mean weight of 58 kilogram with s.d. of 4 kg. Test the hypothesis that the mean weight in the population is 60 kg.

[Hint :  $H_0 : \mu = 60$  kgs,  $H_1 : \mu \neq 60$  kgs (Two tailed test)]

$$\text{Test Statistic} : Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{58 - 60}{4/\sqrt{100}} = -5 \Rightarrow |Z| = 5.$$

Value of  $Z_\alpha$  at  $\alpha = 0.05$  is  $1.96 < |Z| = 5 \Rightarrow H_0$  is rejected  $\Rightarrow$  the mean weight of the population is not 60 kgs.]

3. The mean I.Q. of a sample of 1600 children was 99. It is likely that this was a random sample from a population with mean I.Q. 100 and standard deviation 15?

[Hint : Here  $n = 1600$ ,  $\bar{x} = 99$ ,  $\mu = 100$ ,  $\sigma = 100$ .

Null hypothesis  $H_0 : \mu = 100$ , Alternative hypothesis  $H_1 : \mu \neq 100 \Rightarrow$  Two tailed test.

$$\text{S.E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{1600}} = \frac{15}{40} = \frac{3}{8}$$

$$Z = \frac{\bar{x} - \mu}{\text{S.E.}(\bar{x})} = \frac{99 - 100}{3} \times 8 = -2.67 \Rightarrow |Z| = 2.67.$$

Since  $\alpha$  is not given so we take  $\alpha = 0.05$ .

$\therefore Z_{0.05} = 1.96 < |Z| = 2.67 \Rightarrow H_0$  is rejected  $\Rightarrow$  sample has not been drawn from a population with mean 100 and  $\sigma = 15$ .]

4. A new variety of potato grown in 250 plots gave rise to a mean yield of 82.7 quintals per hectare with a s.d. of 14.6 quintals per hectare. Is it reasonable to assert that the new variety is superior in yield to the standard variety with an established yield of 80.2 quintals per hectare?
5. If a random sample of size 20 from a normal population with standard deviation 5.2 shows a mean of 16.9, test at 5% significance level that the population mean is 15.5. Also calculate 99% confidence limit for mean.

[Hint : Null hypothesis  $H_0 : \mu = 15.5$ . Alternative  $H_1 : \mu \neq 15.5$ .  $Z = \frac{16.9 - 15.5}{(5.2/\sqrt{20})} = 1.20$ .

Also value of  $Z_\alpha$  of  $\alpha = 0.05 = 1.96$ .

Again  $|Z| = 1.20 < Z_\alpha = 1.96 \Rightarrow H_0$  is accepted  $\Rightarrow$  population mean may be 15.5.

99% confidence limits are  $16.9 \pm \frac{5.2}{\sqrt{20}} \times 2.58 = 16.9 \pm 3 = [13.9, 19.9]$

6. In a certain factory there are two independent processes for manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 gms with a standard deviation of 12 gms, while from other process are 124 and 14 in a sample of 400 items. Is the difference between the mean weights significant at 10% level of significance?

[Hint : Null hypothesis –  $H_0 : \mu_1 = \mu_2$ .

Alternative hypothesis –  $H_1 : \mu_1 \neq \mu_2$ .

$$\text{Standard Error (S.E.)} : (\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(12)^2}{250} + \frac{(14)^2}{400}} = 1.0325.$$

$$\text{Test Statistic} : Z = \frac{120 - 124}{1.0325} = -3.87. \quad \therefore |Z| = 3.87$$

$Z_\alpha$  at  $\alpha$  of 10% = 1.645. Again  $|Z| > Z_\alpha$  as  $3.87 > 1.645 \Rightarrow H_0$  is rejected  $\Rightarrow$  there is a significant difference between two sample mean weights].

7. The mean yield of two sets of plots and their variability are as given below. Examine whether the difference in the variability in yields is significant.

	Set of 40 plots	Set of 60 plots
Mean yield per plot	1258 kg	1243 kg
S.D. per plot	34	28

[Hint :  $H_0 : \sigma_1 = \sigma_2$ ,  $H_1 : \sigma_1 \neq \sigma_2$ .

$$Z = \frac{s_1 - s_2}{\sqrt{s_1^2 / 2n_1 + s_2^2 / 2n_2}} = \frac{34 - 28}{\sqrt{(34)^2 / 80 + (28)^2 / 120}} = \frac{6}{\sqrt{4.45 + 6.63}} = 1.31.$$

Take  $\alpha = 0.05$ , then  $Z_\alpha = 1.96$ .

Now  $|Z| < Z_\alpha$  as  $1.31 < 1.96 \Rightarrow H_0$  is accepted  $\Rightarrow$  there is a significant difference between two sample mean weights].

8. A random sample of 50 male employees is taken at the end of a year and the mean number of hours of absenteeism for the year is found to be 63 hours. A similar sample of 50 female employees has mean of 66 hours. Could these samples be drawn from a population with the same mean and s.d. 10 hours? State clearly the assumption you made.

[Hint :  $H_0 : \mu_1 = \mu_2$ .  $H_1 : \mu_1 \neq \mu_2$

$$\text{S.E.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\frac{100}{50} + \frac{100}{50}} = 2.$$

$$\text{Test Statistic } Z = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E.}(\bar{x}_1 - \bar{x}_2)} = \frac{63 - 66}{2} = -1.5. \quad |Z| = 1.5.$$

Take  $\alpha = 0.05$ , then  $Z_\alpha = 1.96$ .

Now  $|Z| < Z_\alpha$  as  $1.5 < 1.96 \Rightarrow H_0$  is accepted  $\Rightarrow$  the two samples have been drawn from the same population.]

9. Intelligence test of two groups of boys and girls gives the following results.

Girls :  $\bar{x}_1 = 84$ , s.d.  $s_1 = 10$ ,  $n_1 = 121$

Boys :  $\bar{x}_2 = 81$ , s.d.  $s_2 = 12$ ,  $n_2 = 81$ .

(a) Is the difference in mean scores significant?

(b) Is the difference between standard deviations significant?

[Hint : (a) Null hypothesis  $H_0 : \mu_1 = \mu_2$ .

$$\text{Test Statistic : } Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}} = \frac{84 - 81}{\sqrt{100 / 121 + 144 / 81}} = 1.863.$$

Take  $\alpha = 0.05$  so that value of  $Z_\alpha = 1.96$ .

Now  $|Z| < Z_\alpha$  as  $1.753 < 1.96 \Rightarrow H_0$  is accepted  $\Rightarrow$  Difference of means scores is not significant.

(b) Null hypothesis :  $H_0 : \sigma_1 = \sigma_2$

Test Statistic :

$$Z = \frac{s_1 - s_2}{\sqrt{s_1^2 / 2n_1 + s_2^2 / 2n_2}} = \frac{10 - 12}{\sqrt{(10)^2 / (2 \times 121) + (12)^2 / (2 \times 81)}} = \frac{-2}{\sqrt{1.302}} = -1.753$$

Take  $\alpha = 0.05$  so that the value of  $Z_\alpha$  at  $\alpha = 0.05$  is 1.96. Again  $|Z| < Z_\alpha$  as  $1.753 < 1.96 \Rightarrow H_0$  is accepted  $\Rightarrow$  the difference between the two standard deviations is not significant.]

10. Random samples drawn from two places gave the following data relating to the heights of children:

:	Place A	Place B
Mean height (in cm)	68.50	68.58
Standard Deviation (in cm)	2.5	3.0
Number of items sample	1200	1500

Test at 5% level that the mean height is the same for children at two places.

[Hint :  $H_0 : \mu_1 = \mu_2$ ,  $H_1 : \mu_1 \neq \mu_2$ ,

$$\text{S.E.} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(2.5)^2}{1200} + \frac{(3)^2}{1500}} = 0.1058$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E.}} = \frac{68.50 - 68.58}{0.1058} = -0.756$$

But  $Z_{0.05} = 1.96 >$  calculated value of  $|Z| = 0.756$

$\Rightarrow$  the null hypothesis is accepted]

## ANSWERS

1. Sample cannot be regarded as having been drawn from the population with mean 4.
2. The mean weight of the population is not 60 kgs.
3. Sample has not been drawn from the population with mean 100 and  $\sigma = 15$ .
4.  $Z = 2.71$ , yes.      5. 13.9 to 19.9
6. There is a significant difference between the two sample means weights.
7.  $|Z| = 5.03$ , Difference is not significant.
8. The two samples have been drawn from the same population.
9. (a) Difference of means scores is not significant.  
(b) The difference between the two standard deviations is not significant.
10. Null hypothesis is accepted.

## 12.23 TEST OF SIGNIFICANCE FOR SINGLE PROPORTION – LARGE SAMPLES

### WORKING RULE

**1. Null hypothesis :** The sample has been drawn from a population with population proportion  $P$  i.e.,  $H_0 : P = P_0$ , where  $P_0$  is a particular value of  $P$ .

**Alternative Hypothesis :**  $H_1 : P \neq P_0 \Rightarrow$  Two tailed test.

**2. Selection of Test Statistic**

$$\text{Standard Error of Proportion : S.E.}(p) = \sqrt{\frac{P}{n}}$$

where  $n$  is the sample size and  $P$  is population proportion.

$$\text{Standard Error of Percentage: S.E.}(P) = \sqrt{\frac{P(100 - P)}{n}}$$

where  $P$  = population percentage and  $n$  is the sample size.

$$\text{Test Statistic : } Z = \frac{p - P}{\text{S.E.}(p)}$$

where  $p$  = sample proportion or percentage.

The other steps such as (3) level of significance; (4) Critical value of  $Z$ , (5) Decision are the same as those discussed in test the significance of a sample mean.

**Example 14 :** A dice was thrown 9000 times and of these 3220 yielded a 3 or 4. Is this consistent with the hypothesis that the dice was unbiased?

**Solution :** Here the sample proportion  $p = \frac{3220}{9000} = 0.358$ .

Also the population proportion  $P = \frac{2}{6} = \frac{1}{3}$ .

**1. Null hypothesis :** The dice is unbiased i.e.,  $H_0 : P = \frac{1}{3}$ .

**Alternative hypothesis**  $H_1 : P \neq \frac{1}{3} \Rightarrow$  It leads to Two Tailed Test.

**2. Selection of Test Statistic :**

$$\text{Standard Error of proportion: S.E.}(p) = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{1}{3} \times \frac{2}{3} \times \frac{1}{9000}} = \sqrt{0.000025} = 0.005$$

$$\therefore \text{Test Statistic : } Z = \frac{p - P}{\text{S.E.}(p)} = \frac{0.358 - 0.333}{0.005} = \frac{0.025}{0.005} = 5.$$

**3. Level of Significance :** Take  $\alpha = 0.05$  as the level of significance is not given.

**4. Critical value :** The critical value of  $Z$  at  $\alpha = 0.05$  is  $Z_\alpha = 1.96$  (from Normal Table).

**5. Decision :** Now  $|Z| > |Z_\alpha|$  as  $5 > 1.96 \Rightarrow$  Null hypothesis  $H_0$  is rejected  $\Rightarrow$  that the dice is a biased one.

**Example 15 :** In a sample of 400 burners there were 12 whose internal diameters were not within tolerances. Is this sufficient evidence for concluding that the manufacturing process is turning more than 2% defective burners. Let  $\alpha = 0.05$ .

**Solution :** It is given that  $P = 0.02$  so that  $Q = 1 - 0.02 = 0.98$ . Also  $p = \frac{12}{400} = 0.03$ .

**1. Null hypothesis :** The process is in control i.e.,  $H_0 : P \leq 0.02$ .

**Alternative hypothesis :**  $H_1 : P > 0.02 \Rightarrow$  it leads to One Tailed Test.

**2. Selection of test Statistic :** The test statistic

$$Z = \frac{p - P}{\text{S.E.}(p)} = \frac{p - P}{\sqrt{PQ/n}} = \frac{0.03 - 0.02}{\sqrt{(0.02 \times 0.98)/400}} = \frac{0.01}{\sqrt{0.000049}} = \frac{0.01}{0.007} = 1.429.$$

**Level of significance :**  $\alpha = 0.05$  or  $\frac{\alpha}{2} = 0.025$ .

**Critical Value :** The critical value  $Z_\alpha$  of  $Z$  for  $\alpha = 0.05$  for One Tailed Test is = 1.645.

**Decision :** Since  $|Z| < Z_\alpha$  as  $1.429 < 1.645$  so the null hypothesis is accepted  $\Rightarrow$  the process is not out of control.

**Example 16 :** In a random sample of 400 persons from a large population 120 are females. Can it be said that males and females are in the ratio 5 : 3 in the population? Use 10% level of significance.

**Solution :** We are given the following data.

$$\text{Population proportion of female } P = \frac{3}{5+3} = \frac{3}{8} = 0.375. \therefore Q = 1 - P = 1 - 0.375 = 0.625$$

$$\text{Sample proportion of female : } p = \frac{120}{400} = 0.30. \text{ Also } n = 400.$$

**1. Null hypothesis :**  $H_1 : P = 0.375$

**Alternative hypothesis :**  $H_1 : P \neq 0.375$ . It leads to Two Tailed Test.

**2. Test Statistic.**

$$Z = \frac{p - P}{\text{S.E.}(p)} = \frac{p - P}{\sqrt{PQ/n}} = \frac{0.30 - 0.375}{\sqrt{0.375 \times 0.625/400}} = \frac{-0.075}{\sqrt{0.000586}} = -\frac{0.075}{0.024} = -3.125.$$

$$\therefore |Z| = 3.125$$

**3. Level of significance :** Here  $\alpha = 0.05$ .

**4. Critical value of  $Z$  at  $\alpha = 0.05$  is  $Z_\alpha = 1.96$ .**

**5. Decision :** Here  $|Z| > |Z_\alpha|$  as  $3.125 > 1.96$  so we reject the null hypothesis  $\Rightarrow$  the males and females in the population are not in the ratio 5 : 3.

## 12.24 TEST OF SIGNIFICANCE OF DIFFERENCE BETWEEN TWO SAMPLE FOR LARGE SAMPLES

### WORKING RULE

**1. Null hypothesis :** The two samples have been drawn from the same population i.e.,

$$H_0 : P_1 = P_2.$$

**Alternative hypothesis :**  $H_1 : P_1 \neq P_2$ . (Two Tailed Test)

**2. Computation of Test Statistic :** We have the following cases:

**Case I :** When the population proportions  $P_1$  and  $P_2$  are known.

In this case  $Q_1 = 1 - P_1$ ,  $Q_2 = 1 - P_2$  and  $p_1$  and  $p_2$  are sample proportions.

$$\therefore \text{Standard Error of Difference} : \text{S.E.}(p_1 - p_2) = \sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}$$

$$\therefore \text{Test Statistic} : Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)}.$$

**Case II :** When the population proportions  $P_1$  and  $P_2$  are not known but sample proportions  $p_1$  and  $p_2$  are known.

In this case we have two methods to estimate  $P_1$  and  $P_2$ .

(a) **Method of substitution :** In this method  $p_1$  and  $p_2$  are substituted for  $P_1$  and  $P_2$ .

$$\text{Standard Error of Difference} : \text{S.E.}(p_1 - p_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}.$$

$$\text{Test Statistic} : Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)}.$$

The other steps such as (i) level of significance; (ii) Critical value, (iii) decision in respect of testing the significance of difference of two proportions are the same as those discussed in testing the significance of sample means.

(b) **Method of Pooling :** In this method the estimated value for the two population proportions is obtained by pooling the two sample proportions  $p_1$  and  $p_2$  into a single proportion  $p$  by the formula given below:

**Sample Proportion of Two Samples or Estimated  $p$**

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad \text{so that } q = 1 - p.$$

$$\text{Standard Error of Difference} : \text{S.E.}(p_1 - p_2) = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$

where  $n_1$  and  $n_2$  are the sizes of the two samples.

$$\text{Test Statistic : } Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)}$$

The other steps such as (i) level of significance, (ii) critical value of  $Z$  at  $\sigma$ , i.e.,  $Z_\alpha$  and (iii) decision making in respect of testing the significance of the difference between two proportions is the same as those discussed the difference of sample means.

**Example 17 :** A machine produced 20 defective articles in a batch of 400. After overhauling, it produced 10 defective in a batch of 300. Has the machine improved.

**Solution :** 1. Null hypothesis : There is no significant difference in the improvement of the machine before and after overhaul i.e.,  $H_0 : P_1 = P_2$ .

Alternative hypothesis :  $H_1 : P_2 > P_1 \Rightarrow$  the machine has improved  $\Rightarrow$  It is a One tailed test.

### 2. Computation of Test Statistic :

$$p_1 = \text{Sample proportion of defective articles before overhaul} = \frac{20}{400} = 0.05$$

$$\therefore q_1 = \text{Sample proportion of defective articles after overhaul} = \frac{10}{300} = 0.033$$

$$\therefore q_2 = 1 - 0.333 = 0.967$$

### Standard Error of difference :

$$\begin{aligned} \text{S.E.}(p_1 - p_2) &= \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{\frac{0.05 \times 0.95}{400} + \frac{0.033 \times 0.967}{300}} \\ &= \sqrt{0.000119 + 0.00011} = \sqrt{0.00023} = 0.015. \end{aligned}$$

$$\text{Test Statistic : } Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)} = \frac{0.05 - 0.033}{0.015} = \frac{0.017}{0.015} = 1.134.$$

$$\therefore |Z| = 1.134$$

3. Level of significance : Take  $\alpha = 0.05$ .

4. Critical value of  $Z$  at  $\alpha = 0.05$  is  $Z_\alpha = 1.645$  for one tailed test.

5. Decision : Since  $|Z| = 0.134 < |Z_\alpha| = 1.645 \Rightarrow$  Null hypothesis  $H_0$  is accepted  $\Rightarrow$  that the machine has not improved after overhauling.

**Example 18 :** On a certain day, 74 trains were arriving on time at Delhi station during the rush hours and 83 were late. At New Delhi there were 65 on time and 107 late. Is there any difference in the proportions arriving on time at the two stations?

**Solution :**

$$p_1 = \text{proportion of trains arriving on time at Delhi Station} = \frac{74}{74 + 83} = \frac{74}{157} = 0.471.$$

$$p_2 = \text{proportion of trains arriving on time at New Delhi station} = \frac{65}{65 + 107} = \frac{65}{172} = 0.378.$$

Mean proportion of trains arriving on time

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{157 \times (74/157) \times (67/172)}{157 + 172} = \frac{74 + 65}{329} = \frac{139}{329} = 0.423$$

$$\therefore q = 1 - p = 0.577.$$

**1. Null hypothesis :** There is no difference of train arriving ‘on time’ at two stations, i.e.,  $H : P_1 = P_2$ .

**Alternative hypothesis :**  $H : P_1 \neq P_2 \Rightarrow$  It leads to **Two Tailed Test**.

**2. Calculation of Test Statistic :** Here  $p = 0.423$ ,  $q = 0.577$ ,  $n_1 = 157$ ,  $n_2 = 172$ .

**∴ Standard Error of Difference :**

$$\begin{aligned} \text{S.E. } (p_1 - p_2) &= \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} = \sqrt{(0.423 \times 0.577) \left( \frac{1}{157} + \frac{1}{172} \right)} \\ &= \sqrt{0.244 \times [0.00637 + 0.00581]} \\ &= \sqrt{0.244 \times 0.012} = \sqrt{0.00293} = 0.054 \end{aligned}$$

$$\therefore \text{Test Statistic : } Z = \frac{P_1 - P_2}{\text{S.E. } (p_1 - p_2)} = \frac{0.471 - 0.378}{0.054} = \frac{0.093}{0.054} = 1.722.$$

**Level of significance :** Take  $\alpha = 0.05$ .

**Critical value :** The critical value of  $Z$  at  $\alpha = 0.05$  is  $Z_{0.05} = 1.96$  (From the normal table).

**Decision :** Since  $|Z| < Z_{0.05}$  as  $1.722 < 1.96 \Rightarrow$  Null hypothesis  $H_0$  is accepted  $\Rightarrow$  that there is no significant difference of trains ‘arriving on time’ at the two stations.

**Example 19 :** In a sample of 600 men from a certain city, 450 men are found to be smokers. In a sample of 900 from another city, 450 are found to be smokers. Do the data indicate that the two cities are significantly different with respect to prevalence of smoking habits among men?

**Solution :** Here we are given for one city  $n_1 = 600$ ,  $p_1 = \text{proportion of smokers} = \frac{450}{600} = 0.75$ .

Also for another city  $n_2 = 900$ ,  $p_2 = \text{proportion of smokers} = \frac{450}{900} = 0.5$ .

**1. Null hypothesis :** There is no significant difference in the smoking habits of two cities, i.e.,  $H_0 : P_1 = P_2$ .

**Alternative hypothesis :**  $H_1 : P_1 \neq P_2 \Rightarrow$  a Two Tailed Test.

**2. Computation of Test Statistic :** Sample mean proportion of the two sample proportions

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{450 + 450}{600 + 900} = \frac{900}{1500} = 0.6.$$

**Standard Error of Difference :**

$$\begin{aligned}\text{S.E. } (p_1 - p_2) &= \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{\frac{0.6 \times 0.4}{600} + \frac{0.6 \times 0.4}{900}} \\ &= \sqrt{0.0004 + 0.000267} = \sqrt{0.000667} = 0.0258.\end{aligned}$$

**Test Statistic :**  $Z = \frac{p_1 - p_2}{\text{S.E. } (p_1 - p_2)} = \frac{0.75 - 0.5}{0.0258} = 9.7$

**3. Level of significance :**  $\alpha = 0.05$ .

**4. Critical value :** The critical value of  $Z$  at  $\alpha = 0.05$  is  $Z_{0.05} = 1.96$  (From the normal table).

**5. Decision :** Since  $|Z| = 9.7 > Z_{0.05} = 1.96 \Rightarrow$  that the null hypothesis is rejected  
 $\Rightarrow$  There is a significant difference in the smoking habits of two cities.

**Example 20 :** In an infantile paralysis epidemic 500 persons contracted the disease. 300 received no serum treatment and of them 75 became paralysed. Of those who received serum treatment 65 became paralysed. Was the serum treatment effective?

**Solution :** Let  $p_1, p_2$  respectively be the proportions of persons who received no serum and who received serum treatment.

$$\therefore p_1 = \frac{75}{300} = 0.25, \quad p_2 = \frac{65}{500 - 300} = \frac{65}{200} = 0.325$$

Sample mean proportion of the two sample proportions

$$= p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{300 \times (75/300) + 200 \times (65/200)}{300 + 200} = \frac{75 + 65}{500} = 0.28.$$

**1. Null hypothesis :** The serum treatment was not effective, i.e.,  $H : P_1 = P_2$ .

**Alternative hypothesis :**  $H_1 : P_1 \neq P_2 \Rightarrow$  A Two Tailed Test.

**2. Computation of Test Statistic :**

**Standard Error of Difference :**

$$\begin{aligned}\text{S.E. } (p_1 - p_2) &= \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{(0.28 \times 0.72) \left[ \frac{1}{300} + \frac{1}{200} \right]} \\ &= \sqrt{\frac{1.008}{600}} = \sqrt{0.00168} = 0.041.\end{aligned}$$

$$3. \therefore \text{Sample statistic } Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)} = \frac{0.25 - 0.325}{0.041} = -\frac{0.075}{0.041} = -1.83.$$

$$\therefore |Z| = 1.83$$

4. **Level of significance :** Let us take  $\alpha = 0.05$ .

5. **Critical value :** The critical value of  $Z$  at  $\alpha = 0.05$  is  $Z_{0.05} = 1.96$ . (From the normal table values)

6. **Decision :** Since the calculated value of  $|Z| = 1.83 < Z_{0.05} = 1.96$  so the **null hypothesis  $H_0$  is accepted**  $\Rightarrow$  **the serum treatment was not effective.**

### EXERCISE 12.3

1. A manufacturer claims at least 95% of the items he produces are failure free. Examination of a random sample of 600 items showed 39 to be defective. Test the claim at a significance of 0.05.
2. A bottle manufacturing process is 'under control' if no more than 1% of the bottles are defective. A random sample of 120 bottles showed 5 to be defective. Do these data indicate that the process is out of control? Use  $\alpha = 0.10$ .
3. In a sample of 400 parts manufactured by a factory, the number of defective parts was found to be 30. The company, however, claimed that only 5% of their product is defective. Is the claim tenable?

[Hint : Here  $n = 400$ ,  $p = \frac{30}{400} = 0.075$ ,  $P = 0.05$ ,  $Q = 0.95$

Null hypothesis  $H_0 : P = 0.05$ .  $H_1 : P > 0.05$  (Right Tailed Test)

$$Z = \frac{p - E(p)}{\text{S.E.}(p)} = \frac{0.075 - 0.050}{\sqrt{(0.05 \times 0.95) / 400}} = 2.27$$

Value of  $Z_\alpha$  at  $\alpha = 0.05$  for one tailed test is 1.645.

Now  $|Z| > Z_\alpha$  as  $2.27 > 1.645$  so we reject the null hypothesis.

$H_0 \Rightarrow$  **Company's claim that only 5% of their product is defective is rejected.**

4. A dice was thrown 400 times and 'six' resulted 80 times. Do the data justify the hypothesis of an unbiased dice?

[Hint :  $H_0 : P = \frac{1}{6}$

$$Z = \frac{p - P}{\text{S.E.}(p)} = \frac{(1/5) - (1/6)}{\sqrt{(1/6) \times (5/6) \times (1/400)}} = 1.79.$$

Also  $Z_\alpha$  at  $\alpha = 0.05$  is 1.96.

Since  $|Z| < Z_\alpha$  as  $1.79 < 1.96$  so we accept the hypothesis that the dice is an unbiased.]

5. A sample of size 600 persons selected at random from a large city shows that the percentage of male in the sample is 53% . It is believed that male to the total population ratio in the city is 1/2. Test whether this belief is confirmed by the observation.

[Hint : Here  $P = 50$ ,  $Q = 100 - 50 = 50$ ,  $p = 53$ ,  $H_0 = P = 50\%$ .

$$\text{S.E.}(p) = \sqrt{\frac{50(100-50)}{600}} = 2.04$$

$$\text{Test Statistic : } Z = \frac{p - P}{\text{S.E.}(p)} = \frac{53 - 50}{2.04} = 1.5.$$

$Z_\alpha$  at  $\alpha = 0.05$  is 1.96.

Now  $|Z| < Z_\alpha$  as  $1.96 > 1.5 \Rightarrow$  Null hypothesis  $H_0$  is accepted  $\Rightarrow$  male to the total population ratio in the city is  $\frac{1}{2}$ . i.e., male population is 50%]

6. A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts 3 imperfect articles in a batch of 100. Has the machine improved?

[Hint : Here  $p_1 = \frac{16}{500} = 0.032 \Rightarrow q_1 = 0.968$ .

$$p_2 = \frac{3}{100} = 0.030 \Rightarrow q_2 = 0.970.$$

$$H_0 : P_1 = P_2. \quad H_1 : P_1 < P_2$$

$$\text{S.E. of difference} = \text{S.E.}(p_1 - p_2) = \sqrt{\frac{0.032 \times 0.968}{500} + \frac{0.030 \times 0.97}{100}} = 0.0187.$$

$$\text{Test Statistic : } Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)} = \frac{0.032 - 0.030}{0.0187} = 0.107.$$

$Z_\alpha$  at  $\alpha = 0.05$  is 1.645. Also  $|Z| < Z_\alpha \Rightarrow H_0$  is accepted  $\Rightarrow$  machine has not improved after over hauling.]

7. In a random samples of 600 and 1000 men from two cities 400 and 600 men are found to be literate. Do the data indicate at 5% level of significance that the population are significantly different in the percentage literacy?

[Hint : Here  $p_1 = \frac{400}{600} = \frac{2}{3}$ ,  $p_2 = \frac{600}{1000} = \frac{3}{5}$ .

$$\therefore p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{5}{8}.$$

Null hypothesis :  $H_0 : P_1 = P_2$ .

Alternative hypothesis :  $H_1 : P_1 \neq P_2$ .

$$\text{Standard Error of Difference : S.E. } (p_1 - p_2) = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} = \frac{1}{40}.$$

$$\text{Test Statistic : } Z = \frac{p_1 - p_2}{\text{S.E. } (p_1 - p_2)} = \frac{(2/3) - (3/5)}{(1/40)} = 2.67$$

Critical value of  $Z$  at  $\alpha = 0.05$  is  $Z_{0.05} = 1.96$ .

Also  $|Z| > Z_{0.05}$  as  $2.67 > 1.96 \Rightarrow H_0$  is rejected  $\Rightarrow$  population are significantly different in the percentage of literacy.]

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

$$[\text{Hint : } H_0 : P_1 = P_2, \quad H_1 : P_1 \neq P_2]$$

$$p = \frac{500 \times 0.62 + 400 \times 0.59}{500 + 400} = 0.607$$

$$Z = \frac{0.62 - 0.59}{\sqrt{0.607 \times 0.393 (1/500) + (1/400)}} = 0.9155$$

$Z_{0.05} = 1.96$ . Also  $|Z| < Z_{0.05} \Rightarrow H_0$  is accepted  $\Rightarrow$  there is no significant difference in the two groups in their attitude towards the new plan].

9. Before an increase in excise duty on tea 400 people out of a sample of 500 persons were found to be tea drinkers. After an increase in the duty, 400 persons were known to the tea drinkers in a sample 600 people. Do you think that there has been a significant decrease in the consumption of tea after the increase in the excise duty?

$$[\text{Hint : } p_1 = \frac{400}{500} = 0.80; \quad p_2 = \frac{400}{600} = 0.67]$$

Null hypothesis :  $H_0 : P_1 = P_2$ .

Alternative hypothesis :  $H_1 : P_2 < P_1 \Rightarrow$  (One tailed test)

$$\text{Also } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{8}{11}, \quad q = 1 - \frac{8}{11} = \frac{3}{11}$$

$$\text{Test Statistic : } Z = \frac{0.80 - 0.67}{\sqrt{(8/11) \times (3/11) [(1/500) + (1/600)]}} = 4.81$$

Also  $Z_{0.05} = 1.645$  for one tailed test.

Now  $|Z| = 4.81 > Z_{0.05} = 1.645 \Rightarrow H_0$  is rejected  $\Rightarrow$  there is significant decrease in the consumption of tea after the increase in excise duty.]

**ANSWERS**

1.  $Z = 1.69$ ;  $Z_{0.05} = 1.64$ ; Reject  $H_0 : p = 0.05$ .
2.  $Z = 1.10$ ;  $Z_{10} = 1.282$ ; No.
3. Company's claim that only 5% of their product is defective is rejected.
4. The die is an unbiased one.
5. Male population is 50%.
6. Machine has not improved after overhauling.
7. Population are significantly different in the percentage of literacy.
8. There is no significant difference in the two groups in their attitude towards the new plan.
9. There is a significant decrease in the consumption of tea after the increase in excise duty.



# 13

## *Students' t-Test*

### 13.1 INTRODUCTION

It will be recalled that, in the development of testing theory, the numerical values of the mean and standard deviation, as calculated from a particular sample, were used as an estimate of the mean and standard deviation of the population of which the sample was drawn.

It can be shown mathematically that the standard deviation calculated from any sample is an underestimation of the standard deviation of the population. This underestimation can be partially compensated for if the standard deviation is calculated from the formula,

$$\text{S.D.} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

rather than from the formula,

$$\text{S.D.} = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

When  $n$  is large, the resultant values from these two formulae will be approximately the same, since division by  $n$  and  $n-1$  will give approximately the same values. When  $n$  is small, however, the resultant values may be different. The **standard deviation for small samples**, therefore, is always calculated from the formula:

$$\text{S.D.} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

To denote that this formula has been used, this value is designated by the small letters  $\hat{s}_x$ . Therefore:

$$\hat{s}_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

where  $\bar{x}$  = mean of the sample.

When dealing with *large* samples, the values of the relative deviates ( $x/s_x$ ) are distributed approximately in a normal distribution and by using a table of areas of a normal curve, the probability (P) of any given difference can be obtained. When the samples are small the normal approximation may not be used. However, if the population distribution is normal, relative deviates may be calculated by using  $\hat{s}$  in the denominator, and these referred to the table of the *t* distribution. These distributions, like the normal distribution, are always symmetrical, but they have more area in the tails than does the normal curve. The particular *t* distribution to be used for any given value will depend on the degrees of freedom. In general, the degrees of freedom are given by the number of independent differences, ( $x - \bar{x}$ ) used in determining the value of  $\hat{s}$  which is  $n - 1$ . When two samples are being considered, it will be the number of independent differences used in computing the common value of  $\hat{s}$  which is  $(n_1 + n_2 - 2)$ .

## EXACT SAMPLE (OR SMALL SAMPLE) TESTS

### 13.2 STUDENT'S *t*-DISTRIBUTION

When the sample is small (i.e.,  $n \leq 30$ ) then the distribution of the standardised variable  $Z$  of the statistic *t* will be far from normality and as a result the **normal test** cannot be applied. In order to deal with **small samples** (when sample size  $n \leq 30$ ) new techniques and tests to significance known as **EXACT SAMPLE TECHNIQUES (TESTS)** have been developed.

It is important to note that "Exact sample technique (test) can be applied even for large samples but the large sample theory cannot be applied for small samples".

The theory of small (or exact) sample was developed by Irish statistician William S. Cosset who used to write under pseudonym (pen name) of student in 1908. Cosset gave his statistic name as **Student's *t*-distribution or *t*-test**. The quantity *t* is defined as

$$t = \frac{\text{Difference of population parameter and the corresponding statistic}}{\text{Standard error of the statistic}}$$

with  $(n - 1)$  degrees of freedom if the sample size is  $n$ .

Note : The degree of freedom for  $t$  = (size of sample) – 1.

**Student's *t*-distribution :** If  $x_1, x_2, \dots, x_n$  is a random sample size of  $n$  drawn from normal population with mean  $\mu$  and variance  $\sigma^2$  (not known) then the student *t* statistic (mean) is defined as

$$t = \frac{\bar{x} - \mu}{(\text{S.E. of mean})} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

with  $(n - 1)$  degrees of freedom.

where  $s$  = standard deviation of the sample.

### 13.3 ASSUMPTIONS FOR *t*-TEST

The *t*-test is applied under following assumptions:

1. Samples are drawn from normal population and are random.

2. For testing the equality of two population means, the population variances are regarded as equal.
3. The population standard deviation may not be known.
4. In case of two samples some adjustments in degrees of freedom for  $t$  are made.

### 13.4 PROPERTIES OF t-DISTRIBUTION

- (i)  $t$ -distribution is asymptotic to  $x$ -axis i.e., it extends to infinity on either side.
- (ii) The shape of the curve or form of  $t$ -distribution varies with the degrees of freedom. The degree of freedom is defined as (size of sample – one).
- (iii)  $t$ -distribution is a symmetrical distribution with mean zero.
- (iv) Its graph is similar to that of normal distribution. There is more area in the tails of the  $t$ -distribution, and the standard normal curve is higher in the middle, i.e.,  $t$ -distribution has a greater spread than normal distribution.
- (v) The larger the number of degrees of freedom the more closely  $t$ -distribution resembles standard normal distribution, i.e.,  $t$ -curve is higher in the middle, i.e.,  $t$ -distribution resembles normal curve as the number of degrees of freedom approaches without limit.
- (vi) Sampling distribution of  $t$  does not depend on population parameter but it depends only  $v$  (the degree of freedom) =  $n - 1$ , i.e., on the sample size.

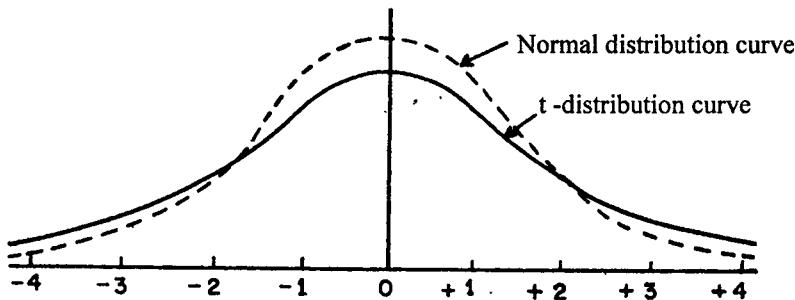


Fig. 13.1

### 13.5 APPLICATION OF t-DISTRIBUTION

The  $t$ -distribution has following three important applications in testing of hypothesis for small sample ( $n < 30$ ).

1. To test the significance of a single mean, when the population variance  $\sigma$  is unknown.
2. To test the significance of difference between two sample means the population variances being equal and unknown.
3. To test the significance of an observed sample correlation coefficient or difference between means of two sample (dependent samples or paired observations).

### 13.6 INTERVAL ESTIMATE OF POPULATION MEAN

Let  $\bar{x}$  be the sample mean, and  $n$  be the size of the sample. Then the interval estimate of the population mean ( $\mu$ ) is given by

$$\bar{x} \pm t_{\alpha} \text{S.E.}(\bar{x})$$

where S.E. ( $\bar{x}$ ) is standard error of mean and is defined as  $\frac{s}{\sqrt{n-1}}$  and  $s^2 = \frac{\sum(x - \bar{x})^2}{n}$ .

The value of  $t$  at different level of  $100(1-\alpha)\%$  are to be obtained from the  $t$ -table given at the end of the book. We shall see the column under 0.05, 0.02, 0.01 for confidence intervals of 95%, 98% and 99% respectively for the given degrees of freedom.

**Example 1 :** A sample of size 9 from a normal population gave  $\bar{x} = 15.8$  and  $s_x^2 = 10.3$ . Find a 99% interval for population mean.

**Solution :** We are given  $\bar{x} = 15.8$ ;  $s_x^2 = 10.3$  and  $n = 9$ .

$$\therefore \text{Degree of freedom} = n - 1 = 9 - 1 = 8.$$

Also  $t_{0.01}$  for 8 d.f. = 3.36. (from table)

99% confidence limits for the population mean  $\bar{x}$  are

$$\begin{aligned}\bar{x} \pm t_{0.01} \frac{s}{\sqrt{n-1}} &= 15.8 \pm 3.36 \times \sqrt{\frac{10.3}{8}} \\ &= 15.8 \pm 3.36 \times 1.135 = 15.8 \pm 3.8136 = 11.9864, 19.6136.\end{aligned}$$

Hence 99% confidence interval = [11.9864, 19.6136]

### 13.7 DETERMINATION OF SAMPLE SIZE

The determination of sample size for estimating a mean or proportion is a crucial question. By selecting a sample size lower than the correct size may affect reliability and a higher one will mean more cost and time. The determination of the size of a sample is the most important factor for the purpose of estimation of the value of population parameters.

#### Sample Size for Estimating Mean

In order to determine the sample size for estimating a population mean, the following factors must be known :

- (i) the desired confidence level
- (ii) the permissible sampling error  $E = \bar{x} - \mu$
- (iii) the standard deviation  $\sigma$ .

After having known the above mentioned three factors, the size of sample mean  $n$  is given by

$$\text{Size of sample: } n = \left( \frac{\sigma Z}{E} \right)^2$$

#### Sample Size for Estimating a Proportion

In this case we must know the following three factors:

- (i) the desired confidence level.

- (ii) the permissible sampling  $E$  = difference between the estimate from the sample  $p$  and the parameter  $P$  to be estimated =  $P - p$ .  
 (iii) the estimated true proportion of success.

The sample size  $n$  is given by

**Sample size:** 
$$n = \frac{Z^2 pq}{E^2}, \text{ where } q = 1 - p.$$

**Example 2 :** It is known that the population standard deviation in waiting time for L.P.G. gas cylinder in Delhi is 15 days. How large a sample should be chosen to be 95% confident, the waiting time is within 7 days of true average.

**Solution :** The required sample size is

$$n = \left( \frac{\sigma Z}{E} \right)^2 = \left( \frac{15 \times 1.96}{7} \right)^2 = 17.64.$$

Hence the size of the sample is 18.

**Example 3 :** A manufacturing concern wants to estimate the average amount of purchase of its product in a month by the customers whose standard error is Rs. 10. Find the sample size if the maximum error is not to exceed Rs. 3 with a probability of 0.99.

**Solution :** It is given that

$$P[|\bar{x} - \mu| < 3] = 0.99. \quad \dots \quad (1)$$

But  $P\left[|\bar{x} - \mu| < 2.58 \frac{\sigma}{\sqrt{n}}\right] = 0.99. \quad \dots \quad (2)$

From (1) and (2), we have

$$2.58 \frac{\sigma}{\sqrt{n}} = 3 \Rightarrow \sqrt{n} = \frac{2.58 \times 10}{3}$$

or  $n = \left( \frac{2.58 \times 10}{3} \right)^2 = (8.6)^2 = 73.96 \approx 74.$

Hence the sample size should be 74.

**Example 4 :** Mr. X wants to determine on the basis of sample study the mean time required to complete a certain job so that he may be 95% confident that the mean may remain with  $\pm 2$  days of the true mean. As per the available records the population variance is 64 days. How large should the sample be for his study?

**Solution :** Here  $s = \sqrt{64} = 8$ . Also  $Z$  is  $N(0, 1)$ .

It is given that

$$P[|\bar{x} - \mu| < 2] = 0.95. \quad \dots \quad (1)$$

But  $P\left[(\bar{x} - \mu) < 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95. \quad \dots \quad (2)$

From (1) and (2), we get

$$\frac{1.96 \times \sigma}{\sqrt{n}} = 2 \Rightarrow n = \left( \frac{1.96 \times 8}{2} \right)^2 = (7.84)^2 = 61.42 \approx 62.$$

Hence the sample size should be 62.

### 13.8 SMALL (OR EXACT) SAMPLE TESTS

In this section we shall discuss the tests of significance for small samples ( $n < 30$ ). We know that the mean of small conforms to the  $t$ -values contained in the  $t$ -distribution discussed in the last section. In order to

(i) Test the significance of a mean (For Small Samples).

(ii) Test the significance of Difference between two means. (For small samples).

we shall use  $t$ -statistic. *The t-values for different degrees of freedom and for different values of the level of significance are given in the table "Value of t for given probability levels", at the end of the book.* The nature of the  $t$ -statistic used for the different tests of significance is as follows:

### 13.9 COMPUTATION OF TEST STATISTIC : $t$ -VALUES

(a) To test the significance of a mean – Small Samples :

$$\text{Standard Error Mean : S.E. } (\bar{x}) = \frac{s}{\sqrt{n-1}}$$

where  $s$  = sample standard deviation and  $n$  = size of sample.

$$\text{Test Statistic : } t = \frac{\bar{x} - \mu}{\text{S.E. } (\bar{x})}.$$

(b) To test the significance of difference between two means – small samples :

The steps that are taken in testing the significance of two means or difference between two means in the case of large samples should also be taken in testing significance of two mean or difference between two means in the case of small samples ( $n < 30$ ), except in respect of the application of test statistic, which are discussed above.

$$\text{Estimated standard deviation of population : } \sigma' = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

where  $s_1$  = Standard deviation of sample I.  $s_2$  = Standard deviation of sample II.

$$\text{Standard Error of Difference} = \text{S.E. } (\bar{x}_1 - \bar{x}_2) = \sigma' \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\text{Test Statistic : } t = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)},$$

where  $\bar{x}_1$  = Mean of sample I,  $n_1$  = Size of Sample I

$\bar{x}_2$  = Mean of sample II,  $n_2$  = Size of Sample II.

### 13.10 TEST OF SIGNIFICANCE OF A SINGLE MEAN – SMALL SAMPLES

#### Working Rule

The various steps that are taken in testing the significance of a single mean in case of small samples ( $n \leq 30$ ) are as under:

**1. Setting up of hypothesis :** These are two types of hypothesis.

(i) Null Hypothesis :  $H_0$

(ii) Alternative Hypothesis :  $H_1$

**2. Computation of test statistic**

(i) Calculate standard Error of mean S.E. ( $\bar{x}$ )

$$\text{S.E.} (\bar{x}) = \frac{S}{\sqrt{n}}, \quad \text{where } S^2 = \frac{1}{(n-1)} \sum (x - \bar{x})^2,$$

and

$$\bar{x} = \frac{\Sigma x}{n}, \quad (\text{when } \bar{x} \text{ is a whole number})$$

However if  $\bar{x}$  is a fraction, then the following formulae are used to calculate  $S^2$ .

$$(i) \quad S^2 = \frac{1}{n-1} \left[ \sum x^2 - \frac{(\Sigma x)^2}{n} \right]$$

$$(ii) \quad S^2 = \frac{1}{n-1} \left[ \sum d^2 - \frac{(\Sigma d)^2}{n} \right], \quad \text{where } d = x - A.$$

In this case  $\bar{x} = A + \frac{\Sigma d}{n}$ ; where  $A$  = assumed mean.

**3. Level of significance :** Let us take  $\alpha = 0.05$ , if  $\alpha$  is not given in the question.

**4. Degree of freedom  $x$  :** Degrees of freedom  $v = n - 1$ .

**5. Critical values :** Find from the 'tables of  $t$ -values' the value of  $t$  for given  $\alpha$  and d.f.  $v$ , i.e., find  $t_{\alpha, v}$ .

**6. Decision :**

(i) If the calculated value of  $|t|$  is less than tabled value  $t_{\alpha, v}$ , then accept the null hypothesis.

(ii) If the calculated value of  $|t|$  is greater than the tabled value  $t_{\alpha, v}$ , then reject the null hypothesis and accept the alternative hypothesis.

**Example 5 :** The pulse rate of a man due to the effect of Amtas AT 25 mg on different days in a month were found to be.

66, 65, 69, 70, 69, 71, 70, 63, 64 and 68.

Discuss whether the mean pulse rate of the man in the month is 65.

**Solution : 1. Null hypothesis :**  $H_0 : \mu = 65$

**Alternative hypothesis :**  $H_1 : \mu \neq 65$

[Two tailed test]

**2. Computation of Test Statistic**

Pulse rate ( $x$ )	66	65	69	70	69	71	70	63	64	68	Total
$d = x - 68$	-2	-3	1	2	1	3	1	-5	-4	0	$\Sigma d = -5$
$d^2 = (x - 68)^2$	4	9	1	4	1	9	1	25	16	0	$\Sigma d^2 = 73$

$$\therefore \bar{x} = A + \frac{\Sigma d}{N} = 68 + \frac{(-5)}{10} = 68 - 0.5 = 67.5 ; \text{ where } A = 68.$$

$$s^2 = \frac{1}{n-1} \left[ \Sigma d^2 - \frac{(\Sigma d)^2}{n} \right] = \frac{1}{9} \left[ 73 - \frac{25}{10} \right] = \frac{1}{9} [70.5] = 7.84.$$

**Standard Error :**  $S.E. (\bar{x}) = \frac{S}{\sqrt{n}} = \sqrt{\frac{7.84}{10}} = 0.89$

**Test Statistic :**  $t = \frac{x - \mu}{S.E. (\bar{x})} = \frac{67.5 - 65}{0.89} = \frac{2.5}{0.89} = 2.81.$

**3. Level of significance :** Let us take  $\alpha = 0.05$ .

**Degrees of freedom :**  $v = n - 1$  or  $v = (10 - 1) = 9$ .

**4. Critical value :** The value of  $t_{0.05}$  for 9 d.f. = 2.262.

**5. Decision :** Since the calculated value of  $|t| = 2.81$  which is greater than the tabulated value  $|t_{0.05, 9}| = 2.262$ , so the null hypothesis is rejected. Hence we conclude that the mean pulse rate of the man cannot be regarded as 65.

**Example 6 :** A random blood sample for the test of fasting sugar of 10 boys give the following data in mg/dl.

70, 120, 110, 101, 88, 83, 95, 107, 100, 98

Do these data support the assumption of population mean of 100 mg/dl? Find a reasonable range in which the most of the mean fasting sugar test of samples of 10 boys lie.

**Solution : Null Hypothesis :**  $H_0 : \mu = 100$ , i.e., the data support the assumption of a mean of 100 mg/dl in the population.

**Alternative hypothesis:**  $H_1 : \mu \neq 100$  [Two tailed test]

Let  $x$  denote the sugar level and let  $d = x - A = x - 90$ , where  $A = 90$ .

Then, we have

$$\Sigma d = (x - 90) = -20 + 30 + 20 + 11 + -2 + -7 + 17 + 10 + 8 = 72$$

$$\Sigma d^2 = (-20)^2 + (30)^2 + (20)^2 + (11)^2 + (-2)^2 + (-7)^2 + (17)^2 + (10)^2 + (8)^2 = 2352.$$

$$\therefore \bar{x} = A + \frac{\Sigma d}{n} = 90 + \frac{72}{10} = 97.2.$$

$$S^2 = \frac{1}{n-1} \left[ \sum d^2 - \frac{(\sum d)^2}{n} \right] = \frac{1}{9} \left[ 2352 - \frac{(72)^2}{10} \right] = \frac{1833.60}{9} = 203.74\%.$$

$$\text{S.E. } (\bar{x}) = \sqrt{\frac{S^2}{n}} = \frac{\sqrt{203.74}}{\sqrt{10}} = \sqrt{20.374}$$

$$\text{Test Statistic } t = \frac{\bar{x} - \mu}{\text{S.E. } (\bar{x})} = \frac{\bar{x} - \mu}{\sqrt{(S^2/n)}}$$

$$\therefore t = \frac{97.2 - 100}{\sqrt{(203.74/10)}} = \frac{-2.8}{\sqrt{20.374}} = \frac{-2.8}{4.51} = -0.621.$$

$$\therefore |t| = |-0.62| = 0.621$$

**Level of significance :** Let us take  $\alpha = 0.05$ .

**Degrees of Freedom :** The degrees of freedom :  $v = n - 1 = 10 - 1 = 9$ .

**Critical values :** The tabled values of  $t_{0.05}$  for 9 d.f. = 2.262.

**Decision :** Since calculated  $|t|$  is less than tabulated value  $t_{0.05, 9}$ , it is significant. Hence  $H_0$  may be accepted at 5% level of significance and we may conclude that the data support the assumption of mean sugar level of 100 mg/dl in the population.

**95% confidence limits for  $\mu$  :** 95% confidence limits within which the sugar level of the sample of 10 boys will lie are given by:

$$\bar{x} \pm t_{0.05} \frac{S}{\sqrt{n}} = 97.2 \pm 2.262 \times 4.51 \quad \left[ \because \frac{S}{\sqrt{n}} = \sqrt{20.374} = 4.51 \right]$$

$$= 97.2 \pm 10.20 = 87.0 \text{ and } 120.40.$$

$\therefore$  95% confidence interval for  $\mu$  is [87.00, 107.40]

i.e.,  $87.0 \leq \mu \leq 107.40$ .

**Example 7 :** A fertiliser mixing machine is set to give 12 kg of nitrate for every quintal bag of fertiliser. Ten 100 kg bags are examined. The percentage of nitrate are as follows:

11, 14, 13, 12, 13, 12, 13, 14, 11, 12

Is there reason to believe that the machine is defective?

[Value of  $t$  for 9 degrees of freedom is 2.262].

**Solution : Null hypothesis :**  $H_0 : \mu = 12 \text{ (kg)}$

**Alternative hypothesis :**  $H_1 : \mu \neq 12$  [Two tailed test]

Let  $x$  : Percentage of nitrate in 100 kg bag of fertilizer (in kgs)

Let  $d = x - A = x - 12$ ; where  $A = 12$ .

$$\therefore \sum d = (-1) + 2 + 1 + 0 + 1 + 0 + 1 + 2 + (-1) + 0 = 5$$

$$\sum d^2 = (-1)^2 + 2^2 + 1^2 + 0 + 1^2 + 0 + 1^2 + 2^2 + (-1)^2 + 0 = 13.$$

$$\bar{x} = A + \frac{\Sigma d}{n} = 12 + \frac{5}{10} = 12.5 \text{ kg}$$

$$S^2 = \frac{1}{n-1} \left[ \Sigma d^2 - \frac{(\Sigma d)^2}{n} \right] = \frac{1}{9} \left( 13 - \frac{25}{10} \right) = \frac{10.5}{9} = 1.1667$$

**Standard Error :** S.E. ( $\bar{x}$ ) =  $\sqrt{(S^2/n)} = \sqrt{(1.1667)/10} = 0.3416$

**Test Statistic :**  $t = \frac{\bar{x} - \mu}{\sqrt{(S^2/n)}} = \frac{12.5 - 12}{0.3416} = \frac{0.5}{0.3416} = 1.4637$

**Level of significance :** Take  $\alpha = 0.05$ .

**Degree of freedom :** d.f. =  $n - 1 = 10 - 1 = 9$ .

**Critical value :** Tabled value of  $t_{0.05}$  for 9 d.f. = 2.262.

**Decision :** Since the calculated value of  $t = 1.4637$  is less than the tabulated value of  $t_{0.05}$  for 9 d.f. = 2.262, so the null hypothesis is accepted. Hence we conclude that  $\mu = 12 \text{ kg}$  and there is no reason to believe that the machine is defective.

**Example 8 :** A sample of 10 television tubes produced by a company showed a mean life time of 1200 hours and a standard deviation of 100 hours. Estimate:

(i) The mean, and

(ii) The standard deviation.

Of the population of all television tubes produced by this company.

Are the estimates obtained by you unbiased?

**Solution :** We are given:

$$n = 10; \bar{x} = 1200 \text{ (hours)}; s = 100 \text{ (hours)}$$

(i) An unbiased estimate of the population mean  $\mu$  is provided by the sample mean. Thus

$$\hat{\mu} = \bar{x} = 1200 \text{ (hours)}$$

(ii) The sample variance  $S^2$ , as an estimate of the population variance  $\sigma^2$  is not unbiased. An unbiased estimate of the population variance  $\sigma^2$  is provided by  $S^2$  given by:

$$\hat{\sigma}^2 = S^2 = \frac{ns^2}{n-1} = \frac{10 \times 10000}{9} = (1.1111) \times (100)^2$$

$$\Rightarrow \hat{\sigma}^2 = \sqrt{1.1111} \times 100 = 105.41 \text{ hours}$$

Thus the estimates obtained are unbiased.

**Example 9 :** A soap manufacturing company was distributing a particular brand of soap through a large number of retail shops. Before a heavy advertisement campaign, the mean sales per shop was 140 dozens. After the campaign a sample of 26 shops was taken and the mean sales figure was found to be 147 dozens with standard 16. Can you consider the advertisement effective?

**Solution :** Here  $n = 26 \Rightarrow$  It is a case of small samples.

Also  $\bar{x} = 147$ ,  $s = 16$ ,  $v = \text{degree of freedom} = 26 - 1 = 25$ .

**1. Null hypothesis :** There is no significant difference between the sample means and population means  $\Rightarrow$  the advertisement is not effective, i.e.,  $H_0 : \mu = 140$ .

**Alternative hypothesis :**  $H_1 : \mu > 140 \Rightarrow$  Single tailed test  
 $\Rightarrow$  the advertisement is effective.

**2. Computation of Test Statistic :**

**Standard Error of mean :**

$$\text{S.E.}(\bar{x}) = \frac{s}{\sqrt{n-1}} = \frac{16}{\sqrt{(26-1)}} = \frac{16}{5} = 3.2$$

$$\text{Test statistic : } t = \frac{\bar{x} - \mu}{\text{S.E.}(\bar{x})} = \frac{147 - 140}{3.2} = \frac{7}{3.2} = 2.19.$$

**3. Level of significance :** Let us take  $\alpha = 0.05$ .

**4. Critical value :** The value of  $t$  for  $\alpha = 0.05$  and  $v = 25$  is  $t_{0.05, 25} = 2.06$  (for single tailed test).

**5. Decision :** The calculated value of  $t = 2.19 > t_{0.05, 25} = 2.06$

$\Rightarrow$  the null hypothesis  $H_0$  is rejected and the alternative hypothesis is  $H_1$  is accepted  
 $\Rightarrow$  the advertisement is effective.

**Example 10 :** A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025 cm. A random sample of 10 washers was found to have an average thickness of 0.024 cm with a standard deviation of 0.002 cm. Test the significance of the deviation of mean.  
[Value of  $t$  for 9 degrees of freedom at 5% level is 2.262.]

**Solution :** Here we are given  $n = 10 \Rightarrow$  Small sample,  $\bar{x} = 0.024$  cm,  $s = 0.002$  cm.

**1. Null hypothesis :**  $H_0 : \mu = 0.025$  cm, i.e., there is no significant deviation between the sample mean and population mean.

**Alternative hypothesis :**  $H_1 : \mu \neq 0.025$  cm. (Two tailed test)

**2. Computation of Test Statistic :**

$$\text{Standard Error of Mean} = \text{S.E.}(\bar{x}) = \frac{s}{\sqrt{n-1}} = \frac{0.002}{\sqrt{9}} = \frac{0.002}{3}.$$

$$\text{Test Statistic } t = \frac{\bar{x} - \mu}{\text{S.E.}(\bar{x})} = \frac{0.024 - 0.025}{0.002/3} = -1.5 \Rightarrow |t| = 1.5.$$

**3. Level of significance :** Let  $\alpha = 0.05$ .

**4. Critical value :** Tabulated value of  $t_{0.05}$  for 9. d.f. = 2.262.

**5. Decision :** Since calculated value  $|t| (= 1.5) < t_{0.05, 9} (= 2.262) \Rightarrow$  Null hypothesis  $H_0$  is accepted  $\Rightarrow$  there is no significant deviation between the sample mean and population mean.

**Example 11 :** Ten objects are chosen at random from a large population and their weights are found to be in gms 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. In the light of the above data, discuss the suggestion that the mean weight in the universe is 65 gms.

**Solution :** Here  $n = 10 \Rightarrow$  a case of small samples,  $\mu = 65$ . Let us calculate the sample mean and sample standard deviation from the following table.

TABLE : Sample mean and S.D. assuming  $A = 66$

x	$d = x - 66$	$d^2$
63	-3	9
63	-3	9
64	-2	4
65	-1	1
66	0	0
69	3	9
69	3	9
70	4	16
70	4	16
71	5	25
$\Sigma d = 10$		$\Sigma d^2 = 98$

$$\therefore \bar{x} = A + \frac{\Sigma d}{n} = 66 + \frac{10}{10} = 67.$$

$$\begin{aligned} \text{Also } \hat{\sigma}^2 &= S^2 = \frac{1}{n-1} \left[ \Sigma d^2 - \frac{(\Sigma d)^2}{n} \right] \\ &= \frac{1}{9} \left[ 98 - \frac{100}{10} \right] = \frac{1}{9} [98 - 10] = \frac{88}{9} \\ &= 9.78 \Rightarrow S = 3.127. \end{aligned}$$

1. Null hypothesis :  $H_0 : \mu = 65$ , i.e., there is no difference between the mean weight of Sample and Universe.

Alternative hypothesis :  $H_1 : \mu \neq 65$  [Two tailed test]

2. Computation of Test statistic :

$$\text{Test Statistic : } t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$\therefore t = \frac{(67 - 65) \times 10}{3.127} = \frac{2 \times 3.162}{3.127} = \frac{6.33}{3.127} = 2.024.$$

3. Level of significance : Take  $\alpha = 0.05$ .

4. Critical value : The tabulated value of  $t$  at  $\alpha = 0.05$  for 9 d.f. is  $t_{0.05, 9} = 2.262$ .

**5. Decision :** Now  $|t| = < t_{0.05, 9}$  as  $2.024 < 2.262$ .

$\Rightarrow$  so the null hypothesis  $H_0$  is accepted.

$\Rightarrow$  the mean weight of the universe is 65 gms.

**Example 12 :** A salesman is expected to effect an average sales of Rs. 3500. A sample test revealed that a particular salesman had made the following sales. Rs. 3700; 3400; 2500; 5200; 3000 and 2000. Using 0.05 level of significance, conclude whether his work is below standard or not.

**Solution :** Mean of the sample is

$$\bar{x} = \frac{3700 + 3400 + 2500 + 5200 + 3000 + 2000}{6} = \frac{19800}{6} = 3,300.$$

TABLE : Computation of Sample S.D.

Sale ('00 Rs.)	$d = x - \bar{x} = x - 33$	$d^2 = (x - \bar{x})^2$
37	4	16
34	1	1
25	-8	64
52	19	361
30	-3	9
20	-13	169
	$\Sigma d = 0$	$\Sigma d^2 = 620$

The unbiased estimate of population standard deviation

$$\hat{\sigma} = S = \sqrt{\frac{(x - \bar{x})^2}{n-1}} = \sqrt{\frac{620}{5}} = 11.13.$$

**1. Null hypothesis :**  $H_0 : \mu \geq 35$ .

$H_1 : \mu < 35$ . (One tailed test)

**2. Computation of test Statistic :**

Test Statistic:  $t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$ .

$$\therefore t = \frac{33 - 35}{(11.13)/\sqrt{6}} = \frac{-2 \times 2.45}{11.13} = -0.44 \Rightarrow |t| = 0.44$$

**3. Level of significance :** Take  $\alpha = 0.05$ .

**4. Critical value :** The critical value or tabulated value of  $t$  at  $\alpha = 0.05$  for 5 d.f. =  $t_{0.05, 5} = 2.015$  for Left Tailed test.

**5. Decision :** Now calculated value of  $|t| = 0.44 <$  Tabulated  $t_{0.05, 5} = 2.015 \Rightarrow H_0$  is accepted  $\Rightarrow$  the salesman is up to standard.

**Example 13 :** The heights of ten children selected at random from a given locality had a mean 63.2 cms and variance 6.25 cms. Test at 5% level of significance the hypothesis that the children of the given locality are on the average less than 65 cms in all. Given the value of for 9 degrees of freedom is 1.83.

**Solution :** Here  $n = 10 \Rightarrow$  Small sample,  $\bar{x} = 63.2$  cms,  $s^2 = 6.25$  cms  $\Rightarrow s = 2.5$  cms,  $\mu = 65$  cms.

1. Null hypothesis : The average height of the children is 65 cms, i.e.,  $H_0 : \mu = 65$ .

Alternative hypothesis :  $H_1 : \mu < 65$  (One Tail Test).

2. Computation of Test Statistic :

Standard error or Mean :

$$\text{S.E.}(\bar{x}) = \frac{s}{\sqrt{n-1}} = \frac{2.5}{\sqrt{9}} = \frac{2.5}{3} = 0.834.$$

$$\therefore \text{Test Statistic } t = \frac{\bar{x} - \mu}{\text{S.E.}(\bar{x})} = \frac{63.2 - 65}{0.834} = \frac{-1.8}{0.824} = -2.158.$$

$$\therefore |t| = 2.158.$$

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

4. Critical value : The tabulated value or critical value of  $t$  for 9. d.f. at  $\alpha = 0.05$  for one tail test is  $t_{0.05, 9} = 1.83$ .

5. Decision : Since  $|t| = 2.158 > t_{0.05, 9} = 1.83 \Rightarrow$  the null hypothesis  $H_0$  is rejected  
 $\Rightarrow$  Alternative hypothesis  $H_1$  is accepted  $\Rightarrow$  the mean height of the children is less than 65 cms.

**Example 14 :** A random sample of size 16 has 53 mean. The sum of the squares of the deviations taken from the mean is 150. Can this sample be regarded as taken from the population having 56 as mean? Obtain 95% and 99% confidence limits of the mean population. (For  $v = 15$ ,  $t_{0.01} = 2.95$ ,  $t_{0.05} = 2.131$ ).

**Solution :** We are given that  $n = 16 \Rightarrow$  Small sample, Sample mean  $\bar{x} = 53$ ;  $\sum (x - \bar{x})^2 = 150$ ,  $\mu = 56$ .

$$\therefore S^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{150}{16-1} = \frac{150}{15} = 10$$

$$\Rightarrow S = \sqrt{10} = 3.1623.$$

1. Null hypothesis :  $H_0 : \mu = 56$ .

After hypothesis :  $H_1 : \mu \neq 56$ . (Two tailed test)

$$2. \text{ Test statistic : } t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{53 - 56}{3.162} \times 16 = -\frac{12}{3.162} = -3.79.$$

$$\therefore |t| = 3.79$$

3. **Level of significance :** (i)  $\alpha = 0.05$ , (ii)  $\alpha = 0.01$ .
4. **Critical value :** (i) The tabulated value of  $t$  for  $\alpha = 0.05$  and 15 d.f. for two tailed test =  $t_{0.05, 15} = 2.131$ . (ii) Also  $t_{0.01, 15} = 2.95$ .
5. **Decisions :** (i) Since  $|t| = 3.79 > t_{0.05, 15} = 2.131 \Rightarrow$  the null hypothesis  $H_0$  is rejected and we conclude that the population mean is not 56 at 5% level of significance. (ii) Also  $|t| = 3.79 > t_{0.01, 15} = 2.95 \Rightarrow H_0$  is rejected at 1% level of significance  $\Rightarrow$  Population mean is not 56 at 1% level of significance.

**95% Confidence Limits for  $\mu$  (d.f. = 15) are**

$$\bar{x} \pm t_{0.05} \times \frac{S}{\sqrt{n}} = 53 \pm 2.131 \times \frac{3.162}{4} = 53 \pm 1.684$$

$$\Rightarrow 50.316 < \mu < 54.684.$$

**99% Confidence Limit for  $\mu$ .**

$$\bar{x} \pm t_{0.05} \times \frac{S}{\sqrt{n}} = 53 \times \frac{3.16}{4} \times 2.95 = 53 \pm 2.33$$

or

$$50.67 < \mu < 55.33.$$

**Example 15 :** A random sample of size 7 from a normal population gave a mean of 977.51 and a standard deviation of 4.42. Find 95% confidence interval for the population mean.

**Solution :** It is given to us that  $n = 7$ ,  $\bar{x} = 977.51$  and  $s = 4.42$ .

**95% confidence limits for  $\mu$  for 6 d.f.**

It is given by

$$\begin{aligned} \bar{x} \pm t_{0.05, 6} \times \text{S.E. of mean} &= \bar{x} \pm t_{0.05, 6} \frac{s}{\sqrt{n-1}} \\ &= 977.51 \pm 2.447 \times \frac{4.42}{\sqrt{6}} = 977.51 \pm \frac{2.447 \times 4.42}{2.449} \quad [\because t_{0.05, 6} = 2.447] \\ &= 977.51 \pm 4.416, \quad \text{i.e., } 973.094 \text{ and } 981.926 \\ \Rightarrow 973.094 < \mu < 981.926. \end{aligned}$$

### 13.11 TEST OF SIGNIFICANCE OF DIFFERENCE BETWEEN TWO – MEANS SMALL SAMPLES

**Example 16 :** The increase in weight in kgs. in particular period of 10 students of a certain age group of a High School, fed with the nourishing food "COMPLAN" were observed as:

$$5, 2, 6, -1, 0, 4, 3, -2, 1, 4$$

Twelve students of the same age-group, but of another High School, were fed with another nourishing food "ASTRA" and the increase in weight in kgs, in the same period were observed as:

$$2, 8, -1, 5, 3, 0, 6, 1, -2, 0, 4, 5$$

Test whether the two foods "COMPLAN" and "ASTRA" differ significantly as regards the effect on the increase in weight.

**Solution :** Let  $x$  and  $y$  respectively denote the increase in weights (in kgs.) of the students fed on the food 'COMPLAN' and 'ASTRA'.

**Null hypothesis :**  $H_0 : \mu_x = \mu_y$ , i.e., there is no significant difference between the foods 'COMPLAN' and 'ASTRA' as regards their effect on increase in weight.

**Alternative hypothesis :**  $H_1 : \mu_x \neq \mu_y$  [Two tailed test]

**Calculations for test statistic  $t$ :**

$$n_1 = 10, \quad \Sigma x = 22, \quad \Sigma x^2 = 112;$$

$$\bar{x} = \frac{\Sigma x}{n_1} = \frac{22}{10} = 2.2$$

$$\begin{aligned} n_1 s_x^2 &= \Sigma (x - \bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n_1} \\ &= 112 - \frac{(22)^2}{10} = 112 - 48.4 \\ &= 63.6. \end{aligned}$$

$$n_2 = 12, \quad \Sigma y = 31, \quad \Sigma y^2 = 185$$

$$\bar{y} = \frac{\Sigma y}{n_2} = \frac{31}{12} = 2.5833$$

$$\begin{aligned} n_2 s_y^2 &= \Sigma (y - \bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n_2} \\ &= 185 - \frac{(31)^2}{12} = 185 - 80.08 \\ &= 104.92. \end{aligned}$$

$$\therefore S^2 = \frac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2} = \frac{63.6 + 104.92}{10 + 12 - 2} = \frac{168.52}{20} = 8.426.$$

**Standard error :**  $S.E. (\bar{x} - \bar{y}) = \sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$

**Test statistic :**

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\begin{aligned} \therefore t &= \frac{2.20 - 2.583}{\sqrt{8.426 \left( \frac{1}{10} + \frac{1}{12} \right)}} = \frac{-0.383}{\sqrt{8.426 (0.10 + 0.08)}} \\ &= \frac{-0.383}{\sqrt{1.51668}} = \frac{-0.383}{1.2315} = -0.3110. \quad \therefore |t| = 0.3110. \end{aligned}$$

**Level of significance :** Let  $\alpha = 0.05$ .

**Also Degrees of freedom :** d.f. =  $(n_1 + n_2 - 2) = 12 + 10 - 2 = 20$ .

**Critical value :** The value of  $t_{0.05}$  for 20, d.f. = 2.09.

**Decision :** Since the value of calculated  $|t| = (0.311)$  is less than tabulated,  $|t_{0.05, 20}|$ ; so we accept the null hypothesis. Hence we conclude foods there is no significant difference between the effect of two foods.

**Example 17 :** A group of seven week-old chickens reared on a high protein diet weigh 12, 15, 11, 16, 14, 14 and 16 ounce. A second group of five chickens similarly treated except that they

receive a low protein diet weigh 8, 10, 14, 10 and 13 ounces. Test whether there is sufficient evidence that additional protein has increased the weight of the chickens.

(The table value of  $t$  for  $v = 10$  at 5% level of significance is 2.33).

**Solution :** Let the weights (in ounces) of the chickens reared on high protein and low protein diets be denoted by the variables  $X$  and  $Y$  respectively.

**Null hypothesis :**  $H_0 : \mu_x = \mu_y$

**Alternative hypothesis :**  $H_1 : \mu_x > \mu_y$  (Right-tailed), i.e., the additional protein increases the weight of the chickens.

TABLE : Computation of Sample S.D.

$X_1$	$(X_1 - \bar{X}_1)$	$(X_1 - \bar{X}_1)^2$	$X_2$	$(X_2 - \bar{X}_2)$	$(X_2 - \bar{X}_2)^2$
12	-2	4	8	-3	9
15	+1	1	10	-1	1
11	-3	9	14	+3	9
16	+2	4	10	-1	1
14	0	0	13	+2	4
14	0	0			
16	+2	4			
$\Sigma X_1$ = 98	$\Sigma(X_1 - \bar{X}_1)$ = 0	$\Sigma(X_1 - \bar{X}_1)^2$ = 22	$\Sigma X_2$ = 55	$\Sigma(X_2 - \bar{X}_2)$ = 0	$\Sigma(X_2 - \bar{X}_2)^2$ = 24

$$\bar{X}_1 = \frac{98}{7} = 14 ; \quad \bar{X}_2 = \frac{55}{5} = 11.$$

$$S^2 = \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{22 + 24}{7 + 5 - 2} = \frac{46}{10} = 4.6.$$

$$\text{S.E. } (\bar{X}_1 - \bar{X}_2) = \sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{4.6 \left( \frac{1}{7} + \frac{1}{5} \right)} = \sqrt{1.5771}.$$

$$\text{Test statistic : } t = \frac{\bar{X}_1 - \bar{X}_2}{\text{S.E. } (\bar{X}_1 - \bar{X}_2)} = \frac{14 - 11}{\sqrt{1.5771}} = \frac{3}{1.2558} = 2.389$$

**Level of significance :** Take  $\alpha = 0.05$ .

**Degree of freedom :** d.f. =  $(n_1 + n_2 - 2) = 7 + 5 - 2 = 10$ .

**Critical value :** Tabled value of  $t_{0.05}$  for 10 d.f. = 1.81 for single tailed test.

**Decision :** Since calculated value of  $t$  is greater than tabulated  $t_{0.05, 10}$ , it is significant at 5% level of significance. Hence, we reject  $H_0$  at 5% level and conclude that the data provide a definite evidence to support the hypothesis  $H_1$  that the additional protein has increased the weight of the chickens.

**Example 18 :** The average number of articles produced by two machines per day are 200 and 250 with standard deviations 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 1% of significance?

**Solution :** In the usual notations we are given

$$n_1 = 25, \bar{x}_1 = 200, s_1 = 20$$

$$n_2 = 25, \bar{x}_2 = 250, s_2 = 25$$

Let  $S^2$  be the unbiased estimate of common population variance based on both samples. Then,

$$\begin{aligned} S^2 &= \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \\ &= \frac{25 \times (20)^2 + 25 \times (25)^2}{25 + 25 - 2} = \frac{10000 + 15625}{48} = \frac{25625}{48} = 533.85 \end{aligned}$$

$$\Rightarrow S = \sqrt{533.85} = 23.1.$$

1. Null hypothesis :  $H_0 : \mu_1 = \mu_2$ , i.e., both the machines are equally efficient.

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test)

2. Computation of test statistic :

Standard Error of Difference

$$\begin{aligned} \text{S.E. } (\bar{x}_1 - \bar{x}_2) &= S \times \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = 23.1 \times \sqrt{\frac{1}{25} + \frac{1}{25}} \\ &= 23.1 \times \sqrt{\frac{2}{25}} = 23.1 \times \sqrt{0.08} = 23.1 \times 0.28 = 6.468. \end{aligned}$$

$$\therefore \text{Test Statistic } t = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} = \frac{200 - 250}{6.468} = \frac{-50}{6.468} = -7.7.$$

$$\therefore |t| = |-7.7| = 7.7.$$

3. Level of significance : Here  $\alpha = 1\% = 0.01$ .

4. Critical value : The value of  $t$  at  $\alpha = 0.01$  and for  $(25 + 25 - 2) = 48$  degrees of freedom for two tailed test is  $t_{0.01, 48} = 2.58$ .

5. Decision : The calculated value of  $|t| = 7.7 > t_{0.01, 48} = 2.58 \Rightarrow$  Null hypothesis is rejected  $\Rightarrow$  that the two machines are not equally efficient at 1% level of significance.

**Example 19 :** Two salesman A and B are working in a certain district from a sample survey conducted by the Head Office, the following results were obtained. State whether there is any significant difference in the average sales between the two salesman?

	<i>A</i>	<i>B</i>
<i>No. of sales</i>	20	18
<i>Average sales (in Rs.)</i>	170	205
<i>Standard deviation (in Rs.)</i>	20	25

**Solution :** In the usual notations we are given

$$\begin{aligned} n_1 &= 20, \quad \bar{x}_1 = 170, \quad s_1 = 20; \\ n_2 &= 18, \quad \bar{x}_2 = 205, \quad s_2 = 25; \end{aligned}$$

- 1. Null hypothesis :**  $H_0 : \mu_1 = \mu_2$ , i.e., there is no significant difference in the average sales between the two salesmen.

**Alternative hypothesis :**  $H_1 : \mu_1 \neq \mu_2$       (Two Tailed Test)

- 2. Computation of Test Statistic:**

Estimation standard deviation.

$$\begin{aligned} S &= \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{20(20)^2 + 18(25)^2}{20 + 18 - 2}} = \sqrt{\frac{8000 + 11250}{36}} \\ &= \sqrt{\frac{19250}{36}} = \frac{138.71}{6} = 23.12. \end{aligned}$$

#### Standard Error of Difference

$$\begin{aligned} \text{S.E. } (\bar{x}_1 - \bar{x}_2) &= S \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 23.12 \times \left[ \sqrt{\frac{1}{20} + \frac{1}{18}} \right] = 23.12 \times \sqrt{0.1060} \\ &= 23.12 \times 0.32 = 7.4. \end{aligned}$$

$$\therefore \text{Test statistic } t = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} = \frac{170 - 205}{7.4} = \frac{-35}{7.4} = -4.73.$$

$$\therefore |t| = 4.73.$$

- 4. Critical value :** The tabulated or critical value of  $t$  at  $\alpha = 0.05$  for  $(n_1 + n_2 - 2) = 20 + 18 - 2 = 36$  degrees of freedom for two tailed test is  $t_{0.05, 36} = 1.96$ .
- 5. Decision :** Since the calculated value of  $|t| = 4.73 >$  tabulated value  $t_{0.05, 36} = 1.96 \Rightarrow$  the Null hypothesis is rejected.  $\Rightarrow$  there is a significant difference in the average sales between the two salesmen.

**Example 20 :** Following table contains the data resulting from a sample of managers trained under different programs.

<i>Program Sampled</i>	<i>Mean sensitivity of the program</i>	<i>Number of Managers observed</i>	<i>Estimated s.d. for sensitivity after the program</i>
<i>Formal</i>	92%	12	15%
<i>Informal</i>	84%	15	19%

Test at 0.05 level of significance whether the sensitivity achieved by the formal program is significantly higher than that achieved under the informal program.

**Solution :** In the usual notations, we are given,

$$\begin{aligned} n_1 &= 12, & \bar{x}_1 &= 92\%, & s_1 &= 15\% \\ n_2 &= 15, & \bar{x}_2 &= 84\%, & s_2 &= 19\% \end{aligned}$$

1. Null hypothesis :  $H_0 : \mu_1 = \mu_2$ , i.e., there is no significant difference between the formal and informal program.

Alternative hypothesis :  $H_1 : \mu_1 > \mu_2$  (One tailed test).

2. Calculation of test statistic :

$$\text{Estimated common standard deviation } S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

$$\sqrt{\frac{12 \times (15)^2 + 15 \times (19)^2}{12 + 15 - 2}} = \sqrt{\frac{2700 + 5415}{25}} = \sqrt{\frac{8115}{25}} = \sqrt{324.6} = 18.02.$$

$$\begin{aligned} \text{S.E. } (\bar{x}_1 - \bar{x}_2) &= S \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 18.02 \times \left( \sqrt{\frac{1}{12} + \frac{1}{15}} \right) = 18.02 \sqrt{\frac{27}{180}} \\ &= 18.02 \times 0.15 = 2.703. \end{aligned}$$

$$\therefore \text{Test statistic } t = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} = \frac{92 - 84}{2.703} = 2.96$$

$$\therefore |t| = 2.96.$$

3. Level of significance : Take  $\alpha = 0.05$ .

4. Critical value : The tabulated or critical value of  $t$  at  $\alpha = 0.05$  for  $12 + 15 - 2 = 25$  degrees of freedom for one tailed test is  $t_{0.05, 25} = 1.708$ .

5. Decision : Since the calculated value of  $|t| = 2.96 >$  tabulated value of  $t_{0.05, 25} = 1.708$   
 $\Rightarrow$  the null hypothesis  $H_0$  is rejected and the alternative hypothesis  $H_1$  is accepted  
 $\Rightarrow$  the sensitivity achieved by the formal program is higher than that achieved under the informal program.

**Example 21 :** The mean life of a sample of 10 electric bulbs was found to be 1456 hours with a standard deviation of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with standard deviation 398 hours. Is there significant difference between the means of the two batches?

**Solution :** In the usual notation, it is given that

$$\begin{aligned} n_1 &= 10, & \bar{x}_1 &= 1456, & s_1 &= 423 \\ n_2 &= 17, & \bar{x}_2 &= 1280, & s_2 &= 398 \end{aligned}$$

1. Null hypothesis :  $H_0 : \mu_1 = \mu_2$  i.e., there is no significant difference between the means of two samples.

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test).

2. Computation of test statistic :

Estimated standard deviation of population

$$\begin{aligned}s &= \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{10 \times (423)^2 + 17(398)^2}{10 + 17 - 2}} \\&= \sqrt{\frac{1789290 + 2692869}{25}} = \frac{2117.11}{5} = 423.42\end{aligned}$$

Standard error of difference

$$\begin{aligned}\text{S.E. } (\bar{x}_1 - \bar{x}_2) &= s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 423.42 \times \sqrt{\frac{1}{10} + \frac{1}{17}} \\&= 423.42 (\sqrt{0.1588}) = 423.42 \times 0.398 = 168.52.\end{aligned}$$

$$\therefore \text{Test Statistic : } t = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} = \frac{1456 - 1800}{168.52} = \frac{176}{168.52} = 1.04.$$

3. Level of significance : Take  $\alpha = 0.05$ .

4. Critical value : The tabulated or critical value of  $t$  at  $\alpha = 0.05$  for  $10 + 17 - 2 = 25$  degrees of freedom for two tailed test is  $t_{0.05, 25} = 2.06$ .

5. Decision : Since the calculated value of  $|t| = 1.04 < t_{0.05, 25} = 2.06 \Rightarrow$  the null hypothesis  $H_0$  is accepted  $\Rightarrow$  there is no significant difference between the means of two samples.

**Example 22 :** Two types of batteries are tested for their lengths of life and the following data are obtained.

	No. of Samples	Mean life	Variance
Type A	9	600 hours	121
Type B	8	640 hours	144

Is there significance difference in the two means? Value of  $t$  for 15 degrees of freedom at 5% level is 2.131.

**Solution :** In the usual notations we are given that

$$\begin{aligned}n_1 &= 9, & \bar{x}_1 &= 600, & s_1^2 &= 121 \\n_2 &= 8, & \bar{x}_2 &= 640, & s_2^2 &= 144.\end{aligned}$$

1. Null hypothesis :  $H_0 : \mu_1 = \mu_2$ , i.e., there is no significant difference between the two means.

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test).

2. Computation of test statistic :

Estimated standard deviation of population

$$\begin{aligned}s &= \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{9 \times 121 + 8 \times 144}{9 + 8 - 2}} \\&= \sqrt{\frac{2241}{15}} = \sqrt{149.4} = 12.22.\end{aligned}$$

Standard Error of Difference

$$\begin{aligned}\text{S.E. } (\bar{x}_1 - \bar{x}_2) &= s \times \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 12.22 \times \sqrt{\frac{1}{9} + \frac{1}{8}} = 12.22 \times \sqrt{0.236} \\&= 12.22 \times 0.486 = 5.94.\end{aligned}$$

$$\therefore \text{Test Statistic : } t = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} = \frac{600 - 640}{5.94} = \frac{-40}{5.94} = -6.73.$$

$$\therefore |t| = |-6.73| = 6.73.$$

3. Level of significance : Take  $\alpha = 0.05$ .

4. Critical value : Value of  $t$  at  $\alpha = 0.05$  for  $(n_1 + n_2 - 2) = 9 + 8 - 2 = 15$  degrees of freedom is  $t_{0.05, 15} = 2.131$ .

5. Decision : The calculated value of  $|t| = 6.73 >$  tabulated value  $t_{0.05, 15} = 2.131 \Rightarrow$  the null hypothesis  $H_0$  is rejected  $\Rightarrow$  there is no significant difference in the two means.

**Example 23 :** A group of five patients treated with medicine 'A' weigh : 42, 39, 48, 60 and 41 kgs. A second group of 7 patients from the same hospital treated with medicine 'B' weigh : 38, 42, 56, 64, 68, 69, 62 kgs. Do you agree with the claim the medicine 'B' increases the weight significantly. (The value of 't' at 5% level of significance for 10 degree of freedom is 2.228).

**Solution :** Let the variables  $x$  and  $y$  denote the weight (in kgs) of the patients treated with medicine A and B respectively.

1. Null hypothesis :  $H_0 : \mu_A = \mu_B$ , i.e., there is no significant difference between the medicines A and B as regards their effect on increase in weight.

Alternative hypothesis :  $H_1 : \mu_A \neq \mu_B$  (Two tailed test)

2. Computation of test statistic : We shall first compute the sample mean and standard deviation by means of the following table.

TABLE : Computation of Sample mean and S.D

## Medicine A

$x$	$d_1 = x - \bar{x} = x - 46$	$d_1^2$
42	-4	16
39	-7	49
48	2	4
60	14	196
41	-5	25
$\Sigma x = 230$	$\Sigma d_1 = 0$	$\Sigma d_1^2 = 290$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{230}{5} = 46.$$

## Medicine B

$y$	$d_2 = y - \bar{y} = y - 57$	$d_2^2$
38	-19	361
42	-15	225
56	-1	1
64	7	49
68	11	121
69	12	144
62	5	25
$\Sigma y = 399$	$\Sigma d_2 = 0$	$\Sigma d_2^2 = 926$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{399}{7} = 57.$$

## Estimated Standard Deviation of the Population :

$$s = \sqrt{\frac{\sum d_1^2 + \sum d_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{290 + 926}{5 + 7 - 2}} = \sqrt{\frac{1216}{10}} = \sqrt{121.6} = 11.03.$$

## Standard Error of Difference:

$$\begin{aligned} \text{S.E. } (\bar{x} - \bar{y}) &= s \times \sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 11.03 \times \sqrt{\frac{5 + 7}{35}} = 11.03 \times \sqrt{0.34} \\ &= 11.03 \times 0.58 = 6.4. \end{aligned}$$

$$\therefore \text{Test Statistic : } t = \frac{\bar{x} - \bar{y}}{\text{S.E. } (\bar{x} - \bar{y})} = \frac{46 - 57}{6.4} = -\frac{11}{6.4} = -1.72$$

$$\therefore |t| = 1.72.$$

3. Level of significance :  $\alpha = 0.05$ .
4. Critical value : The tabulated or critical value of  $t$  at  $\alpha = 0.05$  for  $(n_1 + n_2 - 2) = 5 + 7 - 2 = 10$  degrees of freedom for two tailed test is  $t_{0.05, 10} = 2.2281$ .
5. Decision : The calculated value of  $|t| = 1.72 = 1.72 <$  tabulated value of  $t_{0.05, 10} = 2.2281$   
 $\Rightarrow$  the null hypothesis  $H_0$  is accepted  $\Rightarrow$  There is no significant difference between the medicines A and B as regards the increase in the weight.

**Example 24 :** The following data shows the cost in hundred rupees per square meter of the floor area concerning randomly selected 7 schools and 5 office blocks from those completed during the period 2004 to 2005.

Schools	28	31	26	27	23	38	37
Office blocks :	37	42	34	37	35		

Do the data support the hypothesis that the cost per square metre for the office blocks was greater than that for the schools. Test at 5% level of significance.

**Solution :**

1. Null hypothesis : There is no difference in the costs, i.e.,  $H_0 : \mu_1 = \mu_2$

Alternative hypothesis :  $H_1 : \mu_1 > \mu_2$

( $\mu_1$  = average cost per school,  $\mu_2$  = average cost for office blocks).

2. Computation of Test Statistics :

Computation Table

S. No.	$X_1$	$X_1 - \bar{X}_1$ $X_1 - 30$	$(X_1 - \bar{X}_1)^2$	S. No.	$X_2$	$X_2 - \bar{X}_2$ $X_2 - 37$	$(X_2 - \bar{X}_2)^2$
1	28	-2	4	1	37	0	0
2	31	1	1	2	42	5	25
3	26	-4	16	3	34	-3	9
4	27	-3	9	4	37	0	0
5	23	-7	49	5	35	-2	4
6	38	8	64				
7	37	7	49				
	$\Sigma X_1 = 98$		$\Sigma (X_1 - \bar{X}_1)^2 = 192$		$\Sigma X_2 = 185$		$\Sigma (X_2 - \bar{X}_2)^2 = 38$

$$\bar{X}_1 = \frac{\sum X_1}{n} = \frac{210}{7} = 30 ; \quad \bar{X}_2 = \frac{185}{5} = 37$$

$$\text{S.E. } (\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}} \times \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$= \sqrt{\frac{192 + 38}{7 + 5 - 2}} \times \sqrt{\frac{1}{7} + \frac{1}{5}} = 4.796 \times 0.586 = 2.81.$$

$$\therefore \text{Test Statistic : } t = \frac{\bar{X}_1 - \bar{X}_2}{\text{S.E.}(\bar{X}_1 - \bar{X}_2)} = \frac{30 - 37}{2.81} = -2.491.$$

$$\therefore |t| = 2.491.$$

**3. Level of significance :**  $\alpha = 0.05$ .

**4. Critical value :** The critical value of  $t$  at  $\alpha = 0.05$  for 10 degrees of freedom for one tailed test is 1.812.

**5. Decision :** Since calculated value of  $|t| (= 2.491) >$  tabulated  $t_{0.05, 10} (= 1.812)$   $\Rightarrow$  the null hypothesis is rejected and so we accept Alternative Hypothesis  $H_1$  which concludes that the cost per square metre for office book is greater than for school block.

**Example 25 :** Six guinea pigs injected with 0.5 mg of a medication took on the average 15.4 seconds to fall asleep with an unbiased standard deviation 2.2 seconds while six other guinea pigs injected with 1.5 mg of the medication took on the average 11.2 seconds to fall asleep with an unbiased standard deviation of 2.6 seconds. Use the 5% level of significance to test the null hypothesis that the difference in dosage has no effect.

**Solution :**

**1. Null hypothesis :** Difference is dosage has no effect, i.e.,  $H_0 : \mu_1 = \mu_2$ .

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test)

**2. Calculation of test statistic :** It is given that

$$2.2 = s_1 \sqrt{\frac{n_1}{n_1 - 1}} = s_1 \sqrt{6/5} = s_1 \times 1.095$$

$$\Rightarrow s_1 = 2.2/1.095 = 2.009.$$

$$\text{Also, } 2.6 = s_2 \sqrt{\frac{n_2}{n_2 - 1}} = s_2 \sqrt{6/5} = s_2 \times 1.095$$

$$\Rightarrow s_2 = 2.6/1.095 = 2.374.$$

**Standard Error of Difference :**

$$\text{S.E.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2 + s_2^2}{n}} = \sqrt{\frac{(2.000)^2 + (2.374)^2}{6}} = 1.269.$$

$$\therefore \text{Test Statistic : } |t| = \frac{|\bar{x}_1 - \bar{x}_2|}{\text{S.E.}(\bar{x}_1 - \bar{x}_2)} = \frac{15.4 - 11.2}{1.269} = 3.30.$$

**3. Critical value :** The tabulated value of  $t$  at  $\alpha = 0.05$  for  $(n_1 + n_2 - 2) = 6 + 6 - 2 = 10$  degrees of freedom for a two tailed test is 2.228.

**4. Decision :** Now  $|t| = 3.30 > t_{0.05, 10} = 2.28 \Rightarrow$  Null hypothesis  $H_0$  is rejected

$\Rightarrow$  Difference in dosage has significant effect.

**Example 26 :** Samples of two types of electric bulbs were tested for length of life and the following data were obtained:

	Type I	Type II
Number in the sample	8	7
Mean of the sample (in hours)	1134	1024
Standard deviation of the sample	35	40.

**Solution :** In the usual notation, it is given that

$$\begin{aligned} n_1 &= 8, \quad \bar{x}_1 = 1134, \quad s_1 = 35 \\ n_2 &= 7, \quad \bar{x}_2 = 1024, \quad s_2 = 40. \end{aligned}$$

1. Null hypothesis :  $H_0 : \mu_1 = \mu_2$ .

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test).

2. Computation of test statistic : The unbiased estimate of the Common Population Standard Deviation is given by

$$\begin{aligned} S_p &= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{7 \times (35)^2 + 6 \times (40)^2}{8 + 7 - 2}} \\ &= \sqrt{\frac{7 \times 1225 + 6 \times 1600}{13}} = \sqrt{1398.08} = 37.39. \end{aligned}$$

Standard Error of Difference :

$$\text{S.E. } (\bar{x}_1 - \bar{x}_2) = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 37.39 \times \sqrt{\left(\frac{1}{8} + \frac{1}{7}\right)} = 19.368.$$

$$\therefore \text{Test Statistic: } t = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E. } (\bar{x}_1 - \bar{x}_2)} = \frac{1134 - 1024}{19.368} = 5.679.$$

4. Critical value : The tabulated or critical value of  $t$  at  $\alpha = 0.05$  for  $9 + 7 - 2 = 13$  degrees of freedom for two tailed test is  $t_{0.05, 13} = 2.161$ .

5. Decision : Since the calculated value  $|t| = 5.679 >$  tabulated value  $t_{0.05, 13} = 2.161$ , so the null hypothesis is rejected  $\Rightarrow$  the two types of electric bulbs differ significantly in their mean values.

**Example 27 :** Below are given the gain in weights in kgs of cows fed two diets X and Y:

Diet X :	25	32	30	32	24	14	32			
Diet Y :	24	34	22	30	42	31	40	30	32	35

Test at 5% level whether the two diets differ as regards their effect on mean increase in weight. (Tabulated value of 't' for 15 degrees of freedom at 5% = 2.131).

**Solution :**

1. Null hypothesis :  $H_0 : \mu_1 = \mu_2$ .

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test)

**Table**

$x$	$d_1 = x - \bar{x}$ $= x - 27$	$d_1^2 = (x - \bar{x})^2$	$y$	$d_2 = y - \bar{y}$ $= y - 32$	$d_2^2 = (y - \bar{y})^2$
25	-2	4	24	-8	64
32	5	25	34	2	4
30	3	9	22	-10	100
32	5	25	30	-2	4
24	-3	9	42	10	100
14	-13	169	31	-1	1
32	5	25	40	8	64
			30	-2	4
			32	0	0
			35	3	9
$\Sigma x = 189$	$\Sigma d_1 = 0$	$\Sigma d_1^2 = 266$	$\Sigma y = 320$	$\Sigma d_2 = 0$	$\Sigma d_2^2 = 350$

Now  $\bar{x} = \frac{\Sigma x}{n_1} = \frac{189}{7} = 27$ ;  $\bar{y} = \frac{\Sigma y}{n_2} = \frac{320}{10} = 32$ .

Sample variance for  $x$ :  $s_1^2 = \frac{\sum (x - \bar{x})^2}{n_1} = \frac{266}{7} = 38$ .

Sample variance for  $y$ :  $s_2^2 = \frac{\sum (y - \bar{y})^2}{n_2} = \frac{350}{10} = 35$ .

The unbiased estimate of common population s.d. is given by

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{6 \times 38 + 9 \times 35}{7 + 10 - 2}} = \sqrt{\frac{543}{15}} = \sqrt{36.2} = 6.017.$$

Standard Error Difference :

$$\text{S.E. } (\bar{x} - \bar{y}) = s_p \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = 6.017 \times \sqrt{\left( \frac{1}{7} + \frac{1}{10} \right)} = 6.017 \times 0.493 = 2.966$$

Test Statistic :  $t = \frac{\bar{x} - \bar{y}}{\text{S.E. } (\bar{x} - \bar{y})} = \frac{27 - 32}{2.966} = \frac{-5}{2.966} = -1.686$

$\therefore |t| = |-1.686| = 1.686$ .

2. **Critical value :** The tabulated value of  $t$  at  $\alpha = 0.05$  for 15 degrees of freedom for a two tailed test is  $t_{0.05, 15} = 2.131$ .
3. **Decision :** Since tabulated  $|t| = 1.686 < t_{0.05, 15} = 2.131$ , so the null hypothesis is accepted  $\Rightarrow$  The two diets do not differ significantly as regards their effect on mean increase in weights.

### 13.12 PAIRED t-TEST FOR DIFFERENCE OF MEANS (WHEN THE SAMPLE OBSERVATIONS ARE NOT COMPLETELY INDEPENDENT)

In this case we are given that

- (i) the sizes of the samples are equal, i.e.,  $n_1 = n_2 = n$  (say).
- (ii) the samples are independent, and the sample observations  $(x_1, x_2, x_3, \dots, x_n)$ ,  $(y_1, y_2, y_3, \dots, y_n)$  are dependent in pairs. In other words the pairs of observation  $(x_1, y_1)$ ,  $(x_2, y_2)$ ,  $(x_3, y_3)$ , ...,  $(x_n, y_n)$  correspond to 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, ...  $n^{\text{th}}$  unit respectively. Although we may find two sets of sample values corresponding to two observations  $(x_1, x_2, \dots, x_n)$  and  $(y_1, y_2, y_3, \dots, y_n)$  for the same elementary unit under different situations, the tests discussed earlier are not applicable and we proceed as follows.

#### Working Rule for Estimated Standard Error of Difference

**Step 1.** Let  $D = (x_i - y_i)$ , ( $i = 1, 2, \dots, n$ ) denote the difference in observation for the unit.

**Step 2.** Calculate  $\bar{D} = \frac{\sum D}{n}$ .

**Step 3.** Calculate  $\sum(D - \bar{D})^2$  and  $S = \sqrt{\frac{(\sum(D - \bar{D})^2)}{n-1}}$ .

$$\text{or } S = \sqrt{\frac{\sum D^2 - n(\bar{D})^2}{(n-1)}} = \sqrt{\frac{n \sum D^2 - (\sum D)^2}{n(n-1)}}$$

**Step 4.** Calculate estimated standard error of difference

$$\text{S.E. } (\bar{x}_1 - \bar{x}_2) = \frac{S}{\sqrt{n}}$$

**1. Null hypothesis :**  $H_0 : \mu_1 = \mu_2$ .

**Alternative hypothesis :**  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test)

**2. Calculation of test statistic :**

**Estimated Standard Error of Difference**

$$\text{S.E. } (\bar{x}_1 - \bar{x}_2) = \frac{S}{\sqrt{n}}$$

where

$$S^2 = \frac{1}{n-1} \left[ \sum D^2 - n(\bar{D})^2 \right] = \frac{1}{n-1} \left[ \sum D^2 - \frac{(\sum D)^2}{n} \right]$$

**Test statistic :**  $t = \frac{\bar{D}}{\text{S.E.}(\bar{D})}$

where  $D$  = difference between each pair of observations.

3. **Level of significance :** Let  $\alpha = 0.05$ , or  $\alpha = 0.01$  or  $\alpha = 0.1$  as the case may be.
4. **Critical value :** Find from the table the value of  $t$  at the level of significance  $\alpha$  for  $(n - 1)$  degrees of freedom for one tailed or two tailed test (for the case given by alternative hypothesis), i.e., find  $t_{\alpha, (n-1)}$ .
5. **Decision :** If the calculated value of  $|t| = t_{\alpha, (n-1)}$ , then the null hypothesis is accepted and in case of calculated value of  $|t| > t_{\alpha, (n-1)}$ , then the null hypothesis is rejected and alternative hypothesis is accepted.

**Remarks :** 
$$\frac{\sum(D - \bar{D})^2}{n-1} = \frac{\sum D^2 - n(\bar{D})^2}{(n-1)} \quad \dots (1)$$

$$= \frac{n \sum D^2 - \sum D^2}{n-1} \quad \dots (2)$$

Its proof is beyond the scope of this book.

**Example 28 :** Memory capacity of students was tested before and after giving the nourishing food (CHAVANPRASH). State whether CHAVANPRASH was effective or not from the following scores.

Roll No :	1	2	3	4	5	6	7	8	9	10
Before :	12	14	11	8	7	10	3	0	5	6
After :	15	16	10	7	5	12	10	2	3	8

(For  $v = 9$ ,  $t_{0.05} = 2.26$ )

**Solution :** Since the memory scores before training ( $x$ ) and after training ( $y$ ) are not independent but are paired together, (one pair corresponding to each roll number), we shall apply paired t-test.

1. **Null hypothesis :**  $H_0 : \mu_x = \mu_y$ , i.e., the mean scores before training and after training are same. In other words, the training was not effective.

**Alternative hypothesis :**  $H_1 : \mu_x \neq \mu_y$ , (Two tailed test)

Calculation table

$x$	12	14	11	8	7	10	3	0	5	6	
$y$	15	16	10	7	5	12	10	2	3	8	
$d = x - y$	-3	-2	1	1	2	-2	-7	-2	2	-2	$\Sigma d = -12$
$d^2$	9	4	1	1	4	4	49	4	4	4	$\Sigma d^2 = 84$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-12}{10} = -1.2$$

$$S^2 = \frac{1}{n-1} \left[ \sum d^2 - \frac{(\sum d^2)}{n} \right] = \frac{1}{9} \left[ 84 - \frac{144}{10} \right] = \frac{84 - 14.4}{9} = 7.7333.$$

$$\text{S.E. } (\bar{x} - \bar{y}) = \frac{S}{\sqrt{n}} = \frac{\sqrt{7.7333}}{\sqrt{10}}$$

Test Statistic :  $t = \frac{\bar{d}}{\sqrt{S^2/n}}$

$$\therefore t = \frac{-1.2}{\sqrt{(7.7333/10)}} = \frac{-1.2}{\sqrt{0.7733}} = \frac{-1.2}{0.8794} = -1.3646.$$

$$\therefore |t| = |-1.3646| = 1.3646$$

2. Level of significance : Let  $\alpha = 0.05$ . Also degrees of freedom =  $10 - 1 = 9$ .

3. Critical value : The tabled value of  $t_{0.05}$  for 9 d.f. = 2.26.

4. Decision : Since the calculated value of  $|t| = 1.3646$  is less than the tabled value of  $t_{0.05} = 2.26$ , so the null hypothesis is accepted and we conclude that the effect of CHAVANPRASH is not effective.

**Example 29 :** A certain diet newly introduced to each of the 12 pigs resulted in the following increase in body weight:

$$6, 3, 8, -2, 3, 0, -1, 1, 6, 0, 5 \text{ and } 4.$$

Can you conclude that the diet is effective in increasing the weight of pigs?

[Given  $t_{0.05}$  (II) = 2.20].

**Solution :** It is a case of paired  $t$ -test.

We are given the increments  $d = y - x$ , in the body weight of 12 pigs, where  $x$  and  $y$  denoted the weights of the pigs before and after the introduction of the new diet respectively.

1. Null hypothesis :  $\mu_x = \mu_y$ , i.e., there is no significant difference in the body-weights of the pigs before and after the introduction of the new diet.

Alternative hypothesis :  $H_1 : \mu_x \neq \mu_y$  (Two Tailed Test)

Calculation table

$d$	6	3	8	-2	3	0	-1	1	6	0	5	4	$\Sigma d = 33$
$d^2$	36	9	64	4	0	0	1	1	36	0	25	16	$\Sigma d^2 = 201$

$$\therefore \bar{d} = \frac{\Sigma d}{n} = \frac{33}{12} = 2.75$$

$$S^2 = \frac{1}{n-1} \left[ \sum d^2 - \frac{(\sum d)^2}{n} \right] = \frac{1}{11} \left[ 201 - \frac{(33)^2}{12} \right]$$

$$= \frac{1}{11} \left[ 201 - \frac{1089}{12} \right] = \frac{201 - 90.75}{11} = \frac{110.25}{11} = 10.0227.$$

**Standard Error :** S.E.  $(\bar{x} - \bar{y}) = \frac{S}{\sqrt{n}}$

**Test Statistic :**  $t = \frac{\bar{d}}{\sqrt{S^2/n}}$

$$\therefore t = \frac{2.75}{\sqrt{10.0227/12}} = \frac{2.75}{\sqrt{0.8352}} = \frac{2.75}{0.9139} = 3.0091.$$

**2. Level of significance :** Let  $\alpha = 0.05$ .

**3. Critical value :** Tabulated  $t_{0.05}$  for 11 d.f. = 2.201 (given).

**4. Decision :** Since the calculated ' $t$ ' greater than tabulated ' $t$ ' it is significant. Hence  $H_0$  is rejected at 5% level of significance and we conclude that the new diet is effective in increasing body weight of the pigs.

**Example 30 :** Two laboratories carry out independent estimates of particular chemical in a medicine produced by a certain firm. A sample is taken from each batch, halved and the separate halves sent to the two laboratories. The following data is obtained:

No of samples	10
Mean value of differences of estimates	0.6
Sum of the squares of the differences (from their mean)	20
Is the difference significant?	
(Value of $t$ at 5% level for 9 degrees of freedom is 2.262).	

**Solution :** It is a case of paired  $t$ -test. We are given that:

$$n = 10, \quad \bar{d} = 0.6, \quad \Sigma(d - \bar{d})^2 = 20.$$

1. Null hypothesis :  $H_0 : \mu_1 = \mu_2$  i.e., there is no significant difference in the results of the two laboratories.

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2$ .

**Test statistics :**  $t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{\bar{d} \times \sqrt{n}}{S}$

where  $S = \sqrt{\frac{1}{n-1} \sum (d - \bar{d})^2} = \sqrt{\frac{20}{9}} = \sqrt{2.2222} = 1.4907$ .

**Standard Error :** S.E.  $(\bar{x} - \bar{y}) = S/\sqrt{n} = 1.4907/\sqrt{10}$ .

$$\therefore t = \frac{0.6 \times \sqrt{10}}{1.4907} = \frac{0.6 \times 3.1623}{1.4907} = \frac{1.8974}{1.4907} = 1.2728.$$

- 2. Critical value :** Tabulated value of  $t$  for 9 d.f. = 2.262 which follows students's  $t$ -distribution with d.f. =  $10 - 1 = 9$ .
- 3. Decision :** Since calculated value of  $t$  is less than tabulated value of  $t$ , it is not significant. This implies that the difference in the results of the two laboratories is just due to the fluctuations of sampling. Hence we accept the null hypothesis and conclude that the given difference is not significant.

**Example 31 :** In a certain experiment to compare two types of pig foods A and B, the following results of increase in weights were observed in pigs:

Pig number →		1	2	3	4	5	6	7	8	Total
Increase in weight in lb.	Food A	49	53	51	52	47	50	52	53	407
	Food B	52	55	52	53	50	54	54	53	423

Assuming that the two samples of pigs are independent, can we conclude that food B is better than food A.

**Solution :**

- 1. Null hypothesis :** If the increase in weights due to foods A and B are denoted by variables  $X$  and  $Y$  respectively. Then  $H_0 : \mu_x = \mu_y$ , i.e., there is no significant difference in increase in weights due to diets A and B.

**Alternative hypothesis :**  $H_1 : \mu_x < \mu_y$  (Left-tailed test)

**Test statistic :**

Let  $d = x - A = x - 50$ ;  $D = y - B = y - 53$ . [Take  $A = 50$  and  $B = 53$ ]

We have,  $\sum d = 7$ ,  $\sum d^2 = 37$ ;  $\sum D = -1$ ,  $\sum D^2 = 17$ ,  $n_1 = n_2 = 8$

$$\therefore \bar{x} = 50 + \frac{7}{8} = 50.875; \quad \bar{y} = 53 + \frac{(-1)}{8} = 52.875.$$

$$\Sigma (x - \bar{x})^2 = \Sigma d^2 - \frac{(\sum d)^2}{n_1} = 37 - \frac{49}{8} = 30.825.$$

$$\Sigma (y - \bar{y})^2 = \Sigma D^2 - \frac{(\sum D)^2}{n_2} = 17 - \frac{1}{8} = 16.875.$$

$$\therefore S^2 = \left[ \frac{\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2}{n_1 + n_2 - 2} \right] = \frac{30.875 + 16.875}{14} = 3.41.$$

**Test statistics :**  $t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$

$$\therefore t = \frac{50.875 - 52.875}{\sqrt{3.41 \left( \frac{1}{8} + \frac{1}{8} \right)}} = \frac{-2}{\sqrt{3.41 \times 0.25}} = \frac{-2}{\sqrt{0.8525}} = \frac{-2}{0.9233} = -2.166.$$

$$\therefore |t| = 2.166.$$

**3. Level of significance :** Let  $\alpha = 0.05$ .

**4. Critical value :** Tabulated value of  $t$  for d.f. =  $8 + 8 - 2 = 14$  for one tailed test (left tailed) is  $-1.76$  or  $|t_{0.05}(14)| = 1.76$ .

**5. Decision :** Since calculated  $|t| = 2.166$  is greater than tabulated  $|t_{0.05; 14}|$ , it is significant at 5% level of significance. Hence we reject  $H_0$  and conclude that the foods A and B differ significantly as regards their effect on increase in weight. Further since  $\bar{y} > \bar{x}$ , food B is superior to food A.

**Example 32 :** An I.Q. Test was administrated to 5 persons before and after they were given the nourishing food COMPLAN. The results are given below.

Candidates :

	I	II	III	IV	V
I.Q. Before COMPLAN :	110	120	123	132	125
I.Q. After COMPLAN :	120	118	125	136	121

Test whether there is any change in I.Q. after the COMPLAN. It is given that  $t_{0.01} = 4.6$  for d.f.

**Solution :**

**1. Null hypothesis :**  $H_0 : \mu_1 = \mu_2$  or  $\bar{D} = 0$  i.e., there is no change in I.Q. after the training.

**Alternative hypothesis :**  $H_1 : \mu_1 \neq \mu_2$

**2. Computation of test statistic :** Let us compute mean and standard deviation from the following table.

Candidates	I.Q. before $y$	I.Q. after : $x$	Difference $D = x - y$	$D^2$
I	110	120	10	100
II	120	118	-2	4
III	123	125	2	4
IV	132	136	4	16
V	125	121	-4	16
			$\Sigma D = 10$	$\Sigma D^2 = 140$

$$\text{Mean Difference : } \bar{D} = \frac{\Sigma D}{n} = \frac{10}{5} = 2.$$

$$\text{Estimated Variance : } S^2 = \frac{\Sigma D^2}{n-1} - \frac{(\Sigma D)^2}{n(n-1)}$$

$$= \frac{140}{5-1} - \frac{(10)^2}{5(5-1)} = \frac{140}{4} - \frac{100}{20} = 35 - 5 = 30.$$

$$\therefore S = \sqrt{30}.$$

$$\text{Standard Error of Difference } S.E. (\bar{D}) = \frac{S}{\sqrt{n}} = \frac{\sqrt{30}}{\sqrt{5}} = \sqrt{6}.$$

$$\text{Test statistic : } t = \frac{\bar{D}}{S.E. (\bar{D})} = \frac{2}{\sqrt{6}} = \frac{2}{2.45} = .816.$$

4. **Critical value :**  $\alpha = 0.01$  the tabulated or critical value of  $t$  at  $\alpha = 0.01$  for  $5 - 1 = 4$  degrees of freedom is  $t_{0.01, 4} = 4.6$ .
5. **Decision :** The calculated value of  $|t| = 0.816 < t_{0.01, 4} = 4.6 \Rightarrow$  the null hypothesis is accepted  $\Rightarrow$  There is no significant difference in I.Q. after the COMPLAN was given.

**Example 33 :** A drug was administered to 10 patients and the increments in their blood pressure were recorded to be 6, 3, -2, 4, -3, 4, 6, 0, 0, 2. Is it reasonable to believe that the drug has no effect on change of blood pressure? Use 5% significance level and assume that for 9 degrees of freedom,  $t_{0.5, 9} = 2.26$ .

**Solution :** Let  $D$  denote the increment in blood pressure. We shall compute the mean difference and standard deviation from the following table.

Table

Patient :	1	2	3	4	5	6	7	8	9	10	Total
$D :$	6	3	-2	4	-3	4	6	0	0	2	$\Sigma D = 20$
$D^2 :$	36	9	4	16	9	16	36	0	0	4	$\Sigma D^2 = 130$

$$\text{Mean Difference : } \bar{D} = \frac{\Sigma D}{n} = \frac{20}{10} = 2.$$

$$\begin{aligned} \text{Estimated Variance of Population : } S^2 &= \frac{\Sigma D^2}{n-1} - \frac{(\Sigma D)^2}{n(n-1)} \\ &= \frac{130}{9} - \frac{(20)^2}{10 \times 9} = \frac{130}{9} - \frac{40}{9} = \frac{90}{9} = 10. \end{aligned}$$

$$\text{Now } S^2 = 10 \Rightarrow S = \sqrt{10}.$$

1. **Null hypothesis :**  $H_0 : \mu_1 = \mu_2$ , i.e., there is no change in blood pressure due to the effect of drug.

**Alternative hypothesis :**  $H_1 : \mu_1 \neq \mu_2$ . (Two Tailed Test)

**2. Computation of test statistic :**

$$\text{Standard Error of Differences : } S.E. (\bar{D}) = \frac{S}{\sqrt{n}} = \frac{\sqrt{10}}{\sqrt{10}} = 1.$$

$$\text{Test Statistic : } t = \frac{\bar{D}}{S.E. (\bar{D})} = \frac{2}{1} = 2.$$

$$\therefore |t| = 2$$

**3. Level of significance :**  $\alpha = 0.05$ .

**4. Critical value :** The tabulated critical value of  $t$  at  $\alpha = 0.05$  for  $10 - 1 = 9$  degrees of freedom for two tailed test is  $t_{0.05, 9} = 2.26$ .

**5. Decision :** Since the calculated value  $|t| = 2 <$  tabulated value  $t_{0.05, 9} = 2.26$ , so the null hypothesis  $H_0$  is accepted  $\Rightarrow$  there is no change in blood pressure due to the effect of drug.

**Example 34 :** Ten students were given intensive coaching for a month in statistic. The scores obtained in tests 1 and 5 given below :

Sl. No. of students :	1	2	3	4	5	6	7	8	9	10
Marks in 1st test :	50	52	53	60	65	67	48	69	72	80
Marks in 5th test :	65	55	65	65	60	67	49	82	74	86

does the score from test 1 to test 5 show an improvement? Test at 5% level of significance. (The values of  $t$  for 9 degrees of freedom at 5% level for one tailed test is 1.833 and for two tailed test is 2.262.)

**Solution :** 1. Null hypothesis :  $H_0 : \mu_1 = \mu_2$

Alternative hypothesis :  $H_1 : \mu_1 < \mu_2$ . (One tailed test)

**2. Calculation of test statistic :**

S. No.	Marks in 1st Test = $X_1$	Marks in 5th Test = $X_5$	Difference $D = X_1 - X_5$	$D^2 = (X_1 - X_5)^2$
1	50	65	-15	225
2	52	55	-3	9
3	53	65	-12	144
4	60	65	-5	25
5	65	60	5	25
6	67	67	0	0
7	48	49	-1	1
8	69	82	-13	169
9	72	74	-2	4
10	80	86	-6	36
			$\Sigma D = -52$	$\Sigma D^2 = 638$

$$\bar{D} = \frac{\Sigma D}{n} = -\frac{52}{10} = -5.2.$$

$$S = \sqrt{\frac{\Sigma D^2 - n(\bar{D})}{n-1}} = \sqrt{\frac{638 - 10 \times 27.04}{9}} = 6.391.$$

$$\text{S.E. } (\bar{D}) = \frac{S}{\sqrt{n}} = \frac{6.391}{\sqrt{10}} = 2.021.$$

**Test Statistic :**  $t = \frac{\bar{D}}{\text{S.E. } (\bar{D})} = \frac{-5.2}{2.021} = -2.573.$

$$\therefore |t| = 2.573.$$

**3. Level of significance :**  $\alpha = 0.05$ .

**4. Critical value :** The tabulated critical value of  $t$  at  $\alpha = 0.05$  for  $10 - 1 = 9$  degrees of freedom for two tailed test is  $t_{0.05, 9} = 1.833$ .

**5. Decision :** Since calculated  $|t| = 2.573 >$  tabulated  $t_{0.05, 9} = 1.833$ , so the null hypothesis  $H_0$  is rejected  $\Rightarrow$  the coaching has improved the standard.

### EXERCISE

1. Ten cartons are taken at random from an automatic filling machine. The mean net weight of the 10 cartons is 11.8 Kg and standard deviation is 0.15 kg. Does the sample mean differ significantly from the intended weight of 12 kg? You are given for  $v = 9$ ,  $t_{0.05} = 2.26$ .

[Hint : Null hypothesis :  $H_0 : \mu = 12$ .

Alternative hypothesis :  $H_1 : \mu \neq 12$ .

$$\text{Also } t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{11.8 - 12}{0.15/\sqrt{9}} = -4.0 \Rightarrow |t| = 4.$$

Also  $t_{0.05, 9} = 2.26 < 4 = \text{calculated } |t| \Rightarrow H_0$  is rejected  $\Rightarrow$  the sample mean differ significantly].

2. The mean weekly sales of the chocolate bar in candy stores was 140.3 bars per store. After an advertisement campaign, the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign not successful?
3. Certain pesticide is packed into bags by a machine. A random sample of 10 bags is drawn and their contents are found to weight (in kgs) as follows:

$$50, 49, 52, 44, 45, 48, 46, 45, 49, 45.$$

Test if the average packing can be taken to be 50 kg.

[Hint : Null hypothesis.  $H_0 : \mu = 50$ .  $H_1 : \mu \neq 50$  (Two tailed test)

Here  $\bar{x} = 47.3$ ,  $S^2 = 7.12$ .

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{-2.7}{\sqrt{0.712}} = -3.2 \Rightarrow |t| = 3.2$$

$$t_{0.05, 9} = 2.262 < |t| = 3.2.$$

$\Rightarrow H_0$  is rejected  $\Rightarrow$  Average packing cannot be taken as 50 kgs]

4. A fertiliser mixing machine is set to give 12 kg of nitrate for every quintal bag of fertiliser. Ten 100 kg bags are examined. The percentages of nitrate are as follows.

11, 14, 13, 12, 13, 12, 13, 14, 11, 12

Is there reason to believe that the machine is defective?

[Value of  $t$  for 9 degrees of freedom is 2.262].

[Hint : Here  $\bar{x} = \frac{125}{10} = 12.5$  kgs. Also  $s = \sqrt{\frac{(x - \bar{x})^2}{n-1}} = \sqrt{\frac{10.5}{9}} = 1.08$ .

Null hypothesis.  $H_0 : \mu = 12$ .  $H_1 : \mu \neq 12$  (Two tailed test)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{12.5 - 12}{(1.08)/\sqrt{10}} = 1.464.$$

Now  $|t| = 1.464 < t_{0.05, 9} = 2.262$

$\Rightarrow H_0$  is accepted  $\Rightarrow$  there is no reason to believe that machine is defective].

5. Ten students are selected at random from a college and their heights are found to be 100, 104, 108, 110, 118, 120, 122, 124, 126 and 128 cms. In the light of these data, discuss the suggestion that the mean height of the students of the college is 110 cms. Use 5% level of significance.

[Hint :  $\bar{x} = \frac{1160}{10} = 116$  cm. Also  $s = \sqrt{\frac{(x - \bar{x})^2}{n-1}} = \sqrt{\frac{864}{9}} = 9.798$ .

Test statistic  $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{(116 - 110)}{9.798} \times \sqrt{10} = 1.937$ .

Also  $t_{0.05, 9} = 2.262 > |t| = 1.937$ .

$\Rightarrow H_0$  is accepted  $\Rightarrow$  the mean height of the students can be taken as 110 cms.]

6. A random sample of 10 tins of oil filled in by an automatic machine gave the following weights, in kg?

2.05, 2.01, 2.04, 1.91, 1.96, 2.01, 1.97, 1.99, 2.04, 2.02

Can we accept at 5% level of significance, the claim that the average weight of the tin is 2 kg?

7. In the past a machine has produced washers having a mean thickness of 0.05 cm. To determine whether the machine is in proper order, a sample of 10 washers is taken of which mean thickness is 0.053 cm and S.D = 0.003. Test the hypothesis that the machine is working in proper order.

[Hint :  $H_0 : \mu = 0.05$ ,  $H_1 : \mu \neq 0.05$  (Two tailed test)]

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{0.053 - 0.05}{0.003/\sqrt{9}} = \frac{.003}{.001} = -3$$

Also  $t_{0.05, 9}$  for two tail test is  $2.262 < |t| = 3$

$\Rightarrow H_0$  is rejected  $\Rightarrow$  machine is not working in proper order]

8. Talcum powder is packed into tins by a machine. A random sample of 11 tins is drawn and their contents are found to weigh in kgs as follows:

0.44, 0.51, 0.49, 0.45, 0.48, 0.46, 0.45, 0.47, 0.45 and 0.47

Test if the average packing can be taken to be 0.5 kgs.

9. Samples of the types of electric bulbs were tested for the length of life and the following data were obtained.

	Type I	Type II
Sample No.	8	7
Sample Mean.	1234 hours	1036 hours
Sample S.D.	36 hours	40 hours

Is the difference in the mean sufficient to warrant an inference that type I is superior to type II regarding the length of life?

10. Strength tests carried out on samples of two yarns spun to the same count gave the following results:

	Yarn A	Yarn B
No. of Sample.	4	9
Sample of Mean.	50	42
Sample of variance.	42	56

The strengths are expressed in kgs. Is the difference in mean strengths significant of real difference in the mean strength of the sources from which the samples are drawn?

[Hint : Null hypothesis :  $H_0 : \mu_1 = \mu_2$ .

Alternative hypothesis :  $H_1 : \mu_1 \neq \mu_2$ .

$$S^2 = \frac{4 \times (42)^2 + 9 \times (56)^2}{4+9-2} = \frac{35280}{11} = 3207.27.$$

$$\begin{aligned} \therefore \text{Test statistic } t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2[(1/n_1) + (1/n_2)]}} \\ &= \frac{50 - 42}{\sqrt{3207.27[(1/4) + (1/9)]}} = \frac{8}{34.03} = 0.235. \end{aligned}$$

Also

$$t_{0.05, 11} = 2.201$$

Now

$$|t| = 0.235 < t_{0.05, 11} = 2.201 \Rightarrow H_0 \text{ is accepted}$$

⇒ the difference in the mean strength is insignificant of real difference in the mean strength of the sources].

11. Two samples of 6 and 5 gave the following data.

	Sample I	Sample II
Mean	40	50
S.D.	8	10

Is the difference of the mean significant? The value of  $t$  for  $v = 9$  at 5% level of significance is 2.26.

[Hint :  $H_0 : \mu_1 = \mu_2$ ,  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test)

$$t = 1.84, \quad t_{0.05, 9} = 2.26$$

Since  $|t| < t_{0.05, 9}$  ⇒  $H_0$  is accepted ⇒ the difference in means is not significant].

12. Two batches of the same product are tested for their mean life. Assuming that the lives of the product follow a normal distribution with an unknown variance, test the hypothesis that the mean life is the same for both the branches, given the following information.

Batch	Sample Size	Mean life in hours	S.D.
I	10	750	12
II	8	820	14.

Take  $\alpha = 5\%$ .

[Hint :  $H_0 : \mu_1 = \mu_2$ ,  $H_1 : \mu_1 \neq \mu_2$ . (Two tailed test)]

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 11.43.$$

Also  $t_{0.05, 16} = 2.120$

Since  $|t| = 11.43 > t_{0.05, 16} = 2.120$

⇒  $H_0$  is rejected ⇒ the mean life differs significantly both the batches].

13. Two kinds of manure were applied to sixteen one-hectare plots, other conditions remaining the same. The yields in quintals are given below:

Manure I :	18	20	36	50	49	36	34	49	41
Manure II :	29	28	26	35	30	44	46		

Is there any significant difference between the mean yields? Use 5% significance level.

14. The means of two random samples of size 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the sample be considered to have been drawn from the same normal population?

[Hint : Null hypothesis.  $H_0 : \mu_1 = \mu_2$ . Alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$ .

Estimated variance

$$S^2 = \frac{1}{n_1 + n_2 - 2} \left[ \sum (x - \bar{x})^2 + \sum (y - \bar{y})^2 \right] = \frac{26.94 + 18.73}{9 + 7 - 2} = 3.26.$$

$$\text{Test statistic : } t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left[ \left( 1/n_1 \right) + \left( 1/n_2 \right) \right]}} = \frac{196.42 - 198.82}{\sqrt{3.26 \left( \frac{1}{9} + \frac{1}{7} \right)}} = -2.638.$$

$|t| = 2.368$ . But  $t_{0.05, 14} = 2.15$ .

Now  $|t| = 2.638 > t_{0.05, 14} = 2.15 \Rightarrow \text{Null hypothesis } H_0 \text{ is rejected} \Rightarrow \text{the samples cannot be considered to have come from the same normal population}.$

15. Two sets  $A$  and  $B$  of ten students selected at random from a college were taken : one was given the memory test as they were and the set was given a memory test after two weeks' training and the scores were given below:

Set $A$ :	10	8	7	9	8	10	9	6	7	8
Set $B$ :	12	8	8	10	8	11	9	8	9	9

Do you think that there is any significant effect due to training? (Given  $t = 2.10$  at d.f. = 18,  $\alpha = 0.05$ ).

16. The income of a random sample of labourers in the industry I are Rs. 630, 650, 680, 690, 710 and 720 per month. The incomes of a similar sample from industry II are Rs. 610, 620, 650, 680, 690, 700, 720 and 730 p.m. Discuss the validity of the suggestion that the industry I pays labourers much better industry II.

$$[\text{Hint : } \bar{x} = \frac{4080}{6} = 680, \bar{y} = \frac{6780}{10} = 678.$$

$$S^2 = \frac{1}{n_1 + n_2 - 2} \left[ \sum (x - \bar{x})^2 + \sum (y - \bar{y})^2 \right] = \frac{6000 + 15360}{6 + 10 - 2} = 1525.71$$

Let  $H_0 : \mu_1 = \mu_2$ , and  $H_1 : \mu_1 \neq \mu_2$ . (Two Tailed Test)

$$\text{Test statistic } t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left[ \left( 1/n_1 \right) + \left( 1/n_2 \right) \right]}} = \frac{680 - 678}{\sqrt{1525.71 \left[ \left( 1/6 \right) + \left( 1/10 \right) \right]}} = 0.1$$

Also,  $t_{0.05, 14} = 1.761 > |t| = 0.1 \Rightarrow H_0 \text{ is accepted.}$

$\Rightarrow$  there is no significant difference between the salaries of Industry I and Industry II].

17. The following data present the yields in quintals of corn on ten subdivisions of equal area of two agricultural plots :

Plot 1 :	6.2	5.7	6.5	6.0	6.3	5.8	5.7	6.0	6.0	5.8
Plot 2 :	5.6	5.9	5.6	5.7	5.8	5.7	6.0	5.5	5.7	5.5

Find out whether difference between the mean yields of two plots is significant?

18. Of the two salesman,  $X$  claims that he has made larger yields sales than  $Y$ . For the accounts examined which are comparable for the two men, the results were:

	X	Y
Number of Sales	10	17
Average size	Rs. 6200	Rs. 5600
Standard deviation	Rs. 690	Rs. 600

Do these two average sizes of the sale figures differ significantly? Explain your result.

19. Ten individuals are chosen at random from a population and their heights in inches are found to be : 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. In the light of these data, mentioning the null hypothesis, discuss the suggestion that the mean height in the population is 66 inches.

[Hint :  $H_0 : \mu = 66$ " ;  $H_1 : \mu \neq 66$ " (Two tailed)]

$$d = x - A = x - 66 ; \Sigma d = 18, \Sigma d^2 = 114 ; n = 10.$$

$$\bar{x} = 66 + \frac{18}{10} = 67.8 ; S^2 = \frac{1}{9} \left[ 114 - \frac{324}{10} \right] = 9.067.$$

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{67.8 - 66}{\sqrt{9.067/10}} = 1.89.$$

$t_{0.05}$  (9 d.f.) = 2.262 ;  $t$  not significant.  $H_0$  : that the mean height in the population is 66 inches, may be accepted at 5% level of significance.

20. The yield of alfalfa from six test plots is 2.75, 5.25, 4.50, 4.25 and 3.25 tonnes per hectare. Test at 5 percent level of significance whether this supports the contention that the true average yield for this kind of alfalfa is 3.50 tonnes per hectare.

[Hint :  $H_0 : \mu = 3.50$  ;  $H_1 : \mu \neq 3.50$  (Two tailed test)]

$$\bar{x} = \frac{\Sigma x}{n} = \frac{22.50}{6} = 3.75 ; S^2 = \frac{1}{n-1} \sum (x - \bar{x})^2 = \frac{5.8750}{5} = 1.175$$

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{3.75 - 3.50}{\sqrt{1.175/6}} = 0.5650.$$

Calculated  $t = 0.5650 <$  Tabulated  $t_{0.05}$  (5 d.f.) = 2.57. Not significant.

$H_0$  is accepted at 5% level of significance and we conclude that true average yield for this kind of alfalfa is 3.50 tonnes per hectare.]

21. What do you understand by the paired t-test? Under what conditions will you apply it?

Ten soldiers visit a rifle range for two consecutive weeks. For the first week their scores are

67, 24, 57, 55, 63, 54, 56, 68, 33, 43

and during the second week they score in the same order

70, 38, 58, 58, 56, 67, 68, 72, 42, 38

Examine if there is any significant difference in their performance.

22. To test the effect of a fertilizer on rice production, 24 plots of land having equal areas were chosen. Half of these plots were treated with fertilizer and the other half were untreated. Other conditions were the same. The mean yield of rice on the untreated plots was 4.8 quintals with a standard deviation of 0.4 quintal, while the mean yield on the treated plots

was 5.1 quintals with a standard deviation of 0.36 quintal. Can we conclude that there is significant improvement in rice production because of the fertilizer at 5% level of significance?

[Hint : Let  $\mu_0$  : There is no significant difference in rice production because of the fertilisers. Applying  $t$ -test of difference of means:

$$\text{Here, } S = \sqrt{\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{11(4)^2 + 11(36)^2}{12 + 12 - 2}} = 0.51.$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$t = \frac{5.1 - 4.8}{0.51} \sqrt{\frac{12 \times 12}{12 + 12}} = \frac{3}{51} \times 2.45 = 1.44 ; \text{ d.f.} = 22.$$

For  $v = 22$ ,  $t_{0.05} = 2.07$ . The calculated value of  $t$  is less than the table value. The hypothesis holds true. Hence there is no significant difference in the rice production because of the fertilisers.]

23. Experience shows that a fixed dose of a certain drug causes an average increase of pulse rate 10 beats per minute with a standard deviation of 4. A group of 9 patients given the same dose showed the following increase:

13, 15, 14, 10, 8, 12, 16, 9, 20

Test at 5% level of significance whether this group is different in response to the drug?

Extract from  $t$ -table

Degrees of freedom	7	8	9	10
5%	2.36	2.31	2.26	2.23
1%	3.50	3.36	3.25	3.17

[Hint :  $n = 9$ ,  $\Sigma x = 117$ ,  $\bar{x} = 13$ ,  $\Sigma (x - \bar{x})^2 = 114$

$$S^2 = \frac{1}{n-1} \Sigma (x - \bar{x})^2 = \frac{114}{8} = 14.25$$

$H_0 = 10$  (beats per minute) ;  $H_1 : \mu \neq 10$  (Two-tailed test)

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{13 - 10}{\sqrt{14.25/9}} = 2.3867 \sim t(8 \text{ d.f.})$$

$t_{0.05}(8 \text{ d.f.}) = 2.31$  ; Calculated  $t$  is significant.

$H_0$  rejected  $\Rightarrow$  the given group is different in response to the drug.]

24. A survey proposed to be conducted to know the annual earnings of old commerce graduates of Denn University. How large should the sample size be taken in order to estimate the mean annual earnings within plus and minus Rs. 1000 at 95% confidence level? The standard deviation of the annual earnings of the entire population is known to be Rs. 3000.

[Hint :  $P[|\bar{x} - \mu| < 1000] = 0.05$  .... (1) Also,  $P\left[\frac{\bar{x} - \mu}{(\sigma/\sqrt{n})} \leq 1.96\right] = 0.095$  .... (2)

$$(1) \text{ and } (2) \Rightarrow \frac{1.96 \sigma}{\sqrt{n}} = 1000 \Rightarrow n = \left(\frac{1.96 \times 1000}{3000}\right)^2 = (5.88)^2 = 34.57$$

Hence, sample size should be 35].

25. A researcher wishes to estimate the mean of a population by using sufficiently large sample. The probability is 0.95 that the sample mean will not differ from the true mean by more than 25% of the standard deviation. How large sample should be taken?

[Hint :  $P|\bar{x} - \mu| \leq 25\% \text{ of s.d.}] = 0.95 \Rightarrow P\left[|\bar{x} - \mu| \leq \frac{\sigma}{4}\right] = 0.95$  .... (1)

Also  $P\left[|\bar{x} - \mu| \leq 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$  .... (2)

$$\text{From (1) and (2)} \Rightarrow 1.96 \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{4} \Rightarrow n = 62]$$

26. A firm wishes to estimate with an error of not more than 0.03 and a level of confidence of 98%, the proportion of consumers that prefers its brand of household detergent. Sales reports indicate that about 0.20 of all consumers prefer the firm's brand. What is the requisite sample size?

[Hint :  $n = \frac{Z^2 PQ}{E^2} = \frac{(2.33)^2 \times 0.2 \times 0.8}{(0.03)^2} = 965.14$   $[\because Z = 2.33 \text{ from table}]$

Hence, sample size should be  $n = 965$ ].

27. It is known that population standard deviation in waiting time for new gas connection in a particular town is 20 days. How large sample should be chosen to the 95% confident that the average waiting time is within 3 days of true average?

[Hint :  $n = \left(\frac{\sigma Z}{E}\right)^2 = \left(\frac{1.96 \times 20}{5}\right)^2 = 61.46 \text{ or } 62]$

28. Measurements performed on random samples of two kinds of cigarettes yielded the following results on their nicotine content (in milligrams) :

Brand A : 21.4    23.6    24.8    22.4    26.3

Brand B : 22.4    27.7    23.5    29.1    25.8

Use the 1 present level of significance to check on the claim that brand B has a higher average nicotine contents than brand A.

[Hint :  $H_0 : \mu_A = \mu_B$  ;  $H_1 : \mu_A < \mu_B$  (left tailed test). Also  $\bar{x} = 23.7$  and  $\bar{y} = 31.30$ .

$$S^2 = \frac{\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2}{n_1 + n_2 - 2} = \frac{14.96 + 31.30}{5 + 5 - 2} = 5.7825.$$

$$t = \frac{23.7 - 25.7}{\sqrt{5.7825 \left( \frac{1}{5} + \frac{1}{5} \right)}} = -1.315 \Rightarrow |t| = 1.315.$$

A  $t_{0.05}$  for 8 d.f. = 1.86. Since calculated  $t < t_{0.05, 8}$   $\Rightarrow H_0$  is accepted].

29. Two salesmen  $A$  and  $B$  are working in a certain district. From a sample survey conducted by the head office, the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen.

	<i>A</i>	<i>B</i>
No. of sales	10	18
Average sales (in Rs.)	170	205
Standard deviation (in Rs.)	20	25

[Hint :  $H_0$  : there is no significant difference in the average sale of the two salesman.

$$\therefore S^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2} = \frac{(10 - 1) 400 + (18 - 1) 825}{10 + 18 - 2} = \frac{14225}{26}$$

$$= 547.12 \Rightarrow S = \sqrt{547.12} = 23.39.$$

$$\text{S.E. } (\bar{x}_1 - \bar{x}_2) = S \div \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$\therefore \text{Test statistic } t = \frac{\bar{x}_1 - \bar{x}_2}{S.E. (\bar{x}_1 - \bar{x}_2)} = \frac{170 - 205}{23.39} \times \left( \frac{10 \times 18}{10 + 18} \right) = 3.79.$$

Also, d.f. =  $n_1 + n_2 - 2 = 26$ . Also  $t_{0.05}$  for 26 d.f. = 2.056.

Since the calculated value of  $t$  is greater than the table value of  $t_{0.05, 26}$   $\Rightarrow H_0$  is rejected  
 $\Rightarrow$  there is a significant difference in the average sales of the two salesman].

30. The nicotine contents in milligrams in two sample of tobacco were found to be as follows:

Sample <i>A</i> :	24	27	26	21	25
Sample <i>B</i> :	27	30	28	31	2

Can it be said that two samples come from same normal population?

31. Measurement of the fat contents of two kinds of ice cream : Brand *A* and Brand *B*, yielded the following sample data:

Brand <i>A</i> :	13.5	14.0	13.6	12.9	and	13.0	per cent
Brand <i>B</i> :	12.9	13.0	12.4	13.5	and	12.7	per cent.

Discuss how you will test the null hypothesis :  $\mu_a = \mu_b$  (where  $\mu_a$  and  $\mu_b$  are the respective true average fat contents of the two kinds of ice cream) against the alternative hypothesis  $\mu_a \neq \mu_b$  at the level of significance  $\alpha = 0.05$ . [Outline the procedure without doing numerical calculations].

[Hint :  $x$  - values : Brand *A* readings ;  $y$  - values : Brand *B* readings.

$$\text{Compute : } \bar{x} = \frac{\Sigma x}{n_1} = \frac{67}{5} = 13.4 ; \quad \bar{y} = \frac{\Sigma y}{n_2} = \frac{64.5}{5} = 12.9$$

$$\text{and } S^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma (x - \bar{x})^2 + \Sigma (y - \bar{y})^2]$$

Under  $H_0 : \mu_a = \mu_b$  against  $H_1 : \mu_a \neq \mu_b$  (two tailed)

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = t_0, (\text{say}) \sim t_{n_1 + n_2 - 2} = t_8.$$

32. Slim-Trim, an agency conducting a weight reduction programme, claims that participants in their programme achieve a weight reduction of at least 5 kg. after two weeks of the programme.

In evidence thereof they have given the following data on 10 participants who had undergone this programme, about their weights in kg prior to the programme and two weeks after the programme.

On the basis of this sample evidence, can the claim of the agency on weight reduction be sustained?

Test a significance level of 5%.

Before (kg.) : 86 92 100 93 88 80 88 92 95 106

After (kg.) : 77 84 92 87 80 74 80 85 95 96

The calculated value of  $t$  is higher than the table value. The hypothesis is rejected. Hence there is a significant difference in the weight before and after the programme.

33. A certain stimulus administered to each of 12 patients resulted in the following change in blood pressure.

$$5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6$$

Can it be concluded that the stimulants will in general be accompanied by an increase in blood pressure?

$$[\text{Hint : Here } \Sigma D = 31, \therefore \bar{D} = \frac{\Sigma D}{12} = \frac{31}{12} = 2.58.]$$

$$\text{Also, } S^2 = \frac{\Sigma D^2}{n-1} - \frac{(\Sigma D)^2}{n(-1)} = \frac{185}{11} - \frac{31 \times 31}{12 \times 11} = 9.53.$$

$$\therefore S = 3.08$$

Null hypothesis :  $H_0 : \mu_1 = \mu_2$  ;

$$\text{Standard Error } \text{S.E.}(\bar{D}) = \frac{S}{\sqrt{n}} = \frac{3.08}{\sqrt{12}} = 0.89.$$

$$\text{Test Statistic : } t = \frac{\bar{D}}{\text{S.E.}(\bar{D})} = \frac{2.58}{0.89} = 2.89$$

Now,  $t_{0.05, 11} = 2.2 < \text{Calculated } |t| = 2.89$

$\Rightarrow$  Null hypothesis is rejected.

$\Rightarrow$  the stimulus will in general be accompanied by an increase in blood pressure.]

34. The sales data of an item in six shops before and after a special promotional campaign are as under:

Shops :	A	B	C	D	E	F
Before campaigns :	53	28	31	48	50	42
After campaign :	58	29	30	55	56	45

Can the campaign be judged to be a success? Test at 5% of significance.

35. Point out the various applications and properties of  $t$ -distribution.

### ANSWERS

1. The sample mean differs significantly.
2.  $t = 1.9716$ ; the advertisement was successful at 5% level of significance.
3.  $t = 3.2$ ; the average packing cannot be taken as 50 kgs.
4. There is no reason to believe that the machine is defective.
5. The mean height of the students can be taken as 110 cm.
6.  $t = 0.707$ ; the claim that the average weight of the tin is 2 kg is accepted.
7. The machine is not in proper order.
8.  $t = -3.43$ ; No.
9.  $t = 9.35$ . Type I bulb is superior to type II.
10. The difference in the mean strength is insignificant of the real difference in the mean strength of the sources.
11. The difference in means is not significant.
12. The mean life differs significantly for both the batches.
13. The mean yields do not differ significantly.
14. The sample cannot be considered to have come from the same normal population.
15. There is no significant difference due to training.
16. There is no significant difference between the salaries of industry I and industry II.
17. The mean yields of the two plots differ significantly;  $t = 3.03$ .
18.  $t = 2.28$ ,  $X$ 's claim is justified.
19. Mean height is 66 inches.
20. The average yield of alfalfa is 3.50 tonnes per hectare.
21. There is no significant difference in their performance.
22. There is no significant difference in the rice production because of the fertiliser.

23. The given group is different in response to the drug.
24.  $n = 35$       25.  $n = 62$ .      26.  $n = 965$       27.  $n = 62$
28. It does not support the claim that brand *B* has higher contents than the brand *A*.
29. There is a significant difference in the average sale of the two salesmen.
30. The two samples have come from the same population.
31. There is a significant difference in the weight before and after the programme.
32. The stimulus will be accompanied by an increase in blood pressure.
33.  $|t| = 2.78$ . The sales campaign has been a success.



# 14

# Chi-Square Test

## 14.1 CHI-SQUARE TEST

The chi-square test, written as  $\psi^2$ -test, is a useful measure of comparing experimentally obtained results with those expected theoretically and based on the hypothesis. It is used as a **test statistic in testing a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared.** In general Chi-square test is applied to those problems in which we study whether the frequency with which a given event has occurred is significantly different from the one as expected theoretically. The measure of *Chi-square enables us to find out the degree of discrepancy between observed frequencies and theoretical frequencies and thus to determine whether the discrepancy so obtained between observed frequencies and theoretical frequencies is due to error of sampling or due to chance.*

The Chi-square is computed on the basis of frequencies in a sample and thus the value of chi-square so obtained is a statistic. Chi-square is not a parameter as its value is not derived from the observations in a population. Hence Chi-square test is a Non-parametric test. Chi-test is not concerned with any population distribution and its observations.

The  $\psi^2$ -test was first used in testing statistical hypothesis by Karl Pearson in the year 1900. It is defined as

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

where  $O_i$  = observed frequency of  $i$ th event.

$E_i$  = expected frequency of  $i$ th event.

We require the following steps to calculate  $\psi^2$ .

- Step 1. Calculate all the expected frequencies, i.e.,  $E_i$  for all values of  $i = 1, 2, \dots, n$ .
- Step 2. Take the difference between each observed frequency  $O_i$  and the corresponding expected frequency  $E_i$  for each value of  $i$ , i.e., and  $(O_i - E_i)$ .

**Step 3.** Square the difference for each value of  $i$ , i.e., calculate  $(O_i - E_i)^2$  for all values of  $i = 1, 2, \dots, n$ .

**Step 4.** Divide each square difference by the corresponding expected frequency i.e.,

Calculate  $\frac{(O_i - E_i)^2}{E_i}$  for all values of  $i = 1, 2, 3, \dots, n$ .

**Step 5.** Add all these quotients obtained in Step 4, then

$$\Psi^2 = \sum_{i=1}^n \left[ \frac{(O_i - E_i)^2}{E_i} \right]$$

is the required value of Chi-square.

It should be noted that

- (a) *the value of  $\Psi^2$  is always positive as each pair is squared one.*
- (b)  *$\Psi^2$  will be zero if each pair is zero and it may assume any value extending to infinity, when the difference between the observed frequency and expected frequency in each pair is unequal. Thus,  $\Psi^2$  lies between 0 and  $\infty$ .*
- (c) *The significance test on  $\Psi^2$  is always based on One Tail Test of the right hand side of standard normal curve as  $\Psi^2$  is always non-negative.*
- (d) *As  $\Psi^2$  is a statistic and not a parameter, so it does not involve any assumption about the form of original distribution from which the observation come.*

## 14.2 DEGREES OF FREEDOM

The number of data that are given in the form of a series of variables in a row or column or the number of frequencies that are put in cells in a contingency table, which can be calculated independently is called the **degrees of freedom** and is denoted by  $v$ .

**Case I.** If the data is given in the form of a series of variables in a row or column, then the **degrees of freedom** = (number of items in the series – 1), i.e., =  $n - 1$ , where  $n$  is the number of the variables in the series in a row or column.

**Case II.** When the number of frequencies are put in cells in a contingency table, the degrees of freedom will be the product of (number of rows less one) and the (number of columns less one), i.e.,  $v = (R - 1)(C - 1)$ , where **R** is the number of rows and **C** is the number of columns.

## 14.3 CHI-SQUARE DISTRIBUTION

$\Psi^2$ -distribution is a continuous distribution whose probability density function is given by

$$P(\Psi^2) = y_0 (\Psi^2)^{\frac{1}{2}(v-2)} e^{-\frac{1}{2}\Psi^2}$$

where

$y_0$  = constant depending on the degrees of freedom  
 $v$  = degrees of freedom =  $n - 1$ .

#### 14.4 PROPERTIES OF $\chi^2$ -DISTRIBUTION

1. Chi-square curve is always positively skewed.
2. The mean of distribution is the number of degrees of freedom.
3. The standard deviation of  $\chi^2$  distribution =  $\sqrt{2v}$ , where  $v$  is the degrees of freedom.
4. Chi-square values increase with the increase in degrees of freedom.
5. The value of  $\chi^2$  lies between zero and infinity i.e.,  $0 \leq \chi^2 < \infty$ .
6. The sum of two  $\chi^2$  distribution is again a  $\chi^2$  distribution i.e., if  $\psi_1^2$  and  $\psi_2^2$  are two independent and they have a  $\psi^2 = \psi_1^2 + \psi_2^2$  is a distribution with  $n_1$  and  $n_2$  degrees of freedom respectively, i.e.,  $(\psi_1^2 + \psi_2^2)$  is also a  $\chi^2$  distribution with  $(n_1 + n_2)$  degrees of freedom.
7. For different degrees of freedom, the shape of the curve will be different. See figure 14.1.
8. Like t-distribution its shape depends on the degree of freedom but it is not a symmetrical distribution.

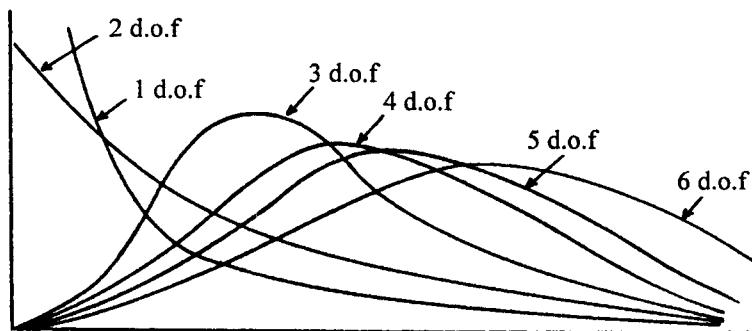


Fig. 14.1

#### 14.5 $\chi^2$ -TEST

The Chi-square test is widely used to test the independence of attributes. It is applied to test the association between the attributes when the sample data is presented in the form of a contingency table with any number of rows or columns.

##### Working Rule

- Step 1.** Set up the Null hypothesis  $H_0$  : No association exists between the attributes.  
**Alternative hypothesis  $H_1$**  : An association exists between the attributes.
- Step 2.** Calculate the expected frequency  $E$  corresponding to each cell by the formula

$$E_{ij} = \frac{R_i \times C_j}{n}$$

where

$R_i$  = Sum total of the row in which  $E_{ij}$  is lying

$C_j$  = Sum total of the column in which  $E_{ij}$  is lying

$n$  = Total sample size

**Step 3.** Calculate  $\psi^2$ -statistic by the formula

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

The characteristics of this distribution are completely defined by the number of degrees of freedom  $n$  which is given by  $v = (R - 1)(C - 1)$ .

where  $R$  = number of rows and

$C$  = number of columns in the contingency table.

**Step 4.** Find from the table the value of  $\psi^2$  for a given value of the level of significance  $\alpha$  and for the degrees of freedom  $v$ , calculated in Step 2.

In no value for  $\alpha$  is mentioned, then take  $\alpha = 0.05$ .

**Step 5.** Compare the computed value of  $\psi^2$ , with the tabled value of  $\psi_{\alpha}^2$  found in step 4.

- (a) If calculated value of  $\psi^2 <$  tabulated value of  $\psi_{\alpha}^2$ , then accept null hypothesis  $H_0$ .
- (b) If calculated value of  $\psi^2 >$  tabulated value of  $\psi_{\alpha}^2$ , then rejected null hypothesis  $H_0$  and accept the alternative hypothesis  $H_1$ .

#### 14.6 USES OF $\psi^2$ TEST

The  $\psi^2$  test is a very powerful test for testing the hypothesis of a number of statistical problems. The important uses of  $\psi^2$  test are:

**1. Test of Goodness of Fit.** If the two curves, viz., (i) Observed frequency curve and (ii) the expected frequency curve are drawn, then the Chi-square statistic may be used to determine whether the two curves so drawn are fitted good or not. Thus the term goods of fit is used to test the concordance of the fitness of these two curves. Under this test there is only one variable, i.e., the degrees of freedom, i.e.,  $v = n - 1$ .

**2. Test of Independence of Attributes.** The Chi-square test is used to see that the principles of classification of attributes are independent. In this test the attributes are classified into a two way table or a contingency table as the case may be. The observed frequency in each cell (square) is known as **Cell frequency**. The total frequency in each row or column of the two way contingency table is known as **Marginal frequency**. The degrees of freedom is:

$$v = (R - 1)(C - 1)$$

where  $R$  = number of rows,  $C$  = number of columns in the two way contingency table. This test discloses whether there is any association or relationship between two or more attributes.

**3. Test of Homogeneity or a Test for a Specified Standard Deviation.** The Chi-square test may be used to test the homogeneity of the attributes in respect of a particular characteristic or it may also be used to test the population variance. In the case of a specified standard deviation the test statistic is given to  $\psi^2 = (n - 1)s^2/\sigma_0^2$ , where  $s^2$  = sample variance and  $\sigma_0^2$  is the hypothesized value of population variance.

## 14.7 CONDITIONS FOR USING THE CHI-SQUARE TEST

1. Each of the observations making up the sample for this test should be independent of each other.
2. The expected frequency of any items or cell should not be less than 5. If it is less than 5, then frequencies from the adjacent items or cells be pooled together in order to make it 5 or more than 5.
3. The total number of observations used in this test must be large, i.e.,  $n > 30$ .
4. This test is used only for drawing inferences by testing hypothesis. It cannot be used for estimation of parameter or any other value.
5. It is wholly dependent on the degrees of freedom.
6. The frequencies used in  $\chi^2$ -test should be absolute and not relative in terms.
7. The observation collected for  $\chi^2$ -test should be on random basis of sampling.

## 14.8 $\chi^2$ -TEST FOR GOODNESS OF FIT

$\chi^2$ -test is a measure of probabilities of association between the attributes. It gives us an idea about the divergence between the observed and expected frequencies. Thus the test is also described as the test of "Goodness of Fit". *If the curves of these two distributions, when superimposed do not coincide or appear to diverge much we say that the fit is poor. On the other hand, if they don't diverge much, then the fit is less poor.*

This concept is illustrated by the following examples.

**Example 1 :** In a sample survey of public opinion, answers to the questions.

- (i) Do you drink?    (ii) Are you favour of local option on sale of liquor?

Question (1)

	Yes	No	Total
Yes	56	31	87
No	18	6	24
Total	74	37	111

Can you infer whether or not the local option on the sale of liquor is dependent on individual drink.

**Solution.** Null hypothesis  $H_0$  : The option on the sale of liquor is independent or not associated with individual drinking.

The theoretical frequencies are tabulated as below.

Question (1)

	Yes	No
Yes	$\frac{87 \times 74}{111} = 58$	$\frac{37 \times 87}{111} = 29$
No	$\frac{76 \times 24}{111} = 16$	$\frac{24 \times 37}{111} = 8$
Total	74	37

**Computation of Test Statistic:**

Applying the formula :

$$\Psi^2 = \sum \frac{(O - E)^2}{E}, \text{ we get}$$

$$\begin{aligned}\Psi^2 &= \frac{(56 - 58)^2}{58} + \frac{(18 - 16)^2}{16} + \frac{(31 - 29)^2}{29} + \frac{(6 - 8)^2}{8} \\ &= \frac{4}{58} + \frac{4}{16} + \frac{4}{29} + \frac{4}{8} = \frac{111}{116} = 0.957.\end{aligned}$$

**3. Degrees of freedom :**  $v = (R - 1)(C - 1) = (2 - 1)(2 - 1) = 1.$

**4. Decision :** The tabulated value of  $\Psi^2$  at  $\alpha = 0.05$  and one degree of freedom is  $\Psi_{0.05, 1}^2 = 3.411.$

Since calculated  $\Psi^2 = 0.957 <$  Tabulated value  $\Psi_{0.05, 1}^2 = 3.411 \Rightarrow$  the null hypothesis  $H_0$  is accepted  $\Rightarrow$  Sale of liquor is independent or not associated with the individual drinking.

**Example 2 :** From the table given below, whether the colour of son's eyes is associated with that of father's eyes.

*Eyes colour in Sons*

		Not light	Light
Eyes colour in fathers	Not light	230	148
	Light	151	471

**Solution :** The observed frequencies are give by the following table.

*Eyes colour in Sons*

		Not light	Light	Total
Eyes colour in fathers	Not light	230	148	378
	Light	151	471	622
	Total	381	619	1000

The theoretical frequencies of expected frequencies are given by

*Eyes colour in Sons*

		Not light	Light
Eyes colour in fathers	Not light	$\frac{381 \times 378}{1000} = 144$	$\frac{619 \times 378}{1000} = 234$
	Light	$\frac{381 \times 622}{1000} = 237$	$\frac{619 \times 622}{1000} = 385$

**1. Null hypothesis  $H_0$  :** The colour of the son's eyes is not associated with the colour of father eyes.

**2. Calculation of test statistic.**

$$\begin{aligned}\psi^2 &= \sum \left[ \frac{(O - E)^2}{E} \right] = \frac{(230 - 144)^2}{144} + \frac{(151 - 217)^2}{237} + \frac{(148 - 234)^2}{234} + \frac{(471 - 385)^2}{385} \\ &= \frac{7396}{144} + \frac{7396}{237} + \frac{7396}{234} + \frac{7396}{385} \\ &= 51.36 + 31.21 + 31.61 + 19.21 = 133.39.\end{aligned}$$

**3. Degrees of freedom.**  $v = (R - 1)(C - 1) = (2 - 1)(2 - 1) = 1$

**4. Decision.** The tabulated value of  $\psi^2$  for  $\alpha = 0.05$  and 1 degree of freedom is  $\psi^2_{0.05, 1} = 3.84$ .

Now calculated  $\psi^2 = 133.39 > \psi^2_{0.05, 1} = 3.84 \Rightarrow$  the null hypothesis is rejected  $\Rightarrow$  there is an association between the colours of eyes of son's and colours of eyes of father's.

**Example 3 :** A survey of 320 families with 5 children each revealed the following distribution.

No. of boys	:	5	4	5	2	1	0
No. of girls	:	0	1	2	3	4	5
No. of families	:	14	56	110	88	40	12

Is this result consistent with the hypothesis that the male and female births are equally probable?

**Solution :** Null hypothesis  $H_0$  : Male and female births are equally probable.

Alternative hypothesis  $H_1$  : Male and female births are not equally probable.

**2. Calculation of Test Statistic.** On the basis of null hypothesis the expected frequencies are

$$320 \left( \frac{1}{2} + \frac{1}{2} \right)^5 = 320 \left( \frac{1}{32}, \frac{5}{32}, \frac{10}{32}, \frac{10}{32}, \frac{5}{32}, \frac{1}{32} \right) = 10, 50, 100, 100, 50, 10$$

Table for  $\psi^2$

O	E	O - E	(O - E) <sup>2</sup>	$\frac{(O - E)^2}{E}$
14	10	4	16	1.60
56	50	6	36	1.72
110	100	10	100	1.00
88	100	-12	144	1.44
40	50	10	100	2.00
12	10	2	4	0.40
$\psi^2 = \sum \frac{(O - E)^2}{E} = 7.16$				

**3. Level of significance.** Take  $\alpha = 0.05$ .

**4. Critical value.** The table value of  $\psi^2$  at  $\alpha = 0.05$  for  $(6 - 1) = 5$  degrees of freedom is  $\psi^2_{0.05} = 11.07$ .

**5. Decision.** Since the calculated value of  $\psi^2 = 7.16 <$  table value  $\psi^2_{0.05}$  for 5 d.f = 11.07, so the null hypothesis  $H_0$  is accepted  $\Rightarrow$  the male and female births are equally probable.

**Example 4 :** In 60 throws of a dice, face one turned up 6 times, face two or three 18 times, face four or five 24 times and face six, 12 times. Test at 10% significance level, if the dice is honest, it being given that  $P(\psi^2 > 6.25) = 0.1$  for 3 degrees of freedom.

**Solution :** 1. Null hypothesis  $H_0$  : The dice is honest.

Alternative Hypothesis  $H_1$  : The dice is not honest.

2. Computation of Test Statistics. Under the assumption of null hypothesis that the dice is honest, the expected frequency for each face is  $60 \times \frac{1}{6} = 10$ .

[ $\because$  Probability of turning up any one of the numbers 1, 2, 3, 4, 5, 6 is  $\frac{1}{6}$ ].

We prepare the following table.

Computation table for  $\psi^2$

Face of the die	Observed frequency $O$	Expected frequency $E$	$O - E$	$\frac{(O - E)^2}{E}$
1	6	10	-4	1.6
2	18	10 } 20	-2	0.2
3		10 }		
4	24	10 } 20	4	0.8
5		10 }		
6	12	10 10	2	0.4
Total	60	60		$\psi^2 = \sum \frac{(O - E)^2}{E} = 3.0$

$$\therefore \psi^2 = 3.0.$$

**3. Critical value.** The table value of  $\psi^2$  at  $\alpha = 0.1$  for  $4 - 1 = 3$  degrees of freedom is 6.25.

**4. Decision.** The tabulated value of  $\psi^2 = 3.0 < \psi^2_{0.1}$  for 3 d.f = 6.25, therefore, the null hypothesis that the die is honest is accepted.

**Example 5 :** Out of a sample of 120 persons in a village, 76 were administered a new drug for preventing influenza and out of them 24 persons were attacked by influenza. Out of those who were not administered the new drug, 12 persons were not affected by influenza. (a) Prepare

*2 × 2 tables showing the actual and expected frequencies, (b) Use Chi-square test for finding out whether the new drug is effective or not. (At 5% level for one degree of freedom, the value of chi-square is 3.84).*

**Solution :** Let the null hypothesis  $H_0$  : The influenza and drug be independent.

**O : 2 × 2 table**

24	32	56
52	12	64
76 (A)	44	120 = N

**E : Expected frequency table**

$\frac{76 \times 56}{120} = 35.5$	$\frac{56 \times 44}{120} = 20.5$	56
$\frac{76 \times 64}{120} = 40.5$	$\frac{64 \times 44}{120} = 23.5$	64
76	44	120 = N

**Computation table for  $\psi^2$**

O	E	O - E	$(O - E)^2$	$\frac{(O - E)^2}{E}$
24	35.5	-11.5	132.25	3.725
52	40.5	11.5	132.25	3.265
32	20.5	11.5	132.25	6.451
12	23.5	-11.5	132.25	5.628
			$\psi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 19.069$	

$$d.f = (2 - 1)(2 - 1) = 1; \text{ From table } \psi^2_{0.05} \text{ for } 5 \text{ d.f.} = 3.84.$$

Calculated value of  $\psi^2$  is 19.069 which is much higher than the table value. Therefore the hypothesis is rejected. Hence we conclude that the drug is undoubtedly effective in controlling the influenza.

**Example 6 :** A certain drug was administered to 500 people out of a total of 800 included in the sample to test its efficacy against typhoid. The results are given below :

	Typhoid	No typhoid	Total
Drug	200	300	500
No drug	280	20	300
Total	480	320	800

On the basis of these data can it be concluded that the drug is effective in preventing the typhoid. [Given for 1 d.f., the value of  $\psi^2_{0.05} = 3.84$ ].

**Solution :** Null hypothesis  $H_0$  : The drug is not effective in preventing typhoid.

Observed frequency Table ( $O$ )

200	300	500
280	20	300
480	320	$800 = N$

E : Expected frequency table

$\frac{480 \times 500}{800} = 300$	$\frac{320 \times 500}{800} = 200$	500
$\frac{480 \times 300}{800} = 180$	$\frac{320 \times 300}{800} = 120$	300
480	320	800

Computation table for  $\chi^2$ 

$O$	$E$	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
200	300	-100	10,000	33.33
280	180	+100	10,000	55.56
300	200	+100	10,000	50.00
20	120	-100	10,000	83.33
800	800		$\chi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 222.22$	

Degree of freedom :  $v = (R - 1)(C - 1) = (2 - 1)(2 - 1) = 1$

$\chi^2_{0.05}$  for 1 d.f. = 3.84

(Given)

The computed value of  $\chi^2$  is much greater than the table value. Therefore the null hypothesis is rejected. Hence we conclude that the drug is effective in preventing the typhoid.

**Example 7 :** In an experiment on the immunization of goats from anthrax the following results were obtained. Derive your inference on the vaccine.

	Died of anthrax	Survived	Total
Inoculated with vaccine	2	10	12
Not inoculated	6	6	12
Total	8	16	24

**Solution :** Let the hypothesis be :

Null hypothesis  $H_0$  : The vaccine is effective i.e., in controlling the disease.

Observed Frequency Table ( $O$ )

2	10	12
6	6	12
8	16	$N = 24$

Expected Frequency Table ( $E$ )

$\frac{8 \times 12}{24} = 4$	8	12
4	8	12
8	16	24

Computation table for  $\psi^2$ 

$O$	$E$	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
2	4	-2	4	1.0
10	8	2	4	0.5
6	4	2	4	1.0
6	8	-2	4	0.5
$\psi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 3.0$				

Degree of freedom :  $v = (R - 1)(C - 1) = (2 - 1)(2 - 1) = 1$

The table value of  $\psi^2$  for 1 d.f. at 5% level of significance is 3.841. The calculated value of  $\psi^2$  is 3, which is less than the table value. Hence, the null hypothesis may be accepted. Thus we conclude that the vaccine is ineffective in controlling the disease.

**Example 8 :** A tobacco company claims that there is no relationship between smoking and lung ailments. To investigate the claims random sample of 300 males in the age group of 40 to 50 is given medical test. The observed sample results are tabulated below :

	Lung ailment	No lung ailment	Total
Smokers	75	105	150
Non-smokers	25	95	120
Total	100	200	300

On the basis of this information, can it be concluded that smoking and lung ailments are independent?

(At 5% level of significance the value of  $X^2 = 3.841$  for 1 d.f.)

**Solution :** Let the null hypothesis be :  $H_0$  = The smoking and lung ailment are not associated.

Expectation of  $(AB) = \frac{180}{300} \times 100 = 60$ . Similarly, the other frequencies are 40, 120 and 80.

Observed Frequency Table ( $O$ )

	(A)	( $\alpha$ )	Total
(B)	75	105	180
( $\beta$ )	25	95	120
Total	100	200	$N = 300$

Expected Frequency Table ( $E$ )

60	120	180
40	80	120
100	200	$N = 300$

Computation table for  $\psi^2$ 

$O$	$E$	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
75	60	15	225	3.750
25	40	-15	225	5.625
105	120	-15	225	1.875
95	80	15	225	2.813
$\psi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 14.063$				

Degree of freedom :  $v = (R - 1)(C - 1) = (2 - 1)(2 - 1) = 1$

The table value of  $\psi_{0.05, 1}$  for 1 d.f. at 5% level of significance is 3.841. The calculated value of  $\psi^2$  is 14.063, which is more than the table value. Hence the hypothesis may be rejected. Thus we conclude that smoking and lung ailments are not independent, i.e., they are associated attributes.

**Example 9 :** From the data given below about the treatment of 500 patients suffering from a disease state whether the new treatment is superior to the conventional treatment.

	No of patients		
	Favourable	Not favourable	Total
New	280	60	340
Conventional	120	40	160
Total	400	100	500

(Given, for  $v = 1$ ,  $\psi^2_{0.05} = 3.84$ ).

**Solution :** Let the hypothesis be :

**Null Hypothesis  $H_0$  :** There is no difference between the new and the conventional treatment.

$$\text{Expectation of } (AB) = \frac{(A) \times (B)}{N} = \frac{340 \times 400}{500} = 272.$$

Similarly, the other expected frequencies are : 128, 68 and 32.

Observed Frequency Table ( $O$ )

	(A)	(B)	Total
(B)	280	60	340
(β)	120	40	160
Total	400	100	$N = 500$

Expected Frequency Table ( $E$ )

272	68	340
128	32	160
400	100	$N = 500$

Computation table for  $\psi^2$ 

$O$	$E$	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
280	272	8	64	0.235
120	128	-8	64	0.500
60	68	-8	64	0.941
40	32	8	64	2.000
$\psi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 3.676$				

Degree of freedom :  $v = (R - 1)(C - 1) = (2 - 1)(2 - 1) = 1$

Table value of  $\psi^2_{0.05}$  for 1 d.f. = 3.84.

The table value of  $\psi^2$  is 3.676, which is less than the table value = 3.84. Hence the hypothesis may be rejected. Thus we conclude that there is no significant difference between the new and conventional treatment.

**Example 10 :** A certain drug was administered to 456 males, out of a total 720 in a certain locality, to test the efficacy against typhoid. The incidence of typhoid is shown below. Find out the effectiveness of the drug against the disease. (The table value of chi square for 1 d.f. at 5% level of significance is 3.84).

	Infection	No infection	Total
Administering the drug	144	312	456
Without administering the drug	192	72	264
Total	336	384	720

**Solution :** Let the hypothesis be :

Null Hypothesis  $H_0$  : The drug is not effective in controlling the typhoid.

2 × 2 Contingency Table ( $O$ )

	Infection	No infection	Total
Administering the drug	144	312	456
Without administration of drug	192	72	264
Total	336	384	$N = 720$

Expected Frequency Table ( $E$ )

$\frac{336 \times 456}{720} = 212.8$	243.2	456
123.2	140.8	264
336	384	720

Computation table for  $\psi^2$ 

$O$	$E$	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
144	212.8	-68.8	4733.44	22.24
192	123.2	+68.8	4733.44	38.42
312	243.2	+68.8	4733.44	19.46
72	140.8	-68.8	4733.44	33.62
$\psi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 113.74$				

Computed value of  $\psi^2$  is 113.74, which, is much greater than the table value of the  $\psi^2_{0.05}$  at 1 d.f. = 3.841. Therefore, it is highly significant. The null hypothesis is rejected. Therefore, the drug is definitely effective in controlling the typhoid.

**Example 11 :** Calculate the expected frequencies for the following data, presuming the two attributes viz., conditions of home and condition of child as independent.

		Condition of home	
		Clean	Dirty
Condition of child	Clean	70	50
	Forty clean	80	20
	Dirty	35	45

Use chi-square test at 5% level of significance to state whether the two attributes are independent.

(Table value of  $\psi^2$  at 5% for d.o.f is 5.991 and for 3 d.o.f is 7.815 and 4 d.o.f is 9.488).

**Solution :** The expected frequency  $E$ , corresponding to each cell in the table is given by

$$E = \frac{R \times C}{n}$$

where  $R$  = a row total,  $C$  = a column total and  $n$  = the sample size.

We form the table of expected frequencies, with the help of the above rule and write the observed frequencies in each cell within brackets.

	Condition of home		Total
	Clean	Dirty	
Clean	(70)	(50)	120
	$\frac{185 \times 120}{300} = 74$	$\frac{115 \times 120}{300} = 46$	

Fairly clean	(80)	(20)	100
	$\frac{185 \times 100}{300} = 61.67$	$\frac{115 \times 100}{300} = 38.33$	
Dirty	(35)	(45)	80
	$\frac{185 \times 80}{300} = 49.33$	$\frac{115 \times 80}{300} = 30.67$	
Total	185	115	300

1. Null hypothesis  $H_0$  : There does not exist any association between the attributes.

2. Alternative Hypothesis  $H_1$  : An association exists between the attributes.

3. Calculation of test statistics

$$\begin{aligned}\psi^2 &= \sum \frac{(O - E)^2}{E} = \frac{(70 - 74)^2}{74} + \frac{(50 - 46)^2}{46} + \frac{(80 - 61.67)^2}{61.67} + \frac{(20 - 38.33)^2}{38.33} \\ &\quad + \frac{(35 - 49.33)^2}{49.33} + \frac{(45 - 30.67)^2}{30.67} \\ &= 0.2162 + 0.3478 + 5.4482 + 8.7657 + 4.1627 + 6.6954 = 25.636.\end{aligned}$$

Degree of freedom :  $v = (R - 1)(C - 1) = 2 \times 1 = 2$

Decision. The table value of  $\psi^2$  at  $\alpha = 0.05$  for 2 degrees of freedom is  $\psi^2_{0.05} = 5.991$ .

Also the calculated  $\psi^2 = 25.636 > \psi^2_{0.05}$  for 2 d.f. = 5.991.

$\Rightarrow$  the null hypothesis is rejected  $\Rightarrow$  Alternative hypothesis  $H_1$  is accepted from which we conclude that there exists an association between the attributes.

**Example 12 :** The following figures show the distribution of digits in number chosen at random from a telephone directory :

Digits : 0      1      2      3      4      5      6      7      8      9

Frequency : 1026    1107    997    966    1075    933    1107    972    964    853

Test at 5% level whether the digits may be taken to occur equally frequently in the directory.

(The table value of  $\psi^2$  for 9 degrees of freedom = 16.919).

**Solution :** 1. Null hypothesis  $H_0$  : The digits occur equally frequently in the directory.

Alternative Hypothesis  $H_1$  : The digits do not occur equally frequently.

2. Calculation of Test Statistic. Here the total number of frequencies

$$n = 1026 + 1107 + 997 + 966 + 1075 + 933 + 1107 + 972 + 964 + 853 = 10,000.$$

$\therefore$  Expected frequency for each of the digits 0, 1, 2, 3, ..., 9 is  $E = \frac{10,000}{10} = 1000$ . Thus

the value of  $\psi^2$  is

$$\begin{aligned}
 \psi^2 &= \frac{(1026 - 1000)^2}{1000} + \frac{(1107 - 1000)^2}{1000} + \frac{(997 - 1000)^2}{1000} + \frac{(966 - 1000)^2}{1000} \\
 &\quad + \frac{(1075 - 1000)^2}{1000} + \frac{(933 - 1000)^2}{1000} + \frac{(1107 - 1000)^2}{1000} + \frac{(972 - 1000)^2}{1000} \\
 &\quad + \frac{(964 - 1000)^2}{1000} + \frac{(853 - 1000)^2}{1000} \\
 &= 0.676 + 11.449 + 0.009 + 1.156 + 5.625 + 4.489 + 11.449 + 0.784 + 1.296 \\
 &\quad + 21.609 = \mathbf{58.542}.
 \end{aligned}$$

**3. Critical value.** The tabulated value or critical value of  $\psi^2$  at  $\alpha = 0.05$  for  $10 - 1 = 9$  degrees of freedom  $\psi_{0.05, 9}^2 = 16.919$ .

Since calculated value of  $\psi^2 = 58.542$  is  $>$  Tabulated  $\psi^2 = 16.919$ , so the null hypothesis  $H_0$  is rejected and the alternative hypothesis  $H_1$  is accepted  $\Rightarrow$  the digits do not occur equally frequently.

**Example 13 :** A dice is tossed 120 times with the following results :

Number turned up :	1	2	3	4	5	6	Total
Frequency :	30	25	18	10	22	15	120

Test the hypothesis that the dice is unbiased.

**Solution : 1. Null hypothesis  $H_0$  :** The dice is unbiased one.

**Alternative Hypothesis  $H_1$  :** The dice is a biased one.

**2. Calculation of Test Statistic.** On the hypothesis that the dice is unbiased, the expected frequency is  $120 \times \frac{1}{6} = 20$ . We calculate  $\psi^2$  from the following table.

$O$	$E$	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
30	20	10	100	5.00
25	20	5	25	1.25
15	20	-2	4	0.20
10	20	-10	100	5.00
22	20	2	4	0.20
15	20	-5	25	1.25
$\psi^2 = \sum \frac{(O - E)^2}{E} = 12.90$				

**3. Critical value.** The control value of  $\psi^2$  at  $\alpha = 0.05$  and for  $6 - 1 = 5$  degrees of freedom is  $\psi_{0.05, 5}^2 = 11.070$ .

**4. Decision.** Since the calculated value of  $\psi^2 = 12.90$  is  $> \psi_{0.05, 5}^2 = 11.07$ , so, the null hypothesis  $H_0$  is rejected and the alternative hypothesis  $H_1$  is accepted  $\Rightarrow$  the dice is biased one.

**Example 14 :** From the adult male population of seven large cities random sample giving  $2 \times 7$  contingency table of married and unmarried men, as given below were taken. Can it be said that there is a significant variation among the cities in the tendency of men to marry?

City	A	B	C	D	E	F	G	Total
Married	133	164	155	106	153	123	146	980
Unmarried	36	57	40	37	55	33	36	294
Total	169	221	195	143	208	156	182	1274

(At  $(2 - 1)(7 - 1)$  d.f. Take  $\chi^2_{0.05, 6} = 12.6$ ).

**Solution : 1.** Null hypothesis  $H_0$  : There is no significant variation among the cities in the tendency of men to marry.

Alternative hypothesis  $H_1$  : There is a significant variation among the cities in the tendency of men to marry.

**2. Calculation of test statistic.** On the basis of null hypothesis the expected frequencies are :

$$\text{Expected number of married people in City A} = \frac{980}{1274} \times 169 = 130$$

$$\text{Expected number of married people in City B} = \frac{980}{1274} \times 221 = 170$$

$$\text{Expected number of married people in City C} = \frac{980}{1274} \times 195 = 150$$

$$\text{Expected number of married people in City D} = \frac{980}{1274} \times 143 = 110$$

$$\text{Expected number of married people in City E} = \frac{980}{1274} \times 208 = 160$$

$$\text{Expected number of married people in City F} = \frac{980}{1274} \times 156 = 120$$

$$\text{Expected number of married people in City G} = \frac{980}{1274} \times 182 = 140$$

Similarly, the expected frequencies of unmarried are 39, 51, 45, 33, 48, 36 and 42.

Table for expected frequencies

Married	130	170	150	110	160	120	140
Unmarried	39	51	45	33	48	36	42
Total	169	221	195	143	208	156	182

Table for calculation of  $\psi^2$ 

$O$	$E$	$O - E$	$(O - E)^2$	$\sum \frac{(O - E)^2}{E}$
133	130	3	9	0.069
164	170	-6	36	0.212
155	150	5	25	0.167
106	110	-4	16	0.145
153	160	-7	49	0.306
123	120	3	9	0.075
146	140	6	36	0.257
36	39	-3	0	0.231
57	51	6	36	0.706
40	45	-5	25	0.556
37	33	4	16	0.485
55	48	7	49	1.021
33	36	-3	9	0.250
36	42	-6	36	0.857
$\therefore \psi^2 = \sum \frac{(O - E)^2}{E} = 5.337$				

3. Level of significance : Take  $\alpha = 0.05$ .

4. Critical value : The critical value or table value of  $\psi^2$  at  $\alpha = 0.05$  for (2 1) (7 1) = 6 degrees of freedom is  $\psi^2_{0.05, 6} = 12.6$ .

5. Decision : Since the calculated of  $\psi^2 = 5.337 <$  critical value  $\psi^2_{0.05}$  for 6 d.f. = 12.6, so the null hypothesis is accepted  $\Rightarrow$  that there is no significant variation among the cities in the tendency of men to marry.

Example 15 : 50 students selected at random from 500 students enrolled in a computer crash programme were classified according to age and grade points giving the following data.

Grade points	20 and under	Age (in years) 21 – 30	Above 30
Upto 5.0	3	5	2
5.1 to 7.5	8	7	5
7.6 to 10.0	4	8	8

Test at 5% level of significance the hypothesis that age and grade point are independent.

**Solution :** Null Hypothesis :  $H_0$  : Age and grade points are independent.

Alternative Hypothesis :  $H_1$  : Age and grade points are not independent.

On the basis of null hypothesis  $H_0$ , the expected frequencies obtained are:

Grade points	20 and under	21 – 30	Above 30	Total
Upto 5	3	4	3	10
5.1 to 7.5	6	8	6	20
7.6 to 10.0	6	8	6	20
Total	15	20	15	50

Since the values which are less than 5 are occurring in some cells of the expected frequencies, we have to amalgamate these cells to its neighbours. After amalgamation, the new frequencies are:

**Observed frequencies  
(Age in years)**

Grade points	20 and under	21 – 30	Above 30	Total
Up to 7.5	11	12	7	30
7.6 to 10	4	8	8	20
Total	15	20	15	50

**Observed frequencies  
(Age in years)**

Grade points	20 and under	21 – 30	Above 30	Total
Up to 7.5	9	12	9	30
7.6 to 10	6	8	6	20
Total	15	20	15	50

#### Calculation of Test Statistic :

$$\text{Here : } \chi^2 = \frac{(11-9)^2}{9} + \frac{(12-12)^2}{12} + \frac{(7-9)^2}{9} + \frac{(4-6)^2}{6} + \frac{(8-8)^2}{8} + \frac{(8-6)^2}{8} = 2.222.$$

**Decision :** The table value of  $\chi^2$  at  $\alpha = 0.05$  for  $1 \times 2 = 2$  degrees of freedom is  $\chi^2_{0.05} = 5.991$ . Since the computed value of  $\chi^2 = 2.222$  is less than the table value of  $\chi^2_{0.05}$  for 2 d.f. = 5.991, so we accept the null hypothesis  $H_0 \Rightarrow$  that the grade points and age are independent of each other.

**Example 16 :** In the accounting department of bank 100 accounts are selected at random and examined for errors. The following results have been obtained.

No. of Errors :      0      1      2      3      4      5      6

No. of Accounts :    36     40     19     2      0      2      1

Does this information verify that the errors are distributed according to Poisson probability law?

**Solution : 1. Null hypothesis  $H_0$  :** The errors are distributed according to Poisson probability law.

**Alternative hypothesis  $H_1$  :** The errors are not distributed according to Poisson probability law.

**2. Computation of test statistics :**

Mean of the given data =  $m = \sum (f \times x)/N$

$$= \frac{0 \times 36 + 1 \times 40 + 2 \times 19 + 3 \times 2 + 4 \times 0 + 5 \times 2 + 6 \times 1}{36 + 40 + 19 + 2 + 0 + 2 + 1} = \frac{100}{100} = 1$$

Also  $e^{-m} = 0.367$ .

Poisson probability distribution =  $P(x) = \frac{e^{-m} m^x}{x!}$ ,  $x = 0, 1, 2, 3, \dots$

$\therefore$  Expected frequencies =  $N \times P(x) = N \times \frac{e^{-m} m^x}{x!}$ ,  $x = 0, 1, 2, 3, \dots$

$$= \frac{0.367 \times 100 \times 1^x}{x!}, x = 0, 1, 2, 3, \dots$$

The expected probabilities for the various errors are :

$$\frac{0.367 \times 100 \times 1^0}{0!}, \frac{0.367 \times 100 \times 1^1}{1!}, \frac{0.367 \times 100 \times 1^2}{2!}, \frac{0.367 \times 100 \times 1^3}{3!}$$

$$\frac{0.367 \times 100 \times 1^4}{4!}, \frac{0.367 \times 100 \times 1^5}{5!}, \frac{0.367 \times 100 \times 1}{6!}, \text{ i.e., } 37, 37, 18, 6, 2, 0, 0.$$

We have the following table :

Computation table for  $\chi^2$

No. of errors	Observed frequency $O$	Expected frequency $E$	$O - E$	$\frac{(O - E)^2}{E}$
0	36	37	-1	0.027
1	40	37	3	0.243
2	19	18	1	0.055
3	2	6		
4	0	2		
5	2	0		
6	1	0	-3	1.125
Total	100	100		$\chi^2 = \sum \frac{(O - E)^2}{E} = 1.450$

**Note :** In the above data, the frequency for each of 3, 4, 5 and 6 errors is less than 5, so the frequencies for these errors have been pooled together in order to make the total = 5 and  $E = 8$ .

$\therefore$  the test statistic  $\psi^2 = 1.450$ .

**4. Critical value :** The table value of  $\psi^2$  for  $4 - 1 = 3$  degrees of freedom at  $\alpha = 0.05$  is  $\psi^2_{0.05, 3} = 7.815$ .

**5. Decision :** The calculated value of  $\psi^2 = 1.450$  is less than the table value of  $\psi^2_{0.05, 3} = 7.815$  so the null hypothesis  $H_0$  is accepted  $\Rightarrow$  that the errors are distributed according to Poisson probability law.

## 14.9 $\psi^2$ DISTRIBUTION OF SAMPLE VARIANCE

Let  $\sigma^2$ ,  $s^2$  respectively be the population variance and sample variance. The sampling distribution  $(n - 1)s^2 / \sigma^2$  has a **chi-square ( $\psi^2$ ) distribution** with  $n - 1$  degrees of freedom. It is a very useful distribution in making inference about the population variance  $\sigma^2$  by using sample variance  $s^2$ . It is used in making interval estimate of the population variance, which is given by the result.

$$\frac{(n - 1)s^2}{\psi^2_{\alpha/2}} \leq \alpha < \frac{(n - 1)s^2}{\psi^2_{(1 - \alpha/2)}}$$

where  $\alpha$  is the level of significance and  $\psi^2_{\alpha/2}$  = value for the chi-square distribution giving an area to the right of stated  $\psi^2$  and  $(1 - \alpha)$  confidence coefficient.

It can also be used as hypothesis test for the value of the population variance.

**Example 18 :** A market analyst took a sample of 25 markets in Bangalore in an attempt to determine how much variation is there in the butter prices. The 25 prices that were quoted to him for four samples of butter yielded the sample value  $\bar{x} = 90$  ans  $s = 7$ . The problem now is to find a 95% confidence interval for the standard deviation of all the markets.

**Solution :** Here  $n = 25$ ,  $s = 7 \Rightarrow \sigma^2 = 49$ .

Now  $\alpha = 0.05$ , so  $\psi^2_{\alpha/2}$  for  $(n - 1)$  d.f. is  $\psi^2_{0.025}$  for  $(25 - 1) = 24$  d.f. = 39.36 and  $\psi^2_{.975}$  for 24 d.f. = 12.40 (from  $\psi^2$ -table).

$\therefore$  Required confidence interval is

$$\frac{(n - 1)s^2}{\psi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n - 1)s^2}{\psi^2_{(1 - \alpha/2)}} = \frac{24 \times 49}{39.36} \leq \sigma^2 \leq \frac{24 \times 49}{12.40}$$

or  $\frac{1176}{39.36} \leq \sigma^2 \leq \frac{1176}{12.40} \Rightarrow 28.88 \leq \sigma^2 \leq 94.84$ .

**Example 19 :** A sample of 30 light tubes yielded a standard deviation of 90 hours burning time whereas the long experience with the particular brand showed deviation of 105 hours. Using  $\alpha = 0.05$ , test if there is any difference in the standard deviation.

**Solution :** Null hypothesis :  $H_0 : \sigma_0 = 105$ , i.e., there is no difference in standard deviation.

**Alternate hypothesis :**  $H_1 : \sigma_0^2 \neq 105.$

Degrees of freedom :  $v = 30 - 1 = 29.$

$$\begin{aligned}\text{Test statistic : } \psi^2 &= \frac{(n-1)s^2}{\sigma_0^2} = s = \frac{29 \times (90)^2}{(105)^2} \\ &= \frac{234900}{11025} = 21.31.\end{aligned}$$

**Critical value :** The table value of  $\psi_{0.025}^2$  for 29 d.f. = 45.72 and  $\psi_{0.975}^2$  for 29 d.f. = 16.05.

**Decision :** The calculated value of  $\psi^2 = 21.31$  is less than tabled value  $\psi_{0.025}^2$  for 29 d.f. = 45.72, so the null hypothesis is accepted.

#### 14.10 TESTING A HYPOTHESIS ABOUT THE VARIANCE OF A NORMALLY DISTRIBUTED POPULATION

The decision rule for **One tailed test** as well as **Two tailed test** is stated below:

The one tailed test with null hypothesis  $H_0 : \sigma^2 \geq \sigma_0^2$  is similar to the other one tailed tests mentioned above with the exception that the one tailed rejection region appears in the lower tail at a critical value of  $\psi_{(1-\alpha)}^2$ . Also  $\sigma_0^2$  = hypothesized value for population variance.

One Tailed Test		Two Tailed Test	
Hypothesis	Decision Rule	Hypothesis	Decision Rule
1. $H_0 : \sigma^2 \geq \sigma_0^2$	Accept $H_0$ if computed $\psi^2 < \text{table } \psi_\alpha^2$	1. $H_0 : \sigma^2 = \sigma_0^2$	Accept $H_0$ if computed $\psi^2 \leq \psi_{\alpha/2}^2$
2. $H_1 : \sigma^2 < \sigma_0^2$	Reject $H_0$ if computed $\psi^2 > \text{tabled } \psi_\alpha^2$	2. $H_1 : \sigma^2 \neq \sigma_0^2$	Reject $H_1$ if computed $\psi^2 > \psi_{\alpha/2}^2$

**Example 20 :** Heights in cms of 10 students are given below :

60, 65, 62, 66, 68, 70, 65, 64, 68, 72

Can we say that variance of the distribution of heights of all students from which the above sample of 10 students was drawn is equal to 50.

**Solution :** Mean of the given sample is given by:

$$\begin{aligned}\bar{x} &= \frac{60 + 65 + 62 + 66 + 68 + 70 + 65 + 64 + 68 + 72}{10} \\ &= \frac{660}{10} = 66.\end{aligned}$$

Calculation table for  $s^2$ 

$x_i$	$x_i - \bar{x} = (x_i - 66)$	$(x_i - \bar{x})^2 = (\bar{x}_i - 66)^2$
60	-6	36
65	-1	1
62	-4	16
66	0	0
68	2	4
70	4	16
65	-1	1
64	-2	4
68	2	4
72	6	36
$\Sigma x_i = 660$		$\Sigma (x_i - \bar{x})^2 = 116$

$$\therefore \text{Sample variance } s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{116}{9}.$$

Null hypothesis  $H_0 : \sigma^2 = \sigma_0^2 = 50$ .

Alternative hypothesis  $H_1 : \sigma^2 \neq 50$  (Two Tailed Test)

Calculation of test statistic : The test statistic is given by

$$\Psi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

and the population height is assumed to be approximately normally distributed.

$$\begin{aligned} \therefore \Psi^2 &= \frac{(n-1)s^2}{\sigma_0^2} = \frac{(n-1)}{\sigma_0^2} \times \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum (x_i - \bar{x})^2}{\sigma_0^2} \\ &= \frac{116}{50} = 2.32. \end{aligned}$$

Level of significance : Let the level of significance be  $\alpha = 0.05$ .

Statement of decision rule : At 5% level of significance and  $10 - 1 = 9$  d.f.

$$\Psi^2_{0.025} = 19.02 \text{ and } \Psi^2_{0.975} = 2.70.$$

Now reject  $H_0$  if computed  $\Psi^2 > \Psi^2_{\alpha/2}$  or computed  $\Psi^2 < \Psi^2_{(1-\alpha/2)}$ .

Making a statistical decision : Here computed  $\Psi^2 = 2.32 < \Psi^2_{0.975} = 2.70$  so the null hypothesis is rejected and we conclude that the population variance may not be 50.

Example 20 : Specification for the manufacture of a particular type of ornament state that the variance in weight shall not exceed 0.0015 mgm squared. A random sample of 15 such ornaments yields a variance of 0.027. Can we say that the specifications are not being met? Use  $\alpha = 0.05$ .

**Solution :** Null hypothesis :  $H_0 : \sigma^2 = \sigma_0^2 \leq 0.015$ .

Alternative hypothesis :  $H_1 : \sigma^2 > 0.015$ . (One Tailed Test)

**Calculation of test statistic :** The test statistic

$$\psi^2 = \frac{(n - 1)s^2}{\sigma_0^2} = \frac{(15 - 1) 0.027}{0.015} = 25.20.$$

**Level of significance :** Let the level of significance be  $\alpha = 0.05$ .

**Critical value :** The table value of  $\psi^2$  for 19 d.f. = 23.685 at  $\alpha = 0.05$ .

**Decision :** Since computed value  $\psi^2 = 25.20 >$  Table value of  $\psi_{0.05}^2 = 23.685$  so the null hypothesis  $H_0$  is rejected the data indicate that the specifications are not being met.

### EXERCISE

1. A sample of 300 students of Under-graduate and 300 students of Post-Graduate classes of a University were asked to give their opinion toward the autonomous colleges. 190 of the Under-Graduate and 210 of the Post-Graduate students favoured the autonomous status.

Present the above data in the form of frequency table and test of 5% levels, the opinion of Under-graduate and Post graduate-students on autonomous status of colleges are independent (Table value of  $\psi^2$  at 5% level of 1 d.f. is 3.84)

[Hint : The contingency table of observed frequencies in which expected frequencies are shown in braces is

	Favour	Not in favour	Total
Under Graduate :	190 (200)	110 (100)	300
Post Graduate :	210 (200)	90 (100)	300
	400	200	600

**Null hypothesis :**  $H_0$  : Opinion on autonomous status and level of graduation are independent.

$$\psi^2 = \frac{100}{200} + \frac{100}{200} + \frac{100}{100} + \frac{100}{100} = 3.$$

$$\text{d.f.} = (2 - 1)(2 - 1) = 1.$$

Table value  $\psi^2$  at  $\alpha = 0.05$  and for 1 d.f. is  $\psi_{0.05}^2 = 3.84$ .

Now calculated  $\psi^2 = 3 < \psi_{0.05}^2 = 3.84 \Rightarrow$  Null hypothesis is accepted.

**Opinion on autonomous status and level of Graduation are independent.**

2. A certain drug is claimed to be effective in curing colds. In an experiment of 164 people with cold, half of them were given the drug and half of them given sugar pills. The patient's reactions to the treatment are recorded in the following table. Test the hypothesis that the drug is not better than sugar pills for curing colds.

	Helped	Harmed	No effect
Drug :	52	10	20
Sugar pills :	44	12	26

3. In a survey of 200 boys, of which 75 were intelligent, 40 had skilled fathers, while 85 of the unintelligent boys had unskilled fathers. Do these support the hypothesis that skilled fathers have intelligent boys. Use  $\chi^2$  test. Values of  $\chi^2$  for 1-degree of freedom at 5% level is 3.84.

4. Four dice were thrown 112 times and the number of times 1, 3 or 5 were as under:

Number of dice showing 1, 3 or 5 : 0      1      2      3      4

Frequency :                            10      25      40      30      7

[Hint : Let the probability of showing 1, 3 or 5 be the success =  $\frac{3}{6} = \frac{1}{2}$ .

Probability of  $x$  successes in throw of 4 dice 112 times =  $112^4 C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x} = 112^4 C_x \left(\frac{1}{2}\right)^4$   
;  $x = 0, 1, 2, 3, 4$ .

No. of days	Frequency			
	Observed <i>O</i>	Expected <i>E</i>	$(O - E)^2$	$\frac{(O - E)^2}{E}$
0	10	7	9	1.286
1	25	28	9	0.391
2	40	42	4	0.025
3	30	28	4	0.143
4	7	7	0	0
Total	112	112	$\sum \frac{(O - E)^2}{E} = 1.845$	

Calculated  $\chi^2 = 1.845$ ,  $\chi^2_{0.05}$  for  $5 - 1 = 4$  d.f. = 9.488.

Now calculate  $\chi^2 < \chi^2_{0.05} \Rightarrow$  All the four dice are fair].

5. To test the efficiency of a new drug a controlled experiment was conducted where in 300 patients were administered the new drug and 200 other patients were not given the drug. The patients were monitored and results were obtained as follows:

	Cured	Condition worsens	No effect	Total
Given the drug	200	40	60	300
Not given the drug	120	30	50	200
Total	320	70	110	500

Use  $\chi^2$  test for finding the effect of the drug.

6. The following table gives the classification of 100 workers according to sex and the nature of work. Test whether the nature of work is independent of the sex of the work.

	<i>Skilled</i>	<i>Unskilled</i>
<i>Male :</i>	40	20
<i>Female :</i>	10	30

7. The following data relate to the sales in a time of trade depression of a certain commodity demand:

<i>District where Sales are</i>	<i>District not hit by depression</i>	<i>Districts hit by depression</i>	<i>Total</i>
<i>Satisfactory</i>	250	80	330
<i>Not satisfactory</i>	140	30	170
<i>Total</i>	390	110	500

Do these data suggest that the sales are significantly affected by depression?

[Hint : Let the hypothesis be that the sales are not significantly affected by depression.

<i>O</i>	<i>E</i>	<i>O – E</i>	$(O - E)^2$	$\frac{(O - E)^2}{E}$
250	274.4	-7.4	54.76	0.213
140	132.6	+7.4	54.76	0.413
80	72.6	+7.4	54.76	0.754
30	37.4	-7.4	54.76	1.464

$$\Psi^2 = \sum \frac{(O - E)^2}{E} = 2.844$$

The calculated value 2.844 is less than the table value of  $\Psi^2_{0.05}$  for 1 d.f. = 3.841. Hence the hypothesis may be accepted. Thus we conclude that sales are not significantly affected by depression.

8. 4 coins were tossed 160 times and the following results were obtained:

<i>No. of heads :</i>	0	1	2	3	4
<i>Observed frequencies :</i>	17	52	54	31	6

Under the assumption that coins are balanced, find the expected frequencies of getting 0, 1, 2, 3 or 4 heads and test the goodness of fit.

[Hint : Hypothesis  $H_0$  : That the coins are unbiased. The expected frequencies are given by the formula :  $160 \times {}^4C_x (0.5)^4$ ,  $x = 0, 1, 2, 3, 4$ . These are 10, 40, 60, 40 and 10.

$$\begin{aligned}\therefore \Psi^2 &= \sum \left[ \frac{(O - E)^2}{E} \right] \\ &= \frac{(17 - 10)^2}{10} + \frac{(52 - 40)^2}{40} + \frac{(54 - 60)^2}{40} + \frac{(31 - 40)^2}{60} + \frac{(6 - 10)^2}{10} = 12.725\end{aligned}$$

Also  $\Psi^2_{0.05}$  for 1 d.f. = 9.488

Calculated value of  $\psi^2$  is 12.725 which is greater than the tabled value 9.488. Therefore it is a poor fit.]

9. In a certain sample of 2,000 families, 1,400 families are consumers of tea. Out of 1,800 Hindu families, 1,236 families consume tea. Use Chi-square test and state whether there is any significant difference between consumption of tea among Hindu and Non-Hindu families.

10. A survey of 200 families having three children selected at random gave the following results:

<i>Male births :</i>	0	1	2	3
<i>No. of families :</i>	40	58	62	40

Test the hypothesis that male and female births are equally likely at 5% level of significance.

[Hint : Here O = 40, 50, 60, 40 and E = 50, 50, 50, 50,

$$\therefore \psi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = 8.16. \text{ Also } \psi^2_{0.05} \text{ for 1 d.f.} = 7.82.$$

The calculated value of  $\psi^2$  is 8.16 which is greater than the table value of  $\psi^2_{0.05}$  for 1 d.f. = 7.82. Hence the hypothesis may be rejected. Thus we conclude that the male and female births are not equally likely at 5% level of significance].

11. Out of 8,000 graduates in a city, 800 are females; out of 1600 graduates employees 120 are female. Use chi-square test to determine if any distinction is made in appointment on the basis of sex. Value of chi-square for 5% level for one degree of freedom is 3.84.

[Hint : Hypothesis : There is no distinction in appointment on the basis of sex.

$$\therefore \psi^2 = \sum \left[ \frac{(O - E)^2}{E} \right] = \frac{(1480 - 1440)^2}{1440} + \frac{(120 - 160)^2}{160} + \frac{(5720 - 5760)^2}{5760} + \frac{(680 - 640)^2}{640} \\ = 13.88.$$

Since  $\psi^2_{0.05} = 13.88$  which is more than  $\psi^2_{0.05}$  for 1 d.f. (= 3.84), so the hypothesis is rejected. We conclude that there is a distinction in appointment on the basis of the sex].

12. From the following information, state whether the two attributes, i.e., condition of house and condition of child are independent.

<i>Condition of child</i>	<i>Condition of house</i>	
	<i>Clean</i>	<i>Dirty</i>
<i>Clean</i>	69	51
<i>Fairly clean</i>	81	20
<i>Dirty</i>	35	44

13. Two groups of 100 people each were taken for testing the use of a vaccine. 15 persons contracted the disease out of the inoculated person, while 25 contracted the disease in the other group. Test the efficiency of vaccine using  $\psi^2$ -test.
14. There is a general belief that high income families send their children to public schools and

low income families send their children to Government schools. For this, 1000 families were selected in a city and the following results were obtained.

<i>Income</i>	<i>Public Schools</i>	<i>Govt. Schools</i>	<i>Total</i>
<i>Low</i>	100	200	300
<i>High</i>	500	200	700
<i>Total</i>	600	400	1000

Use  $\chi^2$  test to determine whether income level and the type of schooling were associated.

15. A market analyst took a sample of 20 markets of Mumbai to determine how much variation is there in bread prices. The 20 prices that were quoted to him for the four samples of the bread yielded the sample values  $\bar{x} = 60$  and  $s = 9$ . Find the 95% confidence interval for the standard deviation of the entire market.
16. A sample 101 light bulbs yielded a standard deviation of 80 hours burning time, whereas long experience with the particular brand showed a standard deviation of 90 hours. Using  $\alpha = 0.05$  test if there is any difference in the standard deviation.
17. Weights in kg of 10 students are given as 38, 40, 45, 53, 47, 43, 55, 48, 52, 49. Can we say that variance of distribution of weights of all students from which the above sample of 10 students was drawn, is equal to 20 kgs?

## ANSWERS

1. Opinion on autonomous status and level of graduations are independent.
2. Computed  $\chi^2 = 1.622$ ,  $\chi^2_{0.05} = 5.991$ , Drug is not better than sugar pills.
3.  $\chi^2 = 0.88$ ,  $\chi^2_{0.05} = 1$  d.f. = 3.84; Skilled fathers have intelligent boys.
4. All the four dice are fair.
5.  $\chi^2 = 2.434$ ,  $\chi^2_{0.05} = 5.99$ ; Drug is not effective.
6.  $\chi^2 = 16.666$ ,  $\chi^2_{0.05} = 3.84$ . The nature of work does not seem to be independent of the sex of the worker.
7. Yes. Expected frequencies are 10, 40, 60, 40, 10. It is a poor fit.
9. The two communities differ significantly as far as consumption of tea is concerned.
10. Male and female births are not equally likely.
11. The distinction is made in appointment on the basis of sex.
12. There is an association between the condition of child and the condition of house.
13.  $\chi^2 = 3.125$ ,  $\chi^2_{0.05} = 3.84$ ; Inoculation is effective in contracting the disease.
14. The level and type of schooling are associated.
15.  $6.84 \leq \sigma^2 < 13.4$
16. There is no difference in standard deviation.
17. The population variance is not 20.



# 15

## *F-Test or Fisher's F-Test*

### 15.1 F-TEST OR FISHER'S F-TEST

The *F*-test was first originated by the statistician R.A. Fisher. This test is also known as **Fisher's *F*-test or simply *F*-test.** It is based on the *F*-distribution, which is defined as the ratio of two independent chi-square variates which is derived by dividing each variable by its corresponding degree of freedom, i.e.,

$$F = \frac{\Psi^2}{v_1} \div \frac{\Psi^2}{v_2}$$

The *F*-test refers to a test of hypothesis concerning two variances derived from two samples. The various tests of hypothesis discussed previously are not suitable for the test of hypothesis concerning two or more sample variances.

### 15.2 F-STATISTIC

The *F*-statistic is the ratio of independent estimates of population variances and expressed as

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \text{ with } n_1 - 1 \text{ degree of the freedom for the numerator and } n_2 - 1 \text{ degree of}$$

freedom for the denominator, where  $\hat{\sigma}_1^2$ ,  $\hat{\sigma}_2^2$  are the unbiased estimate on the basis of sample variances  $s_1^2$  and  $s_2^2$  given by the population variances based on sample sizes  $n_1$  and  $n_2$ , where

$$\hat{\sigma}_1^2 = \frac{n_1 s_1^2}{n_1 - 1}, \quad \hat{\sigma}_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$$

Generally,  $\hat{\sigma}_1^2$  is greater than  $\hat{\sigma}_2^2$ , but if  $\hat{\sigma}_2^2$  is greater than  $\hat{\sigma}_1^2$ , then in such cases the two variances should be interchanged so that the value of *F* is always greater than 1.

### 15.3 ASSUMPTIONS IN F-TEST

*The F-test is based on the following assumptions:*

1. **Normality** : The values in each group should be normally distributed.
2. **Independence of error** : The variation of each value around its own group mean, i.e., error should be independent of each value.
3. **Homogeneity** : The variance within each group should be equal for all groups.

i.e.,  $\sigma^2 = \sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2$ .

*If however, the sample sizes are large enough, we don't need the assumption of normality.*

### 15.4 TESTS OF HYPOTHESIS ABOUT THE VARIANCE OF TWO POPULATIONS

We have the following steps:

1. **Null Hypothesis** :  $H_0 : \hat{\sigma}_1^2 = \hat{\sigma}_2^2$ .

**Alternate Hypothesis** :  $H_1 : \hat{\sigma}_1^2 \neq \hat{\sigma}_2^2$  (Two Tailed Test).

2. **Calculation of test statistic** : Calculate the  $F$ -statistic as given below.

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \quad \text{if } \hat{\sigma}_1^2 \geq \hat{\sigma}_2^2 \text{ so that } F \geq 1$$

or  $F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} \quad \text{if } \hat{\sigma}_2^2 \geq \hat{\sigma}_1^2 \text{ so that } F \geq 1$

3. **Level of significance** : Take the level of significance  $\alpha = 0.05$  if  $\alpha$  is not known.

4. **Decision** : Accept  $H_0$  if computed  $F \leq$  tabled  $F_\alpha$ .

Reject  $H_0$  if computed  $F >$  tabled  $F_\alpha$ .

#### Steps for One Tailed Test

1. Set the Null hypothesis :  $H_0 : \hat{\sigma}_1^2 = \hat{\sigma}_2^2$ .

Alternative hypothesis :  $H_1 : \hat{\sigma}_1^2 > \hat{\sigma}_2^2$  (One Tailed Test)

**Note** : For one tailed test  $H_0$  is set in such a way that the rejection region appears in the upper tail. This is done by judiciously numbering the population variances, so that the alternative hypothesis takes the form :  $H_1 : \hat{\sigma}_2^2 > \hat{\sigma}_1^2$ . If the  $H_0$  is of the form  $\hat{\sigma}_2^2 = \hat{\sigma}_1^2$ ,

then we compute  $F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}$  which has  $F$  distribution, but with  $n_2 - 1$  d.f. in the numerator and  $n_1 - 1$  d.f. in the denominator.

2. **Computation of test statistic** : Calculate the  $F$ -statistic  $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$  if  $\hat{\sigma}_1^2 > \hat{\sigma}_2^2$ .

3. **Level of significance** : Set the level of significance  $\alpha = 0.05$  if no value of  $\alpha$  is given.

4. **Decision** : If computed  $F <$  tabled  $F_{\alpha/2}$ , then accept the null hypothesis  $H_0$ . Or Reject  $H_0$  if computed  $F >$  tabled  $F_{\alpha/2}$ .

A Summary of procedure is given below:

Null hypothesis $H_0$	Alternative hypothesis: $H_1$	Type of Test	Rejection	Decision Rule
$H_0 : \hat{\sigma}_1^2 = \hat{\sigma}_2^2$	$H_1 : \hat{\sigma}_1^2 \neq \hat{\sigma}_2^2$	Two tailed	Both tails	Accept $H_0$ if $F_{\text{cal}} > F_{\alpha/2}$ otherwise accept $H_1$ .
$H_0 : \hat{\sigma}_1^2 \geq \hat{\sigma}_2^2$	$H_1 : \hat{\sigma}_1^2 < \hat{\sigma}_2^2$	One tailed	Right tails	Accept $H_0$ if $F_{\text{cal}} \geq F_\alpha$ .

For one tailed test  $H_0$  is set in such a way that the rejection region appears in the upper tail. This is done by judiciously numbering of populations variances, so that alternative hypothesis takes the form  $H_1 : \hat{\sigma}_1^2 < \hat{\sigma}_2^2$ . If  $H_1$  is of the form  $\hat{\sigma}_1^2 < \hat{\sigma}_2^2$ , we compute the ratio  $(\hat{\sigma}_2^2 / \hat{\sigma}_1^2)$  (instead of  $\hat{\sigma}_1^2 / \hat{\sigma}_2^2$ , which also has a  $F$  distribution, but with  $n_2 - 1$  d.o.f. for the numerator and  $n_1 - 1$  d.o.f. for the denominator. The critical value in the left tail of  $F$  distribution is found from the table of critical values in the right tail of  $F$  distribution.

**Example 1 :** The standard deviations calculated from two random samples of blood for their total lipid profile tests of sizes 9 and 13 are 2.1 and 1.8 respectively. May the samples be regarded as drawn from the normal populations with the same standard deviation. The value of  $F$  from the table with degrees of freedom 8 and 12 is 2.85.

**Solution :** With usual notations, we have  $n_1 = 9$ ,  $s_1 = 2.1$ ,  $n_2 = 13$ ,  $s_2 = 1.8$

$$\hat{\sigma}_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{9 \times (2.1)^2}{8} = 4.96.$$

$$\hat{\sigma}_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{13 \times (1.8)^2}{12} = 3.51.$$

Null hypothesis :  $H_0 : \hat{\sigma}_1^2 = \hat{\sigma}_2^2$

Alternative hypothesis :  $H_1 : \hat{\sigma}_2^2 \neq \hat{\sigma}_1^2$  (Two Tailed Test)

$$\therefore F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{4.96}{3.51} = 1.41.$$

**Decision :** Since calculated  $F = 1.41 <$  tabled  $F$  for 8 and 12 d.f. = 2.85, so we except the null hypothesis  $H_0 \Rightarrow$  the samples may be regarded as drawn from the normal population with same standard deviation.

**Example 2 :** Two independent samples of sizes 9 and 8 gave the sum of squares of deviations from their respective means as 160 and 91 respectively. Can the samples be regarded as drawn from the normal populations with equal variance? Given :  $F_{0.05}(8, 7) = 3.73$ ;  $F_{0.05}(7, 8) = 3.50$ .

**Solution :** We set  $H_0 : (\sigma_1^2 = \sigma_2^2)$ .  $H_1 : (\sigma_1^2 \neq \sigma_2^2)$

$$\text{Also, } s_1^2 = \frac{\sum (x - \bar{x})^2}{n_1 - 1} = \frac{160}{8} = 20; \quad s_2^2 = \frac{\sum (y - \bar{y})^2}{n_2 - 1} = \frac{91}{7} = 13.$$

$$\therefore F_{\text{cal}} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

Now,  $\hat{\sigma}_1^2 = \frac{n_1}{n-1} s_1^2 = \frac{9}{9-1} \times 20 = 45$

$$\hat{\sigma}_2^2 = \frac{n_2}{n_2-1} s_2^2 = \frac{8}{8-7} \times 13 = \frac{104}{7}$$

$$\therefore F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{45}{(104/7)} = \frac{45 \times 7}{104} = \frac{315}{104} = 3.029$$

Here  $F_{\text{cal}} = 3.029 < F_{0.05} (8.7) = 3.73$  (given)

$\Rightarrow H_0$  is accepted  $\Rightarrow$  that the samples be regarded as drawn from the normal population with equal variance.

**Example 3 :** Random samples are drawn from two populations and the following results were obtained regarding their blood cholestrokes.

Sample X : 16 17 18 19 20 21 22 24 26 27

Sample Y : 19 22 23 25 26 28 29 30 31 32 35 36

Find the variance of two populations and test whether the two samples have same variance.

**Solution :** Null hypothesis :  $H_0 : \hat{\sigma}_x^2 = \hat{\sigma}_y^2$ , i.e., the two samples have the same variance.

Alternative hypothesis :  $H_1 : \hat{\sigma}_x^2 \neq \hat{\sigma}_y^2$  (Two Tailed Test)

**Calculation of test statistic :** We have the following table to evaluate  $\hat{\sigma}_x^2$  and  $\hat{\sigma}_y^2$ .

Computation Table

x	$x - \bar{x}$ $= x - 21$	$(x - \bar{x})^2 =$ $(x - 21)^2$	y	$y - \bar{y}$ $= y - 28$	$(y - \bar{y})^2$ $= (y - 28)^2$
16	-5	25	19	-9	81
17	-4	16	22	-6	36
18	-3	9	23	-5	25
19	-2	4	25	-3	9
20	-1	1	26	-2	4
21	0	0	28	0	0
22	1	1	29	1	1
24	3	9	30	2	4
26	5	25	31	3	9
27	6	36	32	4	16
			35	7	49
			36	8	64
$\Sigma x = 210$		$\Sigma (x - \bar{x})^2 = 126$	$\Sigma y = 336$		$\Sigma (y - \bar{y})^2 = 298$

$$\text{Now, } \bar{x} = \frac{\Sigma x}{n} = \frac{210}{10} = 21, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{336}{12} = 28.$$

$$\therefore \hat{\sigma}_1^2 = \frac{\sum (x - \bar{x})^2}{n_1 - 1} = \frac{126}{9} = 14$$

$$\hat{\sigma}_2^2 = \frac{\sum (y - \bar{y})^2}{n_2 - 1} = \frac{298}{11} = 27.09.$$

$$\text{Test Statistic } F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{27.09}{14} = 1.94.$$

**Critical value :** The tabled value of  $F$  at  $\alpha = 0.05$  for 14 and 9 degrees of freedom is

$$F_{0.05} = 2.6458.$$

**Decision :** The compound value of  $F = 1.94 <$  Table value  $F_{0.05} = 2.6458$

$\Rightarrow$  the null hypothesis  $H_0$  is accepted  $\Rightarrow$  The two samples have the same variance.

**Example 4 :** In a laboratory experiment, two samples blood gave the following results:

Sample	Size	Sample mean	Sum of the squares of deviation from the mean
1	10	15	90
2	12	14	108

Test the equality of sample variances at 5% level of significance.

**Solution :** We set the Null Hypothesis :  $H_0 : \sigma_1^2 = \sigma_2^2$ .

Alternate Hypothesis :  $H_1 : \sigma_1^2 \neq \sigma_2^2$ . [Two Tailed Test]

An unbiased estimate of the variance for sample I

$$\hat{\sigma}_1^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{90}{9} = 10$$

An unbiased estimate of the variance for sample II

$$\hat{\sigma}_2^2 = \frac{\sum (y - \bar{y})^2}{n - 1} = \frac{108}{11} = 9.82$$

$$\therefore \text{Test statistic } F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{10}{9.82} = 1.018.$$

**Critical value :** The table value of  $F$  at  $\alpha = 0.05$  for 9 and 11 degrees of freedom is

$$F_{0.05} = 2.90.$$

**Decision :** Since the computed value of  $F = 1.018 <$  Tabled value of  $F_{0.05} = 2.90$ , so the null hypothesis  $H_0$  is accepted  $\Rightarrow$  the two populations have the same variance.

**Example 5 :** In a test given to two groups of students drawn from two normal populations, the marks obtained were as follows:

Group A :	18	20	36	50	49	36	34	49	41
Group B :	29	28	26	35	30	44	46		

Examine at 5% level, whether the two populations have the same variance.

**Solution :** Null hypothesis :  $H_0 : \sigma_A^2 = \sigma_B^2$ .

Alternate hypothesis :  $H_1 : \sigma_A^2 \neq \sigma_B^2$  (Two Tailed Test)

**Calculation of Test Statistic :** We have the following table to find the sample variances.

$$\therefore \bar{x} = \frac{\Sigma x}{n} = \frac{333}{9} = 37 ; \quad \bar{x} = \frac{\Sigma y}{n} = \frac{238}{7} = 34.$$

$$\text{Variance of group } A : \sigma_A^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{1134}{8} = 141.75$$

$$\text{Variance of group } B : \sigma_B^2 = \frac{\sum (y - \bar{y})^2}{n-1} = \frac{386}{6} = 64.33.$$

**Table : Calculations for Sample Variances**

Group A			Group B		
x	$(x - \bar{x}) = x - 37$	$(x - \bar{x})^2$	y	$(y - \bar{y}) = y - 34$	$(y - \bar{y})^2$
18	-19	361	29	5	25
20	-7	289	28	6	36
36	-1	1	26	8	64
50	13	169	35	1	1
49	12	144	30	4	16
36	-1	1	44	10	100
34	-3	9	46	12	144
49	12	144			
41	4	16			

$$n_1 = 9; \quad \Sigma(x - \bar{x})^2 = 1134; \quad \Sigma x = 333 \quad n_2 = 7; \quad \Sigma(y - \bar{y})^2 = 386; \quad \Sigma y = 238$$

$$\therefore \text{Test Statistic} \quad F = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_B^2} = \frac{141.75}{64.33} = 2.203.$$

**Critical value :** The table value of  $F$  at 5% level for 8 and 6 degrees of freedom is

$$F_{0.05} = 4.15.$$

**Decision :** The computed value of  $F = 2.203 <$  Tabled value of  $F_{0.05} = 4.15 \Rightarrow$  the null hypothesis  $H_0$  is accepted  $\Rightarrow$  The populations from where the samples have been taken have the same variances.

**Example 6 :** Following information about two samples selected from two normal populations is available:

$$n_1 = 9 \text{ and } s_1 = 2.9, \quad n_2 = 7 \text{ and } s_2 = 6.3$$

Test whether the two samples have come from the population having the same variance.

[Given :  $F_{6.8(0.05)} = 3.58, F_{8.6(0.05)} = 4.15$ ]

**Solution :** Let the Null Hypothesis be  $H_0 : \sigma_1^2 = \sigma_2^2$ .

Alternative Hypothesis is  $H_1 : \sigma_1^2 \neq \sigma_2^2$ . (Two Tailed Test)

Assume that  $H_0$  is true. Then the test statistic is given by

$$F = \frac{s_2^2}{s_1^2} = \frac{(6.3)^2}{(2.9)^2} = \frac{36.69}{8.41} = 4.38.$$

Since the table value  $F(6, 8)$  at 5% level is 3.58 and is in the right-tail of the distribution; which is less than the computed value of  $F$ , we reject the null hypothesis. Hence, the two populations don't have equal variance.

**Example 7 :** In a sample of 8 observations, the sum of squared deviations of items from the mean was 94.5. In another sample of 10 observations, the value was found to be 101.7. Test whether the difference is significant at 5% level.

You are given that at 5% level, critical value of  $F$  for  $v_1 = 7$  and  $v_2 = 9$  degrees of freedom is 3.29 and for  $v_1 = 8$  and  $v_2 = 10$  degrees of freedom, its value is 3.07.

**Solution :** Null Hypothesis :  $\sigma_1^2 = \sigma_2^2$ .

Alternative Hypothesis :  $\sigma_1^2 \neq \sigma_2^2$ .

It is given that:  $n_1 = 8, \sum (X_1 - \bar{X}_1)^2 = 94.5 ; \quad n_2 = 10, \quad \sum (X_2 - \bar{X}_2)^2 = 101.7$

$$s_1^2 = \frac{\sum (X_1 - \bar{X}_1)^2}{n_2 - 1} = \frac{94.5}{7} = 13.5 ; \quad s_2^2 = \frac{\sum (X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{101.7}{9} = 11.3$$

$$\text{Here, } s_1^2 > s_2^2 \quad \therefore \quad F = \frac{s_1^2}{s_2^2} = \frac{13.5}{11.3} = 1.195.$$

For  $v_1 = 7$  and  $v_2 = 9$   $F_{0.05} = 3.29$ .

**Decision :** The calculated value of  $F$  is less than the table value. Hence we accept the hypothesis and conclude that the difference in the variance of two samples is not significant at 5% level.

**Example 8 :** For a random sample of 10 pigs fed on diet A, the increases in weight in pounds in certain period were:

10      6      16      17      13      12      8      14      15      9

For another random sample of 12 pigs fed on diet B the increases in weight in the same period were:

7      13      22      15      12      14      18      8      21      23      10      17

Test whether both the samples come from population having same variance.

(The  $F_{0.05}$  for  $v_2 = 11, v_1 = 9$  is 3.11).

**Solution : 1. Null hypothesis :**  $H_0 : \sigma_1^2 = \sigma_2^2$ .

**Alternative Hypothesis :**  $H_1 : \sigma_2^2 > \sigma_1^2$ .

Table : Calculations for Sample Variances

Diet A			Diet B		
$X_1$	$(X_1 - \bar{X}_1)$	$(X_1 - \bar{X}_1)^2$	$X_2$	$(X_2 - \bar{X}_2)$	$(X_2 - \bar{X}_2)^2$
10	-2	4	7	-8	64
6	-6	36	13	-2	4
16	+4	16	22	+7	49
17	+5	25	15	0	0
13	+1	1	12	-3	9
12	0	0	14	-1	1
8	-4	16	18	+3	9
14	+2	4	8	-7	49
15	+3	9	21	+6	36
9	-3	9	23	+8	64
			10	-5	25
			17	+2	4
$\bar{X}_1 = \frac{120}{10} = 12$		120	$\bar{X}_2 = \frac{180}{12} = 15$		314

2.  $s_1^2 = \frac{1}{n_1 - 1} \sum (X_1 - \bar{X}_1)^2 = \frac{120}{9} = 13.333$

$$s_2^2 = \frac{1}{n_2 - 1} \sum (X_2 - \bar{X}_2)^2 = \frac{314}{11} = 28.545.$$

Here,  $s_2^2 > s_1^2$

3. Statistical Test : Since  $s_2^2 > s_1^2$ , under  $H_0$ , the test statistic is:

$$F = \frac{s_2^2}{s_1^2} = \frac{28.545}{13.333} = 2.14$$

4. Tabulated  $F_{0.05}$  for 13 and 8 d.f. = 4.30.

5. Decision : Since calculated value of  $F$  is less than tabulated  $F$ , it is not significant. Hence  $H_0$  may be accepted at 5% level of significance. We conclude that variability in the two populations is same.

**Example 9 :** In one sample of 10 observations, the sum of the squares of the deviations of the sample values from sample mean was 120 and in the other sample of 12 observations, it was 314. Test whether the difference is significant at 5 per cent level.

**Solution :** In the usual notations we are given:

$$n_1 = 10, \quad \Sigma(x - \bar{x})^2 = 100, \quad n_2 = 12, \quad \Sigma(y - \bar{y})^2 = 314.$$

**1. Null Hypothesis :**  $H_0 : \sigma_x^2 = \sigma_y^2$  i.e., the sample variances do not differ significantly.

**Alternative Hypothesis :**  $H_1 : \sigma_y^2 > \sigma_x^2$ . [ $\because S_y^2 > S_x^2$ ]

**2. Calculation of Sample Variance :**

Now 
$$S_x^2 = \frac{\sum(x - \bar{x})^2}{n_1 - 1} \bullet$$

$$= \frac{120}{9} = 13.33;$$

$$S_y^2 = \frac{\sum(y - \bar{y})^2}{n_2 - 1}$$

$$= \frac{314}{11} = 28.55.$$

**3. Test statistic :** Since  $S_y^2 > S_x^2$ , under  $H_0$ , the test statistic is

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

$$\therefore F = \frac{28.55}{13.33} = 2.1418.$$

**4. Tabulated  $F_{0.05}(11, 9) = 3.11$ .**

**5. Decision :** Since calculated  $F$  is less than tabulated  $F_{0.05}(11, 9)$ , it is not significant at 5% level of significance. Hence data are consistent with the null hypothesis that the sample variances do not differ significantly.

**Example 10 :** The time taken by workers in performing a job by method I and method II is given below:

Method I : 20    16    26    27    23    22

Method II : 27    33    42    35    32    34    38.

Do the data show that the variances of time distribution in a population from which these samples are drawn do not differ significantly.

**Solution :**

**1. Null Hypothesis :**  $H_0 : \sigma_1^2 = \sigma_2^2$ , i.e., there is no significant difference between the variances of the time distribution by the workers in performing a job by method I and method II.

**Alternative Hypothesis :**  $H_1 : \sigma_1^2 > \sigma_2^2$ . (One Tailed Test)

Table : Calculations for Sample Variances

Method I			Method II		
$X_1$	$d_1 = X_1 - 22$	$d_1^2$	$X_2$	$d_2 = X_2 - 36$	$d_2^2$
20	-2	4	27	-9	81
16	-6	36	33	-3	9
26	+4	16	42	+6	36
27	+5	25	35	-1	1
23	+1	1	32	-4	16
22	0	0	34	-2	4
			38	+2	4
Total	$\Sigma d_1 = +2$	$\Sigma d_1^2 = 82$		$\Sigma d_2 = -11$	$\Sigma d_2^2 = 151$

$$S_1^2 = \frac{1}{n_1 - 1} \left[ \Sigma d_1^2 - \frac{(\Sigma d_1)^2}{n_1} \right] = \frac{1}{5} \left( 82 - \frac{4}{6} \right) = 16.226$$

$$S_2^2 = \frac{1}{n_2 - 1} \left[ \Sigma d_2^2 - \frac{(\Sigma d_2)^2}{n_2} \right] = \frac{1}{6} \left( 151 - \frac{121}{7} \right) = 22.286$$

Here,  $S_2^2 > S_1^2$ .

3. Test Statistic : Since  $S_2^2 > S_1^2$ , under  $H_0$ , the test statistic is:

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

$$\therefore F = \frac{12.286}{16.266} = 1.37.$$

4. Tabulated  $F_{0.05}(6, 5) = 4.95$ .

5. Decision : Since calculated  $F$  is less than tabulated  $F_{0.05}(6, 5)$ , it is not significant. Hence  $H_0$  may be accepted at 5% levels of significance and we conclude that variability of the time distribution in the two populations is same.

Example 11 : Two random samples of sizes 8 and 11, drawn from two normal populations, are characterized as follows:

Population from which the sample is drawn	Size of sample	Sum of observations	Sum of squares of observations
I	8	9.6	61.52
II	11	16.5	73.26

You are to decide if the two populations can be taken to have the same variance. What test function would you use? How is it distributed and what value it has in this sampling experiment?

**Solution :** We are given that:

$$n_1 = 8, \quad \Sigma x = 9.6, \quad \Sigma x^2 = 61.52; \quad n_2 = 11, \quad \Sigma y = 16.5, \quad \Sigma y^2 = 73.26$$

**1. Null Hypothesis :**  $H_0 : \sigma_1^2 = \sigma_2^2$  i.e., the two populations have the same variance.

**Alternative Hypothesis :**  $H_1 : \sigma_1^2 > \sigma_2^2$ .

**2. Calculations of Sample Variances :**

$$\begin{aligned} S_1^2 &= \frac{1}{n_1 - 1} \sum (x - \bar{x})^2 = \frac{1}{n_1 - 1} \left[ \Sigma x^2 - \frac{(\Sigma x)^2}{n_1} \right] \\ &= \frac{1}{7} \left[ 61.52 - \frac{(9.6)^2}{8} \right] = \frac{1}{7} \left[ 61.52 - \frac{92.16}{8} \right] \\ &= \frac{1}{7} (61.52 - 11.52) = \frac{50}{7} = 7.1428. \end{aligned}$$

$$\begin{aligned} S_2^2 &= \frac{1}{n_2 - 1} \sum (y - \bar{y})^2 = \frac{1}{n_2 - 1} \left[ \Sigma y^2 - \frac{(\Sigma y)^2}{n_2} \right] \\ &= \frac{1}{10} \left[ 73.26 - \frac{(16.5)^2}{11} \right] = \frac{1}{10} \left[ 73.26 - \frac{272.25}{11} \right] \\ &= \frac{1}{10} [73.26 - 24.75] = \frac{48.51}{10} = 4.851. \end{aligned}$$

**3. Test Statistic :** Since  $S_1^2 > S_2^2$ , the test statistic is

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

$$\therefore F = \frac{S_1^2}{S_2^2} = \frac{7.1428}{4.851} = 1.4724.$$

**4. Tabulated**  $F_{0.05}(7, 10) = 3.14$ .

**5. Decision :** Since calculated  $F <$  tabulated  $F_{0.05}(7, 10)$ , it is not significant. Hence  $H_0$  is accepted at 5% level of significance and we conclude that the two populations can be taken to have the same variance.

### EXERCISE

- The standard deviations calculated from two random samples of size 9 and 13 are 2 and 1.9 respectively. May the sample be regarded as drawn from the normal populations with the same standard deviation.

[Hint :  $H_0 : \sigma_1^2 = \sigma_2^2, H_1 : \sigma_1^2 \neq \sigma_2^2$ .

$$\hat{\sigma}_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{9}{8} \times 4 = 4.5, \quad \hat{\sigma}_2^2 = \frac{13}{12} \times 3.61 = 3.91.$$

$$\therefore F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{4.5}{3.91} = 1.15 < F_{0.05} = 3.51.$$

$\Rightarrow$  the null hypothesis is accepted  $\Rightarrow$  the samples can be regarded as drawn from the populations with the same standard deviation.]

2. Following results were obtained from two samples, each drawn from two different populations *A* and *B*.

Population	<i>A</i>	<i>B</i>
Sample	I	II
Sample size	$n_1 = 25$	$n_2 = 17$
Sample s.d.	$s_1 = 3$	$s_2 = 2$

Test the hypothesis that the variance of brand *A* is more than that of *B*.

[Hint :  $H_0 : \sigma_1^2 \leq \sigma_2^2$ ,  $H_1 : \sigma_1^2 > \sigma_2^2$

$$\hat{\sigma}_1^2 = \frac{25 \times 9}{24} = 9.375, \quad \hat{\sigma}_2^2 = \frac{17 \times 4}{16} = 4.25, \quad F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{9.375}{4.25} = 2.205.$$

Also, from table  $F_{0.05}$  for 24 and 16 d.f. = 2.24.

Now computed  $F = 2.205 < F_{0.05} = 2.24 \Rightarrow$  the null hypothesis is accepted  $\Rightarrow$  variance of brand *A* is not more than the variance of brand *B*].

3. Two sample are drawn from two normal populations. From the following data test whether the two samples have the same variance at 5% level.

Sample I : 60    65    71    74    76    82    85    87  
 Sample II : 61    66    67    85    78    63    85    86    88    91.

[Hint : Here  $\Sigma x = 600$ ,  $\bar{x} = \frac{600}{8} = 75$ ,  $\sum (x - \bar{x})^2 = 636$ ,

$$\Sigma y = 770, \quad \bar{y} = \frac{770}{10} = 77, \quad \sum (y - \bar{y})^2 = 1200,$$

$$\hat{\sigma}_1^2 = \frac{836}{8 - 1} = 90.857, \quad \hat{\sigma}_2^2 = \frac{1200}{10 - 1} = 133.333.$$

Let  $H_0 : \sigma_1^2 = \sigma_2^2$ .  $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{133.33}{90.857} = 1.468$

$< F_{0.05} = 3.68$  for 9 and 7 d.f.  $\Rightarrow H_0$  is accepted  $\Rightarrow$  the samples have the same variance.]

4. The time taken by workers in performing a job by Method I and Method II is given below.

Method I : 20    16    26    27    22  
 Method II : 27    33    42    35    32    34    38

Do the data show that the variances of time distribution is a population from which these samples are drawn do not differ significantly?

5. Two samples were drawn from two normal populations and their values are:

A :	66	67	75	76	82	84	88	90	92
B :	64	66	74	78	82	85	87	92	93

Test whether the two populations have the same variance at 5% level of significance.

6. Two sources of raw materials are under consideration of a company. Both sources seem to have similar characteristics, but the company is not sure about their respective uniformity. A sample of ten lots from source *A* yields a variance of 225, and a sample of eleven lots from source *B* yields a variance of 200. It is likely that the variance of source *A* is significantly greater than the variance of source *B*. Take  $\alpha = 0.01$ .

[Hint :  $H_0$  : variance of source *A* is not significantly greater than the variance of source *B*, i.e.,  $H_0 : \sigma_1^2 = \sigma_2^2$ .       $H_1 : \sigma_1^2 > \sigma_2^2$       (One Tailed Test)]

$$\hat{\sigma}_1^2 = 250, \quad \hat{\sigma}_2^2 = 220 \quad F = \frac{250}{220} = 1.17.$$

Critical value of *F* at 10% level of significance for  $v_1 = 9$ ,  $v_2 = 10$  degrees of freedom is  $F_{(0.01)}(9, 10) = 4.93 >$  computed value of  $F = 1.17 \Rightarrow H_0$  is accepted  $\Rightarrow$  that variance of source *A* is not significantly greater than the variance of source *B*].

7. It is known that the mean diameters of rivets produced by two firms *A* and *B* are practically the same but standard deviations differ. For 22 rivets produced by firm *A*, the standard deviation is 2.9 mm, while for 16 rivets manufactured by firm *B*, the standard deviation is 3.8 mm. Compute the statistic you would use to test whether the product of firm *A* has the same variability as those of firm *B*. Also state how you would proceed further.
8. The following data relate to a random sample of government employees in two states of the Indian Union:

	<i>State I</i>	<i>State II</i>
Sample size	16	25
Mean monthly income (in Rs.)	440	460
of the sample employees		
Sample variance	40	42

First carry out a test of hypothesis that the variance of the two populations are equal.

In the light of the result of the above test, carry out a test of the hypothesis that the means of the two populations are equal.

9. The following data present the yields in quintal of rice on ten subdivision of equal area of two agricultural plots.

Plot I :	6.2	5.7	6.5	6.0	6.3	5.8	5.7	6.0	6.0	5.8
Plot II :	5.6	5.9	5.6	5.7	5.8	5.7	6.0	5.5	5.7	5.5.

Test whether two samples taken from two random populations have the same variance (5% point of *F* for  $v_1 = 9$  and  $v_2 = 9$  is 3.18).

[Hint : Null hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2$ ; Alternative hypothesis :  $H_1 : \sigma_1^2 > \sigma_2^2$ .

$$\bar{X}_1 = 6, \quad \bar{X}_2 = 5.7; \quad \sum(X_1 - \bar{X}_1)^2 = 0.64; \quad \sum(X_2 - \bar{X}_2)^2 = 0.24.$$

$$S_1^2 = \frac{\sum(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{0.64}{9} = 0.071; \quad S_2^2 = \frac{\sum(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{0.24}{9} = 0.027$$

Since  $S_1^2 > S_2^2$  under  $H_0$ , the test statistic is:

$$F = \frac{S_1^2}{S_2^2} = \frac{0.071}{0.027} = 2.63 < F_{0.05}(9, 9) = 3.18. \quad \text{Thus } H_0 \text{ is accepted}.$$

10. In a sample of 8 observations, the sum of the squared deviations of items from the mean was 94.5. In another sample of 10 observations, the value was found to be 101.7. Test whether the difference is significant at 5% level.

You are given that at 5% level, critical value of  $F$  for  $v_1 = 7$  and  $v_2 = 9$  degrees of freedom is 3.29 and for  $v_1 = 8$  and  $v_2 = 10$  degrees of freedom, its value is 3.07.

11. Can the following two samples be regarded as coming from the same normal population?

<i>Sample</i>	<i>Size</i>	<i>Sample mean</i>	<i>Sum of squares of deviations from the mean</i>
1	10	12	120
2	12	15	314

### ANSWERS

1. The sample can be regarded as drawn from the population with the same standard derivation.
  2. The variance of brand  $A$  is not more than the variance of brand  $B$ .
  3. The samples have the same variance.
  4.  $F = 1.37$ , the variance of time distribution in a population from which the samples are drawn do not differ significantly.
  5.  $F = 1.415$ , the two populations have the same variance.
  6. The variance of source  $A$  is not significantly greater than the variance of source  $B$ .
  7.  $F = 2.14$ ; It is not significant at 5% level of significance.
  8.  $F = 1.33$ ; Variances are equal; No significant difference in means,  $t = 1.44$ .
  9.  $F = 2.63$ ; Two random populations have the same variance.
  10.  $H_0 : \sigma_1^2 = \sigma_2^2$ , i.e., sample variances do not differ significantly.  $F = 1.195$ ; not significant :  $H_0$  accepted.
  11.  $H_0 : \sigma_1^2 = \sigma_2^2$ ;  $F = 2.14$ ; Not significant.  $H_0' : \mu_1 = \mu_2$ ;  $t = -1.51$ ; Not significant.
- Since both  $H_0$  and  $H_0'$  are accepted, the two samples may be regarded as coming from the same normal population.



# 16

# *Demography and Vital Statistics*

## **16.1 DEMOGRAPHY**

*Demography is defined as the study of measurement of human population. It is the quantitative study of human population with respect to events such as death, marriage, morbidity and migration.* Demographic methods encompass two different aspects of this quantification – “static” techniques concerned with characteristics of a population at a fixed time. For example, census methods, and “dynamic” techniques concerned with the changing nature of a population for, vital events: birth, death, marriage, divorce and migration. We are concerned with both static and dynamic characteristics, because they are equally necessary in determining rates, the basic measures of vital statistics. Demography also covers the distribution of population by industry, occupation, geographical area and so on. In fact the interests of demography are at least expensive with the population census and in some respects go beyond it. Demography has many facts and vital statistics is one of them.

## **16.2 VITAL STATISTICS**

**Vital events :** Vital events are all-those events which have to do with an individual's entrance into or departure from life altogether with changes in civil status which may occur to him during his life time such as death, birth, marriage, sickness, divorce etc.

**Vital statistics** may be interpreted in two-ways. In broader sense it refers to all types of population statistics collected by whatever mode, which in a narrower sense it refers only to the statistics derived from the registration of births, deaths and marriages.

Nowadays, family planning is one of the most important programmes of the governments of various countries. It has attracted a special attention of the Indian government due to population explosion. The study of vital statistics enables the government to assess the impact of family planning on population growth.

Many statisticians have given various definitions of vital statistics.

According to Benjamin :

**"Vital statistics are conventional numerical records of marriages, births, sickness and deaths by which the health and growth of a community may be studied."**

Arthur Henshalme defined it as :

**"The branch of biometry which deals with data and the laws of human mortality, morbidity and demography."**

### **16.2.1 Uses of Vital Statistics**

- 1. Use of Individual :** Vital statistics is very handy for authentication of some vital events such as births, deaths, marriages etc., to individual.
- 2. Use in Public Administration :** The role of vital statistics in planning and evaluation is of most important use to which this body of data may be placed.
- 3. Use in Insurance Agency :** The life table known as 'biometer' of public health and longevity enables the insurance agency in fixing the life insurance premium rates at various age levels.
- 4. Use in Research :** The data on life expectancy at various age levels is useful in actuarial calculations of risk on life policies. Vital statistics are indispensable in demographic research. The information regarding fertility, mortality, maternity, urban density is indispensable for planning and evaluations of schemes of health, family planning and other civic amenities by the government.

### **16.3 METHODS OF COLLECTION OF VITAL STATISTICS**

The following are the four methods of collecting vital statistics :

- 1. Census Enumeration**
- 2. Registration Method**
- 3. Survey Method**
- 4. Analytical Method—Estimation of vital rates using census data.**

**Census Enumeration :** Census operations are conducted in almost all countries at intervals of ten years. In a census, the enumeration of every individual of all habitational areas is carried out at a specific time. Information is collected on community, sex, age, literacy, economic conditions etc. This information is analysed in the form of various tables based on regions and characteristics.

**Registration Method :** This method entails registration of vital events such as deaths, births, marriages etc. Registration of vital events is done with the proper authorities as appointed by the government of a country. These authorities issue the certificates of the respective vital events. It serves many useful purposes of authentication of these events. For example, death certificate is necessary for life insurance claim, for succession certificate etc. Similarly, birth certificates required for school admission, job etc.

**Survey Method :** Surveys are, generally conducted in areas, or regions where the system of registration or recording of births, deaths or marriages etc., has not been functioning properly and efficiently. The survey records make vital statistics available for that area or region.

**Analytical Method :** The populations estimates at a given time can be obtained without ad-hoc survey by mathematical methods. These estimates are based on the assumption that the population grows at a constant rate during the intercensal years. Analytical methods are algebraic methods which make use of available data without requiring any fresh survey.

## 16.4 RATES AND RATIOS

### 16.4.1 Rate

A **rate** is a fraction such that the numerator  $c$  is included within the denominator  $c + d$  so that a rate has the form  $c/(c + d)$ . The numbers  $c$  and  $d$  are never negative, and thus a rate is a number lying between 0 and 1. Notice that this is a specialized use of the word rate. This definition, appropriate to vital statistics and demography, does not include such concepts as velocities, growth rates and the like.

We are interested in rates so that comparisons between groups of unequal size can be made. If we look only at the absolute number of events in several groups, we have no meaningful way of relating these groups. However, considering events per population at risk removes this difficulty. We shall illustrate this below, but first we shall define several of the basic vital rates.

These rates fall into one of two categories: **crude and specific rates**. The word “**crude**” implies that the total number of events are used in the computation, whereas “**specific**” means that only the events in a particular category of age, sex, race, particular disease, or other classification variable are used. Although we have noted that rates are proportions always lying between 0 and 1, in vital statistics these rates are usually expressed per some convenient base, such as 1000 or 10,00,000 a large enough multiple of 10 being chosen as base so that the resulting rate is greater than 1. This may not always be possible when several rates are to be compared, for certainly they should all expressed to the same base. To illustrate the concept of

base, consider percentage, in which a rate is expressed per 100. For example,  $\frac{37}{925} = 0.04$ , when

expressed as a percentage, is 4 percent or  $\frac{37}{925} \times 100 = 4\%$ . A base is used for two reasons: First, persons are loath to work with the extremely small number that frequently arise in rate computation. Second, an answer is often more easily understood if stated as parts per some round number such as 100 or 1000, rather than per 1.

*In a rate the numerator is a figure representing the collection of data over a period of time while the denominator represents a static figure, a count at a specific instant of time. The rate then describes the rapidity with which a given event is occurring.* For instance, the rate of admissions to the hospital represents how rapidly persons are entering the hospital per year.

### 16.4.2 Ratio

The study of data that are qualitative is in many ways simpler than is the study of measurement data. Qualitative data are ordinarily summarised in the form of rates or ratios: The count of the number of individuals with a given qualitative characteristic is compared with the count of the number having a second qualitative characteristic by forming a ratio. If at the time of the census in 1990 in a country there were 387,814 white persons and 182,631 non-white

persons. It could be said then that the ratio of white to non-white persons was 387,814/182,631 or 2.1 to 1. This says that there were relatively 21 white persons to every 10 non-white persons at that time.

The difference between a **simple ratio**, a **proportion** and a **rate** is not always clearly understood. Infact, 'ratio' are used interchangeably by many persons. A **ratio** is a simple measure of relationship between two numerical quantities. These quantities need not necessarily represent the same entities, although the unit of measure must be the same for both numerator and denominator of the ratio. A physician has 250 patients, 150 of whom are female. The ratio of female to male patients then is 1.5 to 1. Here the common unit of measure is a number. *A* is forty years old and his son is twenty years old, therefore, the ratio of *A*'s age to his son's age is 2 : 1 or *A* is twice as old as his son. Here the common unit is years of age.

The type of ratio commonly designed as a proportion is somewhat different. Here the individuals or entities in the numerator of the ratio must be included also in the denominator. Thus, for the physician with 250 patients, of whom 150 are female, the proportional of patients that are female is:

$$\frac{150 \text{ females}}{150 \text{ females} + 100 \text{ males}} \times 100 = 60 \text{ per cent}$$

In interpreting any ratio, proportion, or rate the most important fact to consider is the source of the numbers that enter into the numerator and the denominator. Do they represent an accurate count of the event under study? To obtain the rate, ratio or proportion in the first place it is necessary to known what is to be measured. When the final value is obtained it is essential to know how accurately it measures what it was supposed to be measured.

In this chapter the formulae are given for the calculation of the rates and ratios of most importance to physicians. Although all these indexes are commonly called rates, several of them are in reality ratios as can be seen by a study of the figures that form the numerator and denominator. A complete study of each of these indexes is beyond the scope of this chapter. Since the completeness of the figures relative to population, births, deaths and cases of various disease entities used in the calculation of these rates influences the accuracy with which the rate measures what it is supposed to measure, a few words should be said about the source of each of these.

Women who have been pregnant and therefore exposed to the risk of dying from causes related to pregnancy, since all women having a stillbirth, a miscarriage, or an abortion are excluded. Miscarriage and absorptions are for the main part not reported unless they result in death. Stillbirths are incompletely reported in many areas. Hence, to include them in an area where they are well reported would give a rate that could not fairly be compared with one from an area in which reporting is poor. For this reason the rate is calculated on the basis of live births. It should be recognized also that in multiple pregnancies, each live born child is included in the denominator, but only one woman is exposed to the risk of dying. This inclusion of multiple live birth in the denominator does not compensate for the omission of pregnancies resulting in other than live births. As stated before, in many areas births are incompletely reported. It is evident then that the maternal mortality ratio is always an overestimation of the risk of dying from pregnancy.

### 16.4.3 Mortality Rate

A third type of rate, which in its ideal form is a percentage, is the case-fatality rate. This rate commonly and erroneously called a **mortality rate**. It represents the proportion of a group of persons who, ill from a particular disease, die from that disease. That is, out of one hundred patients operated on for acute appendicitis one patient dies. The case fatality, or proportion of those dying, is 1 per cent. As calculated from reported statistics, this is not always true. For instance, the annual case fatality rate for appendicitis for a hospital would be calculated by dividing the deaths from appendicitis occurring in the hospital during a given year by the number of patients admitted to the hospital with a diagnosis of appendicitis during the same year. Obviously, a person might be admitted in one year and not die until the next year. Under ordinary circumstances the carry-over from one year to the next year will be about the same and hence the rate will be about the same and hence the rate will approximate fairly close to the case-fatality rate; but this might not be true if the hospital opens a new surgical ward at the end or beginning of the year.

This carry-over factor is especially important in case-fatality rates calculated for certain chronic disease such as tuberculosis and in which the number of both cases and deaths are taken from reports made to the health authorities. In such a disease a case may be reported in a given year but not die for several years. Likewise, a death in a given year may have been reported as a case several years prior to the year of death. For this reason, although called a case-fatality rate, the rate as calculated in the majority of times is merely a ratio of the number of deaths reported in a given year to the number of cases reported in the same year. A marked decrease in this type of rate may occur when new case-finding methods are introduced. For acute diseases such as diphtheria, scarlet fever, smallpox etc., the rate more nearly approximate the true case fatality rates of these diseases.

Data relative to mortality and natality arise through a process of registration conducted by the individual states. These are the data for which the physician is responsible. For every birth and death, a certificate must be made out and filed with a State Registrar. The state registrars send these data to the National Office of Vital Statistics which, in turn, publishes figures relative to mortality and natality for the entire India; for each state, and for some of the data, for smaller subdivisions. The completeness of these figures reflect the interest and co-operation of practising physicians since they attend the majority of both births and deaths. In general, deaths are more completely reported than are births.

Data relative to morbidity are most inadequate. In general, only the communicable diseases are legally reportable; the laws regulating the reporting of specific communicable diseases vary between the several states. Certain of the diseases related to industrial hazards are also reported, although again the list varies with the individual states. In some of the states cancer is today a reportable disease. The completeness reporting varies directly with the importance of the disease. As a result, most data relative to disease processes other than communicable diseases must be obtained either from a study of hospital cases, or from army and navy and institutional records or from hospital records are usually biased by inclusion of more severe illness and army, navy or institutional records are biased by the fact that they relate to selected groups. Data relative to morbidity, therefore, must be examined with care before drawing conclusions from the indexes obtained.

**Crude Mortality Rate or Death Rate :** One of the simplest rates used in vital statistics is the crude mortality or death rate. This rate attempts to measure the force of mortality on the general population, or to measure the rate at which a population is dying. The numerator is the number of deaths reported during a given time period, usually a year. Its accuracy will depend on the completeness with which deaths have been reported. The denominator is the average population; if the rate is calculated on an annual basis this would be the population as of July 1 of the year for which the deaths have been counted. Since counts of population are made only every ten years, counts for intercensal years must be estimates; these are subject to considerable error, especially toward the end of the intercensal period.

**Maternal Mortality Rate :** A somewhat different type is the *maternal mortality rate*. This is a true ratio rather than a rate, as it is commonly called, since it represents a relationship between the number of deaths resulting from puerperal causes and the total number of live births. It would be interesting to measure the risk of death from causes related to pregnancy. This is impossible to do since the number of pregnancies cannot be determined. Therefore, what is done is to determine the cost in maternal deaths per given number of live births, whether that birth be a live birth, a stillbirth, or an abortion. The numerator is influenced only by the completeness and accuracy in the reporting of such deaths.

## 16.5 FOMULAE FOR CALCULATION OF VITAL STATISTICS RATES

In the following rates, figures for both numerator and denominator must be collected from the same population and must relate to the same time period. For example, if a rate for Delhi for 1986 were to be calculated the figures in both the numerator and denominator must be for Delhi for the year 1986. The population assumed to be at risk at the population at the middle of the time period under study. If the rate is **an annual rate, this would be the population as of July 1 of the specific year**.

$$\text{CRUDE BIRTH RATE} = \frac{\text{Number of live births reported during a given year}}{\text{Estimated population as of July 1 of the same year}} \times 1000$$

### SPECIFIC BIRTH RATE

$$= \frac{\text{Number of live births to women of as specific age reported during a given year}}{[\text{Estimated population as of the same age as of July 1 of the same year}]} \times 1000$$

$$\text{CRUDE DEATH RATE} = \frac{\text{Number of deaths reported a given year}}{\text{Estimated population as of July 1 of the same year}} \times 1000$$

### MORTALITY RATE FOR A SPECIFIC DISEASE

$$= \frac{[\text{Deaths assigned to the specific cause during a given year}]}{[\text{Estimated population as of July 1 of the same year}]} \times 1,00,000$$

### INFANT MORTALITY RATE

$$= \frac{[\text{Number of deaths under 1 year of age reported during a given year}]}{[\text{Number live births reported during the same year}]} \times 1000$$

### NEONATAL MORTALITY RATE

$$= \frac{[\text{Number of deaths under 28 days of age reported during a given year}]}{[\text{Number live births reported during the same year}]} \times 1000$$

### MATERNAL MORTALITY RATE

$$= \frac{[\text{Number of deaths assigned to causes related to pregnancy during a given year}]}{[\text{Number live births reported during the same year}]} \times 1000$$

### CASE FATALITY RATE

$$= \frac{[\text{Number of deaths assigned to a specific disease during a given year}]}{[\text{Number of cases reported of the same disease during the same year}]} \times 100$$

### PROPORTIONAL MORTALITY RATE

$$= \frac{[\text{Number of deaths assigned to a specific disease during a given year}]}{[\text{Number of deaths reported from all causes during same year}]} \times 100$$

### MORBIDITY RATES FOR A SPECIAL DISEASE

$$\text{Incidence} = \frac{[\text{Number of new cases of a specific disease reported during a given year}]}{[\text{Estimated population as July 1 of the same year}]} \times 10^x$$

$$\text{Prevalence} = \frac{[\text{Number of cases present at a given time}]}{[\text{Estimated population at the same time}]} \times 10^x$$

The morbidity incidence rate is the more commonly used of the two since this is calculated on the basis of cases reported to a health department. It is expressed as a rate per 100,000. The prevalence rate can only be determined following a survey of the population concerned and hence is done only rarely.

Any of the above rates may be made specific for any factor for which the data are known. To do this the numerator must include the individuals in whom the event under study occurs. The denominator must include the individuals who are at risk of this event. For example, the mortality rate for children of age five to nine years (age specific rate) would be calculated by dividing the deaths in children five to nine years by the number of children of that same age in the population. The mortality rate for white males age five to nine from measles would be calculated as:

$$\frac{\text{Number of deaths in males, (age 5 – 9 years), assigned to measles during given year}}{\text{Number of males age 5 – 9 in population as of July 1 of given year}}$$

**Example 1 :** The population and the number of deaths of two localities, according to age groups, are given below:

Age group	Locality A		Locality B	
	Population	No. of deaths	Population	No. of deaths
Under 5	6000	212	93000	904
5-15	12685	244	154100	523
15-35	13325	295	186200	618
35-50	4600	361	81900	1237
Over 50	6710	463	64800	1475

Compute (i) the crude death rates, (ii) standardised death rates taking population of locality A as standardised population and compare their health conditions.

**Solution :**

**Table for Computation of CDR**

Age group	Locality A		Locality B	
	Population	No. of deaths	Population	No. of deaths
Under 5	6000	212	93000	904
5-15	12685	244	154100	523
15-35	13325	295	186200	618
35-50	4600	361	81900	1237
Over 50	6710	463	64800	1475
Total	43320	1575	580000	4757

$$\text{CDR for locality } A = \frac{\text{Total number of deaths in locality } A}{\text{Total population of locality } A} \times 1000$$

$$= \frac{1575}{43320} \times 1000 = 36.4 \text{ nearly.}$$

$$\text{CDR for locality } B = \frac{\text{Total number of deaths in locality } B}{\text{Total population of locality } B} \times 1000$$

$$= \frac{4757}{580000} \times 1000 = 8.2.$$

Table for Computation of  $S_T$  DR

Standardised Population $S_x$	Locality B		
	No. of deaths	Death Rate per 1000 ( $D_x$ )	$S_v \times D_x$
6,000	904	$\frac{904}{93,000} \times 1000 = 9.720$	58,320.00
12,685	523	$\frac{523}{154,100} \times 1000 = 3.394$	43,052.89
13,325	618	$\frac{618}{186,200} \times 1000 = 3.319$	44,225.68
4,600	1,237	$\frac{1237}{81,900} \times 1000 = 15.104$	69,478.40
6,710	1,475	$\frac{1475}{64,800} \times 100 = 22.762$	152,735.34
$\Sigma S_x = 43,320$	← Total →		$\Sigma S_x D_x = 3,67,812.31$

Since the population of locality A is taken as standardised population, therefore,

$$S_T \text{DR of locality } A = \text{CDR of locality } A = 36.4 \quad \dots (1)$$

$$S_T \text{DR of locality } B = \frac{\Sigma S_x D_x}{\Sigma S_x} = \frac{367812.31}{43320} = 8.49 \quad \dots (2)$$

From (1) and (2), we get

$$S_T \text{DR of locality } B < S_T \text{DR of locality } A.$$

∴ The locality of B is healthier than the locality A.

**Example 2 :** Find the crude and standardised death rates of locality A and B from the following data:

Age group	District A			District B		
	Population $S_A$	No. of deaths $N_A$	Standardised Population $S_x$	Population $S_B$	No. of deaths $N_B$	Standardised Population $S_x$
0-10	2000	50	2160	1000	20	2160
10-55	7000	75	5830	3000	30	5830
55 and over	1000	25	2010	2000	40	2010
Total	10000	150	10000	6000	90	10000

**Solution :** Here we have

$$\begin{aligned}\text{CDR of District } A &= \frac{\text{Total number of deaths in locality } A}{\text{Total population of locality } A} \times 1000 \\ &= \frac{150}{10000} \times 1000 = 15.\end{aligned}$$

$$\begin{aligned}\text{CDR of District } B &= \frac{\text{Total number of deaths in locality } B}{\text{Total population of locality } B} \times 1000 \\ &= \frac{90}{6000} \times 1000 = 15.\end{aligned}$$

Table for  $S_T$  DR

Age group	Standardised Population $S_x$	District A		District B	
		Age specific death rate $D_s = \frac{N_A}{S_A} \times 1000$	$S_x \times D_x$	Age specific death rate $D_s = \frac{N_B}{S_B} \times 1000$	$S_x \times D_x$
0-10	2160	$\frac{50}{2} = 25$	54000.0	20	43200
10-55	5830	$\frac{75}{7} = 10.71$	62439.3	$\frac{30}{3} = 10$	58300
55 and over	2010	25	50250.0	$\frac{40}{2} = 20$	40200
Total	10000		166689.3		141700

$$\therefore S_T \text{DR for district } A = \frac{\sum S_x \times D_x}{\sum S_x} = \frac{166689.3}{10000} = 16.67 \text{ nearly.}$$

$$\therefore S_T \text{DR for district } B = \frac{\sum S_x \times D_x}{\sum S_x} = \frac{141700}{10000} = 14.17.$$

E x a Calculate the crude death rates for the following data:

Age group	Population (in thousands)	Number of deaths	Standardised population
0-9	21	350	24310
10-24	30	102	32780
25-54	37	229	31350
55-64	17	354	16390
65 and over	5	415	5170

**Solution :** We have the following table to calculate CDR and SDR.

**Table for CDR and SDR**

Age group	Population $S$	Number of deaths $N$	Age specific death rate $D_x = \frac{N}{S} \times 1000$	Standardised Population $S_x$	$S_x \times D_x$
0-9	21000	350	$\frac{350}{21} = 16.67$	24310	405247.7
10-24	30000	102	$\frac{102}{30} = 3.40$	32780	111452.0
25-44	37000	229	$\frac{229}{37} = 6.19$	31350	194056.5
45-64	17000	354	$\frac{354}{17} = 20.82$	16390	341239.8
65 and over	5000	415	$\frac{415}{5} = 83.00$	5170	429110.0
<b>Total</b>	<b>110000</b>	<b>1450</b>		<b>110000</b>	<b>1481106.0</b>

$$\therefore \text{CDR} = \frac{\Sigma N}{\Sigma S} \times 10000 = \frac{1450}{110000} \times 1000 = 13.2 \text{ per thousand.}$$

## 16.6 ADJUSTMENT OF RATES

Upto this point the discussion of rates and ratios has been concerned only with the source and accuracy of the figures that enter into their population. In comparing rates or ratios between different geographic areas, different hospital populations or experimental groups, there are other factors which should be considered. It is well recognized that certain easily identified characteristics influence markedly the prognosis inmost diseases. For example, the majority of diseases are more fatal in the infant in the aged person than in childhood, early adolescent, and young adult life. Many diseases are more fatal to the Negro than to the white person. Severity of disease, complications arising from associated disease, duration of diseases before treatment is instituted and many other such factors are also important.

Any type of a ratio may have both numerator and denominator broken into various subgroups. For instance, consider the case-fatality rate for appendicitis. With the terms "deaths" and "cases" referring in each instance to those with diagnosis of appendicitis, the case-fatality rate might be written;

$$\text{Case-Fatality Rate} = \frac{\text{Deaths among Males} + \text{Deaths among Females}}{\text{Cases among Males} + \text{Cases among Females}}$$

$$\begin{aligned}
 &= \frac{\text{Deaths among Males} + \text{Deaths among Females}}{\text{Cases among Males} + \text{Cases among Females}} \\
 &= \frac{\text{Deaths in age } 0-4 + \text{Deaths in age } 5-9 \dots \text{Deaths in age } 75+}{\text{Cases in age } 0-4 + \text{Cases in age } 5-9 \dots \text{Cases in age } 75+} \\
 &= \frac{\text{Deaths in cases first seen} < 24 \text{ hours after onset} + \text{Deaths in cases first seen} > 24 + \text{hours before onset}}{\text{Cases first seen} < 24 \text{ hours after onset} + \text{Cases first seen} > 24 + \text{hours before onset}}
 \end{aligned}$$

Obviously, the total rate will be affected by the proportional distribution of the various subgroups.

As an example of this effect, the mortality rate from tuberculosis for the year 2005 can be compared between the states of U.P. and K.

**Table 16.1 : Mortality from tuberculosis (all forms) in 2005  
in U.P. and K by race**

State	Population			Number of deaths from tuberculosis			Deaths per 100,000 population		
	White	Non-white	Total	White	Non-white	Total	White	Non-white	Total
U.P	4,611,503	79,011	4,690,514	942	63	1005	20.4	79.7	21.4
K	1,188,632	990,282	2,178,914	191	382	573	16.1	38.6	26.3

Examination of this table shows that the rate of 26.3 deaths per 100,000 population of K is higher than the rate of 21.4 for U.P. is the more fortunate. However, further examination of the table reveals that for white persons the rate in K was 16.1 as compared with a rate of 20.4 in U.P. For the non-whites, K's rate of 38.6 was lower also than U.P.'s rate of 79.7. How can this apparent paradox be explained? The rates for both white and non-white are lower in K but the rate for the total population is lower in U.P.

A study of the racial distribution of the population in the two states points out that although nearly half of the population in K is non-white, less than 2 per cent of the population of U.P. is non-white. This fact, coupled with a rate in non-white between two and four times as high as in whites, results in an over-all rate that is higher in K.

These data illustrate the influence of the racial distribution on the total rate only. Obviously comparisons on the basis of the rate for the total population cannot be made unless the racial composition of the populations is the same. The rates for whites can be compared and the rates among non-whites can be compared. Such comparisons are of great value but over-all comparisons are invalid unless some method of adjustment is applied to the data.

The simplest method of adjustment involves the selection of a standardised population and then the application of the race-specific rates of the two states to this population. In this example the population of K and U.P. would be combined white and white; non-white and non-white. This would give a population of 6,869,428 of whom 5,800,135 are white and 1,069,293 are non-white. If in this standard population the same proportion of white persons had died as died in K(16.1/100,000) and the same proportion of non-whites had died in K(38.6) there would have been:

$16.1/100,000 \times 5,800,135$  or 933 deaths in whites.

$38.6/100,000 \times 1,069,293$  or 413 deaths in non-whites.

This would have given a total of 1,346 deaths for a rate of  $1,346/6,869,428 \times 100,000$  or 19.6 per 100,000 for K.

If in this identical population the same proportion of white persons died as died in U.P. ( $20.4/100,000$ ) and the same proportions of non-whites died as died in U.P. ( $79.7/100,000$ ) there would have been:

$20.4/100,000 \times 5,800,135$  or 1,183 deaths in whites

$79.7/100,000 \times 1,069,293$  or 852 deaths in non-whites.

This would give a total of 2,035 deaths for a rate of  $2,035/6,869,428 \times 100,000$  or 29.6/100,000 for U.P. These rates of 19.6 and 29.6, commonly called adjusted rates, are now comparable. Had the population of K and U.P. been the same and equal to the population of the two states and had the proportion of white and non-white dying from tuberculosis been the same as actually occurred in each state, the rate in U.P. would have been 29.6 and in K it would have been 19.6. These rates reflect the higher rates for both whites and non-whites in U.P. It should be noted that the number of persons dying from tuberculosis per 100,000 population in K is 26.3 and in U.P. is 21.4. Nothing can change these figures but the figures for the total populations of the two states cannot be compared since the racial composition of the population is radically different. In order to make an over-all comparisons, rates must be estimated from a common population. Such rates will then be based on a population that has the same racial distribution.

In the preceding example the effect of the age distribution of the populations was not considered. For the sake of simplicity it was assumed that the populations are similar in that respect. In many situations age is equally as important as race in determining comparability of rates. It is relatively easy to compare race-specific rates since usually only two rates are involved, white and non-white. However, to compare a series of eight or more age-specific rates for two populations is difficult. Hence, adjustment for age is most important. In Table 16.2 data relative to case-fatality in Rocky Mountain spotted fever are given.

**Table 16.2 : Cases and Deaths from Rocky Mountain Spotted Fever, Eastern and Western type by age\***

Age in years	Western type			Eastern Type		
	Cases	Deaths	Percent dying	Cases	Deaths	Percent dying
Under 15	108	13	12.0	310	40	12.9
15-39	264	40	15.1	189	21	11.1
40 and over	375	157	41.9	162	61	37.7
Total	747	210	28.1	661	122	18.5

Examination of the data given in this table shows that 28.1 per cent of the cases of the Western type died as compared with 18.5 per cent of cases with the Eastern type. This would indicate that the Western type was more fatal disease than the Eastern type. Examination of the age-specific rates shows that the proportions dying in those under fifteen years of age were

about the same in both types, 12.0 vs 12.9. In the other two age groups, the rates were somewhat higher for the Western type; for those fifteen to thirty nine years the rates were 15.1 vs. 11.1; and for those forty years and over the rates were 4.19 vs. 37.7.

It should be noted further that the rate increased with age in both Eastern and Western types. Since this is true the rate for each type as a whole will depend on the proportional age distribution of the cases in each type. Examination shows that for the Western type roughly half of the cases were in the youngest age group. The rates for the two types as a whole are, therefore, not comparable.

**The method of adjustment for age parallels that of adjustment for race.** First an arbitrary or standard population is selected. Again, as in the previous illustration, this may be comprised of the cases of the two types, added age for age. The resulting figures are shown in the second column of Table 16.3. Second, the number of persons in each age group in this standard population is multiplied first by the age-specific by the corresponding age group for the Western type and second by the corresponding age-specific rates for the Eastern type. The resulting numbers of deaths that would have occurred in this population on the basis of these rates are shown in columns 5 and 6. New rates are now calculated by dividing first, the sum of the deaths that would have occurred on the basis of the age-specific rates for the Western type and second, for the Eastern type by the total of the standard population. The age-grouped rate for the Western type is  $343.7/1408 \times 100$  or 24.4; for the Eastern type,  $306.4/1408 \times 100$  or 21.8.

**Table 16.3 : Data for Calculation of Age-adjusted rates for Rocky Mountain Spotted Fever and Eastern and Western type  
(Original data taken from table 16.2)**

Age in years	Standard	Age-specific rate		Deaths expected in standard population on basis of rates of	
		Western	Eastern	Western	Eastern
(1)	(2)	(3)	(4)	(5)	(6)
Less than 15	418	12.0	12.9	50.3	53.9
15-39	453	15.1	11.1	68.6	50.3
40 and over	537	41.9	37.7	224.8	202.2
Total	1408			343.7	306.4

Comparison of the rates before adjustment for age shows that the rate for the Western type is one and one-half times as high as for the Eastern type ( $28.1/18.5 = 1.52$ ). After adjustment for age the rate for the Western type is one and one-tenth times as high as that for the Eastern type ( $24.4/21.8 = 1.12$ ). Thus it, can be seen that the major factor in the higher rate for the total of the Western type is age. Again it should be emphasized that there are more persons per hundred dying from the Western type than from the Eastern type; but the conclusion does not follow that the Western type is more fatal than the Eastern type. The major factor in the apparently higher rate is the fact that proportionately more older persons developed the Western type of the disease.

At this point it should be noted that in calculating adjusted rates, the information available from the study of the age or race-specific rates is lost. This information is the most valuable since it is the most specific. For example, in the race adjustment of the tuberculosis mortality for U.P. and K, the fact that rates for both whites and non-whites are higher in U.P. than in K is lost. In the present problem also the fact is lost that are higher in the Western type except in the youngest age group. For this reason, the specific rates should be shown in any analysis of rate data. However, when an over-all comparison is to be made, adjustment of the rate to a common population must be made.

**The selection of the standard population to be used in the adjustment will vary with the number of comparisons.** The arithmetic value of the adjusted rate will depend on the proportional of the population selected as a standard. For example, in the adjustment of the mortality rates from tuberculosis in U.P. and K the following race adjusted rates were obtained:

**Table 16.4 : Death Rates from Tuberculosis, U.P. and K, 1950, Adjusted for Racial Differences Using Differential Populations**

Standard population (all as of April 1, 1950)	Per cent of population that were white	Adjusted rates		Ratio U.P and K.
		U.P.	K	
United States	89.55	26.6	18.3	1.5
U.P.	98.30	21.4	16.4	1.3
K	54.55	47.4	26.2	1.8
U.P. and K combined	84.43	29.6	19.6	1.5

In each of these adjustments, the rate for U.P. is higher than that for K but the arithmetic level of the rates is different. Those adjustment to a population having the greatest proportion of non-white persons are the higher. Because of the marked differences in the arithmetic level of the rates depending on the proportional differences of the races in the distributions, it is necessary to consider carefully the population to be used as a standard.

*The selection of the standard population to be used in the adjustment will vary also on the number of rates that are to be compared. If rates from only two groups are to be compared, the most common practice is to use as a standard the combined population of the two areas or groups. This is comparable to saying that the rates are adjusted to a population that is the average of the two. If the population in one group is relatively small, it is common practice to use that population as a standard population. This avoids using rates based on small numbers in deriving the expected deaths.*

When a series of groups is to be compared, it is common to select what is frequently known as a "standard million" as a standard population. **This standard million is merely one million persons distributed by age in the same proportion as is some actual population.** Commonly used age distributions are: the distribution of the population according to the census count of a specific year such as 1930, 1940, 1950 etc.; the distribution by age of the so-called stationary population from a life table. The choice of a standard the total of which is one million makes easy the calculation of the final age-adjusted rate since the division of the expected deaths by the standard population involves merely the pointing off of the correct number of decimal places.

The particular standard population chosen as has been shown changes the actual value of the age-adjusted rate. This is not of great importance, however, since it is the ratio of one rate to another that is of importance. Usually the ratios will be approximately the same, unless extremely biased population distributions are chosen as a standard.

In any analysis of rates, ratios or proportions, certain factors should be borne in mind:

1. *What is the index designed to measure?*
2. *Does it actually measure what it is designed to measure or is it an approximation of that measure? If an approximation, is it the best one available?*
3. *What are the inherent elements of bias present in the data from which the index is calculated?*
4. *When comparisons of rates, ratios and proportions are to be made, the relative composition of the groups as to factors influenced the level of the index must be considered. If an index is determined from a non-homogeneous group, it cannot be compared with that another non-homogeneous group until some method of adjustment been applied.*

## 16.7 FOLLOW-UP LIFE TABLE

The fundamental technique of the life table that was discussed in the previous section can be modified so as to become a valuable tool in follow-up studies following inoculation, vaccination or any other treatment. This modified technique is also used in the study of chronic disease where the period of incubation or length of illness may be months or even years. In such studies some consideration must be given to the time intervening between exposure and development of the disease since in this interval the exposed population will be increased through births and marriages and will be decreased through deaths and departures to other residences.

Also, in the study of problems related to the determination of rates of survival following treatment account must likewise be taken of relatively long periods of observation. During such periods individuals will die from causes unrelated to the disease under study and will be lost from observation because of change of residence. For example, at the end of five years it may be known that of 500 patients operated on for cancer of the breast, 350 are alive and free from the disease, 25 are known to have died from causes other than cancer and 25 to have died from cancer. To base the survival rate on the 400 for whom definite history is known is equally an erroneous as to base it on the original 500.

*For the study of such problems the modified life table technique or actuarial method, as it is also called, is most satisfactory. It utilises the experience of each individual for the entire time he is under observation.* If a person has been observed for four years following treatment, he will be included in the group at the risk of dying in the first year following treatment, the second year, the third year, and the fourth year. His entire life experience during the period of the study is translated into "persons year" a person-year being defined as the equivalent of the experience of one individual for one year. This may consist of one person under observation for one year, or two persons each under observation for one-half year, or three persons each for one-third year, or four persons each for one-quarter year. The same fundamental assumption is made as in the regular life table, i.e., any person who enters into or withdraws from the experience at any time during a year is considered to have been included for one-half year.

**Table 16.5 : Calculation of Arrest rate for Patients treated with Promin**

Year following administration of drug	Number under observation at beginning of period	Number Discharge from hospital	Lost from Transfer to oral sulfone	Observation Period ended	Number arrested	Person -years obser-vation	Arrest rate per cent
$x$ to $x + 1$	$I_x$	$w_{x_1}$	$w_{x_2}$	$w_{x_3}$	$a_x$	$L_x$	$m_x$
0-1	235	24	13	1	0	216.0	0.0
1-2	197	17	16	1	4	178.0	0.0
2-3	159	9	8	7	10	142.0	7.0
3-4	125	5	7	12	16	105.0	15.2
4-5	85	6	6	5	12	70.5	17.0
5-6	56	1	3	7	6	47.5	12.6
6-7	39	1	7	7	4	29.5	13.6
7-8	20	1	0	10	5	12.0	41.7
8-9	4	0	0	3	1	2.0	50.0

To illustrate the method, data that formed the basis for the arrest rate in leprosy following the administration of the sulfone drug, Promin will be used.

A total of 235 patients had been treated with promin. Periods of observation varied from one to eight years. Patients had been lost from observation for various reasons: discharged from hospital with or without consent of authorities; transferred to another type of sulfone treatment; and arrest of the disease. For certain numbers, the period of observation ended in each of the nine years. The data together with results of the calculation of the specific values are given in table 42.

### 16.7.1 Calculation of Arrest Rates

#### 1. Number of persons in study at beginning of year or $I_x$

$$I_{(x+1)} = I_x - (w_{x_1} + w_{x_2} + w_{x_3}) - a_x.$$

Substituting in the above formula the values from the table:

$$I_{(0-1)} = 235, \quad w_{x_1} = 24, \quad w_{x_2} = 13, \quad w_{x_3} = 1, \quad a_x = 0$$

$$I_{(1-2)} = 235 - (24 + 13 + 1) = 0 = 197.$$

The value of 197 for  $I_{(1-2)}$  represents the number of individuals under observation at the beginning of the year (1-2). For the next year the corresponding value is

$$197 - (17 + 16 + 1) - 4 = 159.$$

#### 2. Person-years of observation or $L_x$

$$L_x = I_x - \frac{1}{2} (w_{x_1} + w_{x_2} + w_{x_3} + a_x).$$

Substituting the following values in the formula:

$$l_{(0-1)} = 235, w_{x_1} = 24, w_{x_2} = 13, w_{x_3} = 1, a_x = 0$$

$$L_{(0-1)} = 235 - \frac{1}{2}(24 + 13 + 1 + 0) = 216.$$

This means that the equivalent of 216 individuals were under observation for the first year, or in other words there were 216 person-years of observation. Another way of looking at this value is to consider it as the average number of persons under observation. Since there were 235 individuals at the beginning of the next year, the average number of individuals would be  $1/2(235 + 197)$  or 216.

### 3. Arrest rate column or " $m_x$ "

$$m_x = \frac{a_x}{L_x} \times 100.$$

For the year (1-2) there was the equivalent of 178 persons under observation or 178 person-years of observation. During this period there were 4 patients whose disease became arrested. The arrest rate, therefore, is  $4.178 \times 100 = 2.2$  per 100 person-years.

The arrest rate for the entire period would then be the sum of the numbers in the  $a_x$  column (total number arrested) divided by the sum of the numbers in the  $L_x$  column (total person-years of observation) or  $8/801 \times 100 = 7.2$  per 100 person-years. Attention should be called to the fact that rates should be calculated only when the number of persons remaining in the population at risk is of sufficient size to minimize the sampling variation of the rate. In the problem used for illustration the rates after the six to seven-year-period are based on much too small numbers. Even the five to six-year-period has a rate that is subject to a relatively large sampling variation.

This same method could be used for computing the death rate following treatment or attack rate following exposure by substituting for the  $a_x$  column a  $d_x$  column (deaths) or a  $c_x$  column (cases). Whether the rate is computed as arrests per 100 person-years or per 1,000 person-years is entirely a matter of convenience. A base sufficiently large to give whole numbers in the rate is usually preferable.

The method may be applied also to obtaining rates for specific age groups. In such cases, persons will enter and leave the study at all ages; they will not all be present ordinarily at birth. For instance, in a follows-up of persons exposed to tuberculosis there might be ten persons exposed to tuberculosis at age two; twelve persons exposed at age three; nine at age four; fourteen at age fifteen etc. There might likewise be four for whom the period of observation ended at age five; three at age six, etc. In this case there will be an added column in the table labelled  $n_x$  (number entering experience at a given year). The  $l_x$  column will usually start at 0 under these conditions for at the beginning of the study no one will be under observation. With the same assumption of one half year experience for any one entering during a year.

$$l_{x+1} = l_x + n_x - (w_{x_1} + w_{x_2} + w_{x_3}) - a_x$$

$$\text{and } L_x = l_x + \frac{1}{2}n_x - \frac{1}{2}(w_{x_1} + w_{x_2} + w_{x_3}) - \frac{1}{2}a_x$$

$$\text{or } \frac{1}{2}$$

### 16.7.2 Calculation of Probability or arrest

In many instances more is desired than the simple calculation of arrest rates in a specific year. For instance, in the problem of arrest from leprosy, it may be important to know not what is the chance of becoming arrested *in the fourth year*, but what is the chance of arrest *in the first four years* after treatment. To do this the rates just calculated may be used. These rates represent the rate of attack, death or arrest based on the average persons at risk. Hence, these rates must first be transformed to probabilities. This can be done by using the formula given in Chapter 16, converting  $m_x$  to  $q_x$

$$q_x = \frac{2m_x}{2 + m_x}$$

In this formula  $m_x$  must always be expressed per single person-year. For example, an  $m_x$  rate of 2.2 per cent, in this form would be .022. Therefore in the present problem  $q_x$  for the 2-3 year period would be

$$q_{(2-3)} = \frac{2 \times .070}{2 + .070} = .067.$$

A simpler way to do this is to calculate not the arrest rate as was done but to calculate the probability of becoming arrested in a given year. This will be:

$$q_x = \frac{a_x}{l_x - \frac{1}{2} (w_{x_1} + w_{x_2} + w_{x_3})}$$

Except for slight inconsistencies because of rounding off decimals the  $q_x$  rates will be identical whether they are calculated by the first or second method. Now the values may be used to set up a regular life table. After construction of the entire table, however, main interest is in the  $l_x$  column which represents the number not attacked, not dead, or with the disease not arrested at the beginning of a given year. The following *Table 16.6* gives the calculations for the material previously used.

Since in this problem interest centres in the proportion arrested by the end of  $x$  years after beginning of treatment, an additional column (the last one in the table) is useful. This is calculated by subtracting from 1,000 (the arbitrary number with which the table started) the number of persons who were not arrested at the beginning of the year.

If at the beginning of the year (3-4) there were 912 not arrested, 88 arrested out of 1,000 starting treatment. Hence the probability of arrest by the end of the third year was 88/1,000 or 8.8 per cent; by the end of the fourth year 217/1,000 or 21.7 per cent. Again it should be pointed out that the probability of arrest at any given time period as calculated from these data for periods beyond the fifth or at the most the sixth year are very unreliable because they are based on such a small number of patients.

**Table 16.6 : Calculation of probability of arrest by the end of any specified year**

Year following administration of drug	Arrest rate	Probability of arrest	Number at risk of arrest at beginning of year	Number arrested during year	Number arrested at beginning of year
$x$ to $x + 1$	$m_x$	$q_x$	$I_x$	$(I_x \times q_x)$	$1,000 - I_x$
0-1	.000	.000	1000	0	0
1-2	.022	.022	1000	22	0
2-3	.070	.067	978	66	22
3-4	.152	.141	912	129	88
4-5	.170	.157	783	123	217
5-6	.126	.119	660	78	340
6-7	.136	.127	582	74	418
7-8	.417	.345	508	175	492
8-9	.500	.400	333	133	667

This method can be used with many types of problems. Follow-up studies of immunization procedure lend themselves easily to this method of analysis. It is an ideal method for equalizing different average years of observation in groups. For example, take 100 persons treated and observed for an average period of five years and compare this with a group treated and observed for an average period of ten years. Whether considering death, arrest, or attack the number of deaths, arrests or persons attacked will be greater in the group with the ten-year average observation. But on a person-year basis the latter group will have more person-years of observation and even with more cases than the other group may have a lower rate per person-year.

An added advantage is that all the data are used. An individual remains in the study as long as he is under observation. If he is under observation for only two years, he counts in the experience of the group for those two years. This is a marked contrast to the method presently used in the so-called "five year" cure or survival rate in cancer. In the latter method, only individuals who have been observed for at least five years are included.

**Example 4 :** The populations of two towns are given as 876942 and 690272 respectively and their respective death rates as 14.3 and 18.2 per 1000. Find to the nearest whole number, the death rate for the two towns taken together.

**Solution :** Population of first town ( $n_1$ ) = 876942.

Death rate ( $\bar{x}_1$ ) = 14.3 per 1000.

Population of second town ( $n_2$ ) = 690272.

Death rate ( $\bar{x}_2$ ) = 18.2 per thousand.

∴ Death rate for the two towns taken together is

$$\begin{aligned}\bar{x} &= \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \\ &= \frac{12540270.6 + 12562950.4}{1567214} = \frac{25103221}{1567214} = 16.02 \\ &= 16 \text{ per 1000}\end{aligned}$$

Hence the number of death rates for the two towns taken together is 16 per thousand.

### EXERCISE 16.1

1. What do you understand by vital statistics?

2. Define the following term.

Crude death, Specific death rate, Standard death rate, Infant mortality rate and Crude birth rate.

3. Compute the crude and standardised death rates of two towns A and B from the following data:

Age group	Town A		Town B	
	Population	No. of deaths	Population	No. of deaths
Below 5	15000	360	40000	1000
5-30	20000	400	52000	1040
Above 30	10000	280	8000	240
Total	45000	1040	100000	2280

(Take the population of town A as standardised population)

[Ans : CDR for town A = 23.11, S<sub>T</sub>DR for town A = 23.11.

CDR for town B = 22.80, S<sub>T</sub>DR for town B = 23.89]

4. From the following data, make a comparative study of crude and standardised death rates for town A and town B.

Age group	Town A		Town B		Standard population (in '000)
	Population (in '000)	Specific Death rate	Population (in '000)	Specific Death rate	
0-5	6	40	7	40	4
5-25	20	18	40	18	34
25-45	24	12	20	13	25
45-60	12	25	8	20	11
Above 60	8	34	5	44	6

[Ans : CDR for town A = 20.5, CDR for town B = 02.5.

S<sub>T</sub>DR for town A = 19.39, S<sub>T</sub>DR for town B = 19.76]

5. Which of the two localities *A* and *B* is healthier?

Age group	Locality A		Locality B	
	Population	Deaths per 1000	Population	Deaths per 1000
0-10	600	30	400	40
10-20	1000	5	1500	4
20-60	3000	8	2400	10
60 and above	440	50	700	30

(Take locality *A* as standard)

[Ans : Locality *A*]

6. For the data given below:

Estimated population (excluding armed forces overseas)	= 161,183,000
Number of deaths	= 1,481,091
Number of live births	= 4,017,362
Number of deaths under 28 days of age	= 76,724
Number of deaths under 1 year age	= 106,791
Number of deaths from tuberculosis	= 16,392
Number of deaths from causes related to the puerperium	= 2,105

Calculate

(i) Crude death rate ; (ii) Crude birth rate ; (iii) Infant mortality rate ; (iv) Neonatal mortality rate ; (v) Maternal mortality rate ; (vi) Death rate from tuberculosis.

7. In 1950 in New Orleans, Louisiana, with a population of 467,591, there were 5,748 deaths from all causes. In Minneapolis, Minnesota in the same year there were 5,451 deaths from all causes in a population of 521,718.

(a) Compare the crude rates for two cities.

(b) What major factor other than socio-economic might be responsible for the differences in these rates? Discuss.

8. In the state of Kerala in 1980 there were 18,916 deaths of which 7.4 per cent were ascribed to cancer. In 1950 there were 19,123 deaths of which 17.9 per cent were ascribed to cancer. Discuss briefly the most important factors that might be responsible for this increase in the percentage of cancer deaths.

9. The following data taken from *Census Reports and Vital statistics* of a country.

**Population, Deaths from all Causes, and Deaths from Tuberculosis  
for two Age Groups, United States 1950**

Age in years	Population for 1950	Number of deaths		Age - specific rate		Proportional mortality rate (per 100)
		All causes	Tuberculosis	All causes (per 1,000)	Tuberculosis (per 100,000)	
20-24	11,437,305	16,767	1,707			
60-64	6,010,755	141,114	3,071			

- (a) Calculate for each of two groups the age-specific mortality rate from all causes and from tuberculosis.
- (b) Calculate the proportional mortality rate from tuberculosis for each of the two age groups.
- (c) Discuss the interpretation of the two types of mortality indexes calculated in *a* and *b*.
10. Data for 1,564 patients showing their states as the time of last report are given in the following table.

**Chronic Ulcerative Colitis; Data for Calculation of Survival Rates**

<i>Interval from diagnosis</i> <i>x to x + 1</i>	<i>Last report</i>	
	<i>Dead</i>	<i>Living</i>
0 – 1	58	–
1 – 2	48	31
2 – 3	34	44
3 – 4	34	44
4 – 5	17	35
5 – 6	16	31
6 – 7	17	15
7 – 8	20	20
8 – 9	18	17
9 – 10	10	18
10 – 11	21	22
11 – 12	29	59
12 – 13	22	98
13 – 14	21	80
14 – 15	18	74
15 – 16	20	64
16 – 17	9	40
17 – 18	9	45
18 – 19	8	55
19 – 20	8	49
20 – 21	3	47
21 – 22	2	41
22 – 23	7	36
23 – 24	3	39
24 – 25	0	26
25 +	7	75

- (a) From these data, set up a modified life table, showing the  $I_x$  column, i.e., the number in study at beginning of each year of observation.
- (b) Calculate the person-years for each interval.
- (c) Calculate the probability of dying in each interval.
- (d) Calculate the probability of dying in each year and subsequent years up to the sixteenth year. Beyond this year the number of survivors is too small to justify such probability calculations.
11. Making use of the standard 1 million population of India as prescribed below, compile by the direct method the age adjusted death rate.
- |                                |           |            |
|--------------------------------|-----------|------------|
| 1. 23 <i>pingue fed sheep</i>  | 11 males, | 12 males   |
| 2. 13 died                     | 8 males,  | 5 males    |
| 3. 10 lived                    | 3 males,  | 7 females  |
| 4. 6 <i>clinically ill</i>     | 2 males   | 4 females  |
| 5. 4 <i>not clinically ill</i> | 1 male    | 3 females. |



# 17

# Non-Parametric Methods

## 17.1 INTRODUCTION

Most of the hypothesis-testing procedures discussed so far, such as Z-test or *t*-test, are based on the assumption that **the random samples are selected from a normal population**. We have seen that most of these tests are still reasonably reliable for slight departures from normality, particularly, when the sample size is large. Traditionally, these testing procedures have been known as **Parametric methods** as they depend on the parameters, *viz.*, mean or proportion or standard deviation etc. These parametric tests have used the parametric statistic of sample that comes from the population being tested.

The need for technique that apply more broadly has led to the development of **non-parametric methods**. These do not require the assumption that the underlying population is normal — or indeed that they have any single mathematical form and some even apply to non-numerical data. The non-parametric methods are also called **distribution-free methods** that assume no knowledge, what so ever, about the distributions of underlying populations, except perhaps that they are **continuous**.

## 17.2 NON-PARAMETRIC OR DISTRIBUTION FREE METHODS

### Parametric Methods

Before discussing non-parametric techniques, we should consider why the methods we usually use are called *parametric*. **Parameters are indices. They index (or label) individual distributions within a particular family.** For example, there are an infinite number of normal distributions, but each normal distribution is uniquely determined by its mean ( $\mu$ ) and standard deviation ( $\sigma$ ). *If you specify all of the parameters (here,  $\mu$  and  $\sigma$ ), you've specified a unique normal distribution.*

*Most commonly used statistical techniques are properly called parametric because they involve estimating or testing the value(s) of parameter(s)—usually, population means or proportions. It should come as no surprise, then, that non-parametric methods are procedures that work their magic without reference to specific parameters.*

Parametric statistical methods are based on the stringent assumptions about the population from which the sample has been drawn. Particularly, the assumption like the form of probability distribution (whether it is normal or not), accuracy of observations, etc. are more common. Also the parametric methods are applicable primarily to the data which are measured in interval or ratio scale.

## Non-Parametric Methods

If the assumptions (mentioned in parametric methods) do not hold good or the data do not meet the requirement of parametric statistical methods, **non-parametric methods come to our rescue**. **Non-parametric methods entail very mild assumption of continuity and symmetry of the distribution.**

Non-parametric (or Distribution-free) statistical methods are mathematical procedures for statistical testing which unlike parametric statistics, make no assumptions about the frequency distributions of the variables being assessed.

**Non-parametric tests are often used in place of their parametric counterparts when certain assumptions about the underlying population are questionable.** For example, when comparing two independent samples, the Mann-Whitney (**a non-parametric test**) does not assume that the difference between the two samples is normally distributed, whereas its **parametric counterpart, the two sample t-test** does assume the condition of their being normally distributed. Many non-parametric methods are applicable for ordered statistics. **All tests involving ranked data (i.e., data that can be put in order) or involving ordinal data are non-parametric.**

**Ordinal Data:** A set of data is said to be the **ordinal if the values or observations belonging to it can be ranked (put in order) or have a rating scale attached. You can count and order, but not measure in ordinal data).**

Non-parametric tests may be, and often are, more powerful in detecting population differences when certain assumptions are not satisfied.

**Non-parametric models differ from parametric models in that the model structure is not specified a priori, but is instead determined from the data.** The term **non-parametric** is not meant to imply that such models completely lack parameters, rather, the number and the nature of parameters is flexible and not fixed in advance. Non-parametric models are, therefore, also called **Distribution free models**. A histogram is a simple non-parametric estimate of a probability distribution.

The precise definition of non-parametric varies slightly among authors. You'll see the terms **non-parametric** and **distribution-free**. They have slightly different meanings, but are often used interchangeably.

Non-parametric tests are often referred to as **distribution-free tests**. These tests have the obvious advantage of not requiring the assumption of normality or the assumption of homogeneity of variance. They compare medians rather than means and as a result, if the data have one or two outliers, their influence is negated.

**Many non-parametric procedures are based on ranked data.** Data are ranked by ordering them from lowest to highest and assigning them, in order, the integer values from 1 to the sample size. Ties are resolved by assigning tied values the mean of the ranks they would have

received if there were no ties, e.g., 117, 119, 119, 125, 128, becomes 1, 2.5, 2.5, 4, 5. (If the two 119s were not tied, they would have been assigned the ranks 2 and 3. The mean of 2 and 3 is 2.5.).

For *large samples* many non-parametric techniques can be viewed as the usual normal-theory-based procedures applied to ranks. The following table contains the names of some normal-theory-based procedures and their non-parametric counterparts.

**TABLE 17.1 : Some Commonly Used Statistical Tests**

<i>Normal theory based test</i>	<i>Corresponding Nonparametric test</i>	<i>Purpose of test</i>
<i>t</i> test for independent samples	Mann-Whitney U test; Wilcoxon rank-sum test	Compares two independent samples
Paired <i>t</i> test	Wilcoxon matched pairs signed-rank test	Examines a set of differences
One way analysis of variance ( <i>F</i> test)	Kruskal-Wallis analysis of variance by ranks	Compares three or more groups
Two way analysis of variance	Friedman Two way analysis of variance	Compares groups classified by two different factors
Pearson correlation coefficient	Spearman rank correlation coefficient	Assesses the linear association between two variables.

For smaller sample sizes, the same statistic (or one mathematically equivalent to it) is used, but decisions regarding its significance are made by comparing the observed value to special tables of critical values.

### 17.3 TYPES OF NON-PARAMETRIC TESTS

A large number of non parametric tests exist, but in this chapter we shall discuss the following tests which are better known and widely used ones:

- 1. Sign test for paired data and One Sample Sign Test.** In these tests, positive or negative signs are substituted for quantitative values.
- 2. Mann-Whitney U Test or a Rank Sum Test.** This test is used to determine whether two independent samples have been drawn from the same population.
- 3. Kruskal-Wallis Test.** It is another rank sum test which generalizes the analysis of variance to enable us to dispense with the assumption that the populations are normally distributed.
- 4. One Sample Run Test.** It is a method to determine the randomness with which the sample items have been selected.
- 5. Rank Correlation Test.** It is a method for finding correlation coefficient when the data available is not in numerical form but the information is sufficient to rank the data first, second, third, and so on.
- 6. Kolmogorov-Smirnov Test.** It is another method of determining the goodness of fit between an observed and theoretical probability distributions.

7. **Kendal Test of Concordance.** This test is applicable to situations where we want to test the significance of more than two sets of rankings of individuals.
8. **Median Test for Two Independent Samples.**
9. **Wilcoxon Signed-Rank Test.**

#### 17.4 ADVANTAGES OF NON-PARAMETRIC METHODS

1. Non-parametric tests are generally simple to understand, quicker and easier to apply when the sample sizes are small.
2. They do not require lengthy and laborious calculations and hence they are less time consuming. If significant results are obtained no further work is necessary.
3. They do not require the assumption that a population is distributed in the shape of a normal curve or another specified shape.
4. Non-parametric tests can be applied to all types of data — qualitative (nominal scaling) data in rank form (ordinary scaling) as well as data that have been measured more precisely (interval or ratio scaling).
5. The chief advantage of non-parametric tests is their inherently greater range of applicability because of milder assumptions.
6. Non-parametric tests make fewer and less stringent assumptions (that are more easily met) than the parametric tests.
7. Non-parametric test make less stringent demands of the data.
8. Non-parametric methods provide an air of objectivity when there is no reliable (universally recognised) underlying scale for the original data and there is some concern that the results of standard parametric techniques would be criticized for their dependence on an artificial metric.

#### 17.5 DISADVANTAGES OF NON-PARAMETRIC METHODS

Distribution-free or non-parametric methods also have a number of important disadvantages which in practice often outweigh their advantages.

1. Non-parametric tests sometimes pay for freedom from assumptions by a marked lack of sensitivity.
2. Important effects can be entirely missed by a non-parametric technique because of its lack of power or because it produces confidence intervals that are too wide. To put it another way, some of these techniques are very blunt instruments. It may cause it unwise to substitute one of these techniques either because of computational ease or because the data do not quite satisfy some distributional assumption.
3. The competing distribution-tied procedure may be quiet robust to departures from that assumption.
4. An additional related disadvantage is that distribution-free techniques often disregard the actual scale of measurement and substitute either ranks or even a more gross comparison of relative magnitudes. This is not invariably bad, but is often the basis for the loss of power or sensitivity referred to in the preceding paragraph.

5. The another disadvantage is one that is gradually being corrected by research in statistical methodology. At present, we have available a much larger body of techniques in distribution-free hypothesis testing than in distribution-free estimations. To some extent, this is an inevitable accompanying feature of techniques that disregard or weaken the underlying scale of measurement, because estimation is often tied to that scale, whereas hypothesis testing is concerned with selecting one of the two competing decisions or actions.
6. Non-parametric tests are often wasteful of information and thus less efficient than the standard techniques which they replace.
7. The major disadvantage is that non-parametric procedures throw away information. Because information is discarded, non-parametric tests can never be as powerful as their parametric counter parts, when parametric tests can be used. For example sign test uses only the sign of the observations and discard the data as such.

## 17.6 USES OF NON-PARAMETRIC METHODS

There are four situations in which the use of a distribution-free or non-parametric technique is indicated:

1. When quick or preliminary data analysis is needed. May be you just want to see roughly how things are going with the data. May be somebody is about to leave for Atlantic City to present a paper tomorrow, and his or her chief thought it would be a nice idea to include some statistics in the paper, and you are feeling unusually tender hearted or browbeaten today. May be your computer is suffering an acute exacerbation of some rare electronic disease. May be your computer programmer is manifesting that chronic, epidemic, occupational disability: temporal disorientation with gross distortion of task-time relationships. May be you are doing research in a “primitive” setting without access to modern computational aids.
2. When the assumptions of a competing distribution-tied or parametric procedure are not satisfied and the consequences of this are either unknown or known to be serious. Notice that violation of assumptions per se is not a sufficient reason to reject one of the classical procedures. As we have emphasised throughout this book, these procedures are often insensitive to such violations. For example, the  $t$ -tests and intervals for normal means are relatively insensitive to non-normality. We think that one of the most common error made today is the under application of classical statistical methods because of the overzealous adherence to the letter of the law with respect to assumptions. Remember that assumptions will rarely be exactly satisfied. It is, therefore, as important to have a grasp of the consequences of violations of assumptions. A great deal of theoretical research in statistics today is directed toward investigating the robustness of statistical procedures.
3. When data are only roughly scaled; for example, when only comparative rather than absolute magnitudes are available. In dealing with clinical data, perhaps patients can only be classified as better, unchanged, or worse. Perhaps only ranks, that is, largest, second largest, ....., smallest, are available.
4. When the basic question of interest is distribution-free or non-parametric in nature. For example, do we have a random samples, or are these two samples drawn from populations with identical distributions?

## 17.7 THE SIGN TEST FOR PAIRED DATA

The sign test is the easiest non-parametric test. Its name comes from the fact that it is based on the direction (or sign for plus or minus) of a pair of observations and not on their numerical magnitudes.

In such problems, each pair of sample value is replaced by a plus sign if the difference between the paired observation is positive (that is, if the first value exceeds the second value) and by a minus sign if the difference between the paired observation is negative (that is, if the first value is less than the second value) and it is discarded if the difference is zero.

For applying sign test to solve a problem, we count:

**Number of +ve signs (+ signs)**

**Number of -ve signs (- signs)**

**Number of 0's (i.e., which cannot be included either as positive or negative)**

We take the two hypothesis :

**Null Hypothesis:**  $H_0 : p = 0.5$

**Alternative Hypothesis:**  $H_1 : p \neq 0.5$ . It is a case of two tailed test. If you look carefully at the above two hypothesis, you will see that the situation is similar to a fair coin toss. If we toss a fair coin 30 times, then probability ‘ $p$ ’ would be 0.5 and we would expect about fifteen heads and fifteen tails. In that case, we would use the binomial distribution as the appropriate sampling distribution. We also know that when  $np$  and  $nq$  are each atleast 5, we can use the normal distribution to approximate the binomial. Thus, we can apply normal distribution test. We calculate the standard error of the proportion  $p$  as given by:

$$\text{Standard error of proportion } p : \text{S.E.}(p) : \sigma_p = \sqrt{\frac{pq}{n}}.$$

**The two limits of acceptance region at 0.05 level of significance are:**

$$p_{H_0} + 1.96 \sigma_p \text{ and } p_{H_0} - 1.96 \sigma_p$$

It is important to note that if the conditions that both  $np$  and  $nq$  are not greater than 5, then we must use the binomial instead of normal distribution test.

**Example 1 :** Use the sign test to see if there is a difference between the number of days required to collect an account receivable before and after a new collection policy. Use the 0.05 significance level.

*Before:* 33 36 41 32 39 47 34 29 32 34 40 42 33 36 27

*After :* 35 29 38 34 37 47 36 32 30 34 41 38 37 35 28

**Solution :** **Table for Before and After Collection Policy**

Before (1)	:	33	36	41	32	39	47	34	29	32	34	40	42	33	36	27
After (2)	:	35	29	38	34	37	47	36	32	30	34	41	38	37	35	28
Sign of Score [(1) – (2)]	:	-	+	+	-	+	0	-	-	+	0	-	+	-	+	-

If we count the bottom row of above table, we get

$$\begin{aligned}\text{Number of '+' signs} &= 6 \\ \text{Number of '-' signs} &= 7 \\ \text{Number of 0's} &= 2 \\ \text{Total sample size} &= 15.\end{aligned}$$

When we use the sign test, we exclude the evaluation 0's as we are testing perceivable differences. We notice that we have 6 plus signs and 7 minus signs, for a total of 13 usable responses. If there is no difference between the two types of collections,  $p$  (the probability that the first collection exceeds the second collection) would be 0.5 and we would expect to get equal number of plus and minus signs. We set up the hypothesis as :

**1. Null Hypothesis:**  $H_0 : p = 0.5 \Rightarrow$  there is no difference between the two types of collections

**Alternative Hypothesis:**  $H_1 : p \neq 0.5 \Rightarrow$  there is a significant difference between the two types of collections  
 $\Rightarrow$  It is a case of two tailed test.

Sample size :  $n = 13$ .

$$\text{Proportion : } p = \frac{6}{13} = 0.46.$$

$$\therefore q = 1 - p = 1 - 0.46 = 0.54. \text{ Also } q = \frac{7}{13} = 0.54.$$

Here  $np = 13 \times \frac{6}{13} = 6$ , and  $nq = 13 \times \frac{7}{13} = 7$  are both greater than 5, so we use the normal distribution to approximate the binomial.

**2. Level of significance:**  $\alpha = 0.05$

**3.  $Z_\alpha$  for two tailed test = 1.96 (from the tables)**

**4. Standard error of the proportion  $p$  :** S.E. ( $p$ ) =  $\sigma_p = \sqrt{\frac{p\bar{q}}{n}}$

$$\therefore \sigma_p = \sqrt{\frac{0.46 \times 0.54}{13}} = \sqrt{0.2484} = 0.498.$$

**5. Limits of Acceptance Region.**

$$\text{Upper Limit: } = 0.5 + (1.96) \times (0.498) = 1.476$$

$$\text{Lower Limit: } = 0.5 - (1.96) \times (0.498) = -0.476$$

**6. Decision.** The two limits of acceptance region are 1.476 and -0.476. The sample proportion  $p = 0.46$  lies within these two limits, so the sample proportion falls within the acceptance region for this hypothesis. We accept the null hypothesis  $H_0 \Rightarrow$  there is no difference between the two types of collections at 0.05 level significance.

**Example 2 :** To determine the effectiveness of a new traffic-control system, the number of accidents that occurred at 12 dangerous intersections during four weeks before and four weeks after the installation of the new system was observed, and the following data was obtained:

3 and 1,      5 and 2,      2 and 0,      3 and 2,      3 and 2,      3 and 0  
 0 and 2,      4 and 3,      1 and 3,      6 and 4,      4 and 1,      1 and 0

Use the paired — simple sign test at  $\alpha = 0.05$  level of significance to test the null hypothesis that  $(\mu_1)$ , the new traffic-control system is only as effective as  $(\mu_2)$ , the earlier system (the population samples are continuous, but this does not matter so long as zero differences are discarded).

**Solution :**

**1. Setting up of Hypothesis:**

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

Alternative Hypothesis:  $H_1 : \mu_1 > \mu_2 \Rightarrow$  It is a case of one-tailed test.

**2. Level of significance:** Here  $\alpha = 0.05$ .

**3. Test Statistic:** Use the test statistic  $X$  = the observed number of plus signs.

Replacing each positive difference of the two values by a plus sign and each negative difference by a minus sign, we get:

+ + + + + + - - + - + + +

Here, the number of paired values  $n = 12$

and the number of plus signs  $X = 10$

It is a case of Binomial distribution as the number of negative signs is only 2, which is less than 5.

From the table, we find that for  $n = 12$ ,  $\theta = \frac{1}{2}$ ,

$$P(X \geq 10) = 0.0161 + 0.0029 + 0.0002 = 0.0192$$

$$\begin{aligned} [\text{Note: For } \alpha = 0.05, \text{ for } n = 12: P(X \geq 10) &= P(X = 10) + P(X = 11) + P(X = 12) \\ &= 0.0161 + 0.0029 + 0.0002 = 0.0192] \end{aligned}$$

**4. Decision:** Since  $P$ -value = 0.0192 is less than 0.05, so the null hypothesis must be rejected and we conclude that the new traffic-control system is not effective in reducing the number of accidents at dangerous crossings.

## 17.8 ONE SAMPLE SIGN TEST

**One-sample sign test** is a non parametric method which is used as an alternative to the one-sample  $t$ -test, where the null hypothesis is  $\mu = \mu_0$  against a suitable alternative hypothesis. The only assumptions underlying the sign test are: the population sample is continuous and symmetrical. We assume that the population is continuous so that there is zero probability of getting a value equal to  $\mu_0$ , and we do not even need the assumption of symmetry if we change the null hypothesis to  $\mu = \mu_0$ , where  $\mu$  is the population mean or median.

In the sign test, we replace each sample value exceeding  $\mu_0$  with a plus sign and each value less than  $\mu_0$  with a minus sign. Then we test the null hypothesis  $H_0$  that  $X$ , the number of plus sign is a value of a random variable having Binomial distribution with parameters

$n$  (the total number of plus or minus signs) and probability  $\theta = \frac{1}{2}$ . The two-sided alternative

hypothesis  $H_1 : \mu \neq \mu_0$  thus becomes  $\theta \neq \frac{1}{2}$ , and one-sided alternatives  $\mu < \mu_0$  and  $\mu > \mu_0$

becomes  $\theta < \frac{1}{2}$  and  $\theta > \frac{1}{2}$  respectively. If a sample value equals  $\mu_0$  (which can happen when we deal with rounded data even though the population is continuous), we simply discard it.

In order to perform a sign test, when the sample size is very small, we use the Table of Binomial probabilities and when the sample is large we use the normal approximation to the Binomial distribution with

$$\text{Mean} = n\theta, \text{ and variance} = n\theta(1 - \theta)$$

$$\therefore \text{S.E.}(\theta) = \sqrt{n\theta(1 - \theta)}$$

The test statistic in this case is,  $Z = \frac{X - n\theta}{\sqrt{n\theta(1 - \theta)}}$ ,

with  $\theta = \frac{1}{2}$ , and  $X$  the number of plus signs.

**Example 3 :** The following are the measurements of breaking strength of a certain kind of 2-inch cotton ribbon in pounds:

|     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 163 | 165 | 160 | 189 | 161 | 171 | 158 | 151 | 169 | 162 |
| 163 | 139 | 172 | 165 | 148 | 166 | 172 | 163 | 187 | 173 |

Use the sign test to test the null hypothesis  $\mu = 160$  against the hypothesis  $\mu > 160$  at the 0.05 level of significance.

**Solution :**

1. Null Hypothesis:  $H_0 : \mu = 160$ .

Alternative Hypothesis:  $H_1 : \mu > 160 \Rightarrow$  It is a case of one tailed test.

2. Level of significance:  $\alpha = 0.05$ .

3. Test statistic:  $X =$  The observed number of plus signs.

Replacing each value exceeding 160 with a plus sign, each value less than 160 with a minus sign, and discarding the value which equals 160, we get :

+ + 0 + + + - - + + + + - + + + + + + + +

Here  $n =$  the total number of plus and minus signs = 19.

$X =$  the total number of plus signs = 15.

**It is a case of Binomial Distribution.**

From the Binomial Table given at the end, we find

$$P(X \geq 15) = 0.0095 \text{ for } \theta = \text{and } n = 19.$$

$$\begin{aligned} [\text{For } \theta = 0.05, P(X \geq 15) &= P(X = 15) + P(X = 16) + P(X = 17) + P(X = 18) + P(X = 19) \\ &= 0.0074 + 0.0018 + 0.0003 + 0.0000 + 0.0000 = 0.0095] \end{aligned}$$

- 4. Decision:** Since  $P$ -value ( $= 0.0095$ ) is less than 0.05, so the null hypothesis must be rejected and we conclude that the mean breaking strength of given kind of ribbon exceeds 160 pounds.

**Example 4 :** The following data, (in tons) are the amounts of sulfur oxides emitted by a large industrial plant in 40 days

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 24 | 15 | 20 | 29 | 19 | 18 | 22 | 25 | 27 | 9  |
| 17 | 20 | 17 | 6  | 24 | 14 | 15 | 23 | 24 | 26 |
| 19 | 23 | 28 | 19 | 16 | 22 | 24 | 17 | 20 | 13 |
| 19 | 10 | 23 | 18 | 31 | 13 | 20 | 17 | 24 | 14 |

Use the sign test to test the null hypothesis  $\mu = 21.5$  against the alternative hypothesis  $\mu > 21.5$  at the 0.01 level of significance.

**Solution :**

- 1. Setting up the Hypothesis:**

Null Hypothesis:  $H_0 : \mu = 21.5$ .

Alternative Hypothesis:  $H_1 : \mu < 21.5 \Rightarrow$  It is a case of one-tailed test.

- 2. Test Statistic:**

Replacing each value exceeding 21.5 with a plus sign, each value less than 21.5 with a minus sign, and discarding the one value which equals 21.5, we get:

$$\begin{array}{ccccccccccccccccccccccccc} + & - & - & + & - & - & + & + & + & - & - & - & - & - & + & - & - & + & + & + \\ - & + & + & - & - & + & + & - & - & - & - & - & - & + & - & + & - & - & - & + & - \end{array}$$

$$X = \text{the number of plus signs} = 16.$$

$$n = \text{total number of plus and minus signs} = 16 + 24 = 40.$$

As the sample size,  $n = 40$  is very large, so we shall use the Normal approximation to Binomial distribution.

$$\therefore \text{Test Statistic: } Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

$$\text{where } \theta = \frac{1}{2} = 0.5, n = \text{sample size.}$$

$$\therefore Z = \frac{16 - 40(0.5)}{\sqrt{40(0.5)(0.5)}} = \frac{-4}{3.16} = -1.26$$

$$\therefore |Z| = 1.26.$$

$$[\because n = 40, \theta = 0.5]$$

3. **Level of significance:** Here  $\alpha = 0.01$
4. **Critical values:** The critical value  $|Z_\alpha|$  for  $\alpha = 0.01$  for one-tailed test is 2.33.  
**Decision:** Since  $|Z| < |Z_\alpha|$  as  $1.26 < 2.33$ , so we accept the null hypothesis.

### EXERCISE – 17.1

1. Use the sign test to see if there is a difference between the number of days until collection of an account receivable before and after a collection policy. Use 0.05 significance level.

*Before* : 30 28 34 35 40 42 33 38 34 45 28 27 25 41 36  
*After* : 34 29 33 32 47 43 40 42 37 44 27 33 30 38 36.

[Ans. There is no significant difference before and after new collection policy in the accounts receivable.]

2. The following data show employees' rates of defective work before and after a change in the wage incentive plan. Compare the two sets of data given below to see if the change lowered the defective units produced. Use the 0.10 level of significance:

*Before* : 8 7 6 9 7 10 8 6 5 8 10 8  
*After* : 6 5 8 6 9 8 10 7 5 6 9 5.

3. Use the sign test on the data given below to determine whether there is a statistical increase in the values produced by treatment *B* over those produced by treatment *A*:

|                    |   |    |    |    |    |    |    |    |    |    |     |
|--------------------|---|----|----|----|----|----|----|----|----|----|-----|
| <i>Subject</i>     | : | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10  |
| <i>Treatment A</i> | : | 46 | 41 | 37 | 32 | 28 | 43 | 42 | 51 | 28 | 27  |
| <i>Treatment B</i> | : | 52 | 43 | 37 | 32 | 31 | 39 | 44 | 53 | 26 | 31. |

Use 0.05 level of significance.

4. When is the sign test used? The scores under two conditions *X* and *Y* obtained by the respondents are given below:

*X* : 12 16 8 6 4 8  
*Y* : 7 12 17 5 12 11.

Apply the sign test and comment on your findings at 0.05 level of significance.

5. A company claims that if its product is added to an automobile's gasoline tank, the distance travelled in kilometres per litre will improve. To test the claim, 15 different automobiles are chosen and the distances with and without the additive are measured; the results are shown below. Assuming that the driving conditions are the same, determine whether there is a difference due to the additive at significance levels of (a) 0.05 and (b) 0.01.

---

*With additive* : 17.3 14.1 9.8 12.5 7.8 12.2 14.3 11.7 13.8 16.0 24.8 11.2 12.8 14.0 12.1

---

*Without additive* : 15.7 13.6 10.2 12.3 7.4 11.1 13.4 12.0 13.1 15.7 14.4 11.5 12.0 13.6 11.4

[Ans. (a) There is a difference at the 0.05 level of significance

(b) There is no difference at the 0.01 level of significance.]

6. The following data represents the number of hours that a rechargeable hedge trimmer operates before a recharge is required.

1.5      2.2      0.9      1.3      2.0      1.6      1.8      1.5      2.6      1.2      1.7.

Use the sign test to test the hypothesis at 0.05 level of significance that this particular trimmer operates, on the average, 1.8 hours before requiring a recharge.

[Hint:  $H_0 : \mu = 1.8$ ,  $H_1 : \mu \neq 1.8$ , we have  $- + - + - + -$ , so,  $n = 10$  and  $X = 3$ . Apply Binomial Table].

[Ans. Accept the null hypothesis  $H_0 : \mu = 1.8$ ]

7. On 12 visits to a doctor's clinic, a patient had to wait in minutes as under:

17;    32;    25;    15;    28;    25;    20;    12;    35;    20;    26;    24.

before being seen by the doctor. Use the sign test with  $\alpha = 0.05$  to test the doctor's claim that, on the average, his patient do not wait more than 20 minutes before being examined by him.

[Ans.  $H_0 : \mu = 20$  is accepted]

## 17.9 RANK SUM TESTS

Rank sum tests are a whole family of tests. We shall concentrate only on the following two members of this family:

### 1. Mann-Whitney U Test

### 2. Kruskal-Wallis H Test.

**Mann-Whitney tests is used when there are only two populations whereas Kruskal-Wallis test is employed when more than two populations are involved.** The use of these tests will enable us to determine whether independent samples have been drawn from the same population or different populations have the same distribution.

## 17.10 MANN-WHITNEY U-TEST

It is a **non-parametric method used to determine whether two independent samples have been drawn from populations with same distribution**. This test is also known as **U-Test**. This test enables us to test the null hypothesis that both population medians are equal (or that the two samples are drawn from a single population). It requires the two samples to be independent samples of observations measured at least at an ordinal level, i.e., we can at least say, of any two observations, which is greater. This method does not require the assumption that the difference between the two samples are normally distributed.

This method helps us to determine whether the two **samples have come from identical populations**. If it is true that the samples **have come from the same population**, it is reasonable to assume that the **medians of ranks assigned to the values of two samples are more or less the same**. The **alternative hypothesis  $H_1$**  would be: That the **medians of the populations are not equal**. In this case, most of the smaller ranks will go to the values of one sample, while most of the higher ranks will go to the other sample. The test involves the calculation of a statistic usually called **U**, whose distribution under the null hypothesis is known. In case of small samples (i.e., when the sample size is less than 8), the distribution is tabulated but for samples above 8, there is good approximation using the normal distribution.

## WORKING METHOD

**Step 1.** Set the null hypothesis  $H_0$  and alternative hypothesis  $H_1$ :

$H_0$  :- Both the medians are equal.

$H_1$  : Both the population medians are not equal.  $\Rightarrow$  a case of two tailed test

**Step 2.** Combine all sample values in an array from smallest to the largest, and assign ranks to all these values. If two or more sample values are identical (*i.e.*, there are tie scores), the sample values are each assigned a rank equal to the mean of the ranks that would otherwise be assigned.

**Step 3.** Find the sum of the ranks for each of the samples. Let us denote these sums by  $R_1$  and  $R_2$ . Also  $n_1$  and  $n_2$  are their respective sample sizes. For convenience, choose  $n_1$  as the smaller size if they are unequal so that  $n_1 \leq n_2$ .

A significant difference between the rank sums  $R_1$  and  $R_2$ , implies a significant difference between the samples.

**Step 4. Calculation of U-Statistic to test the difference between the rank sums.**

$$\text{U Statistic: } U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1 \quad [\text{Corresponding to Sample 1}]$$

$$\text{or } U = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2 \quad [\text{Corresponding to Sample 2}]$$

The sampling distribution of U is symmetrical and has a mean and variance given by the formulae:

$$\text{Mean: } \mu_U = \frac{n_1 n_2}{2}$$

$$\text{Variance: } = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

If  $n_1$  and  $n_2$  are both atleast equal to 8, it turns out that the distribution of U is nearly normal and one could use the statistic Z, where

$$Z = \frac{U - \mu_U}{\text{S.E.}(U)}$$

is normally distributed with mean 0 and variance 1.

**Step 5. Level of significance:** Take  $\alpha = 0.05$

**Step 6. Critical Region:** Accept  $H_0$  if  $|Z| < |Z_\alpha|$ , where  $Z_\alpha$  is the tabled value of Z for the given level of significance  $\alpha$ . The values of  $Z_\alpha$  are given by the normal Table given at the end.

**Remarks: 1.** In many applications Mann-Whitney test is used in place of two samples. t-test when normality assumption is questionable.

**2.** This test can be applied when the observations in a sample of data are ranks that is ordinal data rather than direct measurements.

**Example 5 :** The nicotine contents of two brands of cigarettes, measured in milligrams, was found to be as follows:

|           |     |     |     |     |     |     |     |     |     |     |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Brand A : | 2.1 | 4.0 | 6.3 | 5.4 | 4.8 | 3.7 | 6.1 | 3.3 |     |     |
| Brand B : | 4.1 | 0.6 | 3.1 | 2.5 | 4.0 | 6.2 | 1.6 | 2.2 | 1.9 | 5.4 |

Test the hypothesis, at the 0.05 level of significance, that the average nicotine contents of the two brands are equal against the alternative that they are unequal.

**Solution :**

**1. Setting up of Hypothesis**

**Null Hypothesis:**  $H_1 : \mu_1 = \mu_2$ , i.e., the average nicotine contents of the two brands are equal.

**Alternative Hypothesis:**  $H_1 : \mu_1 \neq \mu_2$ , i.e., the average nicotine contents of the two brands are not equal.

⇒ It is a case of two tailed test.

**2. Level of significance:** Here  $\alpha = 0.05$ .

**3. Computation of U-statistic:**

The observations are arranged in ascending order and ranks from 1 to 18 are assigned.

| <b>Original</b> |     |     |     |     |     |     |     |     |     |      |      |     |     |      |      |     |     |     |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|-----|-----|------|------|-----|-----|-----|
| <b>Data :</b>   | 0.6 | 1.6 | 1.9 | 2.1 | 2.2 | 2.5 | 3.1 | 3.3 | 3.7 | 4.0  | 4.0  | 4.1 | 4.8 | 5.4  | 5.4  | 6.1 | 6.2 | 6.3 |
| <b>Rank</b>     | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10.5 | 10.5 | 12  | 13  | 14.5 | 14.5 | 16  | 17  | 18  |

The ranks of the observations belonging to the smaller samples are underscored (put in bold form)

$$R_1 = 4 + 8 + 9 + 10.5 + 13 + 14.5 + 16 + 18 = 93$$

$$R_2 = 1 + 2 + 3 + 5 + 6 + 7 + 10.5 + 12 + 14.5 + 17 = 78.$$

Also,  $n_1 = 8$ ;  $n_2 = 10$ . For the calculation of U-statistic we take  $R_1$  = sum of the ranks of smaller groups

$$\therefore \text{U-Statistic:} \quad U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$$

$$= 8 \times 10 + \frac{8 \times 9}{2} - 93 = 80 + 36 - 93 = 23.$$

$$\therefore \text{Mean of } U = \mu_U = \frac{n_1 n_2}{2} = \frac{10 \times 8}{2} = 40.$$

$$\text{Variance of } U = \sigma_U^2 = \frac{(n_1 \times n_2)(n_1 + n_2 + 1)}{12}$$

$$= \frac{10 \times 8 (10 + 8 + 1)}{12} = \frac{380}{3} = 126.1 = 126.67.$$

$$\therefore \sigma_U = \sqrt{126.67} = 11.25.$$

Here  $n_2 = 10$ , so we can use the statistic

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{23 - 40}{11.25} = \frac{-17}{11.25} = -1.51.$$

$$\therefore |Z| = 1.51.$$

The tabled value of  $Z_\alpha$  at  $\alpha = 0.05$  is 1.96.

[From normal table]

Now,  $|Z| < |Z_\alpha|$  as  $1.51 < 1.96$ , so we accept the null hypothesis  $H_0$  and conclude that there is no significant difference in the average nicotine contents of the two brands of cigarettes.

**Example 6 :** The following are the weight gains (in pounds) of two random samples of young Indians fed on two different diets but otherwise kept under identical conditions:

|          |      |      |      |      |      |      |      |       |
|----------|------|------|------|------|------|------|------|-------|
| Diet I:  | 16.3 | 10.1 | 10.7 | 13.5 | 14.9 | 11.8 | 14.3 | 10.2  |
|          | 12.0 | 14.7 | 23.6 | 15.1 | 14.5 | 18.4 | 13.2 | 14.0  |
| Diet II: | 21.3 | 23.8 | 15.4 | 19.6 | 12.0 | 13.9 | 18.8 | 19.2  |
|          | 15.3 | 20.1 | 14.8 | 18.9 | 20.7 | 21.1 | 15.8 | 16.2. |

Use U test at 0.01 level of significance to test the null hypothesis that the two population samples are identical against the alternative hypothesis that on the average the second diet produces a greater gain in weight.

**Solution :**

**1. Setting up of Hypothesis:**

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

Alternative Hypothesis:  $H_1 : \mu_1 < \mu_2$ . It is a case of one tailed test.

**2. Level of significance:**  $\alpha = 0.01$

**3. Test Statistic:** Ranking the data jointly according to size we find the values first sample occupy the ranks: 21, 1, 3, 8, 15, 4, 11, 2, 5.5, 13, 31, 16, 12, 22, 7 and 10 (the fifth and sixth values are both 12, so we assigned each the rank 5.5).

$\therefore R_1 = \text{The sum of the ranks of the first sample}$

$$\begin{aligned} &= 21 + 1 + 3 + 8 + 15 + 4 + 11 + 2 + 5.5 + 13 + 31 + 16 + 12 + 22 + 7 + 10 \\ &= 181.5. \end{aligned}$$

Also  $n_1 = 16$ ,  $n_2 = 16$ .

$$\therefore \text{Statistics } U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$$

$$\text{or } U = 16 \times 16 + \frac{6(16+1)}{2} - 181.5 = 210.5.$$

As the sample sizes  $n_1$  and  $n_2$  are both greater than 8, so the distribution of  $U$  is nearly normal with

**Mean:**  $\mu_U = \frac{n_1 n_2}{2} = \frac{16 \times 16}{2} = 128.$

**Variance:**  $\sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{16 \times 16 (16 + 17 + 1)}{12}$   
 $= \frac{16 \times 16 \times 33}{2} = 704.$

$\therefore S.E. (U) = \sigma_U = \sqrt{704} = 26.53.$

**Test Statistic:**  $Z = \frac{U - \mu_U}{S.E.(U)} = \frac{210.5 - 128}{26.53} = \frac{82.5}{26.53} = 3.11.$

**Critical value.** From the normal tables the value of  $Z_\alpha$  for  $\alpha = 0.01$  is 2.33.

$\therefore |Z_\alpha| = 2.33.$

4. **Decision:** Reject the hypothesis if the calculated value of  $|Z|$  is more than the tabled value of  $|Z_\alpha|$ .

Here  $|Z| > |Z_\alpha|$  as  $3.11 > 2.33$ , so the null hypothesis is rejected and we conclude that on the average the second diet produces a greater gain in weight.

## 17.11 KRUSKAL-WALLIS TEST OR H-TEST

The Kruskal-Wallis test is a generalization of Mann-Whitney U-test to the case of  $k > 2$  samples. This test is also known as Kruskal-Wallis H-test. It is used to test the null hypothesis  $H_0$  that  $k$  independent samples are drawn from the identical population. This test is an alternative non-parametric test to the F-test for testing the equality of means in the one factor analysis of variance when the experimenter wishes to avoid the assumption that the samples were selected from the normal populations. Let  $n_i$  ( $i = 1, 2, 3, \dots, k$ ) be the number of observations in the  $i$ th sample. First we combine all  $k$  samples and arrange them to get  $n = n_1 + n_2 + n_3 + \dots + n_k$  observations in ascending order, substituting the appropriate rank from 1, 2, ...,  $n$  for each observation. In the case of ties (identical observations), we follow the usual procedure of replacing the observations by the means of the ranks that the observations would have if they were distinguishable. The sum of the ranks corresponding to the  $n_i$  observations in the  $i$ th sample is denoted by the random variable  $R_i$ . Now let us consider the statistic.

$$H = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1),$$

which is approximated very well by a chi-square distribution with  $k-1$  degrees of freedom when  $H_0$  is true and if each sample consists of at least five observations.

If  $H$  falls in the critical region  $H > \psi_\alpha^2$  with  $v = k-1$  degrees of freedom, we reject  $H$  at  $\alpha$  level of significance; otherwise, we accept  $H$ .

**Example 7 :** A teacher wishes to test three different leading methods I, II and III. To do this, the teacher chooses at random three groups of five students each and teaches each group by a different method. The same examination is then given to all the students and the marks obtained are given below. Determine at  $\alpha = 0.05$  significance level whether there is a difference between the teaching methods.

|            |   |    |    |    |    |    |
|------------|---|----|----|----|----|----|
| Method I   | : | 78 | 62 | 71 | 58 | 73 |
| Method II  | : | 76 | 85 | 77 | 90 | 87 |
| Method III | : | 74 | 79 | 60 | 75 | 80 |

**Solution :** Null Hypothesis:  $H_0$  : There is difference between the teaching methods.

Alternative Hypothesis:  $H_1$  : There is no difference between the three teaching methods.

Since there are three methods (I, II and III) and each group consists of 5 students, so we have  $N_1 = N_2 = N_3 = 5$ , and  $N = N_1 + N_2 + N_3 = 5 + 5 + 5 = 15$ . Arranging all these marks in increasing order of magnitude and assigning appropriate ranks, we get :

|        |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|--------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Marks: | 58 | 60 | 62 | 61 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 85 | 87 | 90 |
| Ranks: | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 |

Rank Table

|          |    |    |   |    |    |            |
|----------|----|----|---|----|----|------------|
| Method 1 | 10 | 3  | 4 | 1  | 5  | $23 = R_1$ |
| Method 2 | 8  | 13 | 9 | 15 | 14 | $59 = R_2$ |
| Method 3 | 6  | 11 | 2 | 7  | 12 | $38 = R_3$ |

$$\therefore H = \frac{12}{N(N+1)} \left[ \frac{R_1^2}{N_1} + \frac{R_2^2}{N_2} + \frac{R_3^2}{N_3} \right] - 3(N+1)$$

$$= \frac{12}{15 \times 16} \left[ \frac{(23)^2}{5} + \frac{(59)^2}{5} + \frac{(38)^2}{5} \right] - 3 \times 16$$

$$H = \frac{1}{100} [529 + 3481 + 1444] - 48 = \frac{5454}{100} - 48 = 6.54.$$

Here degrees of freedom =  $k - 1 = 3 - 1 = 2$ .

Also level of significance :  $\alpha = 0.05$

$\therefore \chi^2$  (for 2 d.f. and  $\alpha = 0.05$ ) =  $\chi^2_{0.05, 2} = 5.991$ . [From  $\chi^2$ -tables]

Decision: Reject  $H_0$  if  $H > \chi^2_{0.95}$

Now  $6.54 > 5.991 \Rightarrow H > \chi^2_{0.05}$

⇒ the null hypothesis  $H_0$  is rejected and the alternative hypothesis  $H_1$  is accepted.  
We conclude that there is no difference in teaching methods.

**Example 8 :** During one semester a student received in various subjects the marks shown below. Test at 0.05 significance level whether there is a difference between the marks in these subjects.

|             |    |    |    |    |
|-------------|----|----|----|----|
| Mathematics | 72 | 80 | 83 | 75 |
| Science     | 81 | 74 | 77 |    |
| English     | 88 | 82 | 90 | 87 |
| Economics   | 90 | 71 | 77 | 70 |

**Solution :** Null Hypothesis:  $H_0$  : There is a significant difference between the marks in these subjects.

Alternative Hypothesis:  $H_1$  : There is no difference between the marks in these subjects.

$$\therefore n = n_1 + n_2 + n_3 + n_4 = 4 + 3 + 5 + 4 = 16.$$

Arranging these marks in increasing order of magnitude and assigning appropriate ranks, we have

Table of Ranks

|        |    |    |    |    |    |     |     |     |     |    |    |    |    |    |      |      |
|--------|----|----|----|----|----|-----|-----|-----|-----|----|----|----|----|----|------|------|
| Marks: | 70 | 71 | 72 | 74 | 75 | 77  | 77  | 80  | 80  | 81 | 82 | 83 | 87 | 88 | 90   | 90   |
| Ranks: | 1  | 2  | 3  | 4  | 5  | 6.5 | 6.5 | 8.5 | 8.5 | 10 | 11 | 12 | 13 | 14 | 15.5 | 15.5 |

Sum of Ranks

|           |      |     |      |    |              |
|-----------|------|-----|------|----|--------------|
| Maths     | 3    | 8.5 | 12   | 5  | 28.5 = $R_1$ |
| Science   | 10   | 4   | 6.5  |    | 20.5 = $R_2$ |
| English   | 14   | 11  | 15.5 | 13 | 8.5 = $R_3$  |
| Economics | 15.5 | 2   | 6.5  | 1  | 25 = $R_4$   |

$$\text{Here } N_1 = 4, \quad N_2 = 3, \quad N_3 = 5, \quad N_4 = 4$$

$$\therefore N = N_1 + N_2 + N_3 + N_4 = 4 + 3 + 5 + 4 = 16.$$

$$\text{Also } R_1 = 28.5, \quad R_2 = 20.5, \quad R_3 = 62, \quad R_4 = 25.$$

$$\begin{aligned} \therefore H &= \frac{12}{N(N+1)} \left[ \frac{R_1^2}{N_1} + \frac{R_2^2}{N_2} + \frac{R_3^2}{N_3} + \frac{R_4^2}{N_4} \right] - 3(N+1) \\ &= \frac{12}{16 \times 17} \left[ \frac{(28.5)^2}{4} + \frac{(20.5)^2}{3} + \frac{(62)^2}{5} + \frac{(25)^2}{4} \right] - 3 \times 17 \\ &= \frac{12}{16 \times 17} \left[ \frac{812.25}{4} + \frac{420.25}{3} + \frac{3844}{5} + \frac{625}{4} \right] - 51 \\ &= \frac{12}{16 \times 17} [203.06 + 140.08 + 768.80 + 156.25] - 51 \end{aligned}$$

$$= \frac{12}{16 \times 17} (1268.19) - 51 = 55.95 - 51 = 4.95.$$

**Degrees of freedom** =  $k - 1 = 4 - 1 = 3$

**Level of significance** : Here  $\alpha = 0.05$

The value of  $\chi^2$  for 3 degrees of freedom for  $\alpha = 0.05$  is 7.81 [From  $\chi^2$  table]

**Decision:** Since the value of  $H$  is less than the tabled value of  $\chi^2$  at  $\alpha = 0.05$  for 3 degrees of freedom as  $4.95 < 7.81$ , so we accept the null hypothesis  $H_0 \Rightarrow$  there is a significant difference between the marks of these subjects.

**Example 9 :** The following are the final examination of marks of three groups of students who were taught computer by three different methods.

*First Method* : 94, 88, 91, 74, 87, 97

*Second Method* : 85, 82, 79, 84, 61, 72, 80

*Third Method* : 89, 67, 72, 76, 69.

Use the  $H$ -test at the 0.05 level of significance to test the null hypothesis that the three methods are equally effective.

**Solution :** 1. Setting up of Hypothesis:

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

**Alternative Hypothesis:**  $H_1 : \mu_1, \mu_2$  and  $\mu_3$  are not all equal.

2. Level of significance: Here  $\alpha = 0.05$ .

3. Calculation of Test Statistic: Ranking the data jointly from 1 to 18, we find that

$$R_1 = 6 + 13 + 14 + 16 + 17 + 18 = 84$$

$$R_2 = 1 + 4.5 + 8 + 9 + 10 + 12 = 55.5$$

$$R_3 = 2 + 3 + 4.5 + 7 + 15 = 31.5$$

[ $\because$  there is only one tie and the tied marks are each assigned the rank 5.5]

$$\text{Also } N_1 = 6, \quad N_2 = 7, \quad N_3 = 5$$

$$\therefore N = N_1 + N_2 + N_3 = 6 + 7 + 5 = 18.$$

$$\therefore \text{Test statistic : } H = \frac{12}{N(N+1)} \left[ \frac{R_1^2}{N_1} + \frac{R_2^2}{N_2} + \frac{R_3^2}{N_3} \right] - 3(N+1)$$

$$\text{or } H = \frac{12}{18 \times 19} \left[ \frac{(84)^2}{6} + \frac{(55.5)^2}{7} + \frac{(31.5)^2}{5} \right] - 3 \times 19$$

$$= \frac{2}{57} \left[ 1176 + \frac{3080.25}{7} + \frac{992.25}{5} \right]$$

$$= \frac{2}{57} [1176 + 440.04 + 198.45] - 57$$

$$= \frac{2}{57} \times 1814.49 - 57 = 63.67 - 57 = 6.67$$

Degrees of Freedom =  $k - 1 = 3 - 1 = 2$

$\therefore \psi^2$  (for 2 d.f. and  $\alpha = 0.05$ ) = 5.991. [From  $\psi^2$  table]

**Decision:** Since  $H = 6.67$  exceeds = 5.991, so the null hypothesis is rejected. We accept the alternative hypothesis  $H_1$  and conclude that the three methods are not equally effective.

### EXERCISE – 17.2

1. What is Mann–Whitney U test? When is it used?
2. Explain the Mann–Whitney U test with the help of an example.
3. A farmer wishes to determine whether there is a difference in yields between two different varieties of wheat, I and II. The following data shows the production of wheat per unit area using the two varieties. Can the farmer conclude at significance levels of (a) 0.05 and (b) 0.01 that a difference exists?

|                 |      |      |      |      |      |      |      |      |      |      |      |
|-----------------|------|------|------|------|------|------|------|------|------|------|------|
| <i>Wheat I</i>  | 15.9 | 15.3 | 16.4 | 14.9 | 15.3 | 16.0 | 14.6 | 15.3 | 14.5 | 16.6 | 16.0 |
| <i>Wheat II</i> | 16.4 | 16.8 | 17.1 | 16.9 | 18.0 | 15.6 | 18.1 | 17.2 | 15.4 |      |      |

4. Instructors  $A$  and  $B$  both teach a first course in chemistry at XYZ University. On a common final examination, their students received the marks shown as given below. Test at the 0.05 significance level the hypothesis that there is no difference between the two instructors' grades.

|          |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| <i>A</i> | 88 | 75 | 92 | 71 | 63 | 84 | 55 | 64 | 82 | 96 | .  |    |    |    |
| <i>B</i> | 72 | 65 | 84 | 53 | 76 | 80 | 51 | 60 | 57 | 85 | 94 | 87 | 73 | 61 |

5. A company wishes to determine whether there is a difference between two brands of gasoline,  $A$  and  $B$ . The following data shows the distances travelled in kilometres per litre for each brand. Can we conclude at the 0.05 significance level (a) that there is a difference between the brands and (b) that brand  $B$  is better than brand  $A$ ?

|          |      |      |      |      |      |      |      |      |      |      |   |
|----------|------|------|------|------|------|------|------|------|------|------|---|
| <i>A</i> | 15.2 | 14.4 | 14.6 | 16.3 | 15.9 | 14.7 | 15.4 | 15.6 | 15.3 | 16.0 | . |
| <i>B</i> | 16.7 | 14.9 | 15.0 | 15.8 | 16.9 | 15.5 | 15.7 | 14.8 | 16.4 | 16.5 | . |

6. Explain Kruskal–Wallis test with the help of an example.
7. In an experiment to determine which of three different missile systems is preferable, the propellant burning rates was measured. The data after coding is given below.

|                   |      |      |      |      |      |      |      |      |
|-------------------|------|------|------|------|------|------|------|------|
| <i>System I</i>   | 24.0 | 16.7 | 22.8 | 19.8 | 18.9 | .    |      |      |
| <i>System II</i>  | 23.2 | 19.8 | 18.1 | 17.6 | 20.2 | 17.8 |      |      |
| <i>System III</i> | 18.4 | 19.1 | 17.3 | 17.3 | 19.7 | 18.9 | 18.8 | 19.3 |

Use the Kruskal–Wallis test and a significance level of  $\alpha = 0.05$  to test the hypothesis that propellant burning rates are the same for the three missiles.

[Hint. Let  $H_0 : \mu_1 = \mu_2 = \mu_3, \alpha = 0.05$ .

**Table : Ranks for Propellant Burning Rates**

| Missile System I | Missile System II | Missile System III |
|------------------|-------------------|--------------------|
| 19               | 18                | 7                  |
| 1                | 14.5              | 11                 |
| 17               | 6                 | 2.5                |
| 14.5             | 4                 | 2.5                |
| 9.5              | 16                | 13                 |
|                  | 5                 | 9.5                |
|                  |                   | 8                  |
|                  |                   | 12                 |
| $R_1 = 61.0$     |                   | $R_2 = 63.5$       |
|                  |                   | $R_3 = 65.5$       |

Also  $n_1 = 5, n_2 = 6, n_3 = 8$  and  $n = n_1 + n_2 + n_3 = 19$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1) \\ &= \frac{12}{19 \times 20} \left[ \frac{(61.0)^2}{5} + \frac{(63.5)^2}{6} + \frac{(65.5)^2}{8} \right] - 3 \times 20 = 1.66. \end{aligned}$$

Also, the value of **for 18 degrees of freedom at  $\alpha = 0.05$  is 28.869.** [from  $\psi^2$  tables]

Since  $H < 1.66 < 5.991$ , so we accept the null hypothesis  $H_0$ ]

8. The same mathematics papers were marked by three teachers A, B and C. The final marks were recorded as follows :

|           |    |    |    |    |    |    |    |    |    |    |    |    |
|-----------|----|----|----|----|----|----|----|----|----|----|----|----|
| Teacher A | 73 | 89 | 82 | 43 | 80 | 73 | 66 | 60 | 45 | 93 | 36 | 77 |
| Teacher B | 88 | 78 | 48 | 91 | 51 | 85 | 74 | 77 | 31 | 78 | 62 | 76 |
| Teacher C | 68 | 79 | 56 | 91 | 71 | 71 | 87 | 41 | 59 | 68 | 53 | 79 |

Use Kruskal-Wallis test, at the 0.05 level of significance to determine if the marks distributions given by the 3 teachers differ significantly.

9. The following data represent the operating times in hours for 3 types of scientific pocket calculators before a recharge is required.

*Calculator*

| A   | B   | C   |
|-----|-----|-----|
| 4.9 | 5.5 | 6.4 |
| 6.1 | 5.4 | 6.8 |
| 4.3 | 6.2 | 5.6 |
| 4.6 | 5.8 | 6.5 |
| 5.3 | 5.5 | 6.3 |
|     | 5.2 | 6.6 |
|     | 4.8 |     |

Use Kruskal–Wallis test, at the 0.01 level of significance, to test the hypothesis that the operating times for all three calculators are equal.

10. The following are the miles per gallon which a test driver got for 10 tankfuls of each of three kinds of gasoline:

|                     |    |    |    |    |    |    |    |    |    |     |
|---------------------|----|----|----|----|----|----|----|----|----|-----|
| <i>Gasoline A :</i> | 20 | 31 | 24 | 33 | 23 | 24 | 28 | 16 | 19 | 26  |
| <i>Gasoline B :</i> | 29 | 18 | 29 | 19 | 20 | 21 | 34 | 33 | 30 | 23  |
| <i>Gasoline C :</i> | 19 | 31 | 16 | 26 | 31 | 33 | 28 | 28 | 25 | 30. |

Use the Kruskal–Wallis test at the 0.05 level of significance to test whether or not there is a difference in the actual average mileage yield of the three kinds of gasoline.

11. Three groups of guinea pigs were injected, respectively, with 0.5, 1.0 and 1.5 milligrams of a tranquilizer, and the following are the numbers of seconds it took them to fall asleep:

*0.5-mg dose* : 8.2 10.0 10.2 13.7 14.0 7.8 12.7 10.9  
*1.0-mg dose* : 9.7 13.1 11.0 7.5 13.3 12.5 8.8 12.9 7.9 10.5  
*1.5-mg dose* : 12.0 7.2 8.0 9.4 11.3 9.0 11.5 8.5.

Use the  $H$  test at the 0.01 level of significance to test the null hypothesis that the difference in dosage have no effect on the length of time it takes guinea pigs to fall asleep.

12. To compare four bowling balls, a professional bowler bowls five games with each ball and gets the following results:

|                 |     |     |     |     |     |
|-----------------|-----|-----|-----|-----|-----|
| <i>Ball A</i> : | 208 | 220 | 247 | 192 | 229 |
| <i>Ball B</i> : | 216 | 196 | 189 | 205 | 210 |
| <i>Ball C</i> : | 226 | 218 | 252 | 225 | 202 |
| <i>Ball D</i> : | 212 | 198 | 207 | 232 | 221 |

Use the Kruskal–Wallis test at the 0.05 level of significance to test whether or not the bowler can expect to score equally well with the four bowling balls.

## ANSWERS

## 17.12 THE ONE SAMPLE RUNS TEST

It is a non-parametric method to determine the randomness with which the samples items have been selected. The runs test, based on the order in which the sample observations are obtained, is a useful technique for testing the null-hypothesis  $H_0$  that the observations have indeed been drawn at random. The runs test can also be used to detect departures in randomness of a sequence of quantitative measurements over time, caused by trends or periodicities.

**RUN.** A run is a subsequence of one or more identical symbols representing a common property of the data. In other words, a run is defined as a set of identical (or related) symbols contained between two different symbols or no symbol (such as at the beginning or end of the sequence).

To understand what a run is, we consider a sequence made up of two symbols,  $A$  and  $B$  such as;

$$AAA | BB | A | BBB | AAAA | BBBB | AA | \dots \text{ (I)}$$

In tossing a coin, for example,  $A$ 's could represent 'heads' and  $B$ 's represent 'tails'.

Proceeding from left to right in the above sequence (I), the first run indicated by a vertical bar | consists of three  $A$ 's, the second run consists of two  $B$ 's, the third run consists of one  $A$ , the fourth run consists of three  $B$ 's, the fifth run consists of four  $A$ 's, the sixth run consists of four  $B$ 's and the seventh run consists of two  $A$ 's. Thus, there are seven runs in all.

It seems that some relationship exists between randomness and the number of runs. Consider the another sequence

$$A | B | A | B | A | B | A | B | A | B | \dots \text{ (II)}$$

There is a cyclic pattern, in the above sequence in which we go from  $A$  to  $B$ , back to  $A$  again, etc., which we could hardly believe to be random. In such case, we have too many runs (in fact, we have the maximum number possible for the given number of  $A$ 's and  $B$ 's).

On the other hand, the following sequence  $AAAAAA | BBBB | AAAA | BBB | AAAA | BB | AAA | B |$  seems to be a trend pattern, in which  $A$ 's and  $B$ 's are grouped together in a pattern. In such there are too few runs, and we could not consider the sequence to be random.

We notice that: if there are too few runs, a definite grouping, or clustering or trend may be suspected. On the other hand, if there are too many runs some sort of repeated alternating pattern may be suspected." Thus, it may be possible to prove that too many or too few runs in a sample indicate something other than chance, (or randomness), when items were selected.

### Sampling Distribution of V-Statistic

The number of runs ' $V$ ' is a statistic with a special sampling distribution. The mean  $\mu_V$  of  $V$ -statistic is given by  $\mu_V$ , where

Mean of  $V$ -Statistic:  $\mu_V = \frac{2n_1 n_2}{n_1 + n_2} + 1,$

where

$n_1$  = the number of first response;

$n_2$  = the number of second response.

The variance of  $V$ -statistic is  $\sigma_V^2$ , where

$$\text{Variance of } V\text{-statistic: } \sigma_V^2 = \left[ \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} \right].$$

The standard error S.E. ( $V$ ) of the  $V$ -statistic is given by  $\sigma_V$ .

The sampling distribution of  $V$ -statistic can be closely approximated by the normal distribution if either  $n_1$  or  $n_2$  is atleast equal to 8. In that case,

**TEST STATISTIC:**

$$Z = \frac{V - (\text{mean of } V\text{-statistic})}{\text{S.E.}(V)} \quad \text{or} \quad Z = \frac{V - \mu_V}{\sigma_V}$$

is normally distributed with mean 0 and variance 1 and thus normal Table can be used.

**Example 10 :** Twenty five individuals were sampled as to whether they like or did not like a product indicated by  $Y$  and  $N$  respectively. The resulting sample is shown by the following sequence:

$YY \ NNNN \ YYY \ N Y NN \ Y NNNNN \ YYYY \ NN.$

(a) Determine the number of runs,  $V$ .

(b) Test at 0.05 significance level whether the responses are random.

**Solution :** (a) Using a vertical bar to indicate a run, we have

$YY | NNNN | YYY | N | Y | NN | Y | NNNNN | YYYY | NN |$

$\therefore \text{Number of runs } V = 10.$

(b) Let  $n_1$ ,  $n_2$  respectively denote the number of  $Y$ 's and  $N$ 's.

$\therefore n_1 = 11 \quad \text{and} \quad n_2 = 14. \quad \text{Also} \quad V = 10.$

### 1. Setting of Hypothesis:

Null hypothesis :  $H_0$  : The responses are random.

Alternative Hypothesis :  $H_1$  : The response are not random

$\Rightarrow$  It is a case of two tailed test.

### 2. Computation of test statistic:

$$\mu_V = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \times 11 \times 14}{11 + 14} + 1 = \frac{308}{25} + 1 = 12.32 + 1 = 13.32.$$

$$\sigma_V^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} = \frac{2 \times 11 \times 14 (2 \times 11 \times 14 - 11 - 14)}{(11 + 14)^2 (11 + 14 - 1)}$$

$$= \frac{308 (308 - 25)}{(25)^2 \times 23} = \frac{308 \times 283}{14375} = \frac{87164}{14375} = 6.064.$$

$$\therefore \sigma_V = \sqrt{6.064} = 2.462.$$

$$\therefore Z = \frac{V - \mu_V}{\sigma_V} = \frac{10 - 13.32}{2.462} = \frac{-3.32}{2.462} = -1.348.$$

$$\therefore |Z| = 1.348.$$

**3. Level of significance:**  $\alpha = 0.05$ .

**4. Critical value:** The value of  $Z_\alpha$  at 5% level of significance for two tailed test is 1.96  
[From the table]

$$i.e., |Z_\alpha| = 1.96.$$

**5. Decision:** Since, the calculated value of  $|Z|$  is less than the tabled value of  $|Z_\alpha|$  as  $1.348 < 1.96$ , so we accept the null hypothesis  $\Rightarrow$  the responses are random.

**Example 11 :** A machine is used to insert randomly one of the two types of toys A and B in each box. The manufactures choose a samples of 60 successive boxes to see if the machine is properly mixing the two types of toys. Test at 0.05 significance level whether the machine is inserting the toys in random order.

B A BBB AAA BB A BBBB AAAA B A B AA BBB AA B A AAA BB A BB AAAA BB A BBBB AA BB A B AA BB.

**Solution :** Let  $n_1, n_2$  respectively denote the number of boxes containing toy A and toy B.

Using a vertical bar to indicate a run, we have

B | A | BBB | AAA | BB | A | BBBB | AAAA | B | A | B | AA | BBB | AA | B | AAAA | BB | A | BB | AAAA | BB | A | BBBB | AA | BB | A | B | AA | BB |

$\therefore$  The number of Runs:  $V =$  the number of vertical lines = 29.

**1. Setting up of Hypothesis:**

**Null Hypothesis :**  $H_0$  : The toys are being inserted in the boxes in random order.

**Alternative Hypothesis :**  $H_1$  : The toys are not being inserted in the boxes in random order  $\Rightarrow$  It is a case of two tailed.

**2. Computation of test statistic V:**

Here  $n_1 = 29, n_2 = 31$  and  $V = 29$ .

$$\begin{aligned}\text{Mean of } V\text{-Statistic} &= \mu_V = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \times 29 \times 31}{29 + 31} + 1 = \frac{1798}{60} + 1 \\ &= 2997 + 1 = 30.97.\end{aligned}$$

$$\begin{aligned}\text{Variance of } V\text{-Statistic} &= \sigma_V^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} \\ &= \frac{(2 \times 29 \times 31)(2 \times 29 \times 31 - 29 - 31)}{(29 + 31)^2 (29 + 31 - 1)} = \frac{1798 \times 1738}{(60)^2 \times 59}\end{aligned}$$

$$= \frac{3124924}{3600 \times 59} = \frac{3124924}{212400} = 14.712.$$

$$\therefore \sigma_V = \sqrt{14.712} = 3.84.$$

$\therefore$  Stand Error of  $V$  : S.E. ( $V$ ) =  $\sigma_V$  = 3.84.

$$\text{Test Statistic : } Z = \frac{V - \mu_V}{\sigma_V} = \frac{29 - 30.97}{3.84} = -\frac{1.97}{3.84} = -0.513.$$

$$\therefore |Z| = 0.513.$$

**3. Level of significance:**  $\alpha = 0.05$

**4. Critical value:** The tabled value of  $|Z_\alpha|$  at 5% level of significance for two tailed test is 1.96.

**5. Decision:** Since, the calculated value of  $|Z|$  is less than the tabled value of  $|Z_\alpha|$  as  $0.513 < 1.96$ , so we accept the null hypothesis  $H_0 \Rightarrow$  the toys are being inserted in the boxes in random order.

**Example 12 :** In 30 tosses of a coin the following sequence of heads (H) and tails (T) is obtained

H T T H T H H H T H H T T H T  
H T H H T H T T H T H H T H T H T.

(a) Determine the number of runs,  $V$ .

(b) Test at the 0.05 significance level whether the sequence is random.

**Solution :** (a)

Using a vertical bar to indicate a run, we have :

H | T T | H | T | H H H | T | H H | T T | H | T |  
H | T | H H | T | H | T T | H | T | H H | T | H | T |

$\therefore$  The number of Runs:  $V = 22$ .

**Solution :** (b)

**1. Setting of Hypothesis :**

**Null Hypothesis :**  $H_0$  : The sequence is random.

**Alternative Hypothesis :**  $H_1$  : The sequence is not random  $\Rightarrow$  It is a case of two tailed test.

**2. Calculation of Test Statistic:** Let  $n_1, n_2$  denote respectively the number of heads and tails. Then  $n_1 = 16$  heads;  $n_2 = 14$  tails.

**Number of Runs:**  $V = 22$ .

Now, 
$$\mu_V = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \times 16 \times 14}{16 + 14} + 1 = 15.93.$$

$$\begin{aligned}\sigma_V^2 &= \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} = \frac{2 \times 16 \times 14 (2 \times 16 \times 14 - 16 - 14)}{(16 + 14)^2 (16 + 14 - 1)} \\ &= \frac{448 (448 - 30)}{900 \times 29} = \frac{187264}{26100} = 7.175.\end{aligned}$$

$$\therefore \sigma_V = \sqrt{7.175} = 2.679.$$

$$\text{Test Statistic: } Z = \frac{V - \mu_V}{\sigma_V} = \frac{22 - 15.93}{2.679} = \frac{6.07}{2.679} = 2.27.$$

**3. Level of Significance:**  $\alpha = 0.05$ .

**4. Critical value:** The value of  $Z_\alpha$  at 5% level of significance for two tailed test is 1.96.  
[From the table]

**5. Decision:** Since, the calculated value of  $Z$  is greater than the tabled value of  $Z_\alpha$  as  $2.27 > 1.96$ , so we reject the null hypothesis  $H_0$  and accept the alternative hypothesis  $H_1 \Rightarrow$  the sequence is not random.

**Example 13 :** The following is an arrangement of men, M, and women, W, lined up to purchase tickets for a rock concert:

M W M W M M M W M W M M M W W M M M M W W M W M  
M M W M M M W W W M W M M M W M W M M M M W W M

Test for randomness of the arrangement at the 0.05 level of significance.

**Solution :**

**1. Setting up of Hypothesis:**

Null Hypothesis :  $H_0$  : Arrangement is random.

Alternative Hypothesis :  $H_1$  : Arrangement is not random. It is a case of two tailed test.

**2. Level of Significance :**  $\alpha = 0.05$ .

**3. Test Statistic:**  $Z = \frac{V - \mu_V}{\sigma_V}$ , where

$$\mu_V = \frac{2n_1 n_2}{n_1 + n_2} + 1 \quad \text{and} \quad \sigma_V^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}.$$

Since,  $n_1 = 30$ ,  $n_2 = 18$  and  $V = 27$ , we get

$$\mu_V = \frac{2 \cdot 30 \cdot 18}{30 + 18} + 1 = 23.5$$

$$\sigma_V = \sqrt{\frac{2 \cdot 30 \cdot 18 (2 \cdot 30 \cdot 18 - 30 - 18)}{(30 + 18)^2 (30 + 18 - 1)}} = 3.21$$

and hence,  $Z = \frac{27 - 23.5}{3.21} = 1.09.$

**4. Critical values:** The value of  $|Z_\alpha|$  for  $\alpha = 0.05$  is 1.96. [From normal tables]

**5. Decision:** Since,  $Z = 1.09$  falls between  $-1.96$  and  $1.96$ , so the null hypothesis cannot be rejected. In other words, there is no real evidence to indicate that the arrangement is not random.

### 17.13 MEDIAN TEST FOR RANDOMNESS OR RUNS ABOVE AND BELOW THE MEDIAN

The method discussed in the last section is not limited to test the randomness of series of attributes. Any sample which consists of numerical measurements or observations can be treated similarly by using the letters  $a$  and  $b$  to denote respectively, the values falling above and below the median of the sample. (Numbers equaling the median are omitted). The resulting series of  $a$ 's and  $b$ 's can then be tested for randomness on the basis of the total number of runs of  $a$ 's and  $b$ 's, namely, the total number of runs above and below the median.

**Example 14 :** The following are the speeds (in kilometre per hour) at which every fifth passenger car was timed at a certain checkpoint: 46, 58, 60, 56, 70, 66, 48, 54, 62, 41, 39, 52, 45, 62, 53, 69, 65, 65, 67, 76, 52, 52, 59, 59, 67, 51, 46, 61, 40, 43, 42, 77, 67, 63, 59, 63, 63, 72, 57, 59, 42, 56, 47, 62, 67, 70, 63, 66, 69 and 73. Test the null hypothesis of randomness at the 0.05 level of significance. [Given median speed = 59.5 km per hour]

**Solution :** Let the number greater than 59.5 be denoted 'a' and the number less than 59.5 be denoted by 'b'.

#### 1. Setting the Hypothesis:

Null Hypothesis :  $H_0$  : The sample is random.

Alternative Hypothesis :  $H_1$  : The sample is not random.

#### 2. Level of Significance: $\alpha = 0.05$

#### 3. Test Statistic:

Here median speed = 59.5.

Since, the median of the speeds is 59.5, we get the following arrangement of  $a$ 's and  $b$ 's:

$b \ b \ |a| \ b \ |a \ a| \ b \ b \ |a| \ b \ b \ b \ |a| \ b \ |a \ a \ a \ a| \ b \ b \ b \ b \ |a|$   
 $b \ b \ |a| \ b \ b \ b \ |a \ a \ a| \ b \ |a \ a \ a| \ b \ b \ b \ b \ b \ |a \ a \ a \ a \ a|$

Then, since  $n_1 = 25$ ,  $n_2 = 25$  and  $V = 20$ , we get:

$$\mu_V = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \times 25 \times 25}{25 + 25} + 1 = 26.$$

$$\sigma_V^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

$$\Rightarrow \sigma_V^2 = \frac{2 \times 25 \times 25 (2 \times 25 \times 25 - 25 - 25)}{(25 + 25)^2 (25 + 25 - 1)} = 12.2.$$

$$\therefore Z = \frac{V - \mu_V}{\sigma_V} = \frac{20 - 26}{\sqrt{12.2}} = -1.72.$$

$$\therefore |Z| = 1.72$$

**4. Critical value:** The value of  $|Z_\alpha|$  for  $\alpha = 0.05$  from the table is  $|Z_\alpha| = 1.96$ .

**5. Decision:** Since the calculated value of  $|Z| <$  Tabled value of  $|Z_\alpha|$  as  $1.72 < 1.96$ , so the null hypothesis is accepted  $\Rightarrow$  the sample is random.

### EXERCISE – 17.3

1. Determine the number of runs,  $V$ , for each of these sequences. Also determine their mean and variance.

- (i) HH TTT H T HHH TTTT H T H T TT H TT
- (ii) A B AA BBB A BB AA BBBB A B A BB AA BBB
- (iii) M W MM WW MM WWW M WW MM WW MMM
- (iv) A B A BB AAA BB A B.

2. Use runs test on the sequence (iii) and (iv) of question 1 and test their randomness at 5% level of significance.

3. A sample of 48 tools produced by a machine shows the following sequence of good ( $G$ ) and defective ( $D$ ) tools:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $G$ | $G$ | $G$ | $G$ | $G$ | $G$ | $D$ | $D$ | $G$ |
| $G$ | $G$ | $D$ | $D$ | $D$ | $D$ | $G$ | $G$ | $G$ | $G$ | $G$ | $G$ | $D$ | $G$ | $G$ | $G$ |
| $G$ | $G$ | $G$ | $G$ | $G$ | $G$ | $D$ | $D$ | $G$ | $G$ | $G$ | $G$ | $G$ | $D$ | $G$ | $G$ |

Test the randomness of the sequence at the 0.05 significance level.

[Hint. Let  $n_1, n_2$  denote the number of  $D$ 's and  $G$ 's respectively. Here  $V = 11$ .

Let  $H_0$  = the sequence is random.

$$\text{Now, } \mu_V = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \times 10 \times 38}{10 + 38} + 1 = 16.83.$$

$$\sigma_V^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} = \frac{(2 \times 10 \times 38)(2 \times 10 \times 38 - 10 - 38)}{(10 + 38)^2 (10 + 38 - 1)} = 4.997$$

$$\Rightarrow \sigma_V = 2.235.$$

$$\therefore Z = \frac{11 - 16.83}{2.235} = -2.61 \Rightarrow |Z| = 2.61.$$

Also  $|Z_\alpha|$  at  $\alpha = 0.05 = 1.96$  [from tables.]

Now  $|Z| > |Z_\alpha|$  as  $2.235 > 1.96 \Rightarrow H_0$  is rejected  $\Rightarrow$  the sequence is not random.]

4. The following is the arrangement of 25 men ( $M$ ) and 15 women ( $W$ ) lined up to purchase tickets for a premier picture show:

$$M | WW | MMM | W | MM | W | M | W | M | WWW | MMM | \\ W | MM | WWW | MMMMM | WWW | MMMMM |$$

Test for randomness at 5% level of significance.

[Hint. Let  $H_0$  : the arrangement is random.

Here  $n_1 = 25$ ,  $n_2 = 15$ ,  $V = 17$ .

$$\therefore \mu_V = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \times 25 \times 15}{25 + 15} + 1 = 19.75.$$

$$\sigma_V^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} = \frac{(2 \times 25 \times 15)(2 \times 25 \times 15 - 25 - 15)}{(25 + 15)^2 (25 + 15 - 1)} \\ = \frac{750 \times 710}{1600 \times 39} = 8.533.$$

$$\therefore \sigma_V = \sqrt{8.533} = 2.92.$$

$$\text{Test Statistic: } Z = \frac{V - \mu_V}{\sigma_V} = \frac{17 - 19.75}{2.92} = -0.94.$$

$$\therefore |Z| = 0.94.$$

Also  $Z_\alpha$  at 5% = 1.96.

Since,  $|Z| < |Z_\alpha|$  at 5% as  $0.94 < 1.96$ , so we accept the null hypothesis  $H_0 \Rightarrow$  the arrangement is random.]

5. Form all possible sequence consisting of three  $a$ 's and two  $b$ 's and give the number of runs,  $V$ , corresponding to each sequence. Also obtain the sampling distribution of  $V$ .

[Hint. (a) The number of possible sequences consisting of three  $a$ 's and two  $b$ 's is

$${}^5C_2 = \frac{5!}{2! \times 3!} = 10. \text{ These sequences are shown in Table 1, along with the number of runs}$$

corresponding to each sequence.

(b) The sampling distribution of  $V$  is given in Table 2, where  $V$  denotes the number of runs and  $f$  denotes the frequency. For example, Table 2 shows that there is one 5, four 4's, etc.

Table 1

| Sequence |   |   |   |   | Runs ( $V$ ) |
|----------|---|---|---|---|--------------|
| a        | a | a | b | b | 2            |
| a        | a | b | a | b | 4            |
| a        | a | b | b | a | 3            |
| a        | b | a | b | a | 5            |
| a        | b | b | a | a | 3            |
| a        | b | a | a | b | 4            |
| b        | b | a | a | a | 2            |
| b        | a | b | a | a | 4            |
| b        | a | a | a | b | 3            |
| b        | a | a | b | a | 4            |

Table 2

| $V$ | $f$ |
|-----|-----|
| 2   | 2   |
| 3   | 3   |
| 4   | 4   |
| 5   | 1   |

6. Find mean and variance of question number 5.

[Hint. Here  $n_1 = 3$ ,  $n_2 = 2$ .

$$\therefore \mu_V = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \times 3 \times 2}{3 + 2} + 1 = .$$

$$\sigma_V^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} = \frac{2 \times 3 \times 2 (2 \times 3 \times 2 - 3 - 2)}{(3 + 2)^2 (3 + 2 - 1)} = \frac{21}{25} .$$

7. What conclusion can you draw using runs test concerning the randomness of the following digits at  $\alpha = 0.05$  significance level.

(a)  $\sqrt{2} = 1.4142135623730950488 \dots$

(b)  $\sqrt{3} = 1.73205 \ 08075 \ 68877 \ 2935 \dots$

(c)  $\pi = 3.14159 \ 26535 \ 89793 \ 2643 \dots$

8. A driver buys gasoline either at a Texaco station,  $T$ , or at a Mobil station,  $M$ , and the following arrangement shows the order of the stations from which she bought gasoline over a certain period of time:

$T \ T \ T \ M \ M \ T \ M \ M \ T \ T \ M \ T \ M \ T \ M \ T \ M \ M \ T \ M \ T$

Test for randomness at the 0.05 level of significance.

9. The numbers of retail stores that opened for business and also quit business in the same year were 108, 103, 109, 107, 125, 142, 147, 122, 116, 153, 144, 162, 143, 126, 145, 129, 134, 137, 143, 150, 148, 152, 125, 106, 112, 139, 132, 122, 138, 148, 155, 146 and 158 during a period of 33 years. Making use of the fact that the medium is 138, test at the 0.05 level of significance whether there is a real trend.
10. The following are the numbers of students absent from school on 24 consecutive school days: 29, 25, 31, 28, 30, 28, 33, 31, 35, 29, 31, 33, 35, 28, 36, 30, 33, 26, 30, 28, 32, 31, 38 and 27. Test for randomness at the 0.01 level of significance.

## ANSWERS



## 17.14 SPEARMAN'S RANK CORRELATION TEST

## Spearman's Rank Correlation

**Spearman rank correlation** is a measure of the correlation that exists between the two sets of ranks, or it is a measure of degree of association between the variables that we would not have been able to calculate otherwise.

## Rank Correlation Coefficient

The simple correlation coefficient  $r$  measures the linear relationship between two variables  $X$  and  $Y$ . If ranks  $1, 2, \dots, n$  are assigned to the  $X$  observations in order of magnitude and similarly to the  $Y$  observations, and if these ranks are then substituted for the actual numerical values into the formula for  $r$ , we obtain the non-parametric counterpart of the conventional correlation coefficient. A correlation coefficient calculated in this manner is known as the **Spearman's rank correlation coefficient\*** and is denoted by  $r_s$ . When there are no ties among either set of measurements, the formula for  $r_s$  reduces to a much simpler expression, which is given below.

**Rank Correlation Coefficient.** A non-parametric measure of association between two variables  $X$  and  $Y$  is given by the **rank correlation coefficient**

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)}, \quad \dots \text{ (A)}$$

**where**

- $\Sigma$  = notation meaning “the sum of”
- $d_i$  = the difference between the ranks assigned to  $x_i$  and  $y_i$
- $n$  = the number of paired observations
- $r_s$  = coefficient of rank correlation

[Notice that the subscripts, from Spearman distinguishes this  $r$  from sample correlation coefficient discussed in Chapter 7\*]

In practice, the preceding formula is also used when there are ties either among  $x$  or  $y$  observations. The ranks for tied observations are assigned by averaging the ranks that would have been assigned if the observations were distinguished.

**Formula for Tie Rank:** In the formula (A) add the correction factor  $[m(m^2 - 1)]/12$  to  $\Sigma D^2$ , where  $m$  is the number of times an item is repeated. This correction factor is to be added for each repeated items in both the series.

$$\therefore r_s = 1 - \frac{6[\sum D^2 + (1/12)(m_1^3 - m_1) + (1/12)(m_2^3 - m_2) + \dots]}{n(n^2 - 1)}$$

The value of  $r_s$  will usually be close to the value obtained by finding  $r$  based on numerical measurements and is interpreted in much the same way. As before, the values of  $r_s$  will range from  $-1$  to  $+1$ . A value of  $+1$  or  $-1$  indicates perfect association between  $X$  and  $Y$ , the plus sign occurring for identical rankings and the minus sign occurring for reverse rankings. When  $r_s$  is close to zero, we would conclude that the variables are uncorrelated.

### Advantages of Rank Correlation Coefficient

1. We do not assume the underlying relationship between  $X$  and  $Y$  to be linear and, therefore, when the data possess a distinct curvilinear relationship, the rank correlation coefficient will likely be more reliable than the conventional measure.
2. A second advantage in using the rank correlation coefficient is the fact that no assumptions of normality are made concerning the distributions of  $X$  and  $Y$ .
3. The greatest advantage occurs when one is unable to make meaningful numerical measurements but nevertheless can establish rankings. Such is the case, for example, when different judges rank a group of individuals according to some attribute. The rank correlation coefficient can be used in this situation as a measure of the consistency of the two judges.

**Example 15 :** The following figures, released by the Federal Trade Commission, show the milligrams of tar and nicotine found in 10 brands of cigarettes.

| Cigarette Brand | Tar Content | Nicotine Content |
|-----------------|-------------|------------------|
| Viceroy         | 14          | 0.9              |
| Marlboro        | 16          | 1.1              |
| Chesterfield    | 28          | 1.6              |
| Kool            | 17          | 1.3              |
| Kent            | 15          | 1.0              |
| Raleigh         | 13          | 0.8              |
| Old Gold        | 24          | 1.5              |
| Philip Morris   | 25          | 1.4              |
| Oasis           | 18          | 1.2              |
| Players         | 31          | 2.0              |

Calculate the rank correlation coefficient to measure the degree of relationship between tar and nicotine content in cigarettes.

**Solution :** Let  $X$  and  $Y$  respectively represent the tar and nicotine contents. We assign ranks to each set of measurements with the rank of 1 assigned to the lowest number in each set, the rank of 2 to the second lowest number in each set, and so forth, until the rank 10 is assigned to the largest number. The following table shows the individual rankings of the measurements and the differences in ranks for the 10 pairs of observations.

Table : Rankings for Tar and Nicotine Contents

| Cigarette Brand | $x_i$ | $y_i$ | $d_i = (x_i - y_i)$ | $d_i^2$            |
|-----------------|-------|-------|---------------------|--------------------|
| Viceroy         | 2     | 2     | 0                   | 0                  |
| Marlboro        | 4     | 4     | 0                   | 0                  |
| Chesterfield    | 9     | 9     | 0                   | 0                  |
| Kool            | 5     | 6     | -1                  | 1                  |
| Kent            | 3     | 3     | 0                   | 0                  |
| Raleigh         | 1     | 1     | 0                   | 0                  |
| Old Gold        | 7     | 8     | -1                  | 1                  |
| Philip Morris   | 8     | 7     | 1                   | 1                  |
| Oasis           | 6     | 5     | 1                   | 1                  |
| Players         | 10    | 10    | 0                   | 0                  |
|                 |       |       |                     | $\Sigma d_i^2 = 4$ |

Substituting into the formula for  $r_s$ , we get

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{(6)(4)}{(10)(100 - 1)} = 1 - 0.0242 = 0.9758$$

Hence,  $r_s = 0.9758$

Also  $r_s = 0.9758$  indicates a high positive correlation between the amount of tar and nicotine found in cigarettes.

**Example 16 :** Ten competitors in a beauty contest are ranked by three judges in the following order:

|           |   |   |   |   |   |    |    |   |   |   |
|-----------|---|---|---|---|---|----|----|---|---|---|
| 1st Judge | 1 | 5 | 4 | 8 | 9 | 6  | 10 | 7 | 3 | 2 |
| 2nd Judge | 4 | 8 | 7 | 6 | 5 | 9  | 10 | 3 | 2 | 1 |
| 3rd Judge | 6 | 7 | 8 | 1 | 5 | 10 | 9  | 2 | 3 | 4 |

Use the rank correlation coefficient to discuss which pair of judges have nearest approach to common tastes in beauty.

**Solution:**

**Table : Computation for Rank Correlation Coefficient**

| I Judge<br>(R <sub>1</sub> ) | II Judge<br>(R <sub>2</sub> ) | III Judge<br>R <sub>1</sub> – R <sub>2</sub> | D <sub>1</sub> =<br>R <sub>1</sub> – R <sub>3</sub> | D <sub>1</sub> <sup>2</sup><br>R <sub>1</sub> – R <sub>3</sub> | D <sub>2</sub> =<br>R <sub>2</sub> – R <sub>3</sub> | D <sub>2</sub> <sup>2</sup><br>R <sub>2</sub> – R <sub>3</sub> | D <sub>3</sub> =<br>R <sub>3</sub> – R <sub>1</sub> | D <sub>3</sub> <sup>2</sup><br>R <sub>3</sub> – R <sub>1</sub> |
|------------------------------|-------------------------------|--|---|--|---|--|---|--|
| 1                            | 4                             | 6  | -3  | 9  | -5  | 25   | -2  | 4  |
| 5                            | 8                             | 7  | -3  | 9  | -2  | 4  | 1   | 1  |
| 4                            | 7                             | 8  | -3  | 9  | -4  | 16   | -1  | 1  |
| 8                            | 6                             | 1  | 2   | 4  | 7   | 49   | 5   | 25   |
| 9                            | 5                             | 5  | 4   | 16   | 4   | 16   | 0   | 0  |
| 6                            | 9                             | 10   | -3  | 9  | -4  | 16   | -1  | 1  |
| 10                           | 10                            | 9  | 0   | 0  | 1   | 1  | 1   | 1  |
| 7                            | 3                             | 2  | 4   | 16   | 5   | 25   | 1   | 1  |
| 3                            | 2                             | 3  | 1   | 1  | 0   | 0  | -1  | 1  |
| 2                            | 1                             | 4  | 1   | 1  | -2  | 4  | -3  | 9  |
|                              |                               |  |   | $\Sigma D_1^2$<br>= 74   |   |  | $\Sigma D_2^2$<br>= 156                             | $\Sigma D_3^2$<br>= 44   |

$$r_{12} \text{ (I & II Judge)} = 1 - \frac{6\Sigma D_1^2}{N^3 - N} = 1 - \frac{6 \times 74}{990} = 1 - 0.45 = 0.55.$$

$$r_{13} \text{ (I& II Judges)} = 1 - \frac{6\Sigma D_2^2}{N^3 - N} = 1 - \frac{6 \times 156}{990} = 1 - 0.945 = 0.055.$$

$$r_{23} \text{ (II & III Judges)} = 1 - \frac{6\Sigma d_3^2}{N^3 - N} = 1 - \frac{6 \times 44}{990} = 1 - 0.27 = 0.73.$$

Since rank correlation is the maximum between second and third judges, we can say that II and III Judges have nearest approach in common tastes of beauty.

### 17.15 KOLMOGOROV-SMIRNOV TEST

The statisticians A.N. Kolmogorov and N.V. Smirnov developed the Kolmogorov-Smirnov test. It is a simple non-parametric test for testing whether there is a significant difference between an observed frequency distribution and a theoretical frequency distribution. In short, this test is known as K-S test. The K-S test is, therefore, another measure of the goodness of fit of a frequency distribution, as was the chi-square test. Its basic advantages are:

1. *It is a more powerful test.*
2. *It is easier to use, since it does not require that the data be grouped in any way.*

## K-S Statistic

The K-S statistic is the maximum absolute deviation of expected relative frequency  $F_e$  and the observed relative frequency  $F_o$ . It is denoted by  $D_n$ .

$$\therefore \text{K-S statistic: } D_n = \max |F_e - F_o|$$

The K-S test is always a one tailed test for a given the level of significance  $\alpha$ .

The critical values for  $D_n$  can be tabulated by using the table. We compare the calculated value of  $D_n$  with the critical value of  $D_n$  from the table.

If the tabled value for the chosen significance level is greater than the calculated value of  $D_n$ , then we will accept the null hypothesis  $H_0$ .

**Example 17 :** Below is the table of observed frequencies, along with the frequency to the observed under a normal distribution.

(a) Calculate the K-S statistic.

(b) Can we conclude that this distribution does in fact follow a normal distribution? Use 0.10 level of significance.

| Test Score         | 51–60 | 61–70 | 71–80 | 81–90 | 91–100 |
|--------------------|-------|-------|-------|-------|--------|
| Observed Frequency | 30    | 100   | 440   | 500   | 130    |
| Expected Frequency | 40    | 170   | 500   | 390   | 100    |

**Solution :** Null Hypothesis :  $H_0$  : This distribution follows a normal distribution.

Table: Calculation of K-S Statistic

| Test Score | Observed Frequency | Observed Cumulative Frequency | Observed Relative Frequency ( $F_o$ ) | Expected Frequency | Expected Cumulative Frequency | Expected Relative Frequency ( $F_e$ ) | $D =  F_e - F_o $ |
|------------|--------------------|-------------------------------|---------------------------------------|--------------------|-------------------------------|---------------------------------------|-------------------|
| 51–60      | 30                 | 30                            | $\frac{30}{1200} = 0.025$             | 40                 | 40                            | $\frac{40}{1200} = 0.033$             | 0.008             |
| 61–70      | 100                | 130                           | $\frac{130}{1200} = 0.108$            | 170                | 210                           | $\frac{210}{1200} = 0.175$            | 0.067             |
| 71–80      | 440                | 570                           | $\frac{570}{1200} = 0.475$            | 500                | 710                           | $\frac{710}{1200} = 0.592$            | 0.117             |
| 81–90      | 500                | 1070                          | $\frac{1070}{1200} = 0.891$           | 390                | 1100                          | $\frac{1100}{1200} = 0.92$            | 0.029             |
| 91–100     | 130                | 1200                          | $\frac{1200}{1200} = 1$               | 100                | 1200                          | $\frac{1200}{1200} = 1$               | 0                 |

(a). K-S statistic:  $D_n = \max |F_e - F_o| = 0.117$

(b) The tabulated value of  $D_n$  for  $n = 5$  and  $\alpha = 0.01$  is 0.510.

[From table]

Since, the table value of  $D_n$  ( $= 0.510$ ) is greater than the calculated value of  $D_n$  ( $= 0.117$ ), so we accept the null hypothesis  $H_0 \Rightarrow$  the distribution follows a normal distribution.

### EXERCISE – 17.4

- The following table shows how 10 students, arranged in alphabetical order, were ranked according to their achievements in both the laboratory and lecture sections of a computer course. Find the coefficient of rank correlation.

|                     |   |   |    |   |   |    |   |   |   |   |
|---------------------|---|---|----|---|---|----|---|---|---|---|
| <i>Laboratory :</i> | 8 | 3 | 9  | 2 | 7 | 10 | 4 | 6 | 1 | 5 |
| <i>Lecture :</i>    | 9 | 5 | 10 | 1 | 8 | 7  | 3 | 4 | 2 | 6 |

|                                     |    |    |    |   |    |   |   |   |    |    |
|-------------------------------------|----|----|----|---|----|---|---|---|----|----|
| [Hint.] Difference of ranks ( $D$ ) | -1 | -2 | -1 | 1 | -1 | 3 | 1 | 2 | -1 | -1 |
| $D^2$                               | 1  | 4  | 1  | 1 | 1  | 9 | 1 | 4 | 1  | 1  |
| $\Sigma D^2 = 24$                   |    |    |    |   |    |   |   |   |    |    |

$$\therefore r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} = 1 - \frac{6(24)}{10(10^2 - 1)} = 0.8545.$$

- In a contest, two judges were asked to rank eight candidates (numbered 1 through 8) in order of preference. The judges submitted their choices in the following manner.

|                       |   |   |   |   |   |   |   |   |
|-----------------------|---|---|---|---|---|---|---|---|
| <i>First judge :</i>  | 5 | 2 | 8 | 1 | 4 | 6 | 3 | 7 |
| <i>Second judge :</i> | 4 | 5 | 7 | 3 | 2 | 8 | 1 | 6 |

- Find the coefficient of rank correlation.
  - Decide how well the judges agreed in their choices.
- A consumer panel tested nine makes of electric heaters for overall quality. The ranks assigned by the panel and the suggested retail prices were as follows:

| <i>Manufacturer</i> | <i>Panel Rating</i> | <i>Suggested Price in Rs.</i> |
|---------------------|---------------------|-------------------------------|
| <i>A</i>            | 6                   | 480                           |
| <i>B</i>            | 9                   | 395                           |
| <i>C</i>            | 2                   | 575                           |
| <i>D</i>            | 8                   | 550                           |
| <i>E</i>            | 5                   | 510                           |
| <i>F</i>            | 1                   | 545                           |
| <i>G</i>            | 7                   | 400                           |
| <i>H</i>            | 4                   | 465                           |
| <i>I</i>            | 3                   | 420                           |

Is there a significant relationship between the quality and the price of a electric heater?

4. The following data were obtained in a study of relationship between the weight and chest size of infants at birth.

|                          |      |      |      |      |      |      |      |      |      |
|--------------------------|------|------|------|------|------|------|------|------|------|
| <i>Weight (kg) :</i>     | 2.75 | 2.15 | 4.41 | 5.52 | 3.21 | 4.32 | 2.31 | 4.30 | 3.71 |
| <i>Chest size (cm) :</i> | 29.5 | 26.3 | 32.2 | 36.5 | 27.2 | 27.7 | 28.3 | 30.3 | 28.7 |

- (a) Calculate the rank correlation coefficient.  
 (b) Test the hypothesis at the 0.025 level of significance that the rank correlation coefficient is zero against the alternative that it is greater than zero.
5. A plant supervisor ranked a sample of eight workers on the number of hours worked overtime and length of employment. Find the rank correlation between two measures. Is it significant at 0.01 level ?

*Amount of overtime :*    5.0    8.0    2.0    4.0    3.0    7.0    1.0    6.0

*Years of employment :*    1.0    6.0    4.5    2.0    7.0    8.0    4.5    3.0

6. Two judges at a college homecoming parade ranked 8 floats in the following order:

|                | <i>Float</i> |   |   |   |   |   |   |   |
|----------------|--------------|---|---|---|---|---|---|---|
|                | 1            | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| <i>Judge A</i> | 5            | 8 | 4 | 3 | 6 | 2 | 7 | 1 |
| <i>Judge B</i> | 7            | 5 | 4 | 2 | 8 | 1 | 6 | 3 |

- (a) Calculate the rank correlation coefficient.  
 (b) Test the hypothesis that the rank correlation coefficient is zero against the alternative hypothesis that it is greater than zero. Use  $\alpha = 0.05$ .

## ANSWERS

1. 0.845.
2. (a) 0.67      (b) The judges did not agree too well in their choices.
3.  $r_s = -0.47$ ; No significant relationship.
4. (a)  $r_s = 0.72$   
     (b) Reject the null hypothesis that the rank correlation coefficient is zero.
5.  $r_s = 0.185$ . The rank correlation is not significant.
6. (a)  $r_s = 0.71$       (b) Reject the null hypothesis

## 17.16 KENDALL TEST OF CONCORDANCE

The test of significance of the rank correlation is applicable only if two sets of rankings of individuals (or items) are available. The Kendall's test of concordance is applicable to situations where we want to test the significance of more than two sets of rankings of individuals.

Let there be  $k$  sets of rankings of  $n$  individuals. The hypothesis to be tested can be stated as

$H_0$  : There is no agreement among the given sets of rankings.

$H_1$  : There is high level of agreement among the given sets of rankings.

### Test Statistic

Let us assume that 3 judges have ranked the 3 participants of a competition in the following three ways :

**Situation: I**

| Jndge →<br>Participant ↓ | I | II | III | Sum of<br>Ranks |
|--------------------------|---|----|-----|-----------------|
| P                        | 1 | 3  | 2   | 6               |
| Q                        | 2 | 1  | 3   | 6               |
| R                        | 3 | 2  | 1   | 6               |

**Situation : II**

| Jndge →<br>Participant ↓ | I | II | III | Sum of<br>Ranks |
|--------------------------|---|----|-----|-----------------|
| P                        | 2 | 2  | 2   | 6               |
| Q                        | 3 | 3  | 3   | 9               |
| R                        | 1 | 1  | 1   | 3               |

**Situation: III**

| Jndge →<br>participant ↓ | I | II | III | Sum of<br>Rank |
|--------------------------|---|----|-----|----------------|
| P                        | 1 | 1  | 1   | 7              |
| Q                        | 2 | 3  | 2   | 7              |
| R                        | 3 | 2  | 3   | 8              |

We note that there is a complete disagreement, perfect agreement and partial agreement among the three judges in the respective situations I, II, and III

Further, we note that the average of the sum of ranks in each of the three situations is 6.

Also  $\Sigma(R - \bar{R})^2$  in the respective situations are:

- (i)  $(6 - 6)^2 + (6 - 6)^2 + (6 - 6)^2 = 0$
- (ii)  $(6 - 6)^2 + (9 - 6)^2 + (3 - 6)^2 = 18$  and
- (iii)  $(3 - 6)^2 + (7 - 6)^2 + (8 - 6)^2 = 14.$

From the above three situations, we note that the *larger the value of  $\Sigma(R - \bar{R})^2$ , the higher is the level of agreement among the rankings of the judges.*

It can be shown that when there are  $k$  sets of rankings of  $n$  individuals such that  $n$  is

at least 8, the distribution of the statistic  $W = \frac{12 \left[ \sum (R - \bar{R})^2 \right]}{kn(n+1)}$  is approximately

$\chi^2$  with  $(n - 1)$  d.f.

To simplify the computational work, statistic  $W$  can also be written as

$$\text{Test Statistics : } W = \frac{12 \sum R^2 - 3n[k(n+1)]^2}{kn(n+1)}.$$

Thus, if  $W > \chi^2$  with  $(n - 1)$  d.f. and at 5% level of significance,  $H_0$  is rejected. The rejection of  $H_0$  indicates a significant level of agreement among the given sets of rankings

**Example 18 :** The ranks of 9 students of a class in four subjects, viz., English, Mathematics, Economics and Commerce are as follows. Perform a Kendall Concordance tests at 5% level of significance.

|             | Ranks |    |    |   |    |    |    |    |    |  |
|-------------|-------|----|----|---|----|----|----|----|----|--|
| English     | 4     | 7  | 9  | 3 | 8  | 2  | 6  | 5  | 1  |  |
| Mathematics | 3     | 6  | 8  | 1 | 7  | 5  | 9  | 2  | 4  |  |
| Economics   | 5     | 7  | 6  | 1 | 9  | 2  | 8  | 3  | 4  |  |
| Commerce    | 4     | 8  | 7  | 2 | 6  | 1  | 9  | 5  | 3  |  |
|             | 16    | 28 | 30 | 7 | 30 | 10 | 32 | 15 | 12 |  |

*Test of hypothesis that there is no agreement of ranks in the four subjects*

**Solution :**

**Null hypothesis –** $H_0$ : There is no agreement of ranks in the four subjects

**Alternative hypothesis –** $H_1$ : The level of agreement of ranks in the four subjects is significant

From the given data, the sum of the ranks in the three subjects of a students respectively 16, 28, 30, 7, 30, 10, 32, 15, 12.

Sum of the squares of ranks:

$$\Sigma R^2 = 16^2 + 28^2 + 30^2 + 7^2 + 30^2 + 10^2 + 32^2 + 15^2 + 12^2 = 4382.$$

$$\text{Test Statistic : } W = \frac{12 \sum R^2 - 3n[k(n+1)]^2}{k n(n+1)}$$

$$\therefore W = \frac{12 \times 4382 - 3 \times 9[4 \times 10]^2}{4 \times 9 \times 10} = \frac{52584 - 43200}{360} = 28.07.$$

Tabled value of  $\chi^2$  for 8 d.f. at 5% level = 15.507.

Since the calculated value of  $W$  (which is a  $\psi^2$  variate) is greater than the tabled value of  $\psi^2_{0.05; 8} \Rightarrow H_0$  is rejected. Thus the level of agreement of ranks in the four subjects is significant.

**Example 19 :** Ten competitors in a beauty contest are ranked by three judges in the following order

|             |   |   |   |   |   |    |    |   |   |   |
|-------------|---|---|---|---|---|----|----|---|---|---|
| Judge I :   | 1 | 5 | 4 | 8 | 9 | 6  | 10 | 7 | 3 | 2 |
| Judge I :   | 4 | 8 | 7 | 6 | 5 | 9  | 10 | 3 | 2 | 1 |
| Judge III : | 6 | 7 | 8 | 1 | 5 | 10 | 9  | 2 | 3 | 4 |

*Test the hypothesis that there is no agreement among the three judges.*

**Solution :**

**Null hypothesis –**  $H_0$  : There is no agreement among the three judges.

**Alternative hypothesis –**  $H_1$  : The level of agreement among the judges is significant.

**Sum of the Ranks of 10 competitors:**

$$\sum R = 11 + 20 + 19 + 15 + 19 + 25 + 29 + 12 + 8 + 7.$$

$$\therefore \sum R^2 = 11^2 + 20^2 + 19^2 + 15^2 + 19^2 + 25^2 + 29^2 + 12^2 + 8^2 + 7^2 = 3191$$

$$\therefore \text{Test Statistic: } W = \frac{12 \sum R^2 - 3n[k(n+1)]^2}{kn(n+1)}$$

$$\therefore W = \frac{12 \times 3191 - 3 \times 10(3 \times 11)^2}{3 \times 10 \times 11} = 17.04$$

Also  $\chi^2$  at 9 d.f. and 5% level of significance is 16.92. Since  $W$  is greater than this value, so  $H_0$  is rejected. This indicates that the level of agreement among the three judges is significant.

### 17.17 MEDIAN TEST FOR TWO INDEPENDENT SAMPLES

This test is used to test the null hypothesis  $H_0$  that two independent samples have been drawn from identical distribution against the alternative hypothesis that their location parameters (medians) are different. The test can be one or two tailed tests. This test is sensitive to differences in location.

#### Test Statistic

Under the assumption that hypothesis  $H_0$  is true, we can expect roughly 50% of the observations of each sample to be above the median and 50% to be below the median of the combined sample. The sample observations can thus be dichotomised. It is presented in the form of a  $2 \times 2$  contingency table,

**Table:  $2 \times 2$  Contingency Table**

|              | Sample I | Sample II | Total           |
|--------------|----------|-----------|-----------------|
| Above Median | $a$      | $b$       | $a + b$         |
| Below Median | $c$      | $d$       | $c + d$         |
| Total        | $a + c$  | $b + d$   | $n = n_1 + n_2$ |

When  $n > 20$  and no cell frequency is less than 5, then

$$\chi^2 = \frac{n \left( |ad - bc| - \frac{n}{2} \right)^2}{(a+b)(c+d)(a+c)(b+d)}, \text{ with correction for continuity, is a } \chi^2 \text{ variable with 1 d.f.}$$

**Example 20 :** An I.Q test was given to a random sample of 15 male and 20 female students of a University. Their scores were recorded as follow :

Male : 56, 66, 62, 81, 75, 73, 83, 68, 48, 70, 60, 77, 86, 44, 72,

*Female: 63, 77, 65, 71, 74, 60, 76, 61, 67, 72, 64, 65, 55, 89, 45, 53, 68, 73, 50, 81*

*Use median test to determine whether 1.Q. of male and female students is same in the university. (Given the median of combined sample = 68)*

**Solution :** It is given that the median of the combined sample is 68. On the *discarding two observations with value equal to median, we have n = 33*. The dichotomised observations of the two samples are presented in the following  $2 \times 2$  contingency table :

**Table :  $2 \times 2$  contingency Table**

|                     | <i>Sample I</i> | <i>Sample II</i> | <i>Total</i> |
|---------------------|-----------------|------------------|--------------|
| <i>Above Median</i> | 8 (a)           | 8 (b)            | 16           |
| <i>Below Median</i> | 6 (c)           | 11 (d)           | 17           |
| <i>Total</i>        | 14              | 19               | 33           |

$$\text{Thus, we have } \chi^2 = \frac{[|ad - bc| - (n/2)]^2}{(a+b)(c+d)(a+d)(b+d)} = \frac{33(|88 - 48| - 16.5)^2}{16 \times 17 \times 14 \times 19} = 0.252$$

**Tabled value of  $\chi^2$  for 1 d.f. and for  $\alpha = 0.05$  is =3.84.**

Since the tabulated value of  $\chi^2 = 0.252$  is less than the tabled value of  $\chi_{1,0.05}^2 = 3.84$ , so the null hypothesis is accepted.

## 17.18 WILCOXON SIGNED – RANK TEST

**Wilcoxon Signed – Rank test** is useful in comparing two populations (or medians of two populations) for which we have paired observations. Unlike the Sign Test, the Wilcoxon Test accounts for the magnitude of differences between paired values and not only their signs. The test does so by considering the ranks of these differences. The test is, therefore, more efficient than the Sign test when the differences may be quantified rather than just given a positive or negative sign. The Sign Test on the other hand, is easier to carry on.

The **Wilcoxon** procedure may also be adopted for testing whether the location parameter of a single population (its median or its mean) is equal to any given value. There are one-tailed and two tailed version of each test. We shall discuss the paired observation test for equality of two populations distributions (or the equality of location parameters of the two populations.)

**Null hypothesis –  $H_0$ :** The median difference between the population 1 and 2 is zero.

**Alternative hypothesis –  $H_1$ :** The median difference between the population 1 and 2 is not zero

We assume

(i) that the distribution of differences between two populations is symmetric

(ii) that the differences are mutually independent.

(iii) that the measurement scale is at least interval.

By the assumption of symmetry, the hypothesis may be stated in terms of means. The alternative hypothesis may also be directed one : that the mean (or median) of one population is greater than the mean (or median) of the other population.

### Steps for Calculation of Test Statistic

- (i) List the pairs of observations we are given for the two populations ( or on the two variables)
- (ii) For each pair, calculate the difference:  $D_c = X_i - Y_i$
- (iii) Omit all observation (s) with equal values and reduce the sample size accordingly.
- (iv) Rank these differences in ascending order without regard to their signs.
- (v) The cases of tied ranks are assigned ranks by the average method.
- (vi) Find  $\Sigma(+)$  and  $\Sigma(-)$ , where  $\Sigma(+)$  is the sum of ranks with positive  $D_i$  and  $\Sigma(-)$  is the sum of ranks with negative.
- (vii) The Wilcoxon T-statistic is defined as the smaller of the two sums of ranks

$$T = \min (\Sigma(+), \Sigma(-))$$

where  $\Sigma(+)$  = sum of the ranks of positive differences;

$\Sigma(-)$  = sum of the ranks of negative differences

- (viii) **The Decision Rule :** Critical points of the distribution of the test statistic  $T$  (when the null hypothesis  $H_0$  is true) are given by the table “Critical values of Wilcoxon  $T$  – Test” given at the end of the book. We carry out the test on the left tail, i.e., we reject the null hypothesis  $H_0$  if the computed value of the test statistic  $T$  is less than the critical point from the table (tabulated value) for a given level of significance.

For One-tailed test, suppose that the alternative hypothesis  $H_1$  is that mean (or median) of population 1 is greater than that of population 2, i.e.,

$$\text{Null hypothesis} \quad H_0 : \mu_1 = \mu_2$$

$$\text{Alternative hypothesis} \quad H_1 : \mu_1 > \mu_2$$

Here we shall use the sum of the ranks of negative differences  $\Sigma(-)$

If the alternative hypothesis  $H_1$  is reversed (population 1 and 2 are switched), then we shall use the sum of the ranks of positive differences  $\Sigma(+)$  as the statistic  $T$ .

In either case, the test is carried out on the left ‘tail’ of the distribution. Table “ Critical values of Wilcoxon  $T$  – test” gives critical points for both one-tailed and two-tailed tests.

### Wilcoxon Signed Rank Test — Large Samples Test

In the Wilcoxon test ‘ $n$ ’ is defined as the number of pairs of observations from population 1 and 2. As the number of pairs  $n$  gets large (as a rule of thumb,  $n > 25$  or so),  $T$  may be approximated by a normal random variable.

$$\text{Mean of } T : E(T) = \frac{n(n+1)}{4}$$

$$\text{Standard Deviation of } T : \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = S.E.(T)$$

$$\text{Statistic : } Z = \frac{T - E(T)}{\sigma_T}$$

**Tabled Value of  $Z_\alpha$ :** Find the tabled value of  $Z_\alpha$  for a given  $\alpha$ -level of significance from table given at the end of the book.

**Decision:** Accept  $H_0$  if calculated value of  $|Z| <$  Table value of  $Z_\alpha$  otherwise accept the alternative hypothesis  $H_1$

**Example 21 :** Two models of a machine are under consideration for purchase. An organisation has one of each type for trial and each operator, out of the team of 25 operators, uses each machine for a fixed length of time. Their outputs are:

|                          |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
|--------------------------|---|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Operator No.             | : | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 |
| Output from Machine I :  |   | 82 | 68 | 53 | 75 | 78 | 86 | 64 | 54 | 62 | 70 | 51 | 80 | 64 |
| Output from Machine II : |   | 80 | 71 | 46 | 58 | 60 | 72 | 38 | 60 | 65 | 64 | 38 | 79 | 37 |
| Operator No.             | : | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |    |
| Output from Machine I :  |   | 65 | 70 | 55 | 75 | 64 | 72 | 55 | 70 | 45 | 64 | 58 | 65 |    |
| Output from Machine II : |   | 60 | 73 | 48 | 58 | 60 | 76 | 60 | 50 | 30 | 70 | 55 | 60 |    |

Is there any significant difference between the output capacities of the two machines? Test at 5% level of significance.

**Solution :** Let  $D = M_1 - M_2$ , where  $M_1$  and  $M_2$ , denote the outputs of the machines I and II respectively.

Table: Computation of Ranks

| No. | $M_1$ | $M_2$ | $D_i$ | Ranks | No. | $M_1$ | $M_2$ | $D_i$ | Ranks |
|-----|-------|-------|-------|-------|-----|-------|-------|-------|-------|
| 1   | 82    | 80    | 2     | 2     | 14  | 65    | 60    | 5     | 10    |
| 2   | 68    | 71    | -3    | 4.5   | 15  | 70    | 73    | -3    | 4.5   |
| 3   | 53    | 46    | 7     | 15.5  | 16  | 55    | 48    | 7     | 15.5  |
| 4   | 75    | 58    | 17    | 20.5  | 17  | 75    | 58    | 17    | 20.5  |
| 5   | 78    | 60    | 18    | 22    | 18  | 64    | 60    | 4     | 7.5   |
| 6   | 86    | 72    | 14    | 18    | 19  | 72    | 76    | -4    | 7.5   |
| 7   | 64    | 38    | 26    | 24    | 20  | 55    | 60    | -5    | 10    |
| 8   | 54    | 60    | -6    | 13    | 21  | 70    | 50    | 20    | 23    |
| 9   | 62    | 65    | -3    | 4.5   | 22  | 45    | 30    | 15    | 19    |
| 10  | 70    | 64    | 6     | 13    | 23  | 64    | 70    | -6    | 13    |
| 11  | 51    | 38    | 13    | 17    | 24  | 58    | 55    | 3     | 4.5   |
| 12  | 80    | 79    | 1     | 1     | 25  | 65    | 60    | 5     | 10    |
| 13  | 64    | 37    | 27    | 25    |     |       |       |       |       |

$$\Sigma(+) = 2 + 15.5 + 20.5 + 22 + 18 + 24 + 13 + 17 + 1 + 25 + 10 + 15.5 + 20.5 + 7.5 \\ + 23 + 19 + 4.5 + 10 = 268$$

$$\Sigma(-) = 4.5 + 13 + 4.5 + 4.5 + 7.5 + 10 + 13 = 57$$

Null hypothesis  $H_0$  : Output capacities of the two machines are same.

Alternative hypothesis  $H_1$  : Output capacities are not same.

$$T - \text{Statistic} = \min(\Sigma(+), \Sigma(-)) = 57$$

$$\therefore \mu_T = \frac{n(n+1)}{4} = \frac{25 \times 26}{4} = 162.5.$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{25 \times 26 \times 51}{24}} = 37.2.$$

$$\therefore \text{Test Statistics: } Z = \frac{T - E(T)}{\sigma_T} = \frac{57 - 162.5}{37.2} = -2.84 \Rightarrow |Z| = 2.84$$

**Critical Value:** Tabled value of  $Z_\alpha$  for  $\alpha = 5\% = 1.96$ .

Since  $|Z| > Z_\alpha$  ( $\alpha = 5\%$  level), so  $H_0$  is rejected at 5% level of significance.

Hence the output capacities are not same at 5% level of significance.

**Example 22 :** A random sample of 30 students obtained the following marks in a class test. Test the hypothesis that their median score is more than 50.

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 55 | 58 | 25 | 32 | 26 | 85 | 44 | 80 | 33 | 72 | 10 | 42 | 15 | 46 | 64 |
| 39 | 38 | 30 | 36 | 65 | 72 | 46 | 54 | 36 | 89 | 94 | 25 | 74 | 66 | 29 |

**Solution :**

Null hypothesis  $H_0$  :  $M_d = 50$

Alternative hypothesis  $H_1$  :  $M_d > 50$

Let  $D = x - 50$ .

The  $D$  values and their rankings are shown in the following table :

|        |      |     |      |     |      |      |      |      |     |      |
|--------|------|-----|------|-----|------|------|------|------|-----|------|
| $D$ :  | 5    | 8   | -25  | -18 | -24  | 35   | -6   | 30   | -17 | 22   |
| Rank : | 4    | 6.5 | 23.5 | 16  | 21.5 | 26.5 | 5    | 25   | 15  | 19.5 |
| $D$ :  | -40  | -8  | -35  | -4  | 14   | -11  | -12  | -20  | -14 | 15   |
| Rank : | 29   | 6.5 | 26.5 | 2   | 11   | 8    | 9    | 17   | 11  | 13   |
| $D$ :  | 22   | -4  | 4    | -14 | 39   | 44   | -25  | 24   | 16  | -21  |
| Rank : | 19.5 | 2   | 2    | 11  | 28   | 30   | 23.5 | 21.5 | 14  | 18   |

Here  $n = 30$ .

$\Sigma(+) = \text{Sum of ranks with positive } D = 220.5$

$\Sigma(-) = \text{Sum of ranks with negative } D = 244.5$

Let us take T-Statistics as  $T = \Sigma(-) = 244.5$ .

[ $\because$  It is a one tailed test so  $T$  is taken as the sum of ranks with negative  $D$ , i.e.,  $T = (-)$ ]

$$\therefore \text{Mean} : \mu_T = \frac{n(n+1)}{4} = \frac{30 \times 31}{4} = 232.5.$$

$$\text{Standard Error} : \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{30 \times 31 \times 61}{24}} = \sqrt{2363.75} = 48.62.$$

$$\text{Test Statistics: } Z = \frac{T - \mu_T}{\text{S.E}(T)} = \frac{244 - 232.5}{48.62} = \frac{12}{48.62} = 0.247.$$

Tabled value of  $Z_\alpha$  at  $\alpha = 5\%$  level of significance = 1.645.

**Decision.** Since  $0.247 < 1.645 \Rightarrow Z < Z_\alpha \Rightarrow$  the null hypothesis  $H_0$  is accepted at 5% level of significance  $\Rightarrow$  The median score is 50

### 17.19 THE MATCHED- PAIRS SIGN TEST

When the same individual is simultaneously observed with regard to two characteristics, we get a matched pair. For example, a survey of obtaining the opinions of persons, regarding two brands of soaps, say X and Y. Given a sample of matched pairs, we can test which of the two soaps is preferable.

#### Test Statistic

Let there be  $n$  matched pairs  $(X_i, Y_i)$ ,  $i = 1, 2, \dots, n$ , in the sample, where  $X_i$  and  $Y_i$  are the ranks of the respective items X and Y by the  $i$ th individual. We associate a plus sign to a pair if  $X_i > Y_i$ ; a minus sign if  $X_i < Y_i$  and discard it  $X_i = Y_i$ . Let  $p$  be the proportion of plus sign, i.e.,

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

Let  $p$  be the proportion of plus signs in a population, i.e., the proportion of individuals having preference for X. We note that if more individuals have preference for X, then  $\pi > \frac{1}{2}$ .

Similarly, the indifference of the individuals is indicated by  $\pi = \frac{1}{2}$ . In general, we can test  $H_0 : p = p_0$ , where  $p_0$  denotes proportion of individuals having preference for X.

It can be shown that  $p$  is a random variable which approximately follows a normal distribution

(when  $> 25$ ) with mean  $p$  and standard error  $\sqrt{\frac{\pi(1-\pi)}{n}}$ .

**Example 23 :** A random sample of 40 persons was selected to determine their preference regarding the two brands of a new tooth paste, A and B. Each person used the two brands for one month and then was asked to rank them by using arbitrary numbers. Their rankings, A and B, were recorded as follows:

| Person | 1 | 2  | 3 | 4 | 5  | 6 | 7 | 8 | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|--------|---|----|---|---|----|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|
| A      | 4 | 15 | 2 | 4 | 15 | 3 | 8 | 6 | 10 | 8  | 1  | 5  | 12 | 6  | 2  | 7  | 5  | 1  | 3  | 8  |

|               |   |
|---------------|---|
| <i>B</i>      | : 2 15 1 5 16 2 6 5 15 8 2 7 10 4 2 3 10 2 4 9                |
| <i>Person</i> | : 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 |
| <i>A</i>      | : 12 20 3 1 2 1 4 6 28 100 15 4 5 30 9 8 5 10 12 25           |
| <i>B</i>      | : 15 40 2 2 1 4 4 7 25 200 20 3. 6 40 15 15 3 2 18 30         |

Test the hypothesis that more than half the population prefer brand *A* to *B*.

**Solution :** Null hypothesis  $H_0 : \pi \leq \frac{1}{2}$

Alternative hypothesis :  $H_0 : \pi > \frac{1}{2}$

We assign a plus sign to a pair if  $A > B$ , a minus sign if  $A < B$  and discard the pair if  $A = B$ . Proceeding in this manner, we find that the number of positive signs in the given sample is 14. Also the total number of pairs, after discarding the cases where  $A = B$ , is  $n = 36$ . Thus, we have

$$p = \frac{14}{36} \text{ and the Standard Error of } p = \sqrt{\frac{pq}{n}}.$$

$$\text{Standard Error of } p: \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \quad \text{or} \quad \sigma_p = \sqrt{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{36}} = 0.083.$$

$$\text{Test Statistic: } |Z| = \frac{|p - \pi|}{\sigma_p} = \frac{|(14/36) - (1/2)|}{0.083} = 1.339.$$

Tabled value of  $Z_\alpha$  at 5% level of significance = 1.645.

Now  $1.339 < 1.645$ ,  $\Rightarrow$  Null hypothesis  $H_0$  is accepted at 5% level of significance. Hence the sample supports the view that more than half of population prefer *A* to *B*.

### MISCELLANEOUS SOLVED EXAMPLES

**Example 24 :** The following data represent the number of hours that a rechargeable hedge trimmer operator before a recharge is required :

1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 12 and 1.7

Use the sign test to test the hypothesis at the 0.05 level of significance that this particular trimmer operates with a median of 1.8 hours before requiring a recharge.

**Solution :**

(i) Null hypothesis.  $H_0 : M_d = 1.8$ .

Alternative hypothesis:  $H_1 : M_d \neq 1.8$  (Two tailed test)

(ii) Level of Significance.  $\alpha = 0.05$

(iii) Computation Statistic. Replacing each value by plus symbol "+" if it exceeds 1.8 and

by minus symbol “-” if it is less than 1.8 and discarding the measurement that equals 1.8, we get:

- + - - + - - + - -

Here the number of plus symbol “+” is the statistic  $x$

$\therefore x = 3$ . Also,  $n = 10$ .

(iv) **Critical region.**  $x \leq k'_{0.025}$  and  $x \geq k_{0.025}$ , where  $x$  is the number of plus signs.

From the Binomial table  $k'_{0.025} = 1$  and  $k_{0.025} = 9$ .

Since  $n = 3$  falls in the acceptance region, so we accept the null hypothesis  $H_0$  which implies that average operating time is not significantly different from 1.8 hours.

**Example 25 :** The weights (gms) of 31 apples picked from a consignment are as follows:

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

Can this be regarded as a random sample?

**Solution :**

We have to test  $H_0: r = m_r$  against  $H_a: r \neq m_r$ .

Let us denote the increase in the successive observation by a plus (+) sign and the decrease of successive observation by a minus (-) sign. From the given observations, we can write a sequence of plus and minus signs, as given below:

+ - + + + - + - + + - - - + - + - + - + + + -

From the above sequence, we have  $n_1 = 16$  (the number of plus signs),  $n_2 = 14$  (the number minus signs) and  $r = 20$  (the number of runs). Also  $n = 16 + 14 = 30$ .

$$\therefore \mu_r = \frac{2n_1 n_2}{n} + 1 = \frac{2 \times 16 \times 14}{30} + 1 = 15.93$$

and  $\sigma_r = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}} = \sqrt{\frac{2 \times 16 \times 14 (2 \times 16 \times 14 - 30)}{900 \times 29}} = 2.68$ .

Further,  $Z = \frac{|r - \mu_r|}{\sigma_r} = \frac{|20 - 15.93|}{2.68} = 1.52$ .

Since this value is less than 1.96, there is no evidence against  $H_0$  at 5% level of significance. Thus, the given sample may be treated as random.

**Example 26 :** Use the Kruskal-Wallis test to test for differences in mean among the three samples. If  $\alpha = 0.01$ , what are your conclusions?

Sample I : 95 97 99 98 99 99 99 94 95 98

Sample II : 104 102 102 105 99 102 111 103 100 103

Sample III : 119 130 132 136 141 172 145 150 144 135

**Solution :**

**Null hypothesis:**  $H_0 : \mu_1 = \mu_2 = \mu_3$

**Alternative Hypothesis:**  $H_1 : \mu_1, \mu_2$  and  $\mu_3$  are not all equal.

**Level of significance:**  $\alpha = 0.05$ .

**Degrees of freedom:**  $d.o.f. = (3 - 1) = 2$ .

**Calculation of Test Statistic:** Ranking the given data from 1 to 30, of all the three samples and assigning tied ranks, if needed, we get :

|                                  |           |           |           |           |           |           |           |           |           |           |
|----------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| <b>Item :</b>                    | 94        | 95        | 95        | 97        | 98        | 98        | 99        | 99        | 99        | 99        |
| <b>Rank (<math>R_1</math>) :</b> | 1         | 2.5       | 2.5       | 4         | 5.5       | 5.5       | 9         | 9         | 9         | 9         |
| <b>Item :</b>                    | 99        | 100       | 102       | 102       | 102       | 103       | 103       | 104       | 105       | 111       |
| <b>Rank (<math>R_2</math>) :</b> | 11        | 12        | 14        | 14        | 14        | 16.5      | 16.5      | 18        | 19        | 20        |
| <b>Item :</b>                    | 119       | 130       | 132       | 135       | 136       | 141       | 144       | 145       | 150       | 172       |
| <b>Rank (<math>R_3</math>) :</b> | <b>21</b> | <b>22</b> | <b>23</b> | <b>24</b> | <b>25</b> | <b>26</b> | <b>27</b> | <b>28</b> | <b>29</b> | <b>30</b> |

$$\text{Also } R_1 = 1 + 2.5 + 2.5 + 4 + 5.5 + 5.5 + 9 + 9 + 9 + 9 = 57.$$

$$R_2 = 11 + 12 + 14 + 14 + 14 + 16.5 + 16.5 + 18 + 19 + 20 = 155.$$

$$R_3 = 21 + 22 + 23 + 24 + 25 + 26 + 27 + 28 + 29 + 30 = 255.$$

**Note:** The ranks of  $R_1$ , Sample I are circled (O), the ranks of Sample II are crossed (X) and the ranks of Sample III are in bold face.

$$\text{Also } n_1 = n_2 = n_3 = 10. \text{ and } n = n_1 + n_2 + n_3 = 30.$$

$$\text{Test statistic : } H = \frac{12}{n(n+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1)$$

$$\text{or } H = \frac{12}{30 \times 31} \left[ \frac{(57)^2}{10} + \frac{(155)^2}{10} + \frac{(255)^2}{10} \right] - 3(30+1)$$

$$= \frac{12}{930} \times \left[ \frac{3249}{10} + \frac{24025}{10} + \frac{65025}{10} \right] - 93 = \frac{12}{930} \times \frac{92299}{10} - 93$$

$$= 119.10 - 93 = 26.10.$$

**Tabled value of H for  $\alpha = 0.05$  and 2 d.o.f. is  $H_{0.05,2} = 5.991$ .**

Since the calculated value of  $H (=26.10)$  is greater than the tabled value  $H_{0.05,2} (=5.991)$  so the null hypothesis is rejected  $\Rightarrow$  the difference in the means among the samples is significant.

**Example 27 :** Twelve entries in a painting competition were ranked by two judges, as shown below:

|                  |   |   |   |   |   |   |    |   |    |    |    |    |
|------------------|---|---|---|---|---|---|----|---|----|----|----|----|
| <b>Entry :</b>   | A | B | C | D | E | F | G  | H | I  | J  | K  | L  |
| <i>Judge I :</i> | 5 | 2 | 3 | 4 | 1 | 6 | 8  | 7 | 10 | 9  | 12 | 11 |
| <i>Judge II:</i> | 4 | 5 | 2 | 1 | 6 | 7 | 10 | 9 | 11 | 12 | 3  | 8. |

*Test the hypothesis that coefficient of rank correlation in population is positive at 5% level of significance.*

**Solution :**

We have to test  $H_0 : r_s = 0$  against  $H_a : r_s > 0$

From the given data, we can find  $d_i = R_{1i} - R_{2i}$  and the  $\sum d_i^2 = 154$ .

$$\therefore r_s = 1 - \frac{6 \times 154}{12 \times 143} = 0.46 \text{ and } Z = 0.46 \sqrt{11} = 1.53.$$

Since the value of  $Z_\alpha$  (at  $\alpha = 5\%$ ) is less than 1.645, there is no evidence against  $H_0$  at 5% level of significance. Hence, the correlation in population cannot be regarded as positive.

**Example 28 :** *The following is an arrangement of 25 men, M, and 15 women, W lined up to purchase tickets for a premier picture show:*

M / WW / MMM / W / MM / W / M / W / M / WWW / MMM / W /  
MM / WWW / MMMMMM / WWW / MMMMMM /

*Test for randomness at the 5 per cent level of significance.*

**Solution :** Here  $n_1 = 25$ ,  $n_2 = 15$ ,  $r = 17$ .

$$\mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \times 25 \times 15}{25 + 15} + 1 = 19.75.$$

$$\begin{aligned}\sigma_r &= \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} \\ &= \sqrt{\frac{2 \times 25 \times 15(2 \times 25 \times 15 - 25 - 15)}{(25 + 15)^2 (25 + 15 - 1)}} = 2.92.\end{aligned}$$

$$Z = \frac{r - \mu_r}{\sigma_r} = \frac{17 - 19.75}{2.92} = -0.94.$$

Since this value of  $Z$  is less than  $Z_\alpha = 1.96$  (5% level) the null hypothesis is accepted. Hence there is no real evidence to suggest that the arrangement is not random.

**Example 29 :** *A company's trainees are randomly assigned to groups which are taught a certain industrial inspection procedure by three different methods: At the end of the instructing period they are tested for inspection performance quality. The following are their scores:*

*Method A :* 80, 83, 79, 85, 90, 68

*Method B :* 82, 84, 60, 72, 86, 67, 91

*Method C :* 93, 65, 77, 78, 88.

*Use H test to determine at the 0.05 level of significance whether the three methods are equally effective*

**Solution :** Arranging the data jointly according to size and assigning ranks, we get :

Values: 60 65 67 68 72 77 78 79 80 82 83 84 85 86 88 90 91 93

Rank : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Here,  $R_1 = 61$ ,  $R_2 = 62$ ,  $R_3 = 48$ . and  $n = 18$ .

$$\begin{aligned} H &= \frac{12}{N(N+1)} \left[ \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(n+1) \\ &= \frac{12}{18 \times 19} \left[ \frac{(61)^2}{6} + \frac{(62)^2}{7} + \frac{(48)^2}{5} \right] - 3(19) = 0.197 \end{aligned}$$

But,  $\chi^2_{0.5}$  for 2 d.f. = 5.991.

The calculated value of  $H$  is less than the table value  $\chi^2_{0.05, 2}$ , so the null hypothesis is accepted and we conclude that the three months are equally effective.

### MISCELLANEOUS EXERCISE

1. What do you understand by non-parametric test? Describe any four non-parametric tests along with the situations of their use and advantages and disadvantages over the parametric tests.
2. Distinguish between parametric and non-parametric methods for testing a statistical hypothesis.
3. What are the advantages and disadvantages of non-parametric methods as compared to parametric methods in statistics?
4. What are non-parametric tests? In what ways are they different from parametric tests?
5. What is a Wilcoxon signed rank test? Explain with the help of an example.
6. Explain briefly the various types of non-parametric tests known to you and the specific situation in which they are applicable.
7. Explain the working of median test for two independent samples with the help of an example.
8. Differentiate between parametric and non-parametric statistical test.
9. Explain the working of Kendall test for concordance with the help of an example. Under what situations it can be applied?
10. Discuss the usefulness of non-parametric tests in statistical analysis.
11. The following figures are a sample of 35 observations. Find median of the data and mark each measurement as A, if greater than it, or B, if less than it. Use the runs test to find whether the sample is random, at 5% level of significance.  
37, 46, 33, 39, 59, 41, 44, 49, 51, 35, 41, 55, 27, 19, 35, 41, 49, 21, 35, 37, 53, 29, 49, 48, 35, 47, 31, 41, 29, 27, 49, 63, 37, 13, 20.
12. A production manager wishes to conduct an experiment to compare two methods of assembling a certain mechanism. He first pairs the workers according to age and assigns

the method to the workers at random. A sample of 12 pairs of workers gives the following data about the mean number of units assembled per hour :

|                   |    |    |    |    |    |    |    |    |    |    |    |    |
|-------------------|----|----|----|----|----|----|----|----|----|----|----|----|
| <i>Pairs No :</i> | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 |
| <i>Method A :</i> | 20 | 18 | 28 | 32 | 30 | 35 | 40 | 22 | 36 | 39 | 42 | 27 |
| <i>Method B :</i> | 22 | 14 | 30 | 35 | 31 | 36 | 40 | 18 | 35 | 8  | 40 | 20 |

Use sign test to determine at 5% level of significance whether method *A* is superior method *B*.

13. Random samples of three models of scooter were tested for the petrol mileage (the number of kilometres per litre). Use Kruskal Walli test to determine if the average mileage of the three models is same.

|                  |    |    |    |    |    |    |    |    |
|------------------|----|----|----|----|----|----|----|----|
| <i>Model A :</i> | 60 | 54 | 76 | 48 | 66 | 52 | 62 | 56 |
| <i>Model B :</i> | 62 | 58 | 52 | 48 | 70 |    |    |    |
| <i>Model C :</i> | 42 | 64 | 36 | 65 | 42 | 60 | 82 |    |

14. An examination designed to measure the basic I.Q. was given Random samples were taken of 20 boys and 20 girls joining as management trainees in a company. The scores obtained by them out of 50 are given below :

|                |    |    |    |    |    |    |    |    |    |     |
|----------------|----|----|----|----|----|----|----|----|----|-----|
| <i>Boys :</i>  | 28 | 30 | 32 | 34 | 46 | 45 | 39 | 15 | 33 | 23  |
|                | 27 | 29 | 36 | 48 | 47 | 45 | 26 | 28 | 47 | 42  |
| <i>Girls :</i> | 35 | 37 | 31 | 33 | 41 | 44 | 38 | 19 | 42 | 46  |
|                | 13 | 1  | 12 | 10 | 40 | 41 | 43 | 42 | 26 | 13. |

By applying U-test determine whether there is a significant difference in the average I.Q. of boys and girls. (Use 5 per cent level of significance.)

15. The following are the number of units produced by a group of workers for 25 days. Test whether the data can be regarded as a random sample ?

210, 180, 170, 240, 150, 215, 198, 181, 237, 209, 165, 176, 224, 201, 181, 252, 219, 154, 197, 235, 182, 167, 214, 221, 243

16. Calculate the coefficient of rank correlation from the following data and test its significance at 5% level of significance.

|            |    |    |    |    |    |    |    |    |    |    |
|------------|----|----|----|----|----|----|----|----|----|----|
| <i>X :</i> | 32 | 35 | 49 | 60 | 43 | 37 | 43 | 49 | 10 | 20 |
| <i>Y :</i> | 40 | 30 | 70 | 20 | 30 | 50 | 72 | 60 | 45 | 25 |

17. Two professors rated thirteen students in terms of ability. Use the data given here to estimate if the correlation between the ratings is significant :

|                       |   |   |    |   |   |    |    |   |    |   |   |    |    |
|-----------------------|---|---|----|---|---|----|----|---|----|---|---|----|----|
| <i>Student</i>        | A | B | C  | D | E | F  | G  | H | I  | J | K | L  | M  |
| <i>Professor I :</i>  | 1 | 7 | 8  | 3 | 6 | 10 | 9  | 2 | 11 | 4 | 5 | 13 | 12 |
| <i>Professor II :</i> | 4 | 8 | 10 | 1 | 5 | 9  | 11 | 3 | 7  | 2 | 1 | 12 | 13 |

18. A cooperative store is interested in knowing whether there is any significant difference between the buying habits of male and female shoppers. Samples of 14 males and 16 females shoppers gave the following information :

*Male* : 62 38 43 79 77 23 11 52 33 41 70 49 69 43

*Female*: 93 101 72 118 100 45 68 72 47 83 92 106 63 66 85 81

Use median test to verify whether there is any reason to suppose that the two populations are different.

19. On 11 successive trips between two cities, a video-coach carried 25, 28, 28, 26, 26, 27, 27, 27, 30, 30, 30, 29, 29, 20, 20, 20, 20 and 25 passengers . Use the total number of runs above and below the median, and the 0.05 level of significance to test – whether the given data constitutes a random sample?
20. Three different methods of advertising a commodity were used and the respective samples of sizes 9, 10 and 10 identical outlets were taken. The increased sales (in Rs. 1000) were recorded as follows :

*Sample I* : 92 79 77 93 99 93 71 87 98

*Sample II* : 95 76 84 85 89 90 72 82 68 83

*Sample III* : 81 91 75 80 78 94 100 86 88 69

Use Kendal concordance test to test the hypothesis that mean increase in sales due to the three methods of advertising is same at 5% level of significance

21. To compare the effectiveness of three types of weight-reducing diets, a homogeneous groups of 22 women was divided into three sub-groups and each sub-group followed one of these diet plans for a period of two months. The weight reductions, in kgs, were noted as given below:

I 4.3 3.2 2.7 6.2 5.0 3.9

*Diet Plans* II 5.3 7.4 8.3 5.5 6.7 7.2 8.5

III 1.4 2.1 2.7 3.1 1.5 0.7 4.3 3.5 0.3

Use Kruskal - Walli's test to test the hypothesis that the effectiveness of the three weight reducing diet plans are same at 5% level of significance.

[Hint. It is given that  $n_1 = 6$ ,  $n_2 = 7$  and  $n_3 = 9$

The total number observations is  $6 + 7 + 9 = 22$ . These are ranked in their ascending order as given below :

I ( $R_1$ ) 12.5 9 6.5 17 14 11 70

*Diet Plans* II ( $R_2$ ) 15 20 21 16 18 19 22 131

III ( $R_3$ ) 3 5 6.5 8 4 2 12.5 10 1 52

From the above table, we get  $R_1 = 70$ ,  $R_2 = 131$  and  $R_3 = 52$ .

$$\therefore H = \frac{12}{22 \times 23} \left( \frac{70^2}{6} + \frac{131^2}{7} + \frac{52^2}{9} \right) - 3 \times 23 = 15.63.$$

The tabulated value of  $\chi^2$  or  $H$  at 2 d.f. and 5% level of significance is 5.99. Since  $H$  is greater than this value,  $H_0$  is rejected at 5% level of significance]. These four pages will replace article 17.2 on page 2.



# 18

# Factorial Analysis

## 18.1 THE FACTORIAL PRINCIPLE

In the experimental lay-outs, like the completely randomised and randomised block designs, the main object is to compare and estimate the effect of a single set of 'treatments' such as different varieties of wheat. The chief purpose of introducing blocks is to make allowance for unwanted but unavoidable heterogeneity. Although, we can in some cases use a randomised-block type of analysis for an experiment in which the blocks merely represent the levels of a second factor like many factors can be incorporated in a single design, and whether this is or is not a desirable practice.

It is customary in the classical type of scientific experimentation to advocate the investigation of any problem by holding most of the variable factors constant and allowing only one or two to vary in each experiment. If one is largely occupied with fundamental research, where the point is to formulate general laws and test crucial predictions, this procedure of isolating one or two factors at a time has much to recommend it.

If on the other hand, one is dealing with work is of a more general nature, such as the prosecution of a plant-breeding programme, then one is essentially concerned to know what happens with a range of combinations of factors. A series of experiments in which only one factor is varied at a time would be both lengthy and costly, and might still be unsatisfactory because of systematic changes in the general background conditions. As alternative approach is to try to investigate variations in several factors simultaneously. This leads us to the idea of a *factorial experiment* in which the set of experiment units, e.g., cultivated plots, animals, Petri dishes, etc., is made large enough to include all possible combinations of levels of the different factors. Thus, if we had three varieties of wheat and three different levels of a fertiliser, there would be nine combinations in all. This would be called a  $3 \times 3$  factorial or  $3^2$  factorial.

## 18.2 BASIC IDEAS AND NOTATION IN THE $2^n$ FACTORIAL

We shall first consider designs in which there are several factors each at two levels. When there are  $n$  factors, we call this a  $2^n$  factorial. 'Levels' may be quite literally two quantitative

levels or concentrations of, say, a fertiliser, or it may mean merely two qualitative alternatives like the different species of a plant. In some cases one level is simply the absence of the factor and the other its presence.

Suppose we indicate by the capital letters  $A, B, C, \dots$  the names of the factors involved; and by the small letters  $a, b, c, \dots$  one of the two levels of each of the corresponding factors. In a specific trial the test the effect of fertilisers, we might use the ordinary notation of  $K, N$  and  $P$  to indicate potash, nitrogen and phosphate; and the letters  $k, n$  and  $p$  to represent the presence of some specified concentrations. Thus the 'treatment'  $knp$  means that all types were being applied:  $k$  means potash only, the other two being absent; and the absence of all three is indicated by '1'. (Instead of 'absence', we might prefer to consider presence at a second, perhaps lower, concentration.)

With three factors  $A, B$  and  $C$ , there are evidently eight different kinds of treatment, namely

'1',  $a, b, ab, c, ac, bc, abc$

and we could construct a design using several different blocks, each of which contained exactly eight plots, one for each treatment. The yields of the plots can be subjected to any analysis of variance technique, with a final summary in an analysis of variance table and a table showing the average effects of the factors, both separately and in different combinations. We first require a more extended notation in order to present the analysis in a concise form and to explain the meaning of what is being done.

Suppose there are  $r$  blocks, or replicates as they are often called in this context. We write  $T_1, T_a, T_b, T_{ab}$  etc., for the total yields of the  $r$  plots having treatments '1',  $a, b, ab$ , etc., respectively. The corresponding mean values, obtain by dividing there by  $r$ , are  $\bar{1}, \bar{a}, \bar{b}, \bar{ab}$  etc. (Some writers use a notation in which the totals are written as  $[1], [a], [b], [ab]$ , etc., and the means as '1',  $a, b, ab$  etc., without the bars.)

We shall now define *main effects* and *interactions*. In order to do this in the simplest possible way, let us first consider an experiment with only two factors,  $A$  and  $B$ . The effect of factor  $A$  can be represented by the difference between mean yields obtained at each level. Thus the observed effect of  $A$  at the first level of  $B$  is  $\bar{a} - \bar{1}$ ; and the observed effect of  $A$  at the second level  $B$  is  $\bar{ab} - \bar{b}$ . The average observed effect of  $A$  over the two levels of  $B$  is called the *main effect* and this is therefore, defined by

$$A = \frac{1}{2} (\bar{a} - \bar{1} + \bar{ab} - \bar{b})$$

where we are using the symbol  $A$  to represent the main effect of the factor  $A$ . A similar argument gives the main effects of  $B$  as

$$B = \frac{1}{2} \{(\bar{ab} - \bar{b}) - (\bar{a} - \bar{1})\}$$

Now, if the two factors act independently of one another, we should expect the true effect of one to be the same at either level of the other. We should, for example, expect that the two observed quantities  $\bar{a} - \bar{1}$  and  $\bar{ab} - \bar{b}$  were really estimates of the same thing. The difference

of these numbers is therefore a measure of the extent to which the factors interact, and we therefore write the interaction as

$$AB = \frac{1}{2} \left\{ (\bar{ab} - \bar{b}) - (\bar{a} - \bar{1}) \right\}$$

A useful approach, as we shall see, is to assume that an interaction is non-existent unless the observed value departs significantly from zero.

These ideas are easily extended to several factors. With three,  $A$ ,  $B$  and  $C$  we have of course three main effects, which we call  $A$ ,  $B$  and  $C$ ; three 'first-order' interactions,  $AB$ ,  $AC$  and  $BC$ ; and one 'second-order' interaction,  $ABC$ . The 'second order' interaction is a slightly more subtle concept to grasp than the 'first-order' quantity, but it can be thought of in various ways, such as the difference in the interaction  $AB$  calculated at each of the two levels of  $C$ . Similarly, for higher-order interactions in experiments with larger number of factors.

### 18.3 THE ANALYSIS OF VARIANCE FOR A $2^n$ FACTORIAL

We are now in a position to discuss the analysis of variance which must contain contributions to the total sum of squares from each main effect and interaction, as well as from the blocks or replicates. When all these items are subtracted from the total sum of squares calculated as usual, we are left with the residual sum of squares which is needed to obtain the residual variance  $s^2$ .

The actual contribution to the sum of squares for any main effect or interaction can be obtained in terms of certain complicated sums and differences of the observed yields. Fortunately, a simple computational rule enables us to avoid discussing these explicitly, and this is illustrated in the following example. Let us consider the  $2^3$  design shown in Table 18.1. The purpose of the experiment is to determine the effect of different kinds of fertilisers,  $K$ ,  $N$  and  $P$ , on potato-crop yield, and it is laid out in four replicates. Table 18.2 summarises the various block and treatment totals.

**Table 18.1 : A  $2^3$  Factorial Design, Laid out in Four Blocks, to Determine the Effect of Different Kinds of Fertiliser on Potato Crop Yield**

| Block 1 |       |      |       | Block 2 |       |      |       |
|---------|-------|------|-------|---------|-------|------|-------|
| $kn$    | '1'   | $kp$ | $k$   | $kp$    | $knp$ | $p$  | '1'   |
| 284     | 98    | 372  | 257   | 385     | 429   | 317  | 105   |
| Block 3 |       |      |       | $k$     | $n$   | $np$ | $kn$  |
| $n$     | $knp$ | $p$  | $np$  | 283     | 95    | 328  | 311   |
| 111     | 446   | 302  | 361   |         |       |      |       |
| Block 4 |       |      |       | $np$    | $kp$  | $k$  | $knp$ |
| $p$     | $kn$  | 'l'  | $n$   | 351     | 422   | 298  | 452   |
| 305     | 320   | 85   | 133   |         |       |      |       |
|         |       |      |       | $n$     | 'l'   | $kn$ | $p$   |
| $np$    | $k$   | $kp$ | $knp$ | 97      | 126   | 277  | 313   |
| 335     | 267   | 399  | 464   |         |       |      |       |

The first step in the analysis is to remove the block effects. To do this we simply carry out a randomised block analysis for the eight treatment combinations and the four blocks. The initial calculations are therefore:

Total number of observations = 32.

$$\text{Grand total, } G = 9128. \text{ Correction factor} = \frac{1}{32}G^2 = 2603762.0.$$

**Table 18.2 : Summary of block and treatment totals, and grand total**

|   |  |
|---|--|
| Block totals<br>$\left\{ \begin{array}{l} 1 : 2231 \\ 2 : 2253 \\ 3 : 2308 \\ 4 : 2336 \end{array} \right.$ | Treatment totals<br>$\left\{ \begin{array}{l} '1' : 414 \\ k : 1105 \\ n : 436 \\ kn : 1192 \\ p : 1237 \\ kp : 1578 \\ np : 1375 \\ knp : 1791 \end{array} \right.$ |
| <i>Grand Total G : = 9128</i>   |  |

$$\begin{aligned} \text{The sum of squares about the mean} &= 284^2 + 98^2 + \dots + 277^2 + 313^2 - 2603762.0 \\ &= 433618.0 \end{aligned}$$

$$\begin{aligned} \text{Sum of squares for blocks} &= \frac{1}{8} (2231^2 + 2253^2 + 2308^2 + 2336^2) - 2603762.0 \\ &= 879.3 \end{aligned}$$

$$\begin{aligned} \text{Sum of squares for treatments} &= \frac{1}{4} (414^2 + 1105^2 + \dots + 1791^2) - 2603762.0 \\ &= 426723.0 \end{aligned}$$

The first part of the analysis of variance is therefore as set out in Table 18.3. The residual variance is  $s^2 = 286.5$  based on 21 degrees of freedom. Large differences between the eight treatment combinations are clearly evident, as we should expect from inspection of the treatment totals shown in Table 18.2. It so happens that the variation between blocks is about equal to the residuals, so that in this example there is no special contribution from fluctuations in soil fertility, and the division into blocks has not led to any gain in accuracy.

**Table 18.3 : Analysis of Variance for Table 18.1**

| Source of variation | Sum of squares | Degrees of freedom | Mean square | Variance ratio |
|---------------------|----------------|--------------------|-------------|----------------|
| Blocks              | 879.25         | 3                  | 293.1       | 1.02           |
| Treatments          | 426723.00      | 7                  | 60960.4     | 212.8          |
| Residual            | 6015.75        | 21                 | 286.5       | —              |
| Total               | 433618.00      | 31                 | —           | —              |

The next step is to pick out the individual items in the treatment sum of squares corresponding to the various main effects and interactions. This is where we use the special computational rule, referred to above, that enables us to avoid specific algebraic formulae for the sums of squares required. Table 18.4 shows the calculations we need. It is an essential part of the procedure that the treatment combinations be written down in a standard order: each factor is introduced in turn, and is then followed by all combinations of itself with the treatment combinations previously written down. Against each treatment combination we write the corresponding total yield.

**Table 18.4 : Scheme of Calculations for Finding the Treatment Main effects and Interactions**

| Treatment combination | Total yield | (1)   | (2)   | (3)   | Effect        |
|-----------------------|-------------|-------|-------|-------|---------------|
| 'l'                   | 414         | +1519 | +3147 | +9128 | (Grand total) |
| <i>k</i>              | 1105        | +1628 | +5981 | +2204 | <i>K</i>      |
| <i>n</i>              | 436         | +2815 | +1447 | +460  | <i>N</i>      |
| <i>kn</i>             | 1192        | +3166 | +757  | +140  | <i>KN</i>     |
| <i>p</i>              | 1237        | +691  | +109  | +2834 | <i>P</i>      |
| <i>kp</i>             | 1578        | +756  | +351  | -690  | <i>KP</i>     |
| <i>np</i>             | 1375        | +341  | +65   | +242  | <i>NP</i>     |
| <i>knp</i>            | 1791        | +416  | +75   | +10   | <i>KNP</i>    |

We next form the entries in column (1) as follows: The first four entries are the sums of the numbers in the total yield column taken in *successive pairs*. The second four entries are the differences of these pairs, the upper figure always being *subtracted from the lower*. To obtain column (2) the whole process is repeated on column (1) and (3) is derived from (2) in a similar fashion. This 'sum-and-difference' procedure is performed as many times as there are factors, three times in fact in the present example.

The first term in column (3) is the grand total of all the observations, which should check with a straight addition. Other entries in column (3) are the totals for main effects or interactions corresponding to the treatment combinations in the initial column of the table. If we want the actual values of these factorial effects, then the totals must be *divided by*  $r2^{n-1}$ , where  $r$  is the number of blocks and  $n$  the number of factors. In the present example this divisor is  $4 \times 2^2 = 16$ . Thus the main effect of *K*, i.e., the average of all plots with *k* minus the average of all plots without *k*, is  $+2204/16 = + 138$ . The variance of each of these factorial effects is estimated by  $s^2/r2^{n-2}$ . Confidence limits can then be attached in the usual way.

We can test the significance of the factorial effects directly from the totals in column (3) of Table 18.4. Each total has a variance of  $32s^2$  or a standard error of  $\sqrt{32s^2}$ , which in this case is  $\sqrt{32 \times 286.5} = 95.7$ . In general, to obtain the standard error we merely multiply  $s$  by the square root of the total number of observations. Significance tests are now based on the *t*-distribution with 21 degrees of freedom. The 5 per cent points for the totals are therefore,  $\pm 2.080 \times 95.7 = \pm 199$ , and 1% points are  $\pm 2.831 \times 95.7 = \pm 271$ . We then see that the main

effects,  $K$ ,  $N$  and  $P$ , are all highly significant, then see that beyond the 1 per cent point. The interaction  $KP$  is also significant at this level, while  $NP$  is between the 1 and 5 per cent points. The other two interactions are not significant.

One way of interpreting the *positive* interaction  $KN$  is to say that potash and nitrogen do not act independently of one another; when both are present their individual effects are enhanced. Similarly, for the interaction  $NP$ . On the other hand, the  $KP$  interaction is *negative* in this experiment. So when potash and phosphate operate jointly the full benefit of each is not achieved.

A check on the treatment sum of squares in Table 18.3 can be made by direct calculation from Table 18.4. The contribution from each factorial effect is given by the square of the corresponding total divided by the number of observations. The treatment sums of squares should, therefore, be

$$\frac{1}{32} (2204^2 + 460^2 + \dots + 242^2 + 10^2) = 426723,$$

checking with Table 18.3.

These are the basic calculations required for the analysis of simple factorial experiments of the  $2^n$  type. A certain amount of further information can be extracted from the material by, for instance, presenting the mean yields of certain combinations of factors averaged over the rest. The actual use to which one puts the results of a factorial experiment, and the mode of interpretation adopted, depend of course on the precise context. A good deal of experience is required to grasp the full implications of a large factorial experiment, especially when some of the further complications mentioned in the next section are involved.

#### 18.4 THE SCOPE OF MORE ADVANCED DESIGNS

Although in this chapter we have only been able to look at the general principles of factorial experimentation and the way these work out in the  $2^n$  type of design, it is worth while giving some indication of the scope of more advanced applications. This will enable the reader to consider for himself the various possibilities that might be tried in relation to any particular experiment programme. However, the actual execution of a more elaborate design will usually require some assistance from a statistical expert.

In the first place we have looked in detail at designs involving factors having only two levels, for which the analysis is relatively simple. It is quite possible to handle designs for several factors having any number of levels, and we saw how the randomised block design could be used when there were just two factors, each having several levels.

The important departures from a simple  $2^n$  factorial described in this section stem from the requirement that block sizes should be kept reasonably small if the residual variation is not to be unduly large. With a  $2^3$  design there are eight different treatment combinations, and so we naturally tend to use several blocks each of eight plots. Now, the fertility of such blocks is often reasonably homogeneous. But as the number of factors increases, and with it the number of plots in a block, we may be faced with serious heterogeneity. One important device for retaining a small-block size is called *confounding*. We might have a  $2^5$  design with 5 factors. Instead of using blocks of 32 plots, it is possible to arrange that a selection only of the full 32 treatment combinations appears in each block. We could, for example, spread the 32 treatment combinations

over 4 blocks each of 8 plots. If this is done in the right way we can still obtain the information we require on main effects and first-order interactions, at the cost of sacrificing information on certain higher-order interactions.

Another expedient is to use *fractional replication*. In this we use only a fraction of the full set of treatment combinations in a single block. Thus 5 drugs each at two levels could be tested

in a factorial experiment on only 16 mice, using a suitably chosen  $\frac{1}{2}$  - replicate. Again certain information must be sacrificed, but properly handled, such experiments can be very economical and time-saving, and are specially valuable in exploratory investigations.

Subtleties like confounding and fractional replication, which are often combined in the same design, are most readily applied in  $2^n$  factorials, but they can also be adopted with advantage in other designs with all factors having the same number of levels like  $3^n$ ,  $4^n$  and  $5^n$  difficulties are, however, liable to arise if the experiment is 'mixed' in the sense of having factors with different levels, and it is, in general, inadvisable to embark on such intricacies.

A good example of the reduction in the size of an experiment is afforded by an investigation into the consistency of certain foods. There were 7 factors involved relating to the shape and size of the cone in an Adams consistometer, the temperature and general level of consistency in the material, etc. Assuming three basic levels for each factor we should have a  $3^7$  factorial with

$3^7 = 2187$  basic treatment combinations. This is unmanageably large number, and so a  $\frac{1}{9}$  - replicate was devised having only 243 treatment combinations. A block in this context would be a day's work, and 243 experimental determinations was certainly too large for the size of a single block. However, it was possible to employ confounding as well, and to distribute the 243 small experiments over 9 days, with 27 determinations for each day. This reduced the scale of the initial factorial arrangement to manageable proportions, but was still sufficient to provide adequate information's on main effects and first-order interactions.

This remark must suffice to indicate the kind of advantages that can be obtained from advanced experimental designs. Their application usually requires expert statistical assistance, and if this is not available, it is best for the experimenter to use something that is simpler and more easily understood, even if it is theoretically less efficient.



# 19

# *Circular Distributions and Descriptive Statistics*

## 19.1 DATA ON A CIRCULAR SCALE

We know that an **interval scale of measurement** was as a scale with equal intervals but with no true zero point. A special type of **interval scale is a circular scale**, where not only is there no true zero, but any designation of high or low values is arbitrary. A common example of a circular scale of measurement would be compass direction [Fig. 19.1(a)], where a circle is said to be divided into 360 equal intervals, called degrees, and  $90^\circ$  cannot be said to be a 'larger' direction than  $60^\circ$ .<sup>1</sup>

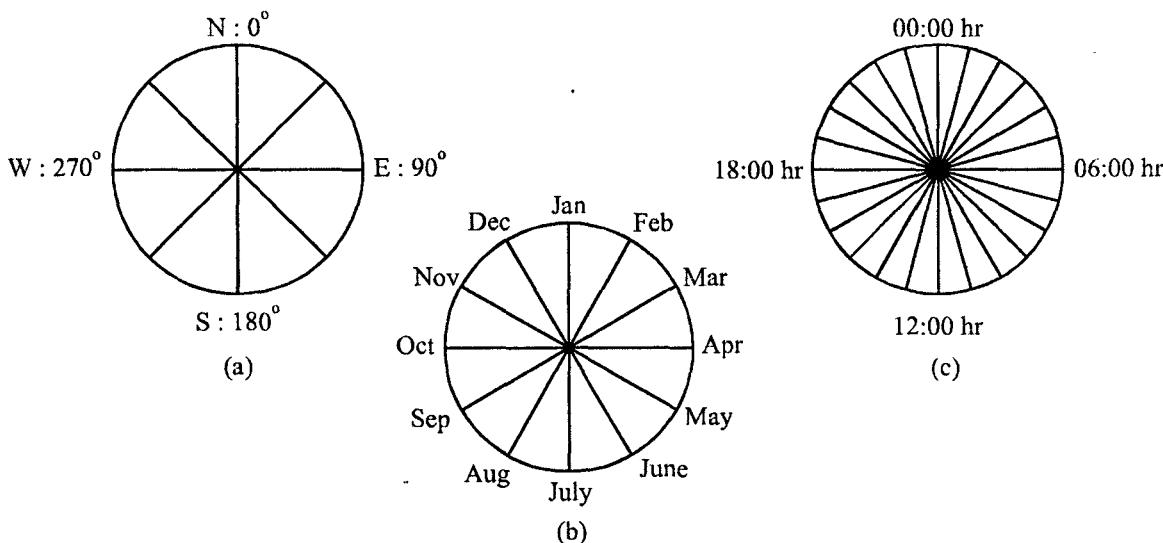
Another common circular scale is time of day [Fig. 19.1(b)], where a day is divided into 24 equal intervals, called hours, but where the designation of midnight as the zero or starting point is arbitrary. One hour of a day corresponds to  $15$  (i.e.,  $360/24$ ) of a circle, and 1 of a circle corresponds to 4 minutes of a day. Other time divisions, such as weeks and years [see Fig. 19.1(c)], also represent circular scales of measurement.

In general, we may convert  $X$  time units to an angular direction ( $a$ , in degrees), where  $X$  has been measured on a circular scale having  $k$  time units in the full cycle:

$$a = \frac{(360^\circ)(X)}{k} \quad \dots (19.1)$$

For example, to convert a time of day ( $X$ , in hours) to an angular direction,  $k = 24$  hr; to convert a day of the week to an angular direction, number the seven days from some arbitrary point (e.g., Sunday = day 1) and use Equation 19.1 with  $k = 7$ ; to convert the  $X$ th. day of the year to an angular direction,  $k = 365$  (or,  $k = 366$  in a leap year); to convert a month of the year,

<sup>1</sup>Occasionally one will encounter angular measurements expressed in radians rather than in degrees. A radian is the angle that is subtended by an arc of a circle equal in length to the radius of the circle. As a circle's circumferences is  $2\pi$  times the radius, a radian is  $360^\circ/2\pi = 180^\circ/\pi = 57.29577951^\circ$  (or 57 deg, 17 min, 44.8062 sec).



**Fig 19.1 : Common circular scales of measurements. (a) Compass directions. (b) Times a day. (c) Days of year (with the first day of each month shown).**

$k = 12$ ; and so on.<sup>2</sup> Such conversions are demonstrated in Example 1. Also note that data from circular distributions generally may not be analyzed using the statistical methods presented earlier in this book.<sup>3</sup> This is so for theoretical reasons as well as for empirically obvious reasons stemming from the arbitrariness of the zero point on the circular scale. For example, consider three compass directions :  $10^\circ$ ,  $30^\circ$  and  $350^\circ$ , for which we wish to calculate an arithmetic mean. The arithmetic mean calculation of  $(10^\circ + 30^\circ + 350^\circ)/3 = 130^\circ$  is clearly absurd, for all data are northerly directions and the computed mean is southeasterly.

Statistical methods for describing and analyzing data from circular distributions are relatively new and are still undergoing development. This chapter will introduce some basic considerations useful in calculating descriptive statistics and Chapter 20 will discuss tests of hypothesis.

<sup>2</sup>Equation 19.1 gives angular directions corresponding to the ends of time periods (e.g., the end of  $X$ th day of the year). If some other point in a time period is preferred, the equation can be adjusted accordingly. (For example, noon can be considered on the  $X$ th of the year by using  $X - 0.5$  in place of  $X$ .) If the same point is used in each time period (e.g., always using either noon or midnight,) then the statistical procedures of this and the following chapter will be unaffected by the choice of point. (However, graphical procedures, as in Section 24.2, will of course be affected, in the form of a rotation of the graph if Equation 19.1 is adjusted. If for example, we considered noon on the  $X$ th day of the year, the entire graph would be rotated about a half a degree counterclockwise.

<sup>3</sup>An exception is the case where the measurement scale is only a portion of a circle. For example, latitude on the earth's surface, even though measured in degrees, is constrained to a range of 0 to  $90^\circ$  on either side of the equator. Such data may be treated as ratio data measured on a linear scale.

**Example 1 :** Conversions of times measured on a circular scale to corresponding angular directions.

**Solution :** We use equation 19.1:

$$a = \frac{(360^\circ)(X)}{k}$$

- Given a time of day of 06:00 hr (which is one-fourth of the 24-hour clock and should correspond, therefore, to one-fourth of a circle):

$$X = 6 \text{ hr}, \quad k = 24 \text{ hr, and}$$

$$a = (360^\circ)(6 \text{ hr})/24 \text{ hr} = 90^\circ.$$

- Given a time of day of 06:25 hr:

$$X = 6.25 \text{ hr}, \quad k = 24 \text{ hr, and}$$

$$a = (360^\circ)(6.25 \text{ hr})/24 \text{ hr} = 93.75^\circ.$$

- Gives the 14th day of February, being the 45th day of the year:

$$X = 45 \text{ days, } k = 365 \text{ days, and}$$

$$a = (360^\circ)(45 \text{ days})/365 \text{ days} = 44.38^\circ.$$

## 19.2 GRAPHICAL PRESENTATION OF DATA

Circular data are often presented as a scatter diagram. Figure 19.2 shows such a graph for the data of Example 19.2. If frequencies of data are:

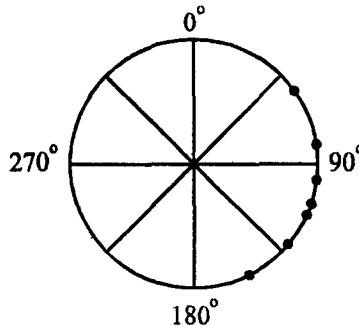


Fig. 19.2 : A circular scatter diagram for the data of Example 19.2

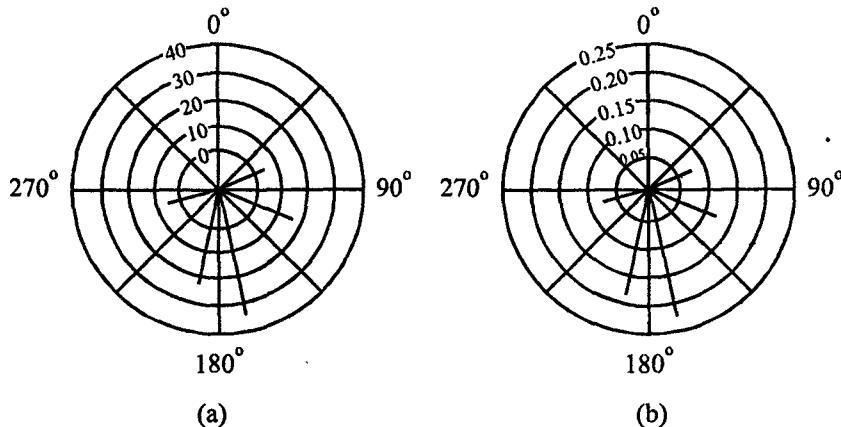
**Example 2 :** A sample of circular data. These data are plotted in Fig. 19.2.

**Solution :** Seven trees are found leaning in the following compass directions:

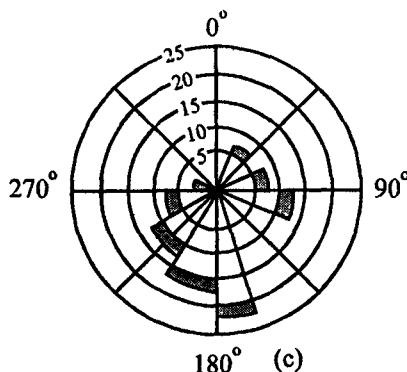
$$55^\circ, 81^\circ, 96^\circ, 109^\circ, 117^\circ, 132^\circ, 154^\circ.$$

are too large to be plotted conveniently on a scatter diagram, then a histogram may be drawn. This is demonstrated in Fig. 19.3, for the data presented in Example 19.3. Recall that in a histogram, the length, as well as the area, of each bar is an indication of the frequency observed at each plotted value of the variable (Section 1.3). Occasionally, as shown in Fig. 19.4, a histogram is seen presented with sectors, rather than bars, comprising the graph; this is sometimes called a *rose diagram*. Here, the radii forming the outer boundaries of the sectors are proportional to

the frequencies being represented, but the areas of the sectors are not. Since it is likely that the areas will be judged by the eye to represent the frequencies, the reader of the graph is being deceived, and this type of graphical presentation is not recommended (An equal-area rose diagram can be obtained by plotting the square roots of frequencies as radii).



**Fig. 19.3 :** (a) Circular histogram for the data of Example 19.3 where the concentric circles represent frequency increments of 5. (b) A relative frequency histogram for the data of Example 3 with the concentric circles representing relative frequency increments of 0.05.



**Fig. 19.4 :** A graphical presentation of the data of Example 3 utilizing sectors instead of bars.

A histogram of circular data can also be plotted as a linear histogram with degrees of the horizontal axis and frequencies (or relative frequencies) on the vertical. But the circular presentation is preferable.

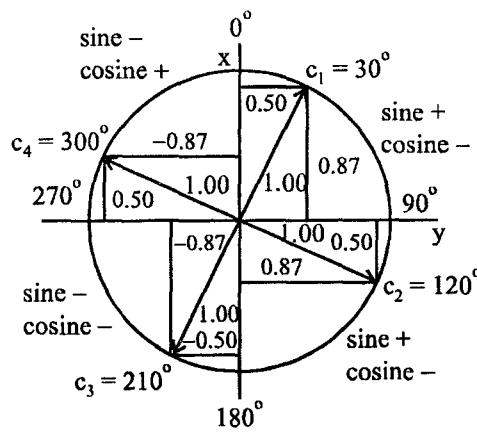
**Example 3 :** A sample of circular data, presented as a frequency table, where  $a_i$  is an angle and  $f_i$  is the observed frequency of  $a_i$ . These data are plotted in Fig. 19.3.

**Solution :**

| $a_i$ (deg)                 | $f_i$ | Relative $f_i$      |
|-----------------------------|-------|---------------------|
| 0–30                        | 0     | 0.00                |
| 30–60                       | 6     | 0.06                |
| 60–90                       | 9     | 0.09                |
| 90–120                      | 13    | 0.12                |
| 120–150                     | 15    | 0.14                |
| 150–180                     | 22    | 0.21                |
| 180–210                     | 17    | 0.16                |
| 210–240                     | 12    | 0.11                |
| 240–270                     | 8     | 0.08                |
| 270–300                     | 3     | 0.03                |
| 300–330                     | 0     | 0.00                |
| 330–360                     | 0     | 0.00                |
| <b><math>n = 105</math></b> |       | <b>Total = 1.00</b> |

### 19.3 SINES AND COSINES OF CIRCULAR DATA

A great many of the procedures that follow in this chapter and the next require the determination of simple trigonometric functions. Let us consider that a circle (perhaps representing a compass face) is drawn on rectangular coordinates (as on common graph paper) with the center as the origin (i.e., zero) of a vertical  $X$ -axis and a horizontal  $Y$ -axis;<sup>4</sup> this is what is done in Fig. 19.5.



*Fig. 19.5 : A unit circle, showing four points and their polar ( $a$  and  $r$ ) and rectangular ( $X$  and  $Y$ ) coordinates.*

<sup>4</sup>This is the opposite of the convention of having the  $X$ -axis horizontal and the  $Y$ -axis vertical. This is done for ease in obtaining trigonometric computations, for standard mathematical notation has angular measurement proceed counterclockwise from zero degrees on the right-hand portion of the horizontal axis (what we call “east” on the compass), rather than clockwise from the upper portion of the vertical axis (what we call “north” on the compass).

There are two methods that can be used to locate any point on a plane (such as a sheet of paper). One is to specify both the angle  $a$ , with respect to some starting direction (say, clockwise from the top of the  $X$ -axis, namely ‘north’) and the straight-line distance,  $r$ , from some reference point (the center of the circle). This pair of numbers,  $a$  and  $r$ , is known as the “polar coordinates” of a point. Thus, for example, in Fig. 19.5, point 1 is uniquely identified by polar coordinates  $a = 30^\circ$  and  $r = 1.00$ , point 2 by  $a = 120^\circ$  and  $r = 1.00$  and so on.<sup>5</sup>

The second method of locating points on the graph is by referring to the  $X$ - and  $Y$ -axes, as we introduced when dealing with regression problems by this method, point 1 in Fig. 19.5 is located by the “rectangular coordinates”<sup>6</sup>  $X = 0.87$  and  $Y = 0.50$ , point 2 by  $X = -0.50$  and  $Y = 0.87$ , point 3 by  $X = -0.87$  and  $Y = -0.50$  and point 4 by  $X = 0.50$  and  $Y = -0.87$ .

The *cosine* (abbreviated “cos”) of an angle is defined as the ratio of the  $X$  and the  $r$  associated with the circular measurement:

$$\cos a = \frac{X}{r}, \quad \dots (19.2)$$

while the *sine* (abbreviated “sin”) of the angle is the ratio of the associated  $Y$  and  $r$ :

$$\sin a = \frac{Y}{r}. \quad \dots (19.3)$$

Thus, for example, the sine of  $a_1$  in Fig. 19.5 is  $\sin 30^\circ = 0.50/1.00 = 0.50$ , and its cosine is  $\cos 30^\circ = 0.87/1.00 = 0.87$ . Also  $\sin 120^\circ = 0.87/1.00 = 0.87$ ,  $\cos 120^\circ = -0.50/1.00 = 0.50$ , and so on. Sines and cosines (two of the most useful “trigonometric”<sup>7</sup> functions) are readily available in published tables and many electronic calculators give them (and sometimes convert between polar and rectangular co-ordinates as well). The sine of  $0^\circ$  and  $180^\circ$  is zero, angles between  $0^\circ$  and  $180^\circ$  have sines that are positive, and the sines are negative for  $180^\circ < a < 360^\circ$ . The cosine is zero for  $90^\circ$  and  $270^\circ$ , with positive cosines obtained for  $0^\circ < a < 90^\circ$  and for  $270^\circ < a < 360^\circ$ , and negative cosines for angles between  $90^\circ$  and  $270^\circ$ .

At most, only sines and cosines are required for the statistical procedures that follow, although brief mention will be made of a third trigonometric functions, the tangent.<sup>8</sup>

$$\tan a = \frac{Y}{X} = \frac{\sin a}{\cos a} \quad \dots (19.5)$$

<sup>5</sup>If we specify that the radius of the circle is 1 unit, our figure is called a “unit circle”.

<sup>6</sup>The familiar system of rectangular coordinates is also known as “Cartesian coordinates,” after the French mathematician and philosopher, Rene Descartes (1596 - 1650), who wrote under the Latinized version of his name, Renatus Cartesius. His other noteworthy introductions were the use of exponents, the square root sign ( $\sqrt{}$ ), and the  $X$  and  $Y$  to denote variables (Asimov, 1972 : 106 - 107).

<sup>7</sup>Trigonometry refers, literally, to the measurement of triangles (such as the triangles that emanate from the center of the circle in Fig. 19.5).

<sup>8</sup>The cotangent is the inverse of the tangent, namely

$$\cot a = \frac{\cos a}{\sin a} \quad \dots (19.4)$$

We shall see later rectangular coordinates,  $X$  and  $Y$  may also be used in conjunction with mean angles just as they are with individual angular measurements.

#### 19.4 THE MEAN ANGLE

If we have a sample consisting of  $n$  angles, denoted as  $a_1$  through  $a_n$ , then the mean of these angles,  $\bar{a}$ , is to be an estimate of the mean angle in the sampled population  $\mu_a$ . To compute the sample mean angle,  $\bar{a}$ , we first consider the rectangular coordinates of the mean angle:

$$X = \frac{\sum_{i=1}^n \cos a_i}{n} \quad \dots (19.6)$$

$$\text{and } Y = \frac{\sum_{i=1}^n \sin a_i}{n} \quad \dots (19.7)$$

$$\text{Then, the quantity } r = \sqrt{X^2 + Y^2} \quad \dots (19.8)$$

is computed;<sup>9</sup> this is the length of the mean vector, which will be further discussed in section 19.5. The value of  $\bar{a}$  is determined as the angle having the following cosine and sine;

$$\cos \bar{a} = \frac{X}{r} \quad \dots (19.9)$$

$$\text{and } \sin \bar{a} = \frac{Y}{r} \quad \dots (19.10)$$

**Example 4 :** Demonstrates these calculations. It is also true that

$$\tan \bar{a} = \frac{Y}{X} \quad \dots (19.11)$$

so we have a check on the calculation of the mean angle  $\bar{a}$ . If  $r = 0$ , the mean angle is undefined and we conclude that there is no mean direction.

If we are dealing with data that are times instead of angles, then the mean time corresponding to the mean angle may be determined by inverting equation 19.1:

$$\bar{X} = \frac{ka}{360^\circ} \quad \dots (19.12)$$

For example, to determine a mean time a day,  $\bar{X}$ , from a mean angle,  $\bar{a}$  :  $\bar{X} = (24 \text{ hr}) (\bar{a})/360^\circ$ .

**Example 5 :** Calculating the mean angle for the data of Example 2.

<sup>9</sup>This quantity,  $r$ , is not to be confused with a sample correlation coefficient (Section 19.1), with which it bears no relationship but which is denoted by the same symbol.

| $a_i$ (deg)                 | $\sin a_i$ | $\cos a_i$                    |
|-----------------------------|------------|-------------------------------|
| 55                          | 0.81915    | 0.57358                       |
| 81                          | 0.98769    | 0.15643                       |
| 96                          | 0.99452    | -0.10453                      |
| 109                         | 0.94552    | -0.32557                      |
| 117                         | 0.89101    | -0.45399                      |
| 132                         | 0.74315    | -0.66913                      |
| 154                         | 0.43837    | -0.89879                      |
| $\Sigma \sin a_i = 5.81941$ |            | $\Sigma \cos a_i = -1.722200$ |

$$n = 7$$

$$Y = \frac{\Sigma \sin a_i}{n} = 0.83134$$

$$X = \frac{\Sigma \cos a_i}{n} = -0.24600$$

$$r = \sqrt{X^2 + Y^2} = \sqrt{(-0.24600)^2 + (0.83134)^2} = \sqrt{0.75164} = 0.86697$$

$$\cos \bar{a} = \frac{X}{r} = -\frac{0.24600}{0.86697} = -0.28375$$

$$\sin \bar{a} = \frac{Y}{r} = \frac{0.83134}{0.86697} = 0.95890.$$

### Grouped Data

Often circular data are recorded in a frequency table. For such data, the following computations are convenient alternatives to Equations 19.6 and 19.7, respectively:

$$X = \frac{\sum f_i \cos a_i}{n} \quad \dots (19.13)$$

$$Y = \frac{\sum f_i \sin a_i}{n} \quad \dots (19.14)$$

(which are analogous to equation 3.3 for linear data). In these equations,  $a_i$  is the midpoint of the measurement interval recorded (e.g.,  $a_2 = 45^\circ$  in Example 19.3, which is the midpoint of the second recorded interval,  $30 - 60^\circ$ ), and  $f_i$  is the frequency of occurrence of data within that interval (e.g.,  $f_2 = 6$  in that example).

There is a bias in computing  $r$  from grouped data, in that the results is too small. A correction for this is available. For data grouped into intervals of  $d$  degrees each,

$$r_c = cr, \quad \dots (19.15)$$

where  $r_c$  is the corrected  $r$ , and  $c$  is a correction factor,

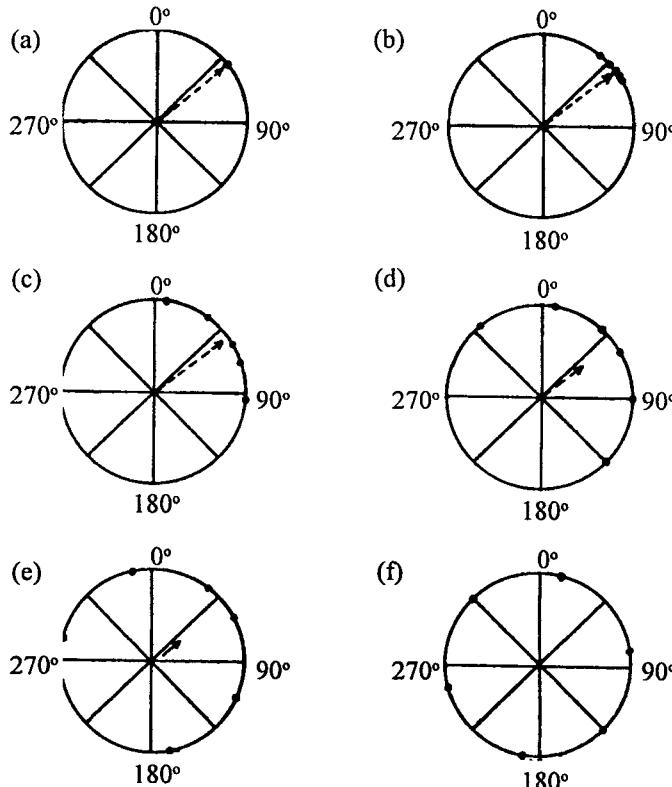
$$c = \frac{d\pi}{\sin\left(\frac{360^\circ}{d}\right)} \quad \dots (19.16)$$

The correction become insignificant as the interval becomes smaller than  $30^\circ$ . This correction is for the quantity  $r$ , and for statistics calculated from it; but the mean angle,  $\bar{a}$ , requires no correction for grouping. The correction may be applied when the distribution is unimodal and does not deviate greatly from symmetry.

## 19.5 ANGULAR DISPERSION

When dealing with circular data, we wish to have a measure, to describe the dispersion of the data.

We can define the *range* in a circular distribution of data as the smallest arc (i.e., the smallest portion of the circle's circumference) that contains all the data in the distribution. For example, in Fig. 19.6 (a), the range is zero; in Fig. 19.6 (b) the shortest arc is from the data point at  $38^\circ$  to the datum at  $60^\circ$ , making the range  $22^\circ$ ; in Fig. 19.6 (c), the data are found from  $10^\circ$  to  $93^\circ$ , with a range of  $83^\circ$ ; in Fig. 19.6 (d), the data run from  $322^\circ$  to  $135^\circ$ , with a range of  $173^\circ$ ; in 19.6 (e), the shortest arc containing all the data is that running clockwise from  $285^\circ$  to  $171^\circ$ , namely an arc of  $246^\circ$ ; and in Fig. 19.6 (f) the range is  $300^\circ$ . For the data of Example 2 the range is  $99^\circ$  (as the data run from  $55^\circ$  to  $154^\circ$ ).



**Fig. 19.6 : Circular distributions with various amounts of dispersion. The broken line indicates the mean angle, which is  $50^\circ$  in each case. Note that the value of  $r$  varies inversely with the amount of dispersion, and that the value of  $s$  varies directly with the amount of dispersion. (a)  $r = 1.00$ ,  $s = 0^\circ$ ,  $s' = 0^\circ$ . (b)  $r = 0.99$ ,  $s = 8.10^\circ$ ,  $s' = 8.12^\circ$ . (c)  $r = 0.90$ ,  $s = 25.62^\circ$ ,  $s' = 26.30^\circ$ . (d)  $r = 0.60$ ,  $s = 51.25^\circ$ ,  $s' = 57.91^\circ$ . (e)  $r = 0.30$ ,  $s = 67.79^\circ$ ,  $s' = 88.91$ . (f)  $r = 0.00$ ,  $s = 81.03^\circ$ ,  $s' = \infty$ .**

Another measure of dispersion is seen by examining Fig. 19.6; the value of  $r$  varies inversely with the amount of dispersion in the data. Therefore,  $r$  is a measure of concentration. It has no units and it may vary from 0 (when there is so much dispersion that a mean angle cannot be described) to 1.0 (when all the data are concentrated at the same direction).

The mean on a linear scale was noted to be the center of gravity of a group of data. Similarly, the length of the mean vector (i.e., the quantity  $r$ ), in the direction of the mean angle ( $\bar{\alpha}$ ) is a center of gravity. (Consider that each circle in Fig. 19.6 is a disc of material of negligible weight. The disc, held parallel to the ground would balance at the tip of the arrow in the Fig 19.6,  $r = 0$  and the center of the circle.)

Since  $r$  is a measure of concentration,  $1 - r$  is a measure of dispersion. Lack of dispersion would be indicated by  $1 - r = 1.0$  a measure called "mean angular deviation," or simply the *angular deviation*, is

$$s = \frac{180^\circ}{\pi} \sqrt{2(1-r)} \quad \dots (19.17)$$

in degrees.<sup>10</sup> This ranges from a minimum of zero to a maximum of  $81.03^\circ$ .

Mardia (1972:24, 74) defines *circular standard deviation* as

$$s' = \frac{180^\circ}{\pi} \sqrt{-2 \ln r} \quad \dots (19.18)$$

degrees; or, employing common, instead of natural, logarithms:

$$s' = \frac{180^\circ}{\pi} \sqrt{-4.60517 \log r} \quad \dots (19.19)$$

degrees. This is analogous to the standard deviation,  $s$ , on a linear scale is that it ranges from zero to infinity (see Fig. 19.6). For a given  $r$ , the values of  $s$  and  $s'$  differ by no more than 2 degrees for  $r$  as small as 0.80, by no more than 1 degree for  $r$  as small as 0.97. It is intuitively reasonable that a measure of angular dispersion should have a finite upper limit, so  $s$  is the measure preferred in this book. Appendix tables convert  $r$  to  $s$  and  $s'$ , respectively. If the data are grouped, then  $s$  and  $s'$ , are biased in being too high, so  $r_c$  (by equation 19.15) can be used in place of  $r$ .

Measures of symmetry and kurtosis on a circular scale, analogous to the  $g_1$  and  $g_2$  that may be calculated on a linear scale (Section 7.1), are discussed by Batschlet (1965: 14–15, 1981: 43–44) and Mardia (1972: 36–38, 74–76).

## 19.6 THE MEDIAN AND MODAL ANGLES

In a fashion analogous to considerations for linear scales of measurement. One can determine the median and the mode of a set of data on a circular scale.

To find the *median angle* we first determine that diameter of the circle which divides the data into two equal-sized groups. The median angle is the angle indicated by the radius on that diameter which is nearer to the majority of the data points. In Example 19.2, a diameter extending

---

<sup>10</sup>Simply delete the constant,  $180^\circ/\pi$ , in this and in the following equations if the measurement is desired in radians than in degrees.

from  $109^\circ$  to  $289^\circ$  (as indicated by the dashed concentrated around  $109^\circ$ , rather than  $289^\circ$ , so the sample median is  $109^\circ$ ). If  $n$  is odd, the median will be one of the data points. If  $n$  is even, the median is midway between two of the data. Mardia shows how the median is estimated, analogously to equation 3.6, when the median is within a group of tied data. If the data distribution in question has the data equally spaced around the circle [as in Fig. 19.6 (f)], then the median as well as the mean, is undefined.

The *modal angle* is defined in the same way as is the mode for linear scale data. Just as with linear data, there may be more than one mode.

## 19.7 CONFIDENCE LIMITS FOR THE MEAN AND MEDIAN ANGLES

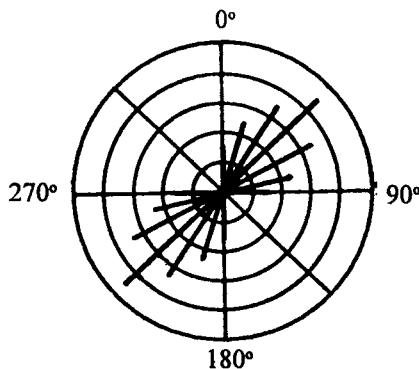
The concept of confidence limits is may be applied to the mean of angles. This Figure consists of two parts: a graph for 95% confidence and one for 99% confidence limits. For a given vector length  $r$ , and sample size,  $n$ , the graph gives the quantity  $d$ , which defines the confidence interval for  $\mu_a$ , as

$$\bar{a} \pm d \quad \dots (19.20)$$

(That is, lower confidence limit is  $L_1 = \bar{a} - d$  and the upper confidence limit is  $L_2 = \bar{a} + d$ ). For example, we may wish to estimate a population mean angle,  $\mu_a$ , when from a sample of 30 data from the population we calculate  $\bar{a} = 186^\circ$  and  $r = 0.80$ . Using the graph of Fig. B.2 for  $\alpha = 0.05$ ,  $d$  is found to be  $14^\circ$ , so the 95% confidence interval for  $\mu_a$  is  $186^\circ \pm 14^\circ$ . (That is, we are 95% confident that  $172^\circ$  and  $200^\circ$  encompass the mean angle in the sampled population.) If the data grouped than (By Equation 19.15) we can use  $r_c$  instead of  $r$ .

## 19.8 DIAMETRICALLY BIMODAL DISTRIBUTIONS

Occasionally populations are encountered having data with two modes lying opposite each other on the diameter of the circle. (Such data are sometimes termed "axial") for example, Fig. 19.7 shows a distribution having opposite modes at  $45^\circ$  and  $225^\circ$ .



*Fig. 19.7 : A bimodal circular distribution. For this distribution,  $H_0 : \rho = 0$  would not be rejected.*

If, as in this Figure the distribution is centrally symmetrical (i.e., each observation is matched by an observation  $180^\circ$  away),  $r$  computes to be zero and no mean angle can be determined. If the diametrically bimodal distribution is not centrally symmetrical,  $r$  will not be zero, but it may

be so small as to have us conclude that there is no significant direction of orientation of the data and the calculate mean may be so far from the diameter along which the bulk of the observations lie. However, we can engage in statistical analysis of such a distribution by a procedure involving the doubling of angles.

Example 19.5 shows the data that are graphed in Fig. 19.7. Each angle,  $a_i$ , is doubled; if the doubled angle is  $< 360^\circ$  it is recorded as  $2a_i$ , and if it is  $\geq 360^\circ$  then  $360^\circ$  is subtracted from it with the result being recorded as  $2a_i$ . (Also note in this examination that Equations 19.13 and 19.14 are used to make the computations easier, since the data are grouped.) Note in Example 19.5 that the angular deviation is one-half the angular deviation of the doubled angles.

**Example 5 :** Descriptive statistics for the centrally symmetric distribution shown in Fig. 19.6.

| $a_i$       | $f_i$ | $2a_i$                                  | $\sin 2a_i$ | $f_i \sin 2a_i$ | $\cos 2a_i$                           | $f_i \cos 2a_i$ |
|-------------|-------|---|-------------|-----------------|---------------------------------------|-----------------|
| $0^\circ$   | 15    | $0^\circ$                               | 0.00000     | 0.00000         | 1.00000                               | 15.00000        |
| $15^\circ$  | 25    | $30^\circ$                              | 0.50000     | 12.50000        | 0.86603                               | 21.65075        |
| $30^\circ$  | 35    | $60^\circ$                              | 0.86603     | 30.31105        | 0.50000                               | 17.50000        |
| $45^\circ$  | 45    | $90^\circ$                              | 1.00000     | 45.00000        | 0.00000                               | 0.00900         |
| $60^\circ$  | 35    | $120^\circ$                             | 0.86603     | 30.31105        | -0.50000                              | -17.50000       |
| $75^\circ$  | 25    | $150^\circ$                             | 0.50000     | 12.50000        | -0.86603                              | -21.65075       |
| $90^\circ$  | 15    | $180^\circ$                             | 0.00000     | 0.00000         | -1.00000                              | -15.00000       |
| $180^\circ$ | 15    | $0^\circ$                               | 0.00000     | 0.00000         | 1.00000                               | 15.00000        |
| $195^\circ$ | 25    | $30^\circ$                              | 0.50000     | 12.50000        | 0.86603                               | 21.65075        |
| $210^\circ$ | 35    | $60^\circ$                              | 0.86603     | 30.31105        | 0.50000                               | 17.50000        |
| $225^\circ$ | 45    | $90^\circ$                              | 1.00000     | 45.00000        | 0.00000                               | 0.00900         |
| $240^\circ$ | 35    | $120^\circ$                             | 0.86603     | 30.31105        | -0.50000                              | -17.50000       |
| $255^\circ$ | 25    | $150^\circ$                             | 0.50000     | 12.50000        | -0.86603                              | -21.65075       |
| $270^\circ$ | 15    | $180^\circ$                             | 0.00000     | 0.00000         | -1.00000                              | -15.00000       |
| $n = 390$   |       | $\Sigma f_i \sin 2a_i$<br>$= 261.24420$ |             |                 | $\Sigma f_i \cos 2a_i$<br>$= 0.00000$ |                 |

$$Y = \frac{261.24420}{390} = 0.66986 ; \quad X = \frac{0}{390} = 0$$

$$r = \sqrt{(0.66986)^2 + (0)^2} = 0.66986$$

$$\cos 2\bar{a} = \frac{0}{0.66986} = 0$$

$$\sin 2\bar{a} = \frac{0.66986}{0.66986} = 1.0000$$

Therefore,  $2\bar{a} = 90^\circ$  and  $\bar{a} = 45^\circ$ , meaning that the bimodal distribution lies along a diameter line oriented at  $45^\circ$  (as can be seen by inspecting Fig. 19.7).

$$s \text{ for doubled angles} = \sqrt{2(1 - 0.66986)} = \sqrt{0.66028} = 0.81258$$

$$s = \frac{0.81258}{2} = 0.41$$

Inasmuch as the data are grouped (in intervals of  $15^\circ$ ) Equation 19.15 may be used, in which case we would find  $s = 0.40$ .

## 19.9 SECOND-ORDER ANALYSIS: THE MEAN OF MEAN ANGLES

If a mean is determined for each of several groups of angles, then we have a set of mean angles. Consider the data in Example 19.6. Here, a mean angle,  $\bar{a}_j$ , has been calculated for each of  $k$  samples of circular data, using procedures. If, now, we desire to determine the grand mean of these several means, it is not appropriate to consider each of the sample means as an angle and employ the method of Section 19.4. To do so would be to assume that each mean had a vector length,  $r$ , of 1.0 (i.e., that an angular deviation,  $s$ , of zero was the case in each of the  $k$  sample), a most unlikely situation. Instead, we shall employ the procedure promulgated by Batschelet<sup>11</sup> whereby the grand mean has rectangular coordinates.

$$\bar{X} = \frac{\sum_{j=1}^k X_j}{k} \quad \dots (19.21)$$

$$\text{and} \quad \bar{Y} = \frac{\sum_{j=1}^k Y_j}{k} \quad \dots (19.22)$$

where  $X_j$  and  $Y_j$  are the quantities  $X$  and  $Y$  respectively, applying Equations 19.2 and 19.3 to sample  $j$ ;  $k$  is the total number of samples. If we do not have  $X$  and  $Y$  for each sample, but we have  $\bar{a}_j$  and  $r_j$  (polar coordinates) for each sample, then

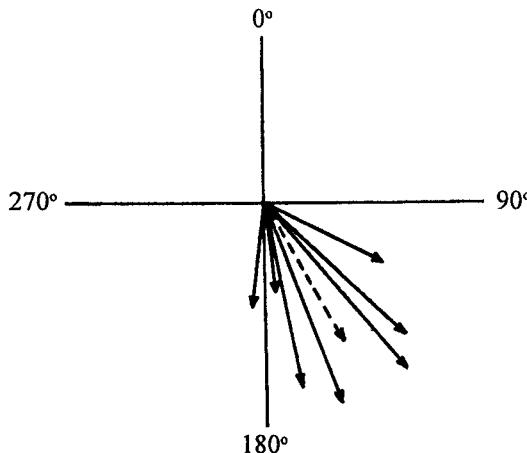
$$\bar{X} = \frac{\sum_{j=1}^k r_j \cos \bar{a}_j}{k} \quad \dots (19.23)$$

$$\text{and} \quad \bar{Y} = \frac{\sum_{j=1}^k r_j \sin \bar{a}_j}{k} \quad \dots (19.24)$$

<sup>11</sup>Batschelet refers to the determination of the mean of a set of angles as a first-order analysis and the computation of the mean of a set of means as a second-order analysis.

Having obtained  $\bar{X}$  and  $\bar{Y}$ , we may substitute them for  $X$  and  $Y$ , respectively, in Equations 19.8, 19.9 and 19.10 (and 19.11, if desired) in order to determine  $\bar{a}$ , which is the grand mean. For this calculation, all  $n_j$ 's (sample sizes) should be equal, although a slight departure from this condition will not severely affect the results.

Figure 19.8 shows the individual means and the grand mean for Example 19.6.



**Fig. 19.8 : The data of Example 19.6. Each of the seven vectors in this sample is itself a mean vector. The mean of these seven means is indicated by the broken line.**

#### **Example 6 : The mean of a set of mean angles.**

Under particular light conditions, each of 7 butterflies is allowed to fly from the center of an experimental chamber 10 times. Using the procedures of Section 19.4, the values of  $\bar{a}$  and  $r$  for each of the samples of data are as follows.

$$k = 7; \quad n = 10$$

| Sample ( $j$ ) | $\bar{a}_j$ | $r_j$  | $X_j = r_j \cos \bar{a}_j$ | $Y_j = r_j \sin \bar{a}_j$ |
|----------------|-------------|--------|----------------------------|----------------------------|
| 1              | 160°        | 0.8954 | -0.84140                   | 0.30624                    |
| 2              | 169         | 0.7747 | -0.76047                   | 0.14782                    |
| 3              | 117         | 0.4696 | -0.21319                   | 0.41842                    |
| 4              | 140         | 0.8794 | -0.67366                   | 0.56527                    |
| 5              | 186         | 0.3922 | -0.39005                   | -0.04100                   |
| 6              | 134         | 0.6952 | -0.48293                   | 0.50009                    |
| 7              | 171         | 0.3338 | -0.32969                   | 0.05222                    |
|                |             |        | -3.69139                   | 1.94906                    |

$$\bar{X} = \frac{\sum_{j=1}^k r_j \cos \bar{\alpha}_j}{k} = \frac{-3.69139}{7} = -0.52734.$$

$$\bar{Y} = \frac{\sum_{j=1}^k r_j \sin \bar{\alpha}_j}{k} = \frac{1.94906}{7} = 0.27844.$$

$$r = \sqrt{\bar{X}^2 + \bar{Y}^2} = \sqrt{0.35562} = 0.59634.$$

$$\cos \bar{\alpha} = \frac{\bar{X}}{r} = \frac{-0.52734}{0.59634} = -0.88429.$$

$$\sin \bar{\alpha} = \frac{\bar{Y}}{r} = \frac{0.27844}{0.59634} = 0.46691.$$

Therefore,  $\bar{\alpha} = 152^\circ$ .

## 19.10 CONFIDENCE LIMITS FOR THE SECOND-ORDER MEAN ANGLE

Section 19.9 explains how to obtain the mean of a set of mean angles. The mean thus computed is a example estimate of a population mean,  $\mu_a$ , and it reasonable to ask how precise an estimate it is. The precision with which we estimate a population mean is typically expressed as a confidence interval for the parameter. For a fist-order circular statistical analysis we may find confidence limits for  $\mu_a$  by the procedure in 19.7. For a second-order analysis we may express confidence limits for  $\mu_a$  if we first conclude (by the method of 2) that there is a significant directionality in the data.

Batschelet (1981:144, 262–265) shows geometrically and analytically how the second-order confidence limits are obtained. Here we shall simply present the arithmetic employed.

$$A = \frac{k-1}{\sum x^2} \quad \dots (19.25)$$

$$B = \frac{(k-1) \sum xy}{\sum x^2 \sum y^2} \quad \dots (19.26)$$

$$C = \frac{k-1}{\sum y^2} \quad \dots (19.27)$$

$$D = \frac{2(k-1) \left[ 1 - \frac{(\sum xy)^2}{\sum x^2 \sum y^2} \right]}{k(k-2)} \quad \dots (19.28)$$

$$H = AC - B^2 \quad \dots (19.29)$$

$$G = A\bar{X}^2 + 2B\bar{X}\bar{Y} + C\bar{Y}^2 - D \quad \dots (19.30)$$

$$U = H\bar{X}^2 - CD \quad \dots (19.31)$$

$$V = \sqrt{DGH} \quad \dots (19.32)$$

$$W = H\bar{X}\bar{Y} + BD \quad \dots (19.33)$$

$$b_1 = \frac{W + V}{U} \quad \dots (19.34)$$

$$b_2 = \frac{W - V}{U} \quad \dots (19.35)$$

The quantities  $b_1$  and  $b_2$  are then examined separately, each yielding one of the confidence limits, as follows:

$$M = \sqrt{1 + b_i^2} \quad \dots (19.36)$$

after which we determine (from trigonometric tables or by calculator) that angle having

$$\text{sine} = \frac{b_i}{M} \quad \dots (19.37)$$

$$\text{and} \quad \text{cosine} = \frac{1}{M} \quad \dots (19.38)$$

The confidence limit is either the angle thus determined or that angle + 180°, whichever is nearer the sample mean angle (and, if the angle +180° is greater than 360°, simply subtract 360°). This procedure is demonstrated in Example 19.7. The confidence interval thus computed is a little conservative (i.e., the confidence coefficient is a little greater than the stated  $1 - \alpha$ ), and the confidence limits are not necessarily symmetrical about the mean.

**Example 7 :** Confidence limits for the mean of a set of mean angles, using the data of example 6, for which  $\bar{a} = 152^\circ$ .

| $j$ | $X_j$    | $X_j^2$ | $Y_j$   | $Y_j^2$  | $X_j Y_j$ |
|-----|----------|---------|---------|----------|-----------|
| 1   | -0.84140 | 0.70795 | 0.30624 | 0.09378  | -0.25767  |
| 2   | -0.76047 | 0.57831 | 0.14782 | 0.02185  | -0.11241  |
| 3   | -0.21319 | 0.04545 | 0.41842 | 0.17508  | -0.08920  |
| 4   | -0.67366 | 0.45382 | 0.56527 | 0.31953  | -0.38080  |
| 5   | -0.39005 | 0.15214 | -0.4100 | 0.001168 | -0.01599  |
| 6   | -0.48293 | 0.23322 | 0.50009 | 0.25009  | -0.24151  |
| 7   | -0.32969 | 0.10870 | 0.05222 | 0.00273  | -0.01722  |
|     | -3.69139 | 2.27959 | 1.94906 | 0.86474  | -1.08282  |

For 95% confidence limits,  $\alpha = 0.05$

$$k = 7$$

$$\bar{X} = \frac{\sum X_j}{k} = \frac{-3.69139}{7} = -0.52734$$

$$\bar{Y} = \frac{\sum Y_j}{k} = \frac{1.94906}{7} = \mathbf{0.27844}.$$

$$\Sigma x^2 = \Sigma X_j^2 - \frac{(\sum X_j)^2}{k} = 2.27959 - \frac{(-3.69139)^2}{7} = \mathbf{0.33297}.$$

$$\Sigma y^2 = \Sigma Y_j^2 - \frac{(\sum Y_j)^2}{k} = 0.86474 - \frac{(1.94906)^2}{7} = \mathbf{0.32205}.$$

$$\Sigma xy = \Sigma x_j y_j - \frac{\Sigma X_j Y_j}{k} = -1.08282 - \frac{(-3.69139)(1.94906)}{7} = \mathbf{-0.05500}.$$

$$A = \frac{k-1}{\Sigma x^2} - \frac{7-1}{0.33297} = \mathbf{18.01964}.$$

$$B = \frac{(k-1) \Sigma xy}{\Sigma x^2 \Sigma y^2} = -\frac{(7-1)(-0.05500)}{(0.33297)(0.32205)} = \mathbf{3.07741}.$$

$$C = \frac{k-1}{\Sigma y^2} = \frac{7-1}{0.32205} = \mathbf{18.63065}.$$

$$F_{\alpha(1), 2, k-2} = F_{0.05(1), 2, 5} = 5.79$$

$$D = \frac{2(k-1) \left[ 1 - \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} \right] F_{\alpha(1), 2, k-2}}{k(k-2)}$$

$$= \frac{2(7-1) \left[ 1 - \frac{(-0.05500)^2}{(0.33297)(0.32205)} \right] (5.79)}{7(7-2)} = \mathbf{1.92914}.$$

$$H = AC - B^2 = (18.01964)(18.63065) - (3.07741)^2 = \mathbf{326.24715}.$$

$$\begin{aligned} G &= A\bar{X}^2 + 2B\bar{XY} + C\bar{Y}^2 - D \\ &= (18.01964)(-0.52734)^2 + 2(3.07741)(-0.52734)(0.27844) \\ &\quad + (18.63065)(0.27844)^2 - 1.92914 = \mathbf{3.62258}. \end{aligned}$$

$$U = H\bar{X}^2 - CD = (326.24715)(-0.52734)^2 - (18.63065)(1.92914) = \mathbf{54.7811}.$$

$$V = \sqrt{DGH} = \sqrt{(1.92914)(3.62258)(326.24715)} = \mathbf{47.748899}.$$

$$\begin{aligned} W &= H\bar{XY} + BD \\ &= (326.24715)(-0.52734)(0.27844) + (3.07741)(1.92914) = \mathbf{-41.96695}. \end{aligned}$$

$$b_1 = \frac{W + V}{U} = \frac{-41.96695 + 47.74899}{54.88411} = 0.10554.$$

$$b_2 = \frac{W - V}{U} = \frac{-41.96695 - 47.74899}{54.88411} = -1.63763.$$

For  $b_1$ :  $M = \sqrt{1 + b_1^2} = \sqrt{1 + (0.10554)^2} = 1.00555$ .

$$\text{sine} = \frac{b_1}{M} = \frac{0.10554}{1.00555} = 0.10496.$$

$$\text{cosine} = \frac{1}{M} = \frac{1}{1.00555} = 0.99448.$$

The angle with this sine and cosine is  $6^\circ$ , so one of the confidence limits is either  $6^\circ$  or  $6^\circ + 180^\circ$ ; of the two possibilities,  $186^\circ$  is closer to the mean ( $152^\circ$ ).

For  $b_2$ :  $M = \sqrt{1 + b_2^2} = \sqrt{1 + (-1.63763)^2} = 1.91881$ .

$$\text{sine} = \frac{b_2}{M} = \frac{-1.63763}{1.91881} = -0.85346.$$

$$\text{cosine} = \frac{1}{M} = \frac{1}{1.91881} = 0.52116.$$

The angle with this sine and cosine is  $301^\circ$ , so the second confidence limit is either  $301^\circ$  or  $301^\circ + 180^\circ = 481^\circ = 121^\circ$ ; of the two possibilities,  $121^\circ$  is closer to the mean ( $152^\circ$ ).

## EXERCISE

1. Five examples of directional data were collected and were as follows:
  - (a) Determine the mean of the five sample means.
  - (b) Determine the 95% confidence limits for the second-order mean.

| <i>Sample</i> | <i>Sample mean</i> | <i>Sample r</i> |
|---------------|--------------------|-----------------|
| 1             | $230^\circ$        | 0.4542          |
| 2             | 245                | 0.6083          |
| 3             | 265                | 0.7862          |
| 4             | 210                | 0.5107          |
| 5             | 225                | 0.8639          |

2. A total of 15 human births occurred as follows:

|          |           |           |          |
|----------|-----------|-----------|----------|
| 1: 15 AM | 4: 40 AM  | 5: 30 AM  | 6: 50 AM |
| 2: 00 AM | 11: 00 AM | 4: 20 AM  | 5: 10 AM |
| 4: 30 AM | 5: 15 AM  | 10: 30 AM | 8: 55 PM |
| 6: 10 AM | 3: 45 AM  | 3: 10 AM  |          |

- (a) Compute the mean time of birth.  
(b) Compute the angular deviation for the data.  
(c) determine 95% confidence limits for the population mean direction.  
(d) Determine the sample direction.
3. Twelve nests of a particular bird species were recorded facing outward from trees at the following directions:

| <i>Directions</i> | <i>Frequency</i> |
|-------------------|------------------|
| N : 0°            | 2                |
| NE: 45            | 4                |
| E: 90             | 3                |
| SE: 135           | 1                |
| S: 180            | 1                |
| SW: 225           | 1                |
| W: 270            | 0                |
| NW: 315           | 0                |

- (a) Compute the sample mean direction.  
(b) Compute the angular deviation for the data.  
(c) determine 95% confidence limits for the population mean.  
(d) Determine the sample median direction.



# 20

## *Tests of Significance for Circular Distributions*

Armed with the procedures in Chapter 19, and the information contained in the basic statistics of circular distributions (e.g.,  $\bar{a}$  and  $r$ ), we can now examine a number of statistical methods for testing hypothesis about populations measured on a circular scale.

### **20.1 GOODNESS OF FIT TESTING**

Either  $\chi^2$  test may be used to test the goodness of fit of theoretical to an observed circular frequency distribution. The procedure is to determine each expected frequency,  $\hat{f}_i$ , corresponding to each observed frequency,  $f_i$ , in each category,  $i$ . For the data of Example 19.3, for instance, we might hypothesize a uniform distribution of data among instance, we might hypothesize a uniform distribution of data among the 12 divisions of the data. The test of this hypothesis is presented in Example 20.1. Batschelet (1981 : 71) recommends grouping the data so that no expected frequency is less than 4 in using chi-square.

Recall that goodness of fit testing by the chi-square, or  $G$ , statistic does not take into account the sequence of categories that occurs in the data distribution. The Kolmogorov-Smirnov test was introduced as an improvement over chi-square when the categories of data are, infact, ordered. Unfortunately the Kolmogorov-Smirnov test yields different results for different starting points on a circular scale; however, a modification of this test by Kuiper (1960) provides a goodness of fit test, the results of which are unrelated to the starting point on a circle.

**Example 1 :** *Chi-square goodness of fit for the circular data:*

$H_0$  : The data in the population are distributed uniformly around the circle.

$H_A$  : The data in the population are not distributed uniformly around the circle.

| $a_i$ (deg) | $f_i$ | $\hat{f}_i$ |
|-------------|-------|-------------|
| 0–30        | 0     | 8.7500      |
| 30–60       | 6     | 8.7500      |
| 60–90       | 9     | 8.7500      |
| 90–120      | 13    | 8.7500      |
| 120–150     | 15    | 8.7500      |
| 150–180     | 22    | 8.7500      |
| 180–210     | 17    | 8.7500      |
| 210–240     | 12    | 8.7500      |
| 240–270     | 8     | 8.7500      |
| 270–300     | 3     | 8.7500      |
| 300–330     | 0     | 8.7500      |
| 330–360     | 0     | 8.7500      |

 $k = 12$  $n = 105$ 

$$\hat{f}_i = 105/12 = 8.7500 \text{ for all } i$$

$$\begin{aligned} \chi^2 &= \frac{(0 - 8.7500)^2}{8.7500} + \frac{(6 - 8.7500)^2}{8.7500} + \frac{(9 - 8.7500)^2}{8.7500} + \dots - \frac{(0 - 8.7500)^2}{8.7500} \\ &= 8.7500 + 0.8643 + 0.0071 + \dots + 8.7500 \\ &= \mathbf{66.543} \end{aligned}$$

$$v = k - 1 = 11$$

$$\chi^2_{0.05, 11} = 19.675$$

Reject  $H_0$ .  $P < 0.001$ .

The Kuiper test is discussed by Batschelet and Mardia.

Another goodness of fit test applicable to circular distributions is that of Watson (1962), often referred to as the *Watson one-sample U<sup>2</sup>* test. To test the null hypothesis of uniformity, we first transform each angular measurement,  $a_i$ , by dividing it by 360°:

$$u_i = \frac{a_i}{360^\circ} \quad \dots (20.1)$$

Then the following quantities are obtained for the set of  $n$  values of  $u_i$ :  $\Sigma u_i$ ,  $\Sigma u_i^2$ ,  $\bar{u}$  and  $\Sigma i u_i$ . The test statistic, called "Watson's  $U^2$ ", is

$$U^2 = \sum u_i^2 \frac{(\Sigma u_i)^2}{n} - \frac{2}{n} \sum i u_i + (n+1) \bar{u} + \frac{n}{12} \quad \dots (20.2)$$

Critical values for this test are  $U^2_{\alpha, n, n}$  in Table at the end.

All the testing procedures for this section are non-parametric.

**Example 2 : Watson's goodness of fit testing using the following data:**

$H_0$  : The sample data come from a population uniformly distributed around the circle.

$H_A$  : The sample data do not come from a population uniformly distributed around the circle.

| $i$     | $a_i$ | $u_i$                 | $u_i^2$                 | $iu_i$                 |
|---------|-------|-----------------------|-------------------------|------------------------|
| 1       | 55°   | 0.1528                | 0.0233                  | 0.1528                 |
| 2       | 81°   | 0.2250                | 0.0506                  | 0.4500                 |
| 3       | 96°   | 0.2667                | 0.0711                  | 0.8001                 |
| 4       | 109°  | 0.3028                | 0.0917                  | 1.2112                 |
| 5       | 117°  | 0.3250                | 0.1056                  | 1.6250                 |
| 6       | 132°  | 0.3667                | 0.1345                  | 2.2002                 |
| 7       | 154°  | 0.4278                | 0.1830                  | 2.9946                 |
| $n = 7$ |       | $\Sigma u_i = 2.0668$ | $\Sigma u_i^2 = 0.6598$ | $\Sigma iu_i = 9.4339$ |

$$\bar{u} = \frac{\sum u_i}{n} = \frac{2.0668}{7} = 0.2953$$

$$\begin{aligned} U^2 &= \sum u_i^2 \frac{(\sum u_i)^2}{n} - \frac{2}{n} \sum iu_i + (n+1) \bar{u} + \frac{n}{12} \\ &= 0.6598 - \frac{(2.0668)^2}{7} - \frac{2}{7}(9.4339) + (7+1)(0.2953) + \frac{7}{12} \\ &= 0.6598 - 0.6102 - 2.6954 + 2.3624 + 0.5833 \\ &= 0.2999. \end{aligned}$$

$$U_{0.05, 7, 7}^2 = 0.1986$$

Therefore, reject  $H_0$ .  $0.01 < P < 0.02$ .

## 20.2 THE SIGNIFICANCE OF THE MEAN AND MEDIAN ANGLES

One can obviously place more confidence in  $\bar{a}$  as an estimate of the population mean angle,  $\mu_a$ , if  $s$  is small than if it is large. This is identical to stating that  $\bar{a}$  is a better estimate of  $\mu_a$  if  $r$  is large than if  $r$  is small. What is desired is a method of asking whether there is, in fact, a mean direction among the population of data which were sampled, for even if there is no mean direction (i.e., the circular distribution is uniform) in the population, a random sample might still display a calculable mean. The test we require is that concerning  $H_0$  : The sampled population is uniformly<sup>1</sup> distributed around a circle vs.  $H_A$  : The population is not a uniform

<sup>1</sup>In dealing with circular statistics, the terms "uniform distribution" and "random distribution" have been used synonymously in the literature.

circular distribution. This may be tested by the non-parametric *Rayleigh test*.<sup>2</sup> As circular uniformity implies no mean direction, the Rayleigh test may be said to test  $H_0: \rho = 0$  vs.  $H_A: \rho \neq 0$ .

The Rayleigh test asks how large a sampler  $r$  must be to indicate confidently a non-uniform population distribution. A quantity referred to as "Rayleigh's  $R$ " is obtainable as

$$R = nr \quad \dots (20.3)$$

and the so-called "Rayleigh's  $z$ " is utilized for testing the null hypothesis of no population mean direction:

$$z = \frac{R^2}{n} \quad \text{or} \quad z = nr^2 \quad \dots (20.4)$$

Table at the end presents critical values of  $z_{\alpha, n}$ . If the data are grouped, then we may substitute  $r_c$  for  $r$  in this test.

If  $H_0$  is rejected by Rayleigh's test, we may conclude that there is a mean population direction (See Example 20.3), but only if the distribution is unimodal. If  $H_0$  is not rejected, we may conclude the population distribution to be uniform around the circle only if we may assume that it does not have more than one mode.

**Example 3 : Rayleigh's test applied to the data of Example 2.**

$H_0: \rho = 0$  (i.e., the population is uniformly distributed around the circle).

$H_A: \rho \neq 0$  (i.e., the population is not uniformly distributed around the circle).

The following were obtained in Example 19.4:

$$n = 7$$

$$r = 0.86697$$

Therefore,

$$R = nr = (7)(0.86697) = 6.06879$$

and

$$z = \frac{R^2}{n} = \frac{(6.06879)^2}{7} = 5.261.$$

Using Table  $z_{0.05, 7} = 2.885$ . Reject  $H_0$ .  $0.001 < P < 0.002$ .

A modification of Rayleigh test (Greenwood and Durand, 1955; Durand and Greenwood, 1958) is available for use when the investigator has reason to expect, *in advance*, a specific mean direction. In Example 20.4 (*a*), 10 birds were released at a site directly west of their home. Therefore, the statistical hypothesis may include the expectation of birds to tend to fly directly east (i.e., at an angle of 90°). The testing procedure considers  $H_0$ : The population angles are randomly distributed (i.e.,  $H_0: p = 0$ ) vs.  $H_A$ : The population angles are not randomly distributed (i.e.,  $H_A: p \neq 0$ ). By using additional information, namely the expected mean angle, this test is more powerful than Rayleigh's test.

---

<sup>2</sup>Named for Lord Rayleigh [John William Strutt, Third Baron Rayleigh (1842-1919)], a physicist and applied mathematician who gained his greatest fame for discovering and isolating the chemical element argon (winning him the Nobel Prize in Physics in 1904), although some of his other contributions to Physics were at least as important (Lindsay, 1976).

**Example 4 : The *V* test.**

$H_0$  : The population is uniformly distributed around the circle (i.e.,  $H_0 : p = 0$ ).

$H_A$  : The population is not uniformly distributed around the circle (i.e.,  $H_A : p \neq 0$ ).

| $a_i$ (deg) | $\sin a_i$                  | $\cos a_i$                   |
|-------------|-----------------------------|------------------------------|
| 66          | 0.91335                     | 0.40674                      |
| 75          | 0.96593                     | 0.25882                      |
| 86          | 0.99756                     | 0.06976                      |
| 88          | 0.99939                     | 0.03490                      |
| 88          | 0.99939                     | 0.03490                      |
| 93          | 0.99863                     | -0.05234                     |
| 97          | 0.99255                     | -0.12187                     |
| 101         | 0.98163                     | -0.19081                     |
| 118         | 0.88295                     | -0.46947                     |
| 130         | 0.76604                     | -0.64279                     |
| $n = 10$    | $\Sigma \sin a_i = 9.49762$ | $\Sigma \cos a_i = -0.67216$ |

$$Y = \frac{9.49762}{10} = 0.94976 ; \quad X = -\frac{0.67216}{10} = -0.06722$$

$$r = \sqrt{(-0.06722)^2 + (0.94976)^2} = 0.95213$$

$$\sin \bar{a} = \frac{Y}{r} = 0.99751$$

$$\cos \bar{a} = \frac{X}{r} = -0.07060$$

$$\bar{a} = 94^\circ.$$

$$R = (10)(0.95213) = 9.5213$$

$$\begin{aligned} V &= R \cos (94^\circ - 90^\circ) = 9.5213 \cos (4^\circ) \\ &= (9.5213)(0.99756) \\ &= 9.498 \end{aligned}$$

$$\begin{aligned} u &= V \sqrt{\frac{2}{n}} = (9.498) \sqrt{\frac{2}{10}} \\ &= 4.248. \end{aligned}$$

Using table B.33,  $u_{0.05, 10} = 1.648$ . Reject  $H_0$ .  $P < 0.0005$ .

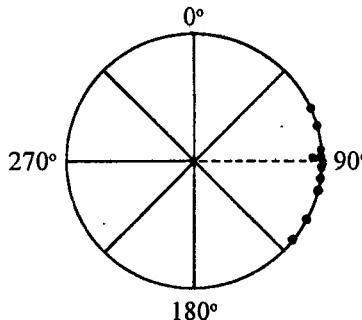
The proceeding hypotheses are tested by what we shall refer to as the *V* test, in which the test statistic is computed as

$$V = R \cos (\bar{a} - \mu_0) \quad \dots (20.5)$$

where  $\mu_0$  is the mean angle predicted. Table at the end gives critical values of  $\mu_{\alpha, n}$ , a statistic which, for large sample sizes, approaches a one-tailed normal deviate,  $Z$ .

$$u = V \sqrt{\frac{2}{n}} \quad \dots (20.6)$$

If the data are grouped, then  $R$  may be determined from  $r_c$  rather than  $r$ .



*Fig. 20.1 : The data for the  $V$  test of Example 20.4. The broken line indicates the expected mean angle ( $94^\circ$ ).*

### One-sample Test for the Mean angle

The Rayleigh test and the  $V$  test are methods for testing for random distribution of a population of data around the circle. (See Batschelet, 1981: Chapter 4, for other tests of the null hypothesis of randomness.) If it is desired to test whether the population mean angle is equal to a specified value, say  $\mu_0$ , then we have a one-sample test situation analogous to that of the one-sample  $t$  test for data on a linear scale (Section 8.1). The hypotheses are

$$H_0 : \mu_0 = \mu_0$$

and

$$H_A : \mu_a = \mu_0,$$

and  $H_0$  is tested simply by observing whether  $\mu_0$  lies within the  $1 - \alpha$  confidence interval for  $\mu_0$ . If  $\mu_0$  lies outside the confidence interval, then  $H_0$  is rejected. Section 19.7 describes the determination of confidence intervals for the population mean angle, and Example 20.4 (b) demonstrates the hypothesis testing procedure.

### Example 5 : The one-sample test for the mean angle.

$H_0$  : The population has mean of  $90^\circ$  (i.e.,  $\mu_a = 90^\circ$ ).

$H_A$  : The population has mean is not  $90^\circ$  (i.e.,  $\mu_a \neq 90^\circ$ ).

The computation of the following is given in Example 20.4 (a):

$$r = 0.95$$

$$\bar{a} = 94^\circ$$

Using appendix Fig. B.2 for  $\alpha = 0.05$  and  $n = 10$ :

$d = 13^\circ$ ; so the 95% confidence interval for  $\mu_a$  is  $94^\circ \pm 13^\circ$ .

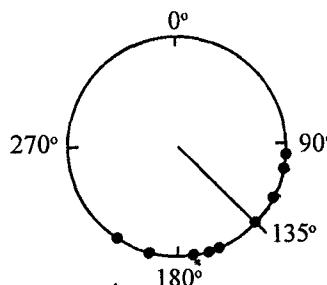
As this confidence interval does contain the hypothesized mean ( $\mu_0 = 90^\circ$ ), we do not reject  $H_0$ .

### One-sample Test for the Median Angle

We can perform a non-parametric test to determine the median angle equals a specified value simply by counting the number of observed angles on either side of a diameter through the hypothesized angle and subjecting these data on the binomial test of Section 20.5. This is demonstrated in Example 20.5.

**Example 5 : The significance of the median angle.**

The sample data consist of the following directions:  $97^\circ, 104^\circ, 121^\circ, 136^\circ, 159^\circ, 164^\circ, 172^\circ, 195^\circ, 213^\circ$ . The median is  $159^\circ$ .



If we wish to test whether the population median is equal to some specific value — say,  $135^\circ$  — we can proceed as follows:

$H_0$  : The population median angle is  $135^\circ$ .

$H_A$  : The population median angle is not  $135^\circ$ .

Employing the two-tailed binomial test, we have  $n = 9$  and  $p = 0.5$ . There are three sample directions  $< 135^\circ$  and six directions  $> 135^\circ$ . It is found that  $P = 0.508$ . Do not reject  $H_0$ .

### 20.3 PARAMETRIC TWO-SAMPLE AND MULTISAMPLE TESTING OF ANGLES

It is common to consider the null hypothesis  $H_0 : \mu_1 = \mu_2$ , where  $\mu_1$  and  $\mu_2$  are the mean angles for each of two circular distributions. Watson and Williams proposed a test that utilizes the statistic

$$F = K \frac{(N - 2)(R_1 + R_2 - R)}{N - R_1 - R_2} \quad \dots (20.7)$$

where  $N = n_1 + n_2$ . In this equation,  $R$  is calculated by Equation 20.3 with the data from the two samples being combined;  $R_1$  and  $R_2$  are the values of Rayleigh's  $R$  for the two samples considered separately.  $K$  is a factor, obtained from Table, that corrects for bias in the  $F$  calculation; in that table one uses for  $r$  the weighted mean of the two vector lengths:

$$r_w = \frac{n_1 r_1 + n_2 r_2}{N} = \frac{R_1 + R_2}{N} \quad \dots (20.8)$$

**Example 6 : The Watson-Williams test for two samples.**

$$H_0 : \mu_a = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

Sample 1

| $a_i$ (deg) | $\sin a_i$                     | $\cos a_i$                     |
|-------------|--------------------------------|--------------------------------|
| 94          | 0.99756                        | -0.06976                       |
| 65          | 0.90631                        | 0.42262                        |
| 45          | 0.70711                        | 0.70711                        |
| 52          | 0.78801                        | 0.61566                        |
| 38          | 0.61566                        | 0.78801                        |
| 47          | 0.73135                        | 0.68200                        |
| 73          | 0.95630                        | 0.29237                        |
| 82          | 0.99027                        | 0.13917                        |
| 90          | 1.00000                        | 0.00000                        |
| 40          | 0.64279                        | 0.76604                        |
| 87          | 0.99863                        | 0.05234                        |
| $n_1 = 11$  | $\Sigma \sin a_i$<br>= 9.33399 | $\Sigma \cos a_i$<br>= 4.39556 |

Sample 2

| $a_i$ (deg) | $\sin a_i$                     | $\cos a_i$                    |
|-------------|--------------------------------|-------------------------------|
| 77          | 0.97437                        | 0.22495                       |
| 70          | 0.93969                        | 0.34202                       |
| 61          | 0.87462                        | 0.48481                       |
| 45          | 0.70711                        | 0.70711                       |
| 50          | 0.76604                        | 0.64279                       |
| 35          | 0.57358                        | 0.81915                       |
| 48          | 0.74314                        | 0.66913                       |
| 65          | 0.90631                        | 0.42262                       |
| 36          | 0.58779                        | 0.80902                       |
| $n_2 = 9$   | $\Sigma \sin a_i$<br>= 7.07265 | $\Sigma \cos a_i$<br>= 512160 |

$$Y = 84854, \quad X = 0.3960$$

$$r_1 = 0.93792$$

$$\sin \bar{a}_1 = 0.90470$$

$$\cos \bar{a}_1 = 0.42605$$

$$\bar{a}_1 = 65^\circ$$

$$R_1 = 10.31712$$

$$Y = 0.78585, \quad X = 0.56907$$

$$r_2 = 0.97026$$

$$\sin \bar{a}_2 = 0.80994$$

$$\cos \bar{a}_2 = 0.58651$$

$$\bar{a}_2 = 54^\circ$$

$$R_2 = 8.73234$$

By combining the 20 data from both samples:

$$\Sigma \sin a_i = 9.33399 + 7.07265 = 16.40664$$

$$\Sigma \cos a_i = 4.39556 + 5.12160 = 9.51716$$

$$N = 11 + 9 = 20$$

$$Y = \frac{16.40664}{20} = 0.82033$$

$$X = \frac{9.51716}{20} = 0.47586$$

$$r = 0.94836$$

$$R = 18.96720$$

$$r_w = \frac{10.31712 + 8.73234}{20} = 0.952$$

$$F = K \frac{(N - 2)(R_1 + R_2 - R)}{N - R_1 - R_2}$$

$$= (1.0351) \frac{(20 - 2)(10.31712 + 8.73234 - 18.96720)}{20 - 10.31712 - 8.73234}$$

$$= (1.0351) \frac{1.48068}{0.95054} = 1.61.$$

$$F_{0.05, 1, 18} = 4.41$$

Therefore, do not reject  $H_0$ .

$$0.10 < \rho < 0.25$$

Therefore, we conclude that the two sample means estimate the same population mean, and the best estimate of this population mean is obtained by:

$$\sin \bar{a} = \frac{Y}{r} = 0.86500$$

$$\cos \bar{a} = \frac{X}{r} = 0.50177$$

$$\bar{a} = 60^\circ.$$

The critical value for this test is  $F_{\alpha(1), 1, N-2}$ . Alternatively,

$$t = \sqrt{K \frac{(N - 2)(R_1 + R_2 - R)}{N - R_1 - R_2}} \quad \dots (20.9)$$

may be compared with  $t_{\alpha(2), N-2}$ . This test may be used for  $r_w$  as small as 0.75, if  $5 \leq N/2 < 10$ ; or for  $r_w$  as low as 0.70, if  $N/2 \geq 10$ . The data may be grouped as long as the grouping interval is  $\leq 10^\circ$ .

### Multiple Testing

The Watson-Williams test can be generalized to a multisample test for testing  $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ , a hypothesis reminiscent of analysis of variance considerations for linear data. In multisample tests (Example 7),

$$F = K \frac{(N - K) \left( \sum_{j=1}^k R_j - R \right)}{(k - 1) \left( N - \sum_{j=1}^k R_j \right)} \quad \dots (20.10)$$

Here,  $k$  is the number of samples,  $R$  is the Rayleigh's  $R$  for all  $K$  samples combined, and  $N = \sum_{j=1}^k n_j$ . The correction factor,  $K$ , is obtained from Table a, using

$$r_w = \frac{\sum_{j=1}^k n_j r_j}{N} = \frac{\sum_{j=1}^k R_j}{N} \quad \dots (20.11)$$

**Example 7 : The Watson-Williams test for three samples.**

$H_0$  : All three samples are from populations with the same mean angle.

$H_A$  : All three samples are not from populations with the same mean angle.

Sample 1

| $a_i$ (deg) | $\sin a_i$                       | $\cos a_i$                        |
|-------------|----------------------------------|-----------------------------------|
| 135         | 0.70711                          | -0.70711                          |
| 145         | 0.57358                          | -0.81915                          |
| 125         | 0.81915                          | -0.57358                          |
| 140         | 0.64279                          | -0.76604                          |
| 165         | 0.25882                          | -0.96593                          |
| 170         | 0.17365                          | -0.98481                          |
| $n_1 = 6$   | $\Sigma \sin a_i$<br>$= 3.17510$ | $\Sigma \cos a_i$<br>$= -4.81662$ |

$$\bar{a}_1 = 147^\circ$$

$$r_1 = 0.96150$$

$$R_1 = 5.76894$$

Sample 2

| $a_i$ (deg) | $\sin a_i$                       | $\cos a_i$                        |
|-------------|----------------------------------|-----------------------------------|
| 150         | 0.50000                          | -0.86603                          |
| 130         | 0.76604                          | -0.64279                          |
| 175         | 0.08716                          | -0.99619                          |
| 190         | -0.17365                         | -0.98481                          |
| 180         | 0.00000                          | -1.00000                          |
| 220         | -0.64279                         | -0.76604                          |
| $n_2 = 6$   | $\Sigma \sin a_i$<br>$= 0.53676$ | $\Sigma \cos a_i$<br>$= -5.25586$ |

$$\bar{a}_2 = 174^\circ$$

$$r_2 = 0.88053$$

$$R_2 = 5.28324$$

Sample 3

| $a_i$ (deg) | $\sin a_i$                  | $\cos a_i$                   |
|-------------|-----------------------------|------------------------------|
| 140         | 0.64279                     | -0.76604                     |
| 165         | 0.25882                     | -0.96593                     |
| 185         | -0.08715                    | -0.99619                     |
| 180         | 0.00000                     | -1.00000                     |
| 125         | 0.81915                     | -0.57358                     |
| 175         | 0.08716                     | -0.99619                     |
| 140         | 0.64279                     | -0.76604                     |
| $n_3 = 7$   | $\Sigma \sin a_i = 2.36356$ | $\Sigma \cos a_i = -6.06397$ |

$$\bar{a}_3 = 159^\circ$$

$$r_3 = 0.92976$$

$$R_3 = 6.50832$$

$$k = 3$$

$$N = 6 + 6 + 7 = 19$$

For all 19 data:

$$\sum \sin a_i = 3.17510 + 0.53676 + 2.36356 = 6.07542$$

$$\sum \cos a_i = -4.81662 - 5.25586 - 6.06397 = -16.13645$$

$$Y = 0.31976$$

$$X = -0.84929$$

$$r = 0.90749$$

$$R = 17.24231$$

$$r_w = \frac{5.76894 + 5.28324 + 6.50832}{19} = 0.924.$$

$$F = K \frac{(N - K) \left( \sum_{j=1}^k R_j - R \right)}{(k - 1) \left( N - \sum_{j=1}^k R_j \right)}$$

$$= (1.0546) \frac{(19 - 3)(5.76894 + 5.28324 + 6.50832 - 17.24231)}{(3 - 1)(19 - 5.76894 - 5.28324 - 6.50832)}$$

$$= (1.0546) \frac{5.09104}{2.87900} = 1.86.$$

$$v_1 = k - 1 = 2$$

$$v_2 = N - k = 16$$

$$F_{0.05(1), 2, 10} = 3.63$$

Therefore do not reject  $H_0$ .

$$0.10 < P < 0.25$$

Therefore, we conclude that the three sample means estimate the same population mean, and the best estimate of that population mean is obtained by:

$$\sin \bar{a} = \frac{Y}{r} = 0.35236$$

$$\cos \bar{a} = \frac{X}{r} = -0.93587$$

$$\bar{a} = 159^\circ$$

The critical value for this test is  $F_{\alpha(1), k-1, N-k}$ . Equation 20.11 (and, thus this test) may be used for  $r_w$  as small as 0.45, if  $N/k > 10$ ; for  $r_w$  as small as 0.50, for  $N/k > 6$ ; and for  $r_w$  as small as 0.55, for  $N/k = 5$  or 6 (Mardia, 1972: 163; Batschelet (1981: 321). If the data are grouped, the grouping interval should be no longer than  $10^\circ$ . Upon (1976) presents an alternative to the Watson-Williams procedure is a little simpler to use.

The Watson-Williams test (for two or more samples) is parametric and assumes that each of the samples come from a population conforming to what is known as Von Mises or Circular normal distribution.

### EXERCISE

1. The direction of spring flight of a certain bird species was recorded as follows in eight individuals released in full sunlight and seven individuals released under overcast skies:

| <i>Sunny</i> | <i>Overcast</i> |
|--------------|-----------------|
| 350°         | 340°            |
| 340°         | 305°            |
| 315°         | 255°            |
| 10°          | 270°            |
| 20°          | 305°            |
| 355°         | 320°            |
| 345°         | 320°            |
| 360°         |                 |

Using the Watson-Williams test, test the hypothesis that the mean flight direction in this species is the same under both sunny and cloudy skies.



# 21

# *Association of Attributes*

## **21.1 ATTRIBUTES AND VARIABLES**

Statistics, in general, deals with data which can be expressed in **quantitative form**. For example, weight, height, wages, income, expenditure etc. These characteristics can be measured quantitatively and they are termed as variables. **The characteristics which are capable of being measured quantitatively are called statistics of variables.**

But there are certain phenomena which are not capable of direct quantitative measurement but we can only study the presence or absence of a particular quality or attribute in a group of individuals. For example, blindness, insanity, deafness, honesty, beauty etc. Such qualities are qualitative characteristics in nature and are called **attributes**.

It is not possible to measure the magnitude of the blindness quantitatively, but we can get it quantitatively by counting the number of persons who have this attribute and those who do not have. **Thus in a phenomena, where direct quantitative measurement is not possible but we can study only the presence or absence of a particular characteristic are called statistics of attributes.**

## **21.2 ASSOCIATION OF ATTRIBUTES**

*Association of attributes measure the degree of relationship between two phenomena whose sizes we cannot measure but we can only determine the presence or absence of a particular attribute or quality. If a relation exists between two or more attributes, they are said to be associated. The association of attributes may be positive or negative or independent.*

When we need to study the relationship between **two variables** statistical methods such as correlation, regression, dispersion etc., are applied to find out the relationship between the two variables. But if it is desired to **study the relationship between two attributes or association between two attributes, the methods of association is resorted to.**

## **21.3 CORRELATION AND ASSOCIATION**

The correlation is used to measure the degree of relationship between two phenomena which are capable of direct quantitative measurement. But the **association of attributes** is used

to measure the degree of relationship between two phenomena whose size we cannot measure but we can only determine the presence or absence of a particular attribute.

## 21.4 CLASSIFICATION OF DATA

When the data are arranged in groups according to some attribute, it is called **classification according to that attribute**. The classification is done on the basis of presence or absence of a particular attribute. When we are studying one attribute, say, Blindness then two classes shall be formed, i.e., (i) Blind and (ii) Not blind. When two attributes are studied, four classes shall be formed. When the attribute is classified into only two classes, it is called **division by dichotomy**. If the two classes are further sub-divided it is called **manifold classification**.

## 21.5 TERMS AND NOTATIONS

Capital letters **A** and **B** are used to denote the **presence of the attributes** and the Greek letters ' $\alpha$ ' and ' $\beta$ ' are used to denote the **absence of the attribute**. For example, if '**A**' represents 'Blind', then  $\alpha$  would represent 'non-blind'. Similarly if **B** represents 'Male', then  $\beta$  would mean 'Female'.

### Combination of attributes

Combination of attributes is represented by grouping together the letters A,  $\alpha$ , B,  $\beta$  such as  $AB$ ;  $A\beta$ ;  $\alpha B$ ,  $\alpha\beta$ .

### Illustration 1

- If **A** : denotes 'Male'      then       $\alpha$  : denotes 'Female'
- If **B** : denotes 'Blind'      then       $\beta$  : denotes 'Not-blind'
- $AB$  : denotes 'Male blind'
- $A\beta$  : denotes 'Male not blind'
- $\alpha B$  : denotes 'Female blind'
- $\alpha\beta$  : denotes 'Female not-blind'

The number of observations in any class is called the **frequency of the class or 'class frequency'**. Class frequencies are represented by enclosing the corresponding class symbols by small brackets '()''. Thus

- (**A**) : stand for the number of individuals who possess the attribute **A**. It is the frequency of **A**.
- (**AB**) : stands for the number of individuals who possess both attributes **A** and **B**. It is the frequency of  $AB$ ; and so on.

### Illustration 2

Using the notations of illustration 1, then

- (i) (**A**) = 30      means that there are 30 individuals who are male.
- (ii) (**AB**) = 30      means that there are 30 individuals who are both male and blind.
- (iii) ( $\alpha\beta$ ) = 10      means that there are 10 individuals. Who are both female and not blind.

## 21.6 ORDER OF CLASSES

The order of a class depends upon the number of attributes under study. A class having one attribute is known as the **class of the first order**; a class having two attributes as **class of second order** and so on.  $N$  denotes the total number of observations without any attributes. It is called the **class of zero order**. In other words,

$N$  : is a class of **Zero Order**.

$A, B, \alpha, \beta$  : are **classes of First Order**.

$AB, A\beta, \alpha B, \alpha\beta$  : are **classes of Second Order**.

Similarly, the frequencies of these classes are known as **frequencies of Zero, First and Second Order**. i.e.,

$N$  is **Frequency of the Zero order**.

$(A), (B), (\alpha), (\beta)$  : are **frequencies of the First order**.

$(AB), (A\beta), (\alpha B), (\alpha\beta)$  : are **frequencies of the Second order**.

## 21.7 NUMBER OF CLASSES

If there is **one attribute**, say  $A$ , then the total number of classes will be  $3^1 = 3$ . These classes are  $A, \alpha$  and  $N$  (the total ' $N$ ' is always considered as a class). If there are two attributes  $A$  and  $B$ , the total number of classes including  $N$  would be  $3^2 = 9$ , which are:  $N, A, \alpha, B, \beta, AB, A\alpha, \alpha B$  and  $\alpha\beta$ .

The total number of classes comprising various attributes can be determined by  $3^n$ ; where ' $n$ ' is equal to the number of attributes under study and  $n$  is a positive integer.

## 21.8 ULTIMATE CLASSES

If there are only two attributes under study, then the **second order classes and the frequencies** are called the "**classes and the frequencies of the ultimate order**", since these are the last set of classes and frequencies.

Similarly, the number of ultimate classes is determined by  $2^n$ . Thus for **one attribute**, there will be  $2^1 = 2$  ultimate classes; for **two attributes** there will be,  $2^2 = 4$  ultimate classes; and for **three attributes**, there will be  $2^3 = 8$  ultimate classes.

## 21.9 POSITIVE AND NEGATIVE CLASSES

**Positive classes** : The classes which represent the presence of an attribute or attributes are called **positive classes**.  $N, A, B$  and  $AB$  are **positive classes**.

**Negative classes** : The classes which represent the absence of an attribute or attributes are called **negative classes**.  $\alpha, \beta$  and  $\alpha\beta$  are **negative classes**.

**Pairs of contraries** : The classes in which one attribute is present and the other attribute is absent are called a **pair of contrary**.  $A\beta$  and  $\alpha B$  are **pairs of contrary classes**.

## 21.10 NINE SQUARE TABLE

The total number of observations is equal to the positive and negative frequencies of the same class of the first order, i.e.,

$$N = (A) + (\alpha)$$

Also

$$N = (B) + (\beta)$$

A convenient method of finding the frequencies is to take the help of the following nine square table.

Nine Square table

| B         | A            | $\alpha$             | N(Total)    |
|-----------|--------------|----------------------|-------------|
|           | (AB)         | ( $\alpha$ B)        | (B)         |
| $\beta$   | (A $\beta$ ) | ( $\alpha$ $\beta$ ) | ( $\beta$ ) |
| N (Total) | (A)          | ( $\alpha$ )         | N           |

From the table it is known that

$$(AB) + (\alpha B) = (B) \quad (AB) + (A\beta) = (A)$$

$$(A\beta) + (\alpha\beta) = (\beta) \quad (\alpha B) + (\alpha\beta) = (\alpha)$$

$$(A) + (\alpha) = N \quad (B) + (\beta) = N.$$

If the known values are substituted for the symbols, then remaining values can be found out by adding or subtracting.

The nine-square table can be constructed by multiplying B with A and  $\alpha$ ; also by multiplying  $\beta$  with A and  $\alpha$ .

**Example 1 :** From the following data find out the missing frequencies.

$$(AB) = 250, (A) = 600, N = 2000, (B) = 1200.$$

**Solution :** Putting these values in the nine-square table, we can find the missing value:

Nine Square Table

|         | A            | $\alpha$             | Total       |
|---------|--------------|----------------------|-------------|
| B       | (AB)         | ( $\alpha$ B)        | (B)         |
| $\beta$ | 250          | 950                  | 1200        |
| Total   | (A $\beta$ ) | ( $\alpha$ $\beta$ ) | ( $\beta$ ) |
|         | 350          | 450                  | 800         |
|         | (A)          | ( $\alpha$ )         | N           |
|         | 600          | 1400                 | 2000        |

From the table

$$(\alpha B) = (B) - (AB) = 1200 - 250 = 950$$

$$(A\beta) = (A) - (AB) = 600 - 250 = 350$$

$$(\beta) = N - (B) = 2000 - 1200 = 800$$

$$(\alpha) = N - (A) = 2000 - 600 = 1400$$

$$(\alpha\beta) = (\beta) - (A\beta) = 800 - 350 = 450$$

**Example 2 :** From the following ultimate class frequencies, find the frequencies of the positive and negative classes and the total number of observations.

$$(AB) = 100; \quad (\alpha B) = 8; \quad (A\beta) = 50; \quad (\alpha\beta) = 40$$

**Solution :** Putting these values in the nine-square table we can find out the total of positive and negative frequencies.

Nine Square Table

|         | A            | $\alpha$          | Total       |
|---------|--------------|-------------------|-------------|
| B       | (AB)         | ( $\alpha B$ )    | (B)         |
|         | 100          | 8                 | 108         |
| $\beta$ | (A $\beta$ ) | ( $\alpha\beta$ ) | ( $\beta$ ) |
|         | 50           | 40                | 90          |
| Total   | (A)          | ( $\alpha$ )      | N           |
|         | 150          | 48                | 198         |

We know that

I. Frequencies of positive classes = (A), (B) and (AB).

$$(A) = (AB) + (A\beta) = 100 + 50 = 150$$

$$(B) = (AB) + (\alpha B) = 100 + 8 = 108$$

$$(AB) = 100. \quad (\text{Given})$$

II. Frequencies of negative classes = ( $\alpha$ ), ( $\beta$ ) and ( $\alpha\beta$ ).

$$(\alpha) = (\alpha B) + (\alpha\beta) = 8 + 40 = 48$$

$$(\beta) = (A\beta) + (\alpha\beta) = 50 + 40 = 90$$

$$(\alpha\beta) = 40. \quad (\text{Given})$$

The total number of observations is equal to the positive and negative frequencies of first order of the same class.

∴ Total number of observations :  $N = (A) + (\alpha) = 150 + 48 = 198$

$$\text{Also} \quad N = (B) + (\beta) = 108 + 90 = 198.$$

## 21.11 CONSISTENCY OF DATA

When the frequencies of various classes are counted, it may be positive or zero but can not be negative. If any class-frequency is negative, then the data is said to be inconsistent. This is due to wrong counting or subtractions etc. If none of the class frequencies is negative, then we can say that the data is consistent. Thus to find out whether the given data is consistent or not, we should apply the simple test to verify whether any class-frequency is negative or not.

**Example 3 :** From the following data find out whether the data is consistent or not.

$$(A) = 40 ; (B) = 60 ; (AB) = 56 ; N = 200,$$

where  $A$  denote male and  $B$  denote literate.

**Solution :** We are given that :  $(A) = 40$  ;  $(B) = 60$  ;  $(AB) = 56$  and  $N = 200$ . Putting these values in the nine square table, we shall find the missing values. What will help us to determine whether the given data is consistent or not.

**Nine Square Table**

|         | $A$        | $\alpha$        | Total     |   |
|---------|------------|-----------------|-----------|---|
| $B$     | $(AB)$     | $(\alpha B)$    | $(B)$     | Calculation   |
|         | 56         | 4               | 60        | $(A\beta) = (A) - (AB) = 40 - 56 = -16$                 |
| $\beta$ | $(A\beta)$ | $(\alpha\beta)$ | $(\beta)$ | $(\alpha B) = (B) - (AB) = 60 - 56 = 4$                 |
|         | -16        | 156             | 140       | $(\alpha\beta) = (\alpha) - (\alpha B) = 160 - 4 = 156$ |
| Total   | $(A)$      | $(\alpha)$      | $N$       | $(\alpha) = N - (A) = 200 - 40 = 160$                   |
|         | 40         | 160             | 200       | $(\beta) = N - (B) = 200 - 60 = 140$                    |

Since one of the ultimate class-frequency, i.e.,  $(A\beta)$  is negative, so we can conclude that the data is inconsistent.

## 21.12 ASSOCIATION AND INDEPENDENCE

When it is desired to study the association between two attributes  $A$  and  $B$  it becomes necessary to find out whether the attribute  $A$  is more commonly found with attribute  $B$  than is ordinarily expected. Thus in a study of association, the first thing is to calculate the expected value of  $AB$ . This value is calculated on the basis of simple probability. The expectation of an attribute is the probability multiplied by the number of observations.

Also expectation of  $A$  :  $E(A) = (\text{Probability of } A) \times (\text{No. of observations})$

It is also called the **expected frequency of  $A$** .

For example, if we toss a coin 15 times, the probability of head will be  $1/2$  and the expectation that the head will be  $1/2 \times 15 = 7.5$ .

In the theory of attributes, if two attribute  $A$  and  $B$  are studied and the frequency of  $A$  is represented by  $(A)$  and that of  $B$  by  $(B)$ , then

$$\text{Probability of } A = \frac{(A)}{N}$$

$$\text{Probability of } B = \frac{(B)}{N}.$$

∴ **Expected frequency of  $AB$**  :  $(AB) = [\text{Probability of } A \times \text{Probability of } B] \times N$

$$= \frac{(A)}{N} \times \frac{(B)}{N} \times N = \frac{(A) \times (B)}{N} \Rightarrow (AB) = \frac{(A) \times (B)}{N}$$

Similarly, **Expected frequency of  $\alpha\beta$**  =  $\frac{(\alpha) \times (\beta)}{N}$  and so on.

## 21.13 TYPES OF ASSOCIATION

There are three types of associations.

### 1. Positive Association

When two attributes are present or absent together in the data, they are said to be **positively associated**. Such an association is found between poverty and illiteracy, sex and crime etc. *Two attributes A and B are positively associated if the observed value of AB is greater than the expected value of AB.* Symbolically

$$(AB) = \frac{(A) \times (B)}{N}$$

↓                    ↓  
 (Observed)      (Expected)

### 2. Negative Association

When the presence of one attribute causes the absence of other attribute, it is called **negative association or disassociation**. Example of such as association is cleanliness (*A*) and ill health (*B*).

*Two attributes A and B are negatively associated if the observed value of AB is less than the expected value of AB.* Symbolically,

$$(AB) = \frac{(A) \times (B)}{N}$$

↓                    ↓  
 (Observed)      (Expected)

### 3. Independent Association

When there exists no association between two attributes or when they have no tendency to be present together or the presence of one attribute does not affect the other attribute, the two attributes are said to be independent. *Two attributes A and B are said to be independent if the observed value of AB is equal to the expected value of AB.*

$$(AB) = \frac{(A) \times (B)}{N}$$

↓                    ↓  
 (Observed)      (Expected)

**Example 4 :** Show from the following data whether (*A*) and (*B*) are independent, positively associated or negatively associated:

1.  $N = 400$  ;  $(A) = 50$  ;  $(B) = 160$  ;  $(AB) = 25$
2.  $N = 800$  ;  $(A) = 160$  ;  $(B) = 300$  ;  $(AB) = 50$
3.  $N = 200$  ;  $(A) = 30$  ;  $(B) = 100$  ;  $(AB) = 15$

**Solution :**

$$1. \text{ Expected frequency of } AB = \frac{(A) \times (B)}{N} = \frac{50 \times 160}{400} = 20.$$

In this case the **expected frequency is 20** and the actual frequency  $AB$  is 25, which is greater than the expected frequency. Thus the **attributes  $A$  and  $B$  are positive associated.**

$$2. \text{ Expected frequency of } AB = \frac{(A) \times (B)}{N} = \frac{160 \times 300}{800} = 60.$$

In this case the actual frequency is 50 which is less than the expected frequency. Thus the **attributes  $A$  and  $B$  are negatively associated.**

$$3. \text{ Expected frequency of } AB = \frac{(A) \times (B)}{N} = \frac{30 \times 100}{200} = 15.$$

In this case the expected frequency of  $AB$  = Actual frequency  $AB$  = 15. Thus the **attributes  $A$  and  $B$  are independent.**

## 21.14 METHODS OF DETERMINING ASSOCIATION

Association can be studied by any one of the following methods:

1. Comparison of Observed and Expected Frequencies or Frequency Method.
2. Comparison of Proportions or Proportion Method.
3. Yule's Coefficient of Association.
4. Yule's Coefficient of Colligation.

## 21.15 COMPARISON OF OBSERVED AND EXPECTED FREQUENCIES OR FREQUENCY METHOD

In order to examine the nature of association between two attributes, we have to compare the observed frequency with the expected frequency.

*Two attributes  $A$  and  $B$  are positively associated if their observed frequency is more than the expected frequency; if the observed observation is less than the expected frequency, then they are negatively associated: if the observed frequency is equal to the expected frequency, then they are independent.*

The same is true for attributes  $\alpha$  and  $B$ ;  $\alpha$  and  $\beta$ ; and  $A$  and  $\beta$ . These relations are presented in a table given below.

Table : 21.1 : Frequency Method

| Attribute            | Independent   | Positive Association                                | Negative Association                                |
|----------------------|---|---|---|
| $A$ and $B$          | $(AB) = \frac{(A) \times (B)}{N}$                   | $(AB) > \frac{(A) \times (B)}{N}$                   | $(AB) < \frac{(A) \times (B)}{N}$                   |
| $\alpha$ and $\beta$ | $(\alpha\beta) = \frac{(\alpha) \times (\beta)}{N}$ | $(\alpha\beta) > \frac{(\alpha) \times (\beta)}{N}$ | $(\alpha\beta) < \frac{(\alpha) \times (\beta)}{N}$ |
| $A$ and $\beta$      | $(A\beta) = \frac{(A) \times (\beta)}{N}$           | $(A\beta) > \frac{(A) \times (\beta)}{N}$           | $(A\beta) < \frac{(A) \times (\beta)}{N}$           |
| $\alpha$ and $B$     | $(\alpha B) = \frac{(\alpha) \times (B)}{N}$        | $(\alpha B) > \frac{(\alpha) \times (B)}{N}$        | $(\alpha B) < \frac{(\alpha) \times (B)}{N}$        |

**Limitation**

When we study the association by Frequency Method or Comparison of Actual and Expected Frequencies, we can only determine the nature of association (*i.e.*, positive, negative or no association). But it does not tell us about the degree of association (*i.e.*, high or low).

**Example 5 :** From the following, find whether blindness and baldness are associated:

$$\text{Total population} = 16264000 ; \quad \text{Number of bald headed} = 24441$$

$$\text{Number of blind} = 7623 ; \quad \text{Number of bald headed blind} = 221$$

**Solution :** Let 'A' denote bald headed and let 'B' denote for blindness. We are given that:

$$N = 1,62,64,000 ; (A) = 24,441 ; (B) = 7,623 ; (AB) = 921$$

$$\text{Expected Frequency of } AB = \frac{(A) \times (B)}{N} = \frac{24,441 \times 7693100}{1,62,64,000} = 11.56$$

$$\text{Observed Frequency : } (AB) = 921 \quad [\text{Given}]$$

$$\therefore (AB) > \frac{(A) \times (B)}{N} \text{ as } 921 > 11.56$$

Hence there is a positive association between bald headed and blind persons.

**Example 6 :** The male population of a particular place is 250. The number of literate males is 100 and total number of male criminals is 20. The number of literate male criminals is 5. Do you find any association between literacy and criminality?

**Solution :** Let  $A$  denotes male literate ; Let  $B$  denotes male criminals.

Let  $AB$  denotes literate male criminals.

Thus, the given frequencies are:  $(A) = 100$  ;  $(B) = 20$  ;  $(AB) = 5$  ;  $N = 250$ .

**Expected frequency of male criminals :**

$$= \frac{(A) \times (B)}{N} = \frac{100 \times 20}{250} = 8.$$

**Observed or Actual frequency of  $AB$  :  $(AB) = 5$**  (Given)

*Since the actual frequency (5) is less than the expected frequency (8), therefore, the attributes are negatively associated, i.e., literacy checks criminality.*

**Example 7 : Find whether the attributes  $A$  and  $B$  are independent from the data given below:**

$$(A) = 470; \quad (B) = 620; \quad (AB) = 320; \quad N = 1000.$$

**Solution :** Attributes  $A$  and  $B$  shall be independent if observed frequency = expected frequency.

$$\Rightarrow (AB) = \frac{(A) \times (B)}{N}$$

**Actual frequency of  $AB$  :  $(AB) = 320$ .**

$$\text{Expected frequency of } AB = \frac{(A) \times (B)}{N} = \frac{470 \times 620}{1000} = 291.4$$

$$\text{Since, } (AB) > \frac{(A) \times (B)}{N}, \quad \text{i.e., } 320 > 291.4,$$

therefore, we can conclude that the attributes are positively associated.

## 21.16 COMPARISON OF PROPORTIONS OR PROPORTION METHOD

Under this method, the ratios or the proportions of the concerned variables are compared. If there is no relationship between two attributes  $A$  and  $B$  we expect to find the same proportion of  $A$ 's amongst the  $B$ 's as amongst the  $\beta$ 's. For example, if blindness and deafness are independent, the proportion of the blind people among the deafs and non-deafs must be the same.

The relationship is given below:

**Table 21.2 : Proportion Method**

| Association | Proportion of $B$ in $A$ and $\alpha$            | Proportion of $A$ in $B$ and $\beta$          |
|-------------|--|---|
| Independent | $\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$ | $\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$ |
| Positive    | $\frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$ | $\frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)}$ |
| Negative    | $\frac{(AB)}{(A)} < \frac{(\alpha B)}{(\alpha)}$ | $\frac{(AB)}{(B)} < \frac{(A\beta)}{(\beta)}$ |

**Limitation :** This method only enable us to determine the nature of association but not the degree of association.

**Example 8 : In an anti-biotic campaign in a certain area, quinine was administered to 850 persons out of a total population of 3500. The number of fever cases is shown below:**

| Treatment  | Fever | No fever |
|------------|-------|----------|
| Quinine    | 50    | 800      |
| No quinine | 250   | 2400     |

Discuss the usefulness of quinine in checking malaria.

**Solution :** Let 'A' denote quinine treatment ; Let 'α' denote No quinine.

Let 'B' denote attacked by fever ; Let 'β' denote No fever.

We should prepare the nine square as given below:

Nine Square Table

|       | A    | α    | Total |
|-------|------|------|-------|
| B     | (AB) | (αB) | (B)   |
|       | 50   | 250  | 300   |
| β     | (Aβ) | (αβ) | (β)   |
|       | 800  | 2400 | 3200  |
| Total | (A)  | (α)  | N     |
|       | 850  | 2650 | 3500  |

**Explanation :**

$$\text{Total population } N = 3500$$

$$\text{Quinine treatment (A)} = 850$$

$$\text{No-quinine treatment (α)} = 3500 - 850 = 2650$$

$$\text{Fever attacked (B)} = 300$$

$$\text{No fever (β)} = (3500 - 300) = 3200$$

$$\text{Quinine treatment and fever (AB)} = 50$$

$$\text{Quinine treatment and no fever (Aβ)} = 800.$$

$$\text{No-quinine treatment and fever (αB)} = 250$$

$$\text{No-quinine treatment and no fever (αβ)} = 2400.$$

**Proportion of quinine treatment cases which were attacked by**

$$\text{Fever} = \frac{(AB)}{(A)} = \frac{50}{850} = 0.059 \text{ or } 5.9\%$$

**Proportion of no-quinine treatment cases which were attacked by**

$$\text{Fever} = \frac{(\alpha B)}{(\alpha)} = \frac{250}{2650} = 0.93 \text{ or } 9.3\%.$$

Since, the proportion of quinine treatment cases which were attacked by fever is less than the proportion of no-quinine treatment case which were attacked by fever, so the attributes are negatively associated, i.e., the quinine is effective in checking malaria.

**Example 9 :** Can vaccination be regarded as a preventive measure for Small Pox from the data given below?

(i) Of 2,000 persons in a locality exposed to Small Pox, 450 in all were attacked.

(ii) If 2000 persons, 365 had been vaccinated; of these only 50 were attacked.

**Solution :** Let 'A' denote vaccinated, them 'α' denote not vaccinated.

Let 'B' denote exempted from small pox ; then 'β' denote attack of small pox. We are given that:  $N = 2000$ ;  $(\beta) = 450$ ;  $(A) = 365$ .

The missing values can be obtained from the following nine square table.

Nine Square Table

|       | A    | α    | Total |
|-------|------|------|-------|
| B     | (AB) | (αB) | (B)   |
|       | 315  | 1235 | 1550  |
| β     | (Aβ) | (αβ) | (β)   |
|       | 50   | 400  | 450   |
| Total | (A)  | (α)  | N     |
|       | 365  | 1635 | 2,000 |

$$\text{Percentage of vaccinated, who were not attacked} = \frac{(AB)}{(A)} = \frac{315 \times 100}{365} = 86.3\%.$$

$$\text{Percentage of not vaccinated, but not attacked} = \frac{(\alpha B)}{(\alpha)} = \frac{1235}{1635} \times 100 = 73.2\%.$$

Thus  $\frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$  and therefore, they positively associated. Hence, vaccination is a

good preventive measure for Small-Pox.

**Example 10 :** Out of 3,000 unskilled workers of a factory, 2,000 come from rural areas and out of 1200 skilled workers, 300 come from rural areas. Determine the association between skill and residence by the method of proportions.

**Solution :** Let A denote skilled workers ; Let α denote unskilled workers.

Let B denote workers from rural areas ; Let β denote workers from urban areas.

We are given that:  $(A) = 1200$ ;  $(\alpha) = 3000$ ;  $(B) = 2000$ ;  $(\alpha B) = 2000$ ;  $(AB) = 300$ .

$$\frac{(AB)}{(A)} = \frac{300}{1200} = 0.25. \quad \text{Also} \quad \frac{(\alpha B)}{(\alpha)} = \frac{2000}{3000} = 0.67.$$

$$\text{Now, } \frac{(\alpha B)}{(\alpha)} > \frac{(AB)}{(A)}, \quad \text{as } 0.67 > 0.25.$$

Thus there is negative association between skill and residence.

**Example 11 :** Out of 70,000 of literates in a particular district of India, the number of criminals was 500. Out of 9,30,000 of illiterates in the same area, number of criminals were 15,000. On the basis of these figures, do you find any association between illiteracy and criminality?

**Solution :** Let ' $A$ ' denote illiteracy ; then ' $\alpha$ ' denote literacy.

Let ' $B$ ' denote criminals ; then ' $\beta$ ' denote non-criminals.

We are given that:  $(A) = 9,30,000$  ;  $(\alpha) = 70,000$  ;  $(AB) = 15,000$  ;  $(\alpha B) = 500$ .

$$\begin{aligned}\text{Proportion of criminals illiterates} &= \frac{(AB)}{(A)} \\ &= \frac{15,000}{9,30,000} = 0.016 \text{ or } 1.6\%.\end{aligned}$$

$$\begin{aligned}\text{Proportion of criminals to literates} &= \frac{(\alpha B)}{(\alpha)} \\ &= \frac{500}{70,000} = 0.0071 \text{ or } 0.71\%.\end{aligned}$$

Since the proportion of criminals to illiterates is more than proportion of criminals to literates, so the attributes are positively associated.

## 21.17 YULE'S CO-EFFICIENT OF ASSOCIATION

This is the most popular method, because here we not only determine the nature of association but also measure the degree of association between the two attributes  $A$  and  $B$ . To measure the intensity of association, Andrew Yule has given the formula of co-efficient of association denoted by the symbol  $Q$ , as follows:

$$\text{Yule's coefficient of Association : } Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}.$$

**Limitations for  $Q$  :**

1. *Yule's coefficient of association 'Q' lies between -1 and 1, both inclusive i.e.,  $-1 \leq Q \leq 1$ .*
  - (i)  *$Q = 1$ , implies that there is a perfect positive association between A and B.*
  - (ii)  *$Q = -1$ , implies that there a perfect negative association between A and B.*
  - (iii)  *$Q = 0$ , implies that A and B are independent.*
2. *Yule's coefficient of association 'Q' is independent of the relative proportion of A's or  $\alpha$ 's in the data.*

**Example 12 :** For two attributes  $A$  and  $B$ , we have  $(AB) = 16$ ,  $(A) = 36$ ;  $(\alpha\beta) = 10$  and  $N = 70$ . Calculate co-efficient of association.

**Solution :** Let us prepare nine-square table.

Nine Square Table

|          | <i>A</i>   | $\alpha$        | Total     |
|----------|------------|-----------------|-----------|
| <i>B</i> | $(AB)$     | $(\alpha B)$    | $(B)$     |
|          | 16         | 24              | 40        |
| $\beta$  | $(A\beta)$ | $(\alpha\beta)$ | $(\beta)$ |
|          | 20         | 10              | 30        |
| Total    | $(A)$      | $(\alpha)$      | $N$       |
|          | 36         | 34              | 70        |

**Yule's Co-efficient of Association:**

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{16 \times 10 - 20 \times 24}{16 \times 10 + 20 \times 24} = \frac{160 - 480}{160 + 480} = \frac{-320}{640} = -0.5.$$

**Example 13 :** Investigate the association between eye colour of husbands and eye colour of wives from the data given below:

Husbands with light eyes and wives with light eyes = 309

Husbands with light eyes and wives with not-light eyes = 214

Husbands with not-light eyes and wives with light eyes = 132

Husbands with not-light eyes and wives not-light eyes = 119

**Solution :** Let 'A' denote husbands with light eyes; then 'α' denote husbands with not light eyes.

Let 'B' denote wives with light eyes; then 'β' denote wives with not light eyes.

Thus the given data in terms of symbols is:

$$(AB) = 309 ; (A\beta) = 214 ; (\alpha B) = 132 ; (\alpha\beta) = 119.$$

Using Yule's Co-efficient of Association Method, we have:

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{309 \times 119 - 214 \times 132}{309 \times 119 + 214 \times 132} = \frac{8523}{65019} = 0.131.$$

Thus  $Q = 0.131$  shows that there is a little positive association between the eye colour of husbands and eye colour of wives.

**Example 14 :** Out of 400 candidates appeared for a competitive examinations, the number of candidates succeeded was 120. Out of 70 candidates who received special coaching, 40 candidates were succeeded. On the basis of these figures (a) Do you find any association between success and coaching classes (b) also using Yule's coefficient, discuss whether special coaching is effective or not.

**Solution :** Let 'A' denote candidates who attended coaching class; then 'α' denote the candidates who do not attend coaching class.

Let 'B' denote successful candidates; then 'β' denote unsuccessful candidates.

Nine Square Table

|          | <i>A</i>            | <i>α</i>             | Total               |
|----------|---------------------|----------------------|---------------------|
| <i>B</i> | ( <i>AB</i> )<br>40 | ( <i>αB</i> )<br>30  | ( <i>B</i> )<br>70  |
| <i>β</i> | ( <i>Aβ</i> )<br>80 | ( <i>αβ</i> )<br>250 | ( <i>β</i> )<br>330 |
| Total    | ( <i>A</i> )<br>120 | ( <i>α</i> )<br>280  | <i>N</i><br>400     |

(a) The proportion of specially coached and successful candidates among the successful students

$$\frac{(\text{AB})}{(A)} = \frac{40}{120} = 0.33 \text{ or } 33.33\%.$$

The proportion of not specially coached and successful students amongst the unsuccessful students

$$\frac{(\alpha B)}{(\alpha)} = \frac{30}{280} = 0.107 \text{ or } 10.75.$$

Since the proportion of specially coached and successful candidates is more than the proportion of not coached and failed candidates, so the attributes coaching and success are positively less associated.

(b) Using Yule's Co-efficient of Association, we have:

$$Q = \frac{(\text{AB})(\alpha\beta) - (\text{A}\beta)(\alpha\text{B})}{(\text{AB})(\alpha\beta) + (\text{A}\beta)(\alpha\text{B})} = \frac{40 \times 250 - 80 \times 30}{40 \times 250 + 80 \times 30} = \frac{10000 - 2400}{10000 + 2400} = \frac{7600}{12400} = 0.613.$$

This shows that there is a high degree of positive association. We conclude that the special coaching class was effective.

## 21.18 YULE'S COEFFICIENT OF COLLIGATION

Another method to calculate the coefficient of association given by Yule is known as Coefficient of Colligation ( $\gamma$ ). But  $Q$  is more popular than  $\gamma$ .

$$\text{Coefficient of Colligation } (\gamma) = \frac{1 - \sqrt{\frac{(\text{A}\beta)(\alpha\text{B})}{(\text{AB})(\alpha\beta)}}}{1 + \sqrt{\frac{(\text{A}\beta)(\alpha\text{B})}{(\text{AB})(\alpha\beta)}}}$$

or

$$\gamma = \frac{\sqrt{(\text{AB})(\alpha\beta)} - \sqrt{(\text{A}\beta)(\alpha\text{B})}}{\sqrt{(\text{AB})(\alpha\beta)} + \sqrt{(\text{A}\beta)(\alpha\text{B})}}$$

$$\text{Coefficient of Association : } (Q) = \frac{2\gamma}{1 + \gamma^2}$$

**Example 15 :** From the following data, calculate the coefficient of colligation:

$$(AB) = 6; (\alpha B) = 13; (A\beta) = 8; (\alpha\beta) = 3$$

**Solution :**

$$\gamma = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

$$= \frac{1 - \sqrt{\frac{8 \times 13}{6 \times 3}}}{1 + \sqrt{\frac{8 \times 13}{6 \times 3}}} = \frac{1 - \sqrt{5.8}}{1 + \sqrt{5.8}} = \frac{1 - 2.41}{1 + 2.41} = \frac{-1.41}{3.41} = -0.414.$$

**Example 16 :** Do you find any association between the tempers of brothers and sisters from the following data?

$$\text{Good-natured brothers and good-natured sisters} = 1040$$

$$\text{Good-natured brothers and sullen sisters} = 160$$

$$\text{Sullen brothers and good-natured sisters} = 180$$

$$\text{Sullen brothers and sullen sisters} = 120$$

**Solution :**

**First Method:**

Let 'A' denote good-natured brothers; then 'α' denote sullen brothers.

Let 'B' denote good-natured sisters; then 'β' denote sullen sisters.

It is given that:  $(AB) = 1040$ ;  $(A\beta) = 160$ ;  $(\alpha B) = 180$ ;  $(\alpha\beta) = 120$

**Yule's Coefficient of Association:**

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{1040 \times 120 - 160 \times 180}{1040 \times 120 + 160 \times 180}$$

$$= \frac{124800 - 28800}{124800 + 28800} = \frac{96000}{153600} = 0.625.$$

Hence, there is positive association between the tempers of brothers and sisters.

**Second Method:**

**Coefficient of colligation:**  $Y = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}} = \frac{1 - \sqrt{\frac{160 \times 180}{1040 \times 120}}}{1 + \sqrt{\frac{160 \times 180}{1040 \times 120}}}$

$$= \frac{1 - \sqrt{\frac{28,800}{124800}}}{1 + \sqrt{\frac{28,800}{124800}}} = \frac{1 - 0.48}{1 + 0.48} = \frac{0.52}{1.48} = 0.3513.$$

$$\therefore Q = \frac{2\gamma}{1 + \gamma^2}$$

$$= \frac{2 \times 0.3513}{1 + 0.3513^2} = \frac{0.7026}{1.1234} = 0.625.$$

**Example 17 :** 88 residents of posh colony in New Delhi, who were interviewed during a sample survey, are classified below according to their smoking and tea drinking habits. Calculate Yule's coefficient of Association and comment on its value.

|                  | Smokers | Non-smokers |
|------------------|---------|-------------|
| Tea Drinkers     | 40      | 33          |
| Non-Tea drinkers | 3       | 12          |

**Solution :** Let 'A' denote smokers and 'α' denote non-smokers.

Let 'B' denote tea drinkers and 'β' denote non-tea drinkers.

The given data in terms of these symbols are:

(AB) = 40, i.e., Number of smokers and tea drinkers.

(Aβ) = 3, i.e., Number of smokers and non-tea drinkers.

(αB) = 33, i.e., Number of tea drinkers and non-smokers.

(αβ) = 12, i.e., Number of non-smokers and non-tea drinkers.

**Yule's Coefficient of Correlation:**

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{(40 \times 12) - (3 \times 33)}{(40 \times 12) + (3 \times 33)} = \frac{480 - 99}{480 + 99} = \frac{381}{579} = 0.658.$$

Here, there is a association between the drinking and smoking.

**Example 18 :** From the table compare the intensity of association between literacy and unemployment among the males in urban areas with that in rural areas:

|       | Total        | Literate | Unemployed | Literate unemployed |
|-------|--------------|----------|------------|---------------------|
| Urban | ('00000) 25  | 10       | 5          | 3                   |
| Rural | ('00000) 200 | 40       | 12         | 4                   |

**Solution :** Let 'A' denote literate; then 'α' denote illiterate.

Let 'B' denote unemployment; then 'β' denote employment.

Given values are:

| URBAN |     |     | RURAL |     |     |       |
|-------|-----|-----|-------|-----|-----|-------|
|       | (A) | (α) | Total | (A) | (α) | Total |
| (B)   | 3   |     | 5     | 4   |     | 12    |
| (β)   |     |     |       |     |     |       |
| Total | 10  |     | 25    | 40  |     | 200   |
|       | N   |     |       | N   |     |       |

Completing the nine-square tables, we have:

| URBAN |     |     | RURAL |     |     |       |
|-------|-----|-----|-------|-----|-----|-------|
|       | (A) | (α) | Total | (A) | (α) | Total |
| (B)   | 3   | 2   | 5     | 4   | 8   | 12    |
| (β)   | 7   | 13  | 20    | 36  | 152 | 188   |
| Total | 10  | 15  | 25    | 40  | 160 | 200   |
|       | N   |     |       | N   |     |       |

$$\begin{aligned} \text{Urban } Q_U &= \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \\ &= \frac{(3 \times 13) - (7 \times 2)}{(3 \times 13) + (7 \times 2)} \\ &= \frac{39 - 14}{39 + 14} = \frac{25}{53} = 0.47. \end{aligned}$$

$$\begin{aligned} \text{Rural } Q_R &= \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \\ &= \frac{(4 \times 152) - (36 \times 8)}{(4 \times 152) + (36 \times 8)} \\ &= \frac{608 - 288}{608 + 288} = \frac{320}{896} = 0.36. \end{aligned}$$

As  $Q_U > Q_R$ , so the intensity of association between literate and unemployment is more in urban areas than in rural areas.

## EXERCISE

1. Find the missing frequencies from the following data:

$$(A) = 400; (AB) = 250; (B) = 500; N = 1200.$$

$$[\text{Hint : } (A\beta) = (A) - (AB) = 400 - 250 = 150; (\alpha) = N - (A) = 1200 - 400 = 800]$$

$$(\alpha B) = (B) - (AB) = 500 - 250 = 250; (\alpha\beta) = (\alpha) - (\alpha B) = 800 - 250 = 550;$$

$$(\beta) = N - (B) = 1200 - 500 = 700]$$

2. Is there any inconsistency in the following data:  $(AB) = 200$ ;  $N = 1000$ ;  $(A) = 150$ ;  $(\beta) = 300$ .

$$[\text{Hint : } (A\beta) = (A) - (AB) = 150 - 200 = -50]$$

From the following data, find out whether the attributes  $A$  and  $B$  are positively associated or negatively associated or independent.

3.  $N = 400$ ;  $(A) = 60$ ;  $(B) = 200$ ;  $(AB) = 30$ .

[Hint : 1. Expected frequency of  $AB = \frac{(A) \times (B)}{N} = \frac{60 \times 200}{400} = 30$ . Actual frequency of  $(AB) = 30$ ; As actual frequency (30) is equal to expected frequency (30), the attributes  $(A)$  and  $(B)$  are independent.]

4.  $N = 96$ ;  $(A) = 10$ ;  $(B) = 32$ ;  $(AB) = 5$ .

[Hint : Actual frequency  $(AB) = 5$ .

$$\text{Expected frequency of } AB = \frac{(A) \times (B)}{N} = \frac{10 \times 32}{96} = 3.33.$$

As the actual frequency (5) is more than the expected frequency (3.33), the attributes  $A$  and  $B$  are positively associated.]

5.  $N = 160$ ;  $(A) = 32$ ;  $(B) = 60$ ;  $(AB) = 10$ .

6. Out of 6000 unskilled workers in a factory 4000 come from rural areas and out of 2400 skilled workers 600 come from rural areas. Determine the association between skill and residence in rural areas by the method of proportion.

[Hint : Let  $A$  denote skilled workers and  $\alpha$  unskilled workers.

Let  $B$  denote rural areas and  $\beta$  non-rural areas.

Then  $(A) = 2400$ ;  $(AB) = 600$ ;  $(\alpha) = 6000$ ;  $(\alpha B) = 4000$ .

The proportion of skilled rural workers amongst the skilled workers

$$\frac{(AB)}{(A)} = \frac{600}{2400} = 0.25 \text{ or } 25\%.$$

The proportion of un-skilled rural workers amongst the unskilled workers:

$$\frac{(\alpha B)}{(\alpha)} = \frac{4000}{6000} = 0.75 \text{ or } 75\%.$$

Since the proportion of skilled rural workers amongst the skilled workers is less than the proportion of unskilled rural workers amongst the unskilled workers, the attributes are negatively associated i.e., skilled workers come more from non-rural areas.]

7. Can inoculation be regarded as a preventive measure for cholera from the data given below:

(i) Out of 2000 persons in a locality exposed to cholera, 216 in all were attacked.

(ii) Out of 500 persons inoculated only 31 were attacked.

[Hint : Let ' $A$ ' denote inoculated; then  $\alpha$  denote not-inoculated.

Let ' $B$ ' denote attack of cholera; then ' $\beta$ ' denote not-attacked of cholera.

Complete the nine square table to get:  $(A) = 500$ ,  $(B) = 216$ ;  $N = 2000$ ,  $(AB) = 31$  (Actual frequency)

$$\text{Expected frequency of } AB = \frac{(A) \times (B)}{N} = \frac{500 \times 216}{2000} = 54.$$

Since the actual frequency (31) is less than the expected frequency (54), we may conclude that there exists a negative correlation, i.e., inoculation may be regarded as a preventive measure for cholera.]

8. Calculate Yule's coefficient of association and interpret the result.

$$N = 1500, (\alpha) = 1117, (B) = 360, (AB) = 35.$$

[Hint :  $(\alpha B) = (B) - (AB) = 360 - 35 = 325$ ;  $(\alpha \beta) = (\alpha) - (\alpha B) = 1117 - 325 = 792$ ;  
 $(A \beta) = (\beta) - (\alpha \beta) = 1140 - 792 = 348$ .]

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{(35 \times 792) - (348 \times 325)}{(35 \times 792) + (348 \times 325)} = -0.606.]$$

9. Test for association between extravagance in fathers and in sons, from the following data:

Extravagant fathers with extravagant sons = 347

Miserly fathers with extravagant sons = 741

Extravagant fathers with miserly sons = 545

Miserly fathers with miserly sons = 234

[Hint : Let 'A' extravagant fathers; then 'α' miserly fathers.

Let 'B' extravagant sons; then 'β' miserly sons.

Then  $(AB) = 327$ ,  $(\alpha\beta) = 234$ ,  $(\alpha B) = 741$ ;  $(A\beta) = 545$ .

$$\therefore Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{327 \times 234 - 545 \times 741}{327 \times 234 + 545 \times 741} = -0.681.]$$

10. From the following data, prepare nine-square table. Using Yule's coefficient of association, discuss whether there is association between literacy and unemployment.

Illiterate unemployed = 220 persons; Literate employed = 20 persons;

Illiterate employed = 180 persons; Total number of persons = 500

[Hint : Let 'A' = literacy; 'α' = illiteracy; Let 'B' = unemployment, 'β' = employment.

|            |             |            |
|------------|-------------|------------|
| 80<br>(AB) | 220<br>(αB) | 300<br>(B) |
| 20<br>(Aβ) | 180<br>(αβ) | 200<br>(β) |
| 100<br>(A) | 400<br>(α)  | 500<br>(N) |

$$\begin{aligned} Q &= \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \\ &= \frac{80 \times 180 - 20 \times 220}{80 \times 180 + 20 \times 220} \\ &= \frac{14400 - 4400}{14400 + 4400} = \frac{10,000}{18,800} = 0.532.] \end{aligned}$$

11. Calculate the coefficient of association between intelligence in father and son from the following data:

Intelligent fathers with intelligent sons = 248; Intelligent fathers with dull sons = 81

Dull fathers with intelligent sons = 92; Dull fathers with dull sons = 579.

12. From the data given below, calculate Yule's coefficient of association between weight of children and their economic condition and interpret it.

|  | <i>Poor Children</i> | <i>Rich Children</i> |
|--|----------------------|----------------------|
|--|----------------------|----------------------|

|                     |    |    |
|---------------------|----|----|
| Below normal weight | 75 | 23 |
| Above normal weight | 5  | 42 |

13. Out of 3000 unskilled workers of a factory, 2000 come from rural areas and out of 1200 skilled workers, 300 come from rural areas. Determine the association between skill and residence by the method of proportions.
14. From the data given below test whether there is association between economic status and economic achievement:

|            | <i>Rich</i> | <i>Poor</i> |
|------------|-------------|-------------|
| Educated   | 508         | 1559        |
| Uneducated | 905         | 1114        |

15. According to a survey the following results were obtained:

|  | <i>Boys</i> | <i>Girls</i> |
|--|-------------|--------------|
| No. of candidates appeared at an examination | 800         | 200          |
| Married                                      | 150         | 50           |
| Married and successful                       | 70          | 20           |
| Unmarried and successful                     | 550         | 110          |

Find the association between marital status and the success at the examination both for boys and girls.

16. In an assortive matching study to find whether tall husbands tend to marry tall wives, the following information about the wives of 125 tall and 125 short husbands was published.

|             | <i>Tall Husbands</i> | <i>Short Husbands</i> |
|-------------|----------------------|-----------------------|
| Tall wives  | 56%                  | 13%                   |
| Short wives | 11%                  | 48%                   |

Find the coefficient of association between the stature of wives and husbands ignoring medium sized wives.

[Hint: Let  $A$  and  $B$  denote tall husbands and tall wives respectively.

$\alpha$  and  $\beta$  denote short husbands and short wives respectively.

$$(AB) = 56; (A\beta) = 11; (\alpha B) = 13; (\alpha\beta) = 48.$$

$$\therefore Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{56 \times 48 - 11 \times 13}{56 \times 48 + 11 \times 13} = 0.9.]$$

**ANSWERS**

1.  $(A\beta) = 150$ ;  $(\alpha B) = 250$ ;  $(\alpha\beta) = 550$ ;  $(\alpha) = 800$ ;  $(\beta) = 700$ .
2. The given data is inconsistent as  $(A\beta)$  is negative;  $(A\beta) = -50$ .
3. Attributes  $A$  and  $B$  are independent.
4. Attributes  $A$  and  $B$  are positively associated.
5. Attributes  $A$  and  $B$  are negatively associated.
6. Skilled workers come more from non-rural areas.
7. Inoculation may be regarded as a preventive measure for cholera.
8.  $Q = -0.606$ ; there is a negative association between the attributes  $A$  and  $B$ .
9.  $Q = -0.681$ .
10.  $Q = 0.532$ ; There is a positive association between literacy and unemployment.
11.  $Q = 0.9$ .
12.  $Q = 0.93$ . There is a high degree of positive association.
13. There is a negative association.
14.  $Q = -0.43$ .
15.  $Q_{\text{Boys}} = -0.73$ ;  $Q_{\text{Girls}} = -0.61$ . There is a negative association between marital status and success.



# 22

# *Models of Data Presentation with Special Reference to Biological Samples*

## **22.1 INTRODUCTION**

Biostatistics can refer to different applied mathematics and quantitative models in several different area of application.

Design and analysis of clinical trials is perhaps the most publicly visible application of statistics in medicine. Statistical genetics in populations is another applied area that is closely allied to biostatistics. This analysis attempts to link variation in genotype with a variation in phenotype. This has been used in agriculture to improve crops and farm animals. In biomedical research, this work can assist in finding candidates for gene alleles that can cause or influence predisposition to disease in human genetics.

Ecology, biological sequence analysis, and epidemiology are among other diverse fields that have built upon strong biostatistical components. Statistical methods are beginning to be integrated into medical informatics and bioinformatics.

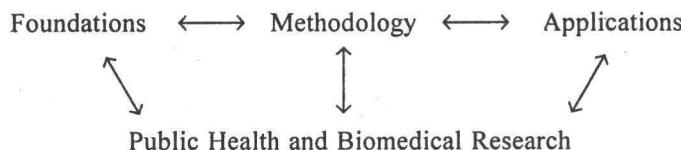
**Spatial modelling** is increasingly prominent in the biological sciences as scientists attempt to characterize variability of processes that are spatially indexed. It shows that the mixed model framework is useful for characterizing spatial statistical methodology. In particular, the classical geostatistical approach known as kriging can be cast as a linear mixed model. Furthermore, the generalized linear mixed model provides a natural framework for extending the methodology to allow modelling of non-Gaussian spatial processes. The mixed model framework is also useful for describing multivariate spatial models and many spatiotemporal models.

## **22.2 HOPKINS PERSPECTIVE (MODEL) ON BIOSTATISTICS**

Biostatistics comprises the reasoning and methods for using data as evidence to address public health and biomedical questions. It is an approach and set of tools for designing studies and for quantifying the resulting evidence, for quantifying what we believe, and for making decisions.

At Johns Hopkins Department of Biostatistics, research is characterized by a commitment to statistical science, its foundations and methods, as well as the application of statistical science

to the solution of public health and biomedical problems. As indicated in the two-way arrows in figure 1, research on foundations informs and is informed by methods research, which in turn benefits and is benefited by statistical applications. To be excellent, biostatistical research must be built on a foundation of first-rate public health and biomedical research.



*Fig. 22.1 : Biostatistics Research Model*

Research on foundations has as its goal the development of better strategies, or ways of reasoning, for empirical research. For example, past chair William Cochran demonstrated how observational studies can be used to draw inferences about the causal effect of a treatment on a health outcome. Jerry Cornfield showed how case control studies can be used to draw valid inferences about parameters in prospective models. Today, Richard Royall is leading a transition in statistical reasoning from decision methods (*p*-values, tests of hypothesis) toward likelihood methods, which quantify scientific evidence.

Research on statistical methodology has as its goal the creation of new strategies for drawing inferences from data. To illustrate, Ron Brookmeyer and Mitch Gail developed the methodology used to monitor and project the size of the US AIDS epidemic; Kung-Yee Liang, Mei-Cheng Wang, and Scott Zeger developed methods for regression analysis with correlated responses.

## 22.3 VARIOUS FIELDS FOR BIOSTATISTICAL MODELS

1. Statistical genetics/genomics/bioinformatics.
2. Integrated analyses of data from longitudinal studies: joint analyses of multivariate
  - Repeated measures
  - Times-to-events
3. Early detection of disease processes – biomarkers
4. Teaching statistical reasoning and methods to health scientists and professionals
5. Non-invasive measurement systems.
  - Biomonitoring: time series on many persons
  - Imaging
  - Gene expression arrays
  - Proteomics
  - Complex questionnaires – latent variable models
6. Internet-based data collection, management, measurement and analysis
7. Post-marketing surveillance of drug treatments
8. Clinical trials
  - Summarizing evidence

- Measuring treatment effects with : partial compliance; drop-outs; treatment efficacy for subgroups
  - Combining evidence from many trials
9. Causal inference from observational and experimental studies
- Statistical models with bias terms
  - Quantifying uncertainty beyond sampling variation
10. Transition to evidence-based statistics
11. Quality assurance for laboratory research
- Pooling data
  - Variance components models
12. Environmental epidemiology
- Time and space risk models
  - Measurement error

## 22.4 COVARIANCE MODELS

The essence of spatial statistics is spatial correlation, and consequently it is important to model this aspect of the problem adequately. Unfortunately, there are many limitations (having to do with both data and covariance models) which make this a difficult task. To guarantee that the covariance matrix is positive definite, the spatial covariance matrix  $\Sigma_\alpha$  is assumed to be of some parametric form, indexed by the parameter  $\theta$  (possibly a vector). To be more precise, the spatial covariance matrix is expressed as  $\Sigma_\alpha(\theta)$ . Much of the detail concerning *implementation* of contemporary spatial statistics focuses on the choice of the covariance function and estimation of its parameters.

The covariance function describes the spatial association between the random effect at any two locations in space, say  $s$  and  $s'$ :

$$\text{cov} [\alpha(s), \alpha(s')] = c_\alpha(s, s'; \theta)$$

If the variance of  $\alpha$  is homogeneous, we may write  $c_\alpha(s, s') = \sigma_\alpha^2 r_\alpha(s, s'; \theta)$  where  $r_\alpha(\cdot)$  is the *correlation function*, being scaled by the variance component  $\sigma_\alpha^2$ . Since elements of  $\alpha$  are indexed by space, the covariance function allows one to “fill-in” the elements of  $\Sigma_\alpha(\theta)$ . Thus, given  $c_\alpha$ , spatial parameters  $\theta$ , and any two locations in space,  $s$  and  $s'$  (sample locations, or not), the covariance between  $\alpha(s)$  and  $\alpha(s')$  may be determined.

Typically, assumptions are imposed on the process to facilitate estimation of parameters (as will be discussed further below), but also because there is a severe shortage of more general covariance models. The two usual assumptions are *second order stationarity* and *isotropy*, the former being *translation invariance* of the second moment structure of  $\alpha$ , and the latter being *rotation invariance*. Normally the stationarity assumption would imply a similar constraint on the first moment structure, but we have assumed  $\alpha$  to have mean 0, accommodating any mean non-stationarity in  $X\beta$ . Thus, under these assumptions, the covariance between any two points is only a function of the *distance* separating them:

$$\text{cov} [\alpha(s), \alpha(s')] = \sigma_\alpha^2 r_\alpha(\|s - s'\|; \theta)$$

where  $\|s - s'\|$  is the distance between points  $s$  and  $s'$ , say Euclidean, geographic distance, etc. This simplified correlation structure conveniently dictates the covariance between observed and unobserved values of  $y$  for which predictions are desired, a quantity required to formulate the predictor (discussed below). A common correlation model is the exponential model is given by

$$r_\alpha(\|s - s'\|; \theta) = \exp\left(\frac{-\|s - s'\|}{\theta}\right)$$

## 22.5 COCK RIGING MODEL

Cokriging, involves direct specification of the joint variance-covariance structure among a set of variables. In the bivariate case, with variables  $y_1(s)$  and  $y_2(s)$ , the cokriging model can be formulated as a joint-normality assumption on the response vectors  $y_1$  and  $y_2$ :

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \text{Gau}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_2 \end{pmatrix}\right)$$

Here, the mean vectors  $\mu_1$  and  $\mu_2$  may be related to regression variables in the usual manner. As before, the marginal covariance matrices are defined as  $\Sigma_1 = \sigma_1^2 R_{\theta 1}$  and  $\Sigma_2 = \sigma_2^2 R_{\theta 2}$ , where the correlation matrices  $R_{\theta 1}$  and  $R_{\theta 2}$  are parameterized by correlation functions  $r_1(s, s'; \theta_1)$  and  $r_2(s, s'; \theta_2)$ . These may be chosen to be of the same parametric form, with different parameters, or of different parametric forms. Similarly, the cross-covariance matrix  $\Sigma_{12} = \sigma_{12} R_{\theta 12}$ , where  $\sigma_{12}$  is the covariance between  $y_1(s)$  and  $y_2(s)$  and  $R_{\theta 12}$  is the cross-correlation matrix. The key to the cokriging model is specification of a cross-correlation function  $\text{corr}[y_1(s), y_2(s')] = r_{12}(s, s'; \theta_{12})$  used to parameterize the matrix  $R_{\theta 12}$ .

In general, estimation and prediction under this model proceeds as in the univariate case. That is, parameter estimation based on MLE (or other methods) from the joint model, and then use of a plug-in procedure where these estimates are substituted into the expressions for the best linear unbiased predictor (BLUP) and its variance.

## 22.6 SPATIAL STATISTICAL MODELS

Statistics, since its inception as discipline, has provided tools by which scientists can better understand complex processes. **Primarily this is because statistics is concerned with the study of variability, and all natural processes exhibit variability.** As scientists seek to answer ever more challenging questions concerning processes that vary over space, the traditional statistics methods that one might learn in introductory statistics courses are not sufficient to adequately account for this variability. However, at least in principle, relatively simple extensions to simple statistical concepts such as linear models and regression, provide the foundation for basic spatial statistical analysis.

Although not universally true, objects in close proximity are more alike. Consequently, one must include the effects of spatial proximity when performing statistical inference on such processes, or at least show that there is no need to do so. Including these **spatial effects is important for efficient estimation of parameters, prediction and the design of sampling networks.** As a simple illustration, consider some spatial process, denoted by  $Y$ , at three locations,  $A, B, C$  such that  $A$  and  $B$  are very close together in space (i.e., adjacent pots in a field trial) and

$C$  is widely separated from both  $A$  and  $B$ . Assume the spatial process has zero mean and variance  $\sigma^2$  at all spatial locations. It is then the case that  $\text{var}[Y(A) - Y(B)] = 2\sigma^2 - 2\text{cov}[Y(A), Y(B)]$  and  $\text{var}[Y(A) - Y(C)] = 2\sigma^2 - 2\text{cov}[Y(A), Y(C)]$ . If the covariance is positive and decreases with distance (that is, things close together are more alike), then  $\text{cov}[Y(A), Y(B)] > \text{cov}[Y(A), Y(C)]$  and thus  $\text{var}[Y(A) - Y(B)] < \text{var}[Y(A) - Y(C)]$ . Clearly, inference on the differences should include the effects of the spatial dependence.

The later twentieth century and beginning of the twenty-first century has seen a **tremendous growth in spatial statistical methodological development** and application. This is primarily a **function of the rapid progression of computational technology, hardware, software and algorithmic, and the need to solve challenging problems**. The corresponding propagation of Bayesian methodology into mainstream statistics has been responsible for a sizeable portion of this development.

Kriging and its derivatives constitute the most common class of spatial models used in diverse disciplines such as crop and soil science, geology, atmospheric science, and more recently in ecology and the biological sciences. Many software packages have “kriging” routines, and kriging is the core of many contemporary graduate level courses on spatial statistics. Much of the terminology common in spatial statistics today first arose within the field of geostatistics.

Kriging can be viewed as arising under a linear mixed model (LMM), which have been intensively studied and have a well developed theory. Thus, understanding the basics of conventional mixed models is helpful for understanding spatial statistical models. In fact, it can be argued that the LMM perspective is natural since LMM's are widely used in biological, medical and epidemiological fields, particular in relation to *longitudinal* data, of which spatial data are a special case. When viewed from a LMM perspective, estimation and prediction of spatially correlated processes poses no additional complexity beyond that required for LMMs. This is in contrast.

Linear mixed models are well-known to both statisticians and practitioners of statistics alike, and so this formulation is often simpler as an introductory framework. An additional benefit of the LMM development is that extension to non-Gaussian problems is straightforward by way of the generalized linear mixed model (GLMM) extension of the normal, linear case. The discipline of disease mapping makes widespread use of GLMMs within a spatial modelling context.

## 22.7 MULTIVARIATE SPATIAL MODELS

In environmental and biological studies, it is seldom the case that measurements are made on a single variable. For example, most air pollution monitoring networks collect data on several pollutants in addition to relevant covariates such as temperature and precipitation.

We consider data collected as part of the annual North American Breeding Bird Survey (BBS). This survey is conducted in May-June of each year. Volunteer observers traverse a roadside sampling route containing 50 stops. At each stop, the observer records the number of birds (by species) seen and heard. There are several thousand BBS routes in North America. We focus on mourning dove (*Zenaida macroura*) counts from 103 routes in the state of Pennsylvania. Let  $y(s)$  be the total count of doves on the BBS route centered at (aggregated over the 50 stops). Our goal is to produce map of dove relative abundance within the state.

The breeding bird survey described in the previous example produces counts on over 200 species of birds, many of which have similar habitat and resource requirements. It stands to reason that relationships exist among these variables, and these relationships should manifest themselves in the joint spatial structure of the variables.

The most obvious benefit of having multiple spatial variables is the use of a covariate to aid in prediction of a primary variable. For example, temperature is informative about ozone concentration. At the same time, temperature is more abundant and cheaper to collect. Consequently, more efficient prediction of ozone concentration may be achieved by inclusion of the relationship between ozone and temperature within a bivariate spatial model. In the context of bird monitoring, some species are difficult to observe and thus it is difficult to construct precise maps of abundance of similar species can serve this purpose.

The need to incorporate inter-dependance among two or more spatial variables gives rise to several strategies for modelling spatial dependance. For the most part, these are straight forward extensions of univariate methods and thus, much of that material applies in the multivariate setting.

## 22.8 MARKOV RANDOM FIELD SPATIAL MODELS

In the previous discussion, it has been assumed that the spatial process can occur at any spatial location in some two-dimensional Euclidean space, a continuous region. Special classes of spatial models known as Markov random fields (MRFs) have been developed to account for the situation where the set of all possible spatial locations is discrete (countable). We sometimes say that the collection of such sites in a *lattice*. Examples include mapping disease in counties and modelling air-pollution on a grid. The first is an example of an *irregular lattice* and the latter is often a *regular lattice*. Regular lattices have neighbourhoods that are often defined by adjoining sites and irregular lattices often have neighbourhoods defined by Euclidean proximity.

Consider a spatial process defined at  $n$  spatial locations  $\{s_1, \dots, s_n\}$ ,  $y = [y(s_1), \dots, y(s_n)]$ . This process has joint distribution  $P[y(s_1), \dots, y(s_n)]$ . From this joint distribution, the conditional distribution of the process at each location  $i$  can be expressed in terms of all other sites ( $j \neq i$ ) as

$$P[y(s_i) | \{y(s_j) : j \neq i\}], \quad i = 1, \dots, n.$$

We then define the neighbourhood  $N_i$  of the  $i$ -th site as the collection of locations such that

$$P[y(s_i) | \{y(s_j) : j \neq i\}] = P[y(s_i) | \{y(s_j) : j \in N_i\}], \quad i = 1, \dots, n$$

That is, the conditional probability at site  $i$  only depends on nearby values of the process  $\{y(s_j) : j \in N_i\}$ . The specification of these conditional, neighbourhood-specific distributions must be made consistency so that the joint distribution is well-defined. Examples of such models include auto-Gamma models for non-negative continuous processes, auto-Poisson models for spatial counts, auto-logistic models for binary spatial random variables, and auto-Gaussian models for spatial Gaussian processes on a lattice.

### 22.8.1 Gaussian Markov Random Field Model

A natural model when the spatial process  $y(\cdot)$  is continuous is Gaussian MRF model. The conditional models (12) are Gaussian with,

$$P[y(s_i) | y(N_i)] = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left[ -\frac{1}{2\sigma_i^2} \left\{ y(s_i) - \mu(s_i) - \sum_{j \in N_i} c_{ij} [y(s_j) - \mu(s_j)] \right\}^2 \right]$$

where  $c_y \sigma_i^2 = c_{ii} \sigma_i^2$ ,  $c_{ii} = 0$ ,  $c_{ik} = 0$  for  $k$  not in  $N_i$ , and  $\mu(s_i) = E[y(s_i)]$ . It can be shown that the joint distribution for  $y$  is then given by

$$y \sim \text{Gau} [\mu, (I - C)^{-1} M]$$

provided  $(I - C)^{-1}$  is positive-definite, and where  $\mu \equiv [\mu(s_1), \dots, \mu(s_n)]'$ ,  $C$  is an  $n \times n$  matrix with  $(i, j)$ th element  $c_{ij}$ , and  $M$  is a diagonal matrix with  $\sigma_1^2, \dots, \sigma_n^2$  on the main diagonal.

The Gaussian MRF can easily be incorporated in the LMM framework described previously. For example, let  $y = X\beta + \alpha + \epsilon$ , where now  $\alpha \sim \text{Gau}(0, (I - C)^{-1} M)$ , a MRF with zero mean. The parameters of the LMM covariance structure  $\lambda$  are now the elements of  $C$  and  $M$ . In practice, the neighbourhood structure of the MRF is often simplified greatly so that the number of unknown parameters is reasonable small. For example, a typical assumption is that only those neighbours that are immediately adjacent to a given location are necessary to specify the conditional distribution; this is often referred to as a first-order dependence model.

## 22.9 GAUSSIAN RANDOM PROCESS MODELS

Consider a spatial process  $Y(s)$  where  $s \in D$ , some domain in d-dimensional Euclidean space. In this model, we will only consider two-dimensional spatial processes. Furthermore, we assume that the process  $Y(s)$  has a Gaussian (normal) distribution with mean  $\mu(s)$  and is correlated so that  $c_y(s, s') = \text{cov}[Y(s), Y(s')]$  for some  $s, s' \in D$  where  $s \neq s'$ . We refer to such a process as a Gaussian random process or Gaussian random field.

### 22.9.1 Linear Mixed Model Framework

The classical linear model generalizes the traditional linear model to include random effects. In the present context, we will equate the random effect to a correlated spatial process. A common statement of the LMM is:

$$y = X\beta + H\alpha + \epsilon$$

where  $y$  is an  $n \times 1$  vector responses,  $X$  and  $H$  are known matrices of independent, explanatory, or regression variables ( $n \times p$  and  $n \times q$ , respectively),  $\beta$  is a  $p \times 1$  vector of regression coefficients or *fixed effects*; and  $\alpha$  and  $\epsilon$  are  $q \times 1$  and  $n \times 1$  random vectors, respectively. Typically, columns of  $H$  are *indicator* variables, so that each observation is associated with a particular element of  $\alpha$ . The usual assumption on these random effects is multivariate normality:

$$\begin{bmatrix} \alpha \\ \epsilon \end{bmatrix} \sim \text{Gau} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_\alpha & 0 \\ 0 & \Sigma_\epsilon \end{bmatrix} \right)$$

In many statistical problems, including spatial statistics, one often assumes independence of the random errors, in which case  $\Sigma_\epsilon = \sigma_\epsilon^2 I_{n \times n}$ , where  $I_{n \times n}$  is the  $n$ -dimensional identity matrix. The variance component  $\sigma_\epsilon^2$  is *measurement error variance*, and may additionally include effects of small-scale spatial variability – that is, anything unexplained by the random effect. In the field of geostatistics,  $\sigma_\epsilon^2$  is called the “nugget effect”.

The application of this model to spatial settings is straightforward. Suppose that the response vector is spatially indexed, so that  $y = [y(s_1), \dots, y(s_n)]'$  for spatial locations  $s_i, i = 1, \dots, n$ . Let the elements of  $\alpha$  represent "spatial effects", then  $\Sigma_\alpha$  is a  $q \times q$  *spatial* covariance matrix where  $q$  is the number of spatial locations. In many spatial statistical problems  $q = n$ ; i.e., there is a single response observation at each site, in which case  $H = I_{n \times n}$ . This is the essence of the model used in conventional kriging applications. The other consideration in the context of spatial applications is that prediction of "unobserved" data is of primary interest. This is in contrast to most mixed-model applications, where the primary interest is in estimation of the vectors  $\beta$ , perhaps the variance components  $\sigma_\alpha^2$  and  $\sigma_\epsilon^2$ , and to a less extent,  $\alpha$ .

For the model (1) note that  $\alpha \sim \text{Gau}(0, \Sigma_\alpha)$  and  $\epsilon \sim \text{Gau}(0, \Sigma_\epsilon)$ . One can think of this model hierarchically, as

$$y|\alpha \sim \text{Gau}(X\beta + H\alpha, \Sigma_\epsilon) \quad \dots (2)$$

$$\alpha \sim \text{Gau}(0, \Sigma_\alpha) \quad \dots (3)$$

The joint distribution is given by  $f(y, \alpha) = f(y|\alpha) f(\alpha)$ . One can obtain the marginal distribution for  $y$  by integrating out the random effects,  $f(y) = \int f(y|\alpha) f(\alpha) d\alpha$ , which is easily shown in this case to be

$$y \sim \text{Gau}(X\beta + H\alpha, \Sigma_\epsilon) \quad \dots (4)$$

Thus, the marginal model (4) follows from the hierarchical formulation (2) and (3). In traditional LMM applications (e.g., longitudinal analysis) it is often convenient (and arguably more general) to proceed in terms of the marginal model, without the need for specific inference or estimation concerning the random effects. That is, one accounts for spatial dependence but is not interested in the underlying process that generates such dependence. However, for most traditional spatial applications, one is interested in performing inference on the random effects (spatial process) and the hierarchical formulation is more appropriate.

## 22.10 NON-GAUSSIAN RANDOM PROCESS MODELS

The linear model presented above may be considered free of formal distributional assumptions (which is the usual kriging development). That is, first and second moment assumptions alone lead to the same estimators and predictors for parameters and random effects as if the processes were assumed to be Gaussian. Thus, such approaches may be viewed as having an implicit assumption of normality.

Such informality leads to ambiguity in applying the basic linear mixed model approach when there is a strong belief as to the distribution of the random variable under study, other than Gaussian. For example, count data are common in many biological, ecological and epidemiological problems. In such problems, there are unusual discernible mean-variance relationships, and the variable is clearly discrete and positive valued. It is natural to consider Poisson and Binomial models in these situations, and treatment in the usual LMM/kriging context will lead to some inefficiency in both estimation and prediction. Similarly, many problems in the atmospheric and environmental sciences involve positive-values variables, often right-skewed, again with strong mean-variance relationships. In this case, the log-normal distribution is a natural candidate for a model-based analysis, though other possibilities exist (e.g., Gamma).

## 22.11 GENERALIZED MIXED MODEL FRAMEWORK

Non-Gaussian spatial problems may be formally analyzed within the context of generalized linear mixed models (GLMM). Formulation of a GLMM requires specification of the likelihood of the random variable  $y(s)$ . As in classical generalized linear models (GLMs), there is a so-called *canonical parameter* corresponding to the distribution, which is nominally a function  $g(\cdot)$  ( $C$  the *link function*) of the location parameter for the distribution. It is this quantity that is assumed to be linear in the explanatory variables. In the classical formulation of GLMs containing only fixed effects,  $g(\mu) = X\beta$ , where  $X$  is the matrix of explanatory variables. To incorporate a spatial process, we assume  $y(s_i|\alpha)$  is conditionally independent for any location  $s_i$  with conditional mean  $E[y(s_i)|\alpha] = \mu(s_i)$ . Then, the spatially correlated random effect is incorporated into the linear predictor:

$$g(\mu) = X\beta + H\alpha + \epsilon$$

where the additional noise term  $\epsilon$  may or may not be included, depending on the application. Nominally, such an error term accommodates *over-dispersion* relative to the mean-variance relationship implied by the distribution under consideration. As before,  $\alpha \sim \text{Gau}(0, \Sigma_\alpha(\theta))$  and  $\epsilon \sim \text{Gau}(0, \sigma_\epsilon^2 I)$ , with spatial correlation parameterized by  $\theta$  in  $\Sigma_\alpha(\theta)$ .

## 22.12 HIERARCHIAL MODELS

Notwithstanding the difficulty in constructing valid joint variance-covariance models, cokriging becomes unwidely when several variables are considered. In addition, specification of the various cross-correlation functions is awkward and typically little scientific basis exists to guide construction of the joint second-moment model. Cokriging is essentially an empirical approach. Alternatively, one may construct a joint model for two or more variables by construction of a sequence of conditional models (the hierarchical approach). Sticking to the bivariate model consists of the following two components.

$$y_1|y_2 \sim \text{Gau}(\mu_1 + \beta y_2, \sigma_{12}^2 R_{\theta_{12}})$$

and

$$y_2 \sim \text{Gau}(\mu_2, \sigma_2^2 R_{\theta_2})$$

As usual, one must specify correlation models for  $R_{\theta_{12}}$  and  $R_{\theta_2}$ , but there is no cross-correlation term. Instead, the cross-correlation between the two variables is accommodated in the *conditional mean* or  $y_1$  via the regression parameter  $\beta$ . General parameterizations of this conditional mean may be considered. The point is that, **under this model, one avoids having to model the cross-correlation function explicitly**, thereby avoiding difficulty in specifying models which produce valid joint variance-covariance structures. It is easy to show that the implied joint variance-covariance is valid, assuming that the correlation models used to construct the conditional correlation matrices are.

The hierarchical formulation is **particularly useful when there is a mechanistic basis for the conditioning**, such as causal relationship between  $y_1$  and  $y_2$ . However, there need not be scientific basis – the model is valid regardless. Consequently, this strategy may be viewed as a convenient framework for building valid joint models.

## 22.13 SPATIOTEMPORAL MODELS

Spatiotemporal processes are ubiquitous in the biological sciences. In principle models for such processes are relatively easy to formulate in the traditional LMM or GLMM framework. However, a lack of understanding of the underlying processes and the “curse of dimensionality” make the implementation of these models challenging.

Consider a spatiotemporal process  $y(s; t)$  defined for  $s \in D_s$ ,  $t \in D_t$ , where  $D_s$  and  $D_t$  are spatial and temporal domains, respectively. The domains may be continuous or discrete, but we typically consider  $D_t$  a discrete collection of times (e.g.,  $D_t = \{t_1, t_2, \dots, t_T\}$ ). In the remainder of this section, we will assume the discrete temporal domain. We may then formulate the spatiotemporal process as a GLMM, where  $f(y(s; t)|\alpha)$  is a conditionally independent distribution for all  $s$  and  $t$ , and where  $\alpha(s; t)$  is a spatiotemporal random process such that  $\alpha \equiv [\alpha_1' \dots \alpha_T']'$ , where  $\alpha_i \equiv [\alpha(s_1; t), \dots, \alpha(s_n; t)]'$  and  $\alpha \sim \text{Gau}(0, \Sigma_\alpha)$ . Note that  $\Sigma_\alpha$  is an  $(n + T) \times (n + T)$  covariance matrix with elements  $c_\alpha(s, s', t, t') = \text{cov}[\alpha(s, t), \alpha(s', t')]$ . Clearly, if  $n$  or  $T$  is large, this matrix will be very large. Thus, although easy to formulate, the limitation with this approach is that the known class of realistic and valid spatiotemporal covariance functions,  $c_\alpha$ , is very small and the dimensionality of the joint spatiotemporal process  $\alpha$  is prohibitively large. One alternative is to further factorize the joint spatiotemporal distribution for  $\alpha$  as a series of conditional models. For many processes, a dynamical factorization based on a Markov assumption in time is appropriate. That is,

$$f(\alpha) = f(\alpha_0) \prod_{t=1}^T f(\alpha_t | \alpha_{t-1}, \alpha)$$

Where the conditional distributions  $f(\alpha_t | \alpha_{t-1}, \theta)$  depend on a collection of parameters  $\theta$  that describe the dynamical evolution and the variance/covariance structure of the associated spatial noise process. For example, a conditional model might follow a first-order vector autoregressive model such as  $\alpha_t = H_\theta \alpha_{t-1} + \eta_t$ , where  $\eta_t \sim \text{Gau}(0, \Sigma_\eta(\theta))$ , is the spatial noise process  $H_\theta$  is a collection of parameters that describe the evolution of the  $\alpha$  process (i.e., vector-autoregression parameter matrix). One can implement such a model in a Kalman filter framework. However, when the number of spatial locations is large, the number of parameters in  $H_\theta$  may prohibit likelihood-based estimation procedures. In some cases, one may have scientific knowledge that suggests relatively simple parameterizations of  $H_\theta$ .

## 22.14 COMPUTERS AND BIOLOGY

The pervasive presence of computers together with their ever-increasing computational power encourages biologists to apply statistical methods to analyze data that is collected in the laboratory or the field. One important software application used by biologists is the spreadsheet. Increasingly, spreadsheet applications contain sophisticated tools sufficient for use with undergraduate biology majors. It is observed that *the graphing calculator is not the tool of choice* for biology students. Technological tools must be capable of producing graphs that can be incorporated into printed and presentation documents.

Implementation of linear mixed models is possible on most of the standard statistical software packages and some specialized programs. These packages can be very useful for relatively simple spatial problems, and have the advantage of synergism with other statistical methods contained

in the particular package. However, these packages typically assume relatively simple spatial models (e.g., isotropic and stationary) and are difficult to implement in dense and /or extensive prediction domains. This is even more pronounced in the context of generalized linear mixed models. Thus, for “complicated spatial processes or high-dimensional prediction applications one often must write custom software for a particular application.

## 22.15 HIGH-LEVEL OR LOW-LEVEL LANGUAGE AND MODELS

Practitioners typically favour either so-called “low-level” programming language implementations (such as C, C++, FORTRAN) or “high-level” languages (such as S, R, MATLAB, GAUSS). Although preference for one over the other is often an issue of familiarity, there are distinct advantages to each. For example, the “high-level” languages are often matrix-oriented and as such are very efficient when it comes to matrix calculations, but inefficient for loop-intensive programs. Conversely, the low level languages are very efficient when it comes to traditional programming structures such as loops but often less efficient for matrix-oriented calculations.

Many practitioners have found that combinations of high- and low-level languages work best for implementation of complicated spatial models in practice. In most cases, prototype algorithms and test cases are most efficiently implemented in a matrix-oriented language. When one is satisfied with the code, one may then translate the matrix-oriented program to a low-level language for “operational” implementation. In addition, many practitioners use the high-level language as a computational “shell” and then call low-level routines for portions of the code that are less efficient.

Of course, there are significant differences between the various matrix language programs as well.

## 22.16 SOME EXAMPLES ON STATISTICAL MODELS

**Example 1 :** Represent the following data by a percentage sub-divided bar diagram model.

| Item of<br>Expenditure | Family A       | Family B       |
|------------------------|----------------|----------------|
|                        | Income Rs. 500 | Income Rs. 300 |
| Food                   | 150            | 150            |
| Clothes                | 125            | 60             |
| Education              | 25             | 50             |
| Miscellaneous          | 190            | 70             |
| Savings or Deficit     | +10            | -30            |

**Solution :** In this problem the incomes of the two families are different (i.e., Rs. 500 and Rs. 300), so the appropriate **bar diagrams** for the given data will be rectangular diagram on percentage basis. The width of the rectangles will be taken in the ratio of their incomes, i.e., 500 : 300 or 5 : 3.

Table : Calculations for Percentage Rectangular Bar Diagram

| Item of Expenditure | Family A          |    |              | Family B          |      |              |
|---------------------|-------------------|----|--------------|-------------------|------|--------------|
|                     | Expenditure (Rs.) | %  | Cumulative % | Expenditure (Rs.) | %    | Cumulative % |
| Food                | 150               | 30 | 30           | 150               | 50   | 50           |
| Clothes             | 125               | 25 | 55           | 60                | 20   | 70           |
| Education           | 25                | 5  | 60           | 50                | 16.7 | 86.7         |
| Miscellaneous       | 190               | 38 | 98           | 70                | 23.3 | 110.0        |
| Savings or Deficit  | +10               | 2  | 100          | -30               | -10  | 100          |
| Total               | 500               |    |              | 300               |      |              |

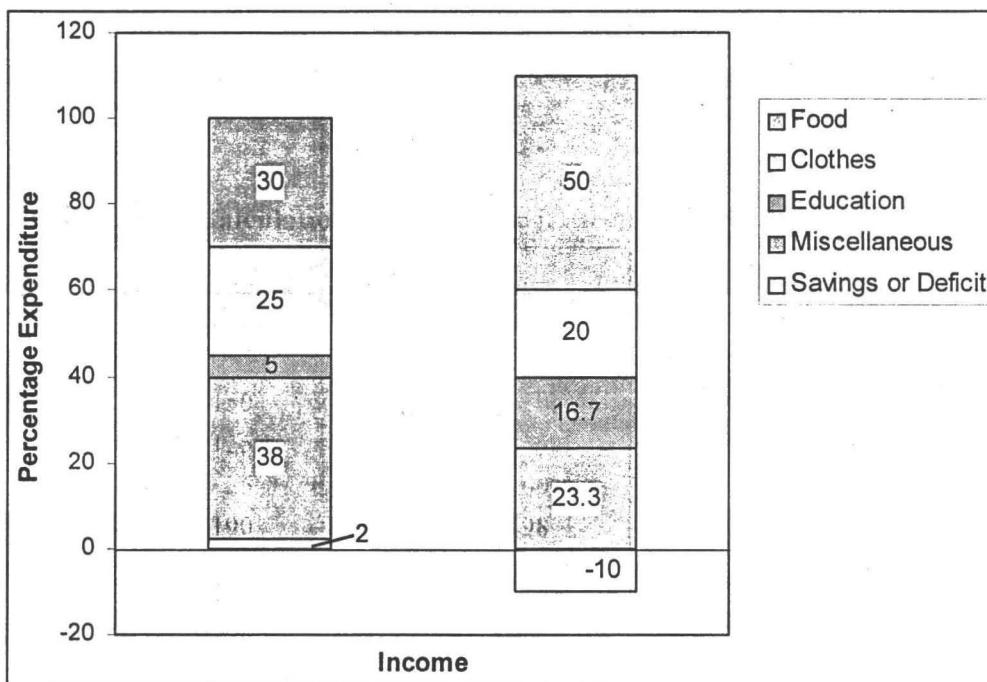


Fig. 12.2 : Percentage diagram showing monthly income and expenditure of two families A and B.

**Example 2 :** Draw a pie diagram model for the following data of a Five Year Plan Public Sector outlays:

|                                    |       |
|------------------------------------|-------|
| Agricultural and Rural Development | 12.9% |
| Irrigation etc.                    | 12.5% |
| Energy                             | 27.2% |
| Industry and Minerals              | 15.4% |

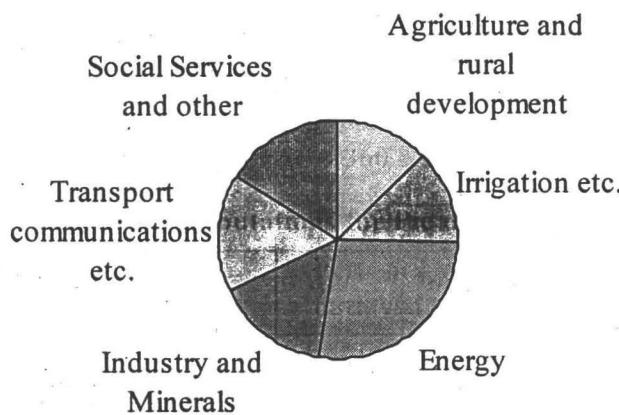
|                                  |       |
|----------------------------------|-------|
| <i>Transport communication</i>   | 15.9% |
| <i>Social services and other</i> | 16.1% |

**Solution :** The angle at the centre is given by

$$\text{Angle} = \frac{\text{Percentage outlay}}{100} \times 360 = \text{Percentage outlay} \times 3.6^\circ$$

#### Computations for Pie diagram

| Sector<br>(1)                     | Percentage<br>outlays (2) | Angle at the centre<br>(3) = (2) $\times$ 3.6° |
|-----------------------------------|---------------------------|--|
| Agriculture and Rural Development | 12.9                      | $12.9 \times 3.6 = 46^\circ$                   |
| Irrigation etc.                   | 12.5                      | $12.5 \times 3.6 = 35^\circ$                   |
| Energy                            | 27.2                      | $27.2 \times 3.6 = 98^\circ$                   |
| Industry and Minerals             | 15.4                      | $15.4 \times 3.6 = 56^\circ$                   |
| Transport Communications etc.     | 15.9                      | $15.9 \times 3.6 = 57^\circ$                   |
| Social Services and Others        | 16.1                      | $16.1 \times 3.6 = 38^\circ$                   |
| Total                             | 100.0                     | 360°   |



**Fig. 12.3 : Pie chart showing Five Year Plan Public Sector Outlay.**

**Example 3 :** Draw the graph of the following:

| Year              | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|-------------------|------|------|------|------|------|------|------|------|
| Yield             | 12.8 | 13.9 | 12.8 | 13.9 | 13.4 | 6.5  | 2.9  | 14.8 |
| (in million tons) |      |      |      |      |      |      |      |      |

**Solution :** Taking the scale along X-axis as 1 cm = 1 year and along Y-axis is 1 cm = 2 million tons, the required histogram or graph is given below.

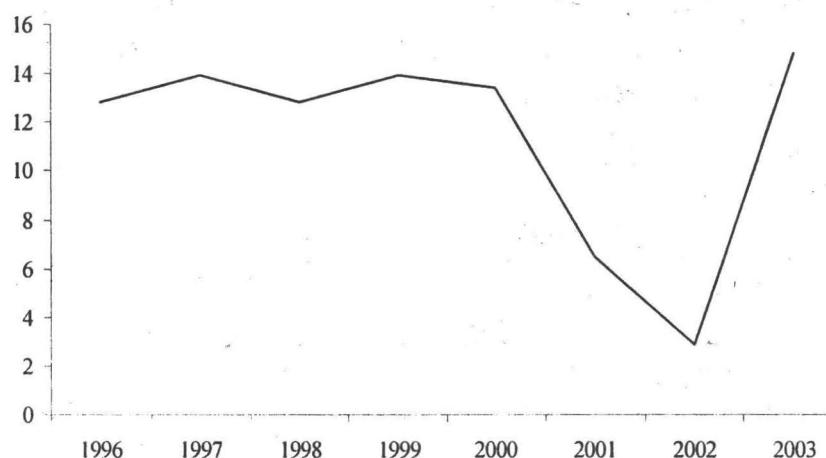


Fig. 22.3 : A graph showing yield for different years

### EXERCISE 22.1

1. The following data relates to the blood samples of two persons A and B represent it by a suitable percentage diagram model.

| <i>Nature of test</i> | <i>Result in mg/dl</i> |              |
|-----------------------|------------------------|--------------|
|                       | <i>Mr. A</i>           | <i>Mr. B</i> |
| Sugar pp              | 160                    | 100          |
| Sugar pasting         | 80                     | 92           |
| HDL                   | 60                     | 38           |
| LDL                   | 20                     | 16           |
| Triglyceride          | 80                     | 14           |
| Total                 | 400                    | 260          |

2. The following data shows the expenditure on various heads of a five year plans (in crores of rupees).

| <i>Subject</i>                    | <i>Expenditure (in crores Rs.)</i> | <i>Five Year Plan</i> |
|-----------------------------------|------------------------------------|-----------------------|
|                                   |                                    |                       |
| Agriculture and C.D.              |                                    | 1068                  |
| Irrigation and power              |                                    | 1662                  |
| Village and small industries      |                                    | 264                   |
| Industry and minerals             |                                    | 1520                  |
| Transport and communication       |                                    | 1486                  |
| Social services and miscellaneous |                                    | 1500                  |
| Total                             |                                    | 7500                  |

Represent the data by angular (pie) diagram model.

[Hint :

### Five year plan

| <i>Rs.</i> | <i>Degrees</i>                        |
|------------|---------------------------------------|
| 1,068      | $\frac{1068}{7500} \times 360 = 51.2$ |
| 1,662      | $\frac{1662}{7500} \times 360 = 79.8$ |
| 264        | $\frac{264}{7500} \times 360 = 12.7$  |
| 1,520      | $\frac{1520}{7500} \times 360 = 73.0$ |
| 1,486      | $\frac{1486}{7500} \times 360 = 71.3$ |
| 1,500      | $\frac{1500}{7500} \times 360 = 72.0$ |
| 7,500      | 360                                   |
| 86.60      |                                       |
| 1.8        |                                       |

3. Plot a graph to represent the following data in a suitable manner.

| Year               | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|--------------------|------|------|------|------|------|------|------|------|
| Imports ('000 mds) | 400  | 450  | 560  | 620  | 580  | 460  | 500  | 540  |
| Imports ('000 Rs.) | 220  | 235  | 385  | 420  | 420  | 380  | 360  | 400  |

4. Draw a percentage bar diagram model to represent the following data.

| <i>Items of expenditure</i> | <i>Income in Rupees</i> |                 |
|-----------------------------|-------------------------|-----------------|
|                             | <i>Family A</i>         | <i>Family B</i> |
| Food                        | 400                     | 480             |
| Clothing                    | 200                     | 400             |
| House rent                  | 160                     | 200             |
| Fuel                        | 80                      | 120             |
| Miscellaneous               | 160                     | 400             |
| Total                       | 1,000                   | 1,600           |

5. The following tables gives the monthly expenditure of two families *A* and *B*. Compare these figures by a suitable diagram.

*Items of expenditure**Income in Rupees*

|                   | <i>Family A</i> | <i>Family B</i> |
|-------------------|-----------------|-----------------|
| Food              | 500             | 800             |
| Clothing          | 140             | 2400            |
| House rent        | 80              | 160             |
| Education         | 30              | 80              |
| Fuel and Lighting | 40              | 40              |
| Miscellaneous     | 40              | 80              |

6. Represent the following data on production of Tea, Cocoa and Coffee by means of a pie diagram model.

| Tea        | Cocoa      | Coffee   | Total      |
|------------|------------|----------|------------|
| 3,260 tons | 1,850 tons | 900 tons | 6,010 tons |

7. Draw a suitable model for the following:

| <i>Expenditure item</i> | <i>Percentage of total expenditure</i> |
|-------------------------|--|
| Food                    | 65                                     |
| Clothing                | 10                                     |
| Housing                 | 12                                     |
| Fuel and Lighting       | 5                                      |
| Miscellaneous           | 8                                      |

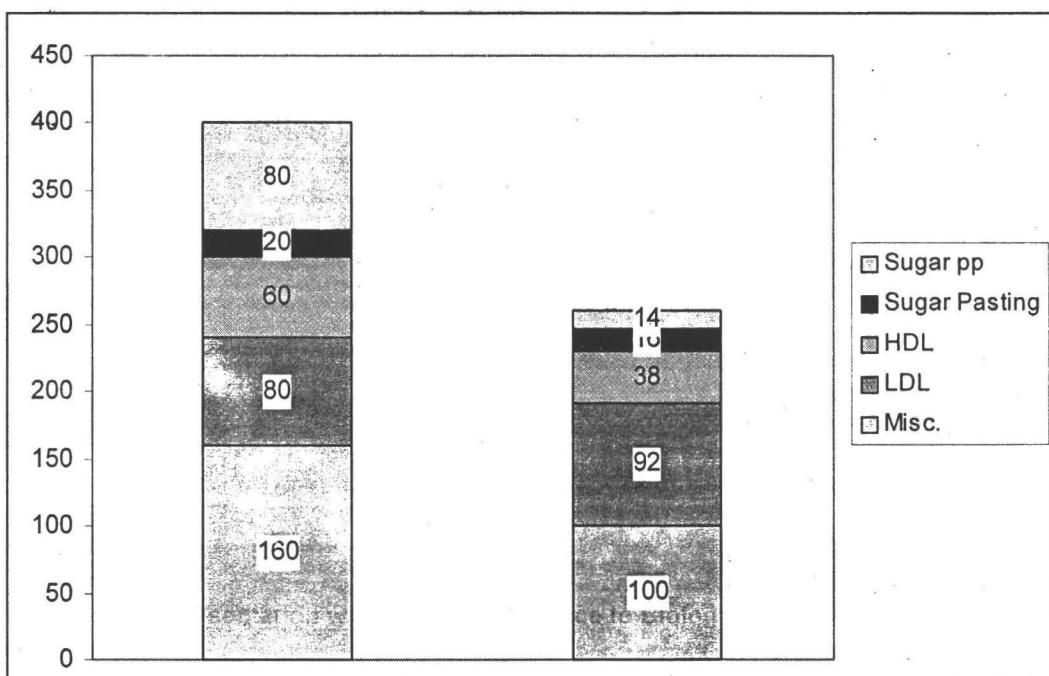
[Hint : Draw pie diagram]

8. The sale proceeds, cost and the profit or loss per wooden chair of a firm are:

| <i>Particulars</i>      | <i>2004</i> | <i>2005</i> | <i>2006</i> |
|-------------------------|-------------|-------------|-------------|
| Sales proceeds          | 190         | 220         | 250         |
| Cost per chair material | 100         | 110         | 130         |
| Wages                   | 40          | 75          | 90          |
| Other costs             | 30          | 50          | 60          |
| Total cost              | <u>170</u>  | <u>235</u>  | <u>280</u>  |
| Profit/loss             | +20         | -15         | -30         |

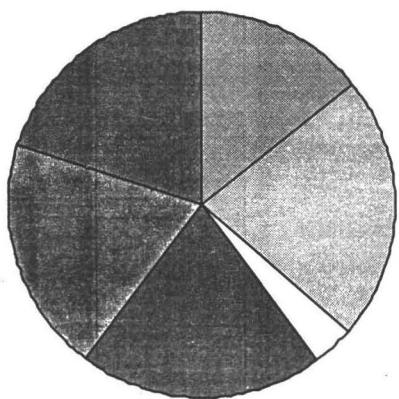
Draw bar diagrams model for the above data.

## ANSWERS



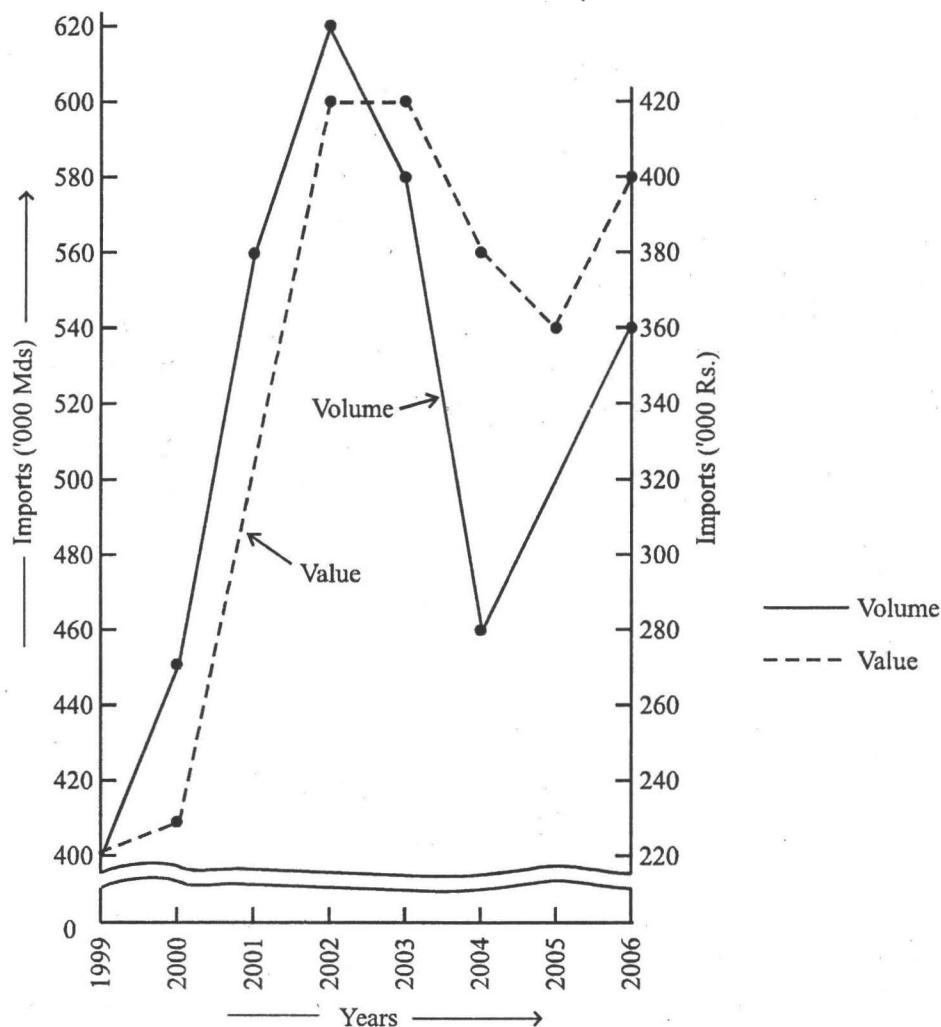
## Third Plan

2.

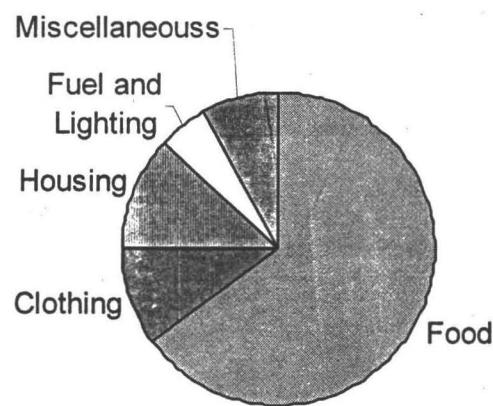


- Agriculture and C.D.
- Irrigation and power
- Village and small industries
- Industry and minerals
- Transport and communication
- Social services and miscellaneous

3.



7.



8.

