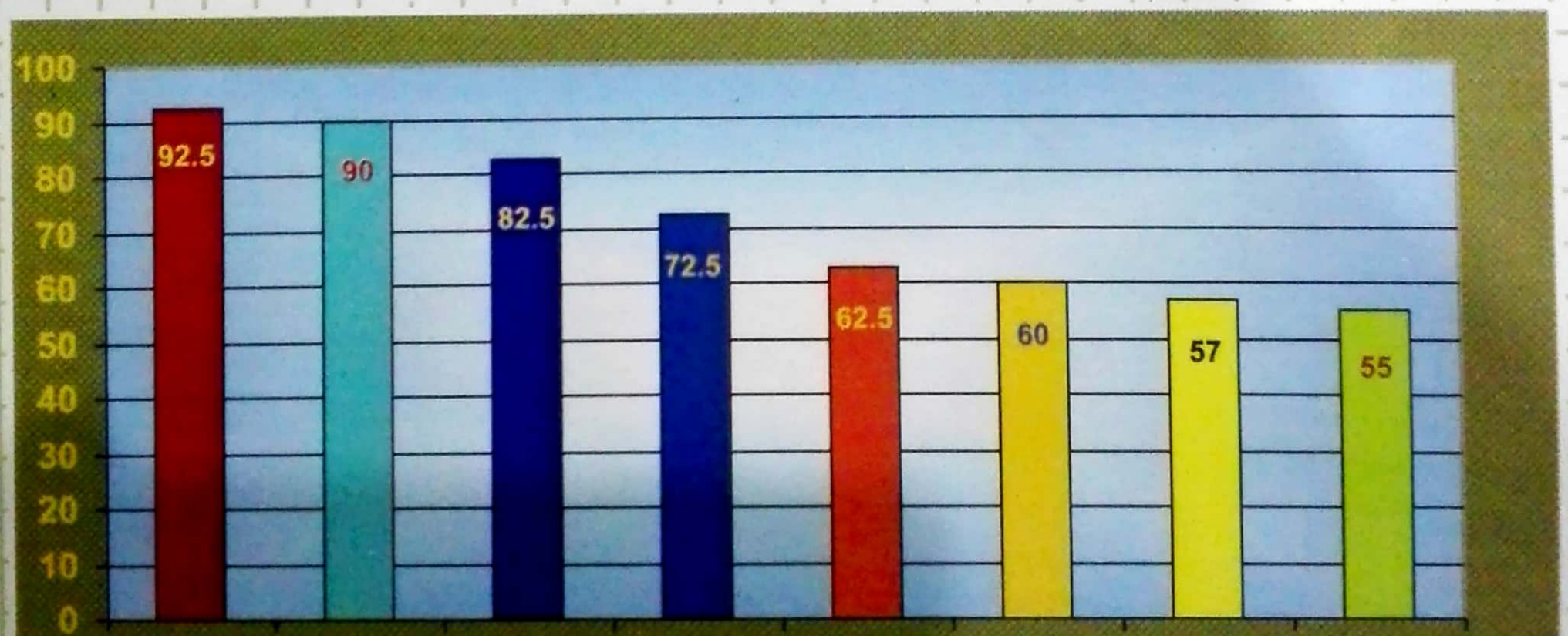
INTRODUCTION TO STATISTICAL THEORY

Part-1

Prof. Sher Muhammad Chaudhry
Prof. Dr. Shahid Kamal

Percentages of Adult Population using Wireless Phones
in Major Cities of 8 Countries in June 2008



INTRODUCTION TO STATISTICAL THEORY

Part 1

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

By

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

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

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

CONTENTS

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

2.7.4 Rectangles and Sub-divided Rectangles 32
2.7.5 Pictograms 33
2.7.6 Pie Diagrams 34
2.7.7 Profit and Loss Chart 35
2.8 **Graphs** 35
2.8.1 Graph of Time Series – Historigram 36
2.8.2 Histogram 37
2.8.3 Frequency Polygon 39
2.8.4 Frequency Curve 40
2.8.5 Cumulative Frequency Polygon or Ogive 40
2.8.6 Ogive for a Discrete Variable 41
2.8.7 Types of Frequency Curves 41
2.8.8 Ratio Charts or Semi-logarithmic Graphs 43
Exercises 44

3. MEASURES OF CENTRAL TENDENCY OR AVERAGES
3.1 **Introduction** 55
3.2 **Criteria of a Satisfactory Average** 55
3.3 **Types of Averages** 55
3.4 **The Arithmetic Mean** 56
3.4.1 The Weighted Arithmetic Mean 57
3.4.2 Properties of the Arithmetic Mean 57
3.4.3 Mean from Grouped Data 59
3.4.4 Change of Origin and Scale 61
3.5 **The Geometric Mean** 62
3.6 **The Harmonic Mean** 64
3.7 **The Median** 67
3.7.1 Quantiles 68
3.8 **The Mode** 72
3.9 **Empirical Relation between Mean, Median and Mode** 74
3.10 **The Box Plots** 74
3.11 **Relative Merits and Demerits of Various Averages** 75
3.11.1 The Arithmetic Mean 75
3.11.2 The Geometric Mean 75
3.11.3 The Harmonic Mean 76
3.11.4 The Median 76
3.11.5 The Mode 76
Exercises 77

4. MEASURES OF DISPERSION, MOMENTS AND SKEWNESS
4.1 **Introduction** 87
4.2 **Range** 87
4.3 **The Semi-Interquartile Range or the Quartile Deviation** 88
4.4 **The Mean (or Average) Deviation** 89

4.5 **The Variance and Standard Deviation** 91
4.5.1 Change of Origin and Scale 96
4.5.2 Interpretation of the Standard Deviation 97
4.5.3 Co-efficient of Variation 98
4.5.4 Properties of Variance and Standard Deviation 100
4.5.5 Standardized Variables 103
4.6 **Trimmed and Winsorized Measures** 104
4.7 **Moments** 105
4.7.1 Moments about the Mean in terms of Moments about an arbitrary origin, say a, and conversely 106
4.7.2 Sheppard's Corrections 108
4.7.3 Moment-Ratios 108
4.7.4 Change of Origin and Scale 110
4.7.5 Charlier Check 111
4.8 **Skewness** 114
4.9 **Kurtosis** 115
4.10 **Describing a Frequency Distribution** 116
Exercises 116

5. **INDEX NUMBERS**
5.1 **Introduction** 131
5.1.1 Simple and Composite Index Numbers 131
5.1.2 Problems Involved in Index Number Construction 131
5.2 **Main Steps in the Construction of Index Numbers of Wholesale Prices** 132
5.2.1 Selection of Commodities for Inclusion 132
5.2.2 Selection of the Base Period 132
5.2.3 Selection of Average 134
5.2.4 Selection of Appropriate Weights 135
5.3 **Unweighted Index Numbers** 135
5.3.1 Simple Aggregative Index 135
5.3.2 Simple Average of Relatives 136
5.4 **Weighted Index Numbers** 140
5.4.1 Weighted Aggregative Price Index Numbers 140
5.4.2 Weighted Average of Relatives Price Index Number 145
5.5 **Quantity Index Numbers** 147
5.6 **Tests for Index Number Formulae** 151
5.6.1 Time Reversal Test 151
5.6.2 Factor Reversal Test 153
5.6.3 Circular Test 155
5.7 **Consumer Price Index Number** 156
5.7.1 Meaning 156
5.7.2 Construction of Consumer Price Index Numbers 156

	5.7.3	Shortcomings or Drawbacks of Consumer Price Index Numbers	159
5.8	**Uses of Index Numbers**		160
5.9	**Limitations of Index Numbers**		160
	Exercises		161
6.	**PROBABILITY**		
6.1	**Introduction**		173
6.2	**An Aside – Sets**		173
	6.2.1	Subsets	174
	6.2.2	Venn Diagram	175
	6.2.3	Operations on Sets	175
	6.2.4	The Algebra of Sets	177
	6.2.5	Partition of Sets	177
	6.2.6	Class of Sets	177
	6.2.7	Cartesian Product Sets	177
	6.2.8	Relation and Function	178
6.3	**Random Experiment**		179
	6.3.1	Sample Space	179
	6.3.2	Events	180
	6.3.3	Events and Symbolic Representations	181
	6.3.4	Counting Sample Points	181
6.4	**Definitions of Probability**		184
	6.4.1	Subjective or Personalistic Probability	187
6.5	**Laws of Probability**		189
6.6	**Conditional Probability**		196
6.7	**Independent and Dependent Events**		205
	Exercises		214
7.	**RANDOM VARIABLES**		
7.1	**Introduction**		227
7.2	**Distribution Function**		227
7.3	**Discrete Random Variables and its Probability Distribution**		228
7.4	**Continuous Random Variable and its Probability Density Function**		233
7.5	**Joint Distributions**		237
	7.5.1	Bivariate Distribution Function	237
	7.5.2	Bivariate Probability Functions	238
	7.5.3	Marginal Probability Functions	238
	7.5.4	Conditional Probability Functions	239
	7.5.5	Independence	239
	7.5.6	Continuous Bivariate Distributions	243
7.6	**Mathematical Expectation of a Random Variable**		248
	7.6.1	Expectation of a Function of a Random Variable	250

7.6.2 Properties of Expected Values 256
7.6.3 Covariance of Two Random Variables 263
7.6.4 Variance of the Sum or Difference of Two Random Variables 263
7.6.5 Correlation Co-efficient of Random Variables 263
7.7 **Medians and Modes of Continuous Random Variables** 266
7.8 **Chebyshev's Inequality** 268
7.9 **Moment Generating Function** 269
7.9.1 Cumulant Generating Function 271
7.9.2 Relation between Cumulants and Moments 272
7.9.3 Characteristic Function 273
Exercises 273

8. **DISCRETE PROBABILITY DISTRIBUTIONS** 285
8.1 **Introduction** 285
8.2 **Binomial Probability Distribution** 285
8.2.1 Derivation of Binomial Probability Distribution 291
8.2.2 Binomial Frequency Distribution 292
8.2.3 Properties of the Binomial Probability Distribution 297
8.2.4 The Recurrence Formula for Binomial Distribution 298
8.2.5 Fitting a Binomial Distribution to Observed Data 299
8.2.6 Moment Generating and Cumulant Generating Functions of the Binomial Distribution 302
8.3 **Hypergeometric Probability Distribution** 302
8.3.1 Derivation of Hypergeometric Distribution 305
8.3.2 Properties of Hypergeometric Distribution 308
8.4 **Poisson Distribution** 309
8.4.1 Derivation of Poisson Approximation to the Binomial 312
8.4.2 Poisson Frequency Distribution 313
8.4.3 Properties of the Poisson Distribution 316
8.4.4 The Recurrence Formula for Poisson Distribution 317
8.4.5 Fitting a Poisson Distribution to Observed Data 318
8.4.6 Poisson Process 320
8.4.7 Moment Generating and Cumulant Generating Functions of the Poisson Distribution 321
8.5 **Negative Binomial Distribution** 322
8.5.1 Derivation of the Negative Binomial Distribution 324
8.5.2 Properties of the Negative Binomial Distribution 325
8.6 **Geometric Distribution** 325
8.6.1 Derivation of the Geometric Distribution 326
8.6.2 Properties of the Geometric Distribution 327
8.6.3 Moment Generating Function of the Geometric Distribution 328
8.7 **Multinomial Distribution** 328
8.7.1 Derivation of the Multinomial Distribution

Exercises 329

9. CONTINUOUS PROBABILITY DISTRIBUTIONS

9.1 **Introduction** 341
9.2 **Uniform Distribution** 341
9.2.1 Properties of the uniform distribution 341
9.2.2 Moment-Generating Function of Uniform Distribution 342
9.3 **Exponential Distribution** 342
9.3.1 Properties 343
9.3.2 Moment-Generating Function of Exponential Distribution 344
9.4 **Gamma and Beta Distributions** 344
9.4.1 Gamma Function 344
9.4.2 Beta Function 346
9.4.3 Gamma Distribution 347
9.4.4 Properties of Gamma Distribution 347
9.4.5 Moment-Generating Function of Gamma Distribution 348
9.4.6 Beta Distribution of the First Kind 349
9.4.7 Properties of $\beta_1(m,n)$ 349
9.4.8 Beta Distribution of Second Kind 350
9.4.9 Properties of $\beta_2(m,n)$ 351
9.5 **Normal Distribution** 351
9.5.1 Standardized Normal Distribution 353
9.5.2 Properties of Normal Distribution 355
9.5.3 Moment Generating and Cumulant Generating Functions of the Normal Distribution 362
9.5.4 Tabulated Areas of Normal Distribution 364
9.5.5 Inverse Use of Table of Areas under Normal Curve 370
9.5.6 Normal Approximation to the Binomial Distribution 374
9.5.7 Normal approximation to the Poisson Distribution 378
9.5.8 Fitting a Normal Distribution 379
9.5.9 Ordinates of the Normal Distribution 382
Exercises 384

10. SIMPLE REGRESSION AND CORRELATION

10.1 **Introduction** 395
10.2 **Deterministic and Probabilistic Relations or Models** 395
10.3 **Scatter Diagram** 396
10.4 **Simple Linear Regression Model** 396
10.4.1 An Aside – The Principle of Least Squares 398
10.4.2 Least-Squares Estimates in Simple Linear Regression 398
10.4.3 Properties of the Least-Squares Regression Line 402
10.4.4 Standard Deviation of Regression or Standard Error of Estimate 402
10.4.5 Co-efficient of Determination 404

10.5 **Correlation** 406
10.5.1 Pearson Product Moment Correlation Co-efficient 406
10.5.2 Correlation and Causation 408
10.5.3 Properties of r 408
10.5.4 Correlation Co-efficient for Grouped Data 411
10.6 **Rank Correlation** 413
10.6.1 Derivation of Rank Correlation 414
10.6.2 Rank Correlation for Tied Ranks 416
10.6.3 Co-efficient of Concordance 417
Exercises 419

11. MULTIPLE REGRESSION AND CORRELATION
11.1 **Introduction** 429
11.2 **Multiple Linear Regression with two Regressors** 429
11.2.1 Expression of Multiple Linear Regression in Deviation Form 431
11.2.2 Standard Error of Estimate 433
11.2.3 Co-efficient of Multiple Determination and Multiple Correlation 434
11.2.4 Subscript Notation 435
11.2.5 Properties of Residuals 436
11.2.6 Multiple Regression in terms of Linear Correlation Coefficients 437
11.3 **Multiple Correlation Co-efficient** 438
11.4 **Partial Correlation** 441
11.4.1 Relationship between Multiple and Partial Correlation Co-efficients 446
11.5 **Curvilinear Regression** 447
Exercises 447

12. CURVE FITTING BY LEAST SQUARES
12.1 **Introduction** 455
12.2 **Approximating Curves and the Principle of Least-Squares** 455
12.2.1 Fitting a Straight Line 455
12.2.2 Fitting of a second degree Parabola 457
12.2.3 Fitting of Higher Degree Parabolic Curves 460
12.2.4 Change of Origin and Unit 461
12.3 **Exponential Curves** 462
12.4 **Other Types of Curves** 465
12.4.1 Modified Exponential Curve 465
12.4.2 The Gompertz Curve 467
12.4.3 The Logistic Curve 467
12.4.4 The Makeham Curve 467
12.5 **Criteria for a Suitable Curve** 468
12.6 **Finding Plausible Values by LS Method** 468

Exercises 470

13. TIME SERIES ANALYSIS

13.1 **Introduction** 477
13.2 **Components of a Time Series** 478
13.2.1 Secular Trend 478
13.2.2 Seasonal Variations 479
13.2.3 Cyclical Fluctuations 479
13.2.4 Irregular or Random Variations 480
13.3 **Time Series Decomposition** 480
13.4 **Analysing the Secular Trend** 481
13.4.1 The Method of Freehand Curve 481
13.4.2 The Method of Semi-Averages 481
13.4.3 The Method of Moving Averages 483
13.4.4 The Method of Least-Squares 487
13.5 **Detrending** 491
13.6 **Analysing the Seasonal Variations** 492
13.6.1 The Percentage of Annual Average method 492
13.6.2 The Ratio-to-Moving-Average Method 493
13.6.3 The Ratio-to-Trend Method 495
13.6.4 The Link-Relative Method 497
13.7 **Deseasonalization of Data** 498
13.8 **Analysing the Cyclical Variations** 499
13.9 **Analysing the Irregular Variations** 500
13.10 **Forecasting** 501
13.10.1 Forecasting by Exponential Smoothing 501
13.11 **Serial Correlation** 501
Exercises 503

Answers to Exercises 515

Tables 539

Index 545

CHAPTER 1

INTRODUCTION

INTRODUCTION

1.1 MEANING OF STATISTICS

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

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

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

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

i) Children who brush their teeth with brand XYZ toothpaste have 60% fewer countries.

ii) The Bureau of census projects the population of Pakistan to be 170.1 million in the year 2010.

iii) Eight out of ten Pakistanis do not have skills.

iv) The prevalence of diabetes is nearly 3 times as high in over weight people as compared to normal weight people.

v) In 1980 it was estimated that 0.1% of people had tried any sort of drug; where as in 2008 it was estimated that 10% had done so.

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

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

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

i) to inform general public;

ii) to explain things that have happened;

iii) to justify a claim;

iv) to provide general comparisons;

v) to predict the decision regarding future outcomes;

vi) to estimate the unknown quantities;

vii) to establish association / relationship between factors.

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

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

i) should we survey every student?

ii) how will the survey be conducted?

iii) how will the data be analyzed?

iv) how will the best teacher be determined? etc.

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

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

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

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

i) Statistics deals with the behaviour of aggregates or large groups of data. It has nothing to do with what is happening to a particular individual or object of the aggregate.

ii) Statistics deals with aggregates of observations of the same kind rather than isolated figures.

iii) Statistics deals with variability that obscure underlying patterns. No two objects in this universe are exactly alike. If they were, there would have been no statistical problem.

iv) Statistics deals with uncertainties as every process of getting observations whether controlled or uncontrolled, involves deficiencies or chance variation. That is why we have to talk in terms of probability.

v) Statistics deals with those characteristics or aspects of things which can be described numerically either by counts or by measurements.

vi) Statistics deals with those aggregates which are subject to a number of random causes, *e.g.* the heights of persons are subject to a number of causes such as race, ancestry, age, diet, habits, climate and so forth.

vii) Statistical laws are valid *on the average* or in the long run. There is no guarantee that a certain law will hold in all cases. Statistical inference is therefore made in the face of uncertainty.

viii) Statistical results might be misleading and incorrect if sufficient care in collecting, processing and interpreting the data is not exercised or if the statistical data are handled by a person who is not well versed in the subject matter of statistics.

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

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

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

Comparison

Descriptive Statistics	Inferential Statistics
i) A cricket player wants to find his score average for the last 20 games.	i) A cricket player wants to estimate his chance of scoring based on his current season average.
ii) Aamir wants to describe the variation in his four test scores in Statistics.	ii) Based on the first four test scores, Aamir would like to predict the variation in his final Statistics test scores.
iii) Mrs. Rashid wants to determine the average weekly amount she spent on groceries in the past 6 months.	iii) Based on last six months grocery bills, Mrs. Rashid would like to predict the average amount she will spend on groceries for the upcoming year.

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

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

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

i) Total number of absentees by all students in a college during the last month.

ii) Number of colour TV sets owned by all families in Lahore.

iii) Monthly salaries of all employees of a company.

iv) Wheat yield per acre for 5 pieces of land.

v) Number of computers sold during the last month at all the computer stores in Lahore.

Solution

i) Population

ii) Population

iii) Population

iv) Sample

v) Population

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

i) Statistics assists in summarizing the larger sets of data in a form that is easily understandable.

ii) Statistics assists in the efficient design of laboratory and field experiments as well as surveys.

iii) Statistics assists in a sound and effective planning in any field of inquiry.

iv) Statistics assists in drawing general conclusions and in making predictions of how much of a thing will happen under given conditions.

v) Statistical techniques being powerful tools for analysing numerical data, are used in almost every branch of learning. In the biological and physical sciences, Genetics, Agronomy, Anthropometry, Astronomy, Physics, Geology, etc. are the main areas where statistical techniques have been developed and are increasingly used.

vi) A businessman, an industrialist and a research worker all employ statistical methods in their work. Banks, Insurance companies and Governments all have their statistics departments.

vii) A modern administrator whether in public or private sector, leans on statistical data to provide a factual basis for decision.

viii) A politician uses statistics advantageously to lend support and credence to his arguments while elucidating the problems he handles.

ix) A social scientist uses statistical methods in various areas of socio-economic life of a nation. It is sometimes said that "*a social scientist without an adequate understanding of statistics, is often like the blind man groping in a dark room for a black cat that is not there*".

1.2 OBSERVATIONS AND VARIABLES

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

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

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

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

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

A variable whether countable or measurable, is generally denoted by some symbol such as X or Y and X_i or Y_j represents that ith or jth value of the variable. The subscript i or j is replaced by a number such as 1,2,3,... when referred to a particular value.

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

a) the hair colour of children;

b) length of time required for a wound to heal;

c) the number of telephone calls arriving at a switch board per 1 hour period;

d) the breaking strength of a given type of a string;

e) the number of questions answered correctly on a test;

f) the number of stop signs in the city of Lahore;

g) the colour of your eye;

h) the number of children in your family.

Solution

a) Attribute

b) Continuous

c) Discrete

d) Continuous

e) Discrete

f) Discrete

g) Attribute

h) Discrete

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

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

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

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

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

Example of Measurement Scales

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

1.2.4 Errors of Measurement. Experience has shown that a continuous variable can never be measured with perfect fineness because of certain habits and practices, methods of measurements, instruments used, etc. The measurements are thus always recorded correct to the nearest units and hence are of limited accuracy. The *actual* or *true* values are, however, assumed to exist. For example, if a student's weight is recorded as 60 kg (correct to the nearest kilogram), his true weight in fact lies between 59.5 kg and 60.5 kg, whereas a weight recorded as 60.00 kg means the true weight is known to lie between 59.995 and 60.005 kg. Thus there is a difference, however small it may be, between the measured value and the true value. This sort of departure from the true value is technically known as the *error of measurement*. In other words, if the observed value and the true value of a variable are denoted by x and $x+\varepsilon$ respectively, then the difference $(x+\varepsilon)-x$, *i.e.* ε is the error. This error involves the unit of measurement of x and is therefore called an *absolute error*. An absolute error divided by the true value is called the *relative error*. Thus the relative error $=\dfrac{\varepsilon}{x+\varepsilon}$, which when multiplied by 100, is *percentage error*. These errors are independent of the units of measurement of x. It ought to be noted that an error has both magnitude and direction and that the word *error* in statistics does not mean mistake which is a chance inaccuracy.

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

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

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

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

1.2.5 Significant Digits. Accuracy in measurements is related to significant digits. The significant digits in a number, are those that represent accurate and meaningful information. For instance, the number 35 representing a continuous variable has two significant digits. In recorded measurements, all digits except zeros are always significant. For zeros, we may state as:

i) Zeros are significant if they follow a decimal point and conclude a number, e.g. the measurement 2.500 has four significant digits.

ii) Zeros are non-significant when they follow a decimal point but commence a number, *e.g.* the measurements .04 and .000237 contain only 1 and 3 significant digits respectively.

iii) Zeros may or may not be significant when they lie entirely to the left of the decimal point, where they may not represent measurement but may be used to simply locate the decimal point. In such a case, a definite specification such as standard notation, becomes necessary. When any number is expressed as a product of a power of 10 and a number between 1 and 10, it is said to be written in standard notation. For example, the number 75400 can have 3 significant digits when written in standard notation as 7.54×10^4. It can also have 5 significant digits if written as 7.5400×10^4.

iv) Zeros are always significant when occur within a series of significant digits, *e.g.* the numbers 20.3, .1001, 4.00507, etc., have 3, 4 and 6 significant digits respectively.

It should be remembered that

a) significant digits in a number are not disturbed by the location of the decimal point, *e.g.* the measurements recorded as 269., 26.9, .269 or .000269 have only 3 significant digits;

b) in case of discrete data which are generated by the process of counting, the number of significant digits is considered indefinite because the level of accuracy cannot be improved, *e.g.* the number 15700 has indefinite significant digits;

c) the rules regarding the determination of the number of significant digits, are applicable to continuous variables;

d) in the operations of addition and subtraction, all digit positions which are not significant in any of the values being added or subtracted, are not significant in the total or difference;

e) in the operations of multiplication and division, the number of significant digits in the result is determined by the value with the smallest number of significant digits that enters into the calculations.

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

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

ii) The last significant digit remains unaltered, if the first digit of the remainder to be dropped is 4 or less, *e.g.* the numbers 2.1548 and 7.3627 are rounded off to three significant digits as 2.15 and 7.36 respectively.

iii) When the digit to be dropped is exactly 5, the accepted practice is to increase the last significant digit by 1, if it is odd and to leave unaltered if it is even *e.g.* the number 4.535 and 2.745 are rounded off to three significant digits as 4.54 and 2.74 respectively.

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

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

1.3 COLLECTION OF DATA

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

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

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

The survey research includes the following important steps:

a) to define the objectives of the survey.

b) to define the variable(s) and the population of interest.

c) to define the data collection and data measuring schemes.

d) to determine the appropriate descriptive and inferential data analysis techniques.

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

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

Survey Research

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

ii) **Indirect Investigation or Personal Interviews.** Sometimes the direct sources do not exist or the informants hesitate to respond for some reasons or other. In such a case, third parties or witnesses having information are interviewed. As some of the informants are likely to deliberately give wrong information, so the reliance is not placed on the evidence of one witness only. Moreover, due allowance is to be made for the personal bias. This method is useful when the information desired is complex or there is reluctance or indifference on the part of the informants. It can be adopted for extensive inquires.

iii) **Collection through Questionnaires.** A questionnaire is an inquiry form comprising of a number of pertinent questions with space for entering information asked. The questionnaires are usually sent by mail and the informants are requested to return the questionnaires to the investigator, after doing the needful within a certain period. This method is cheap, fairly expeditious and good for extensive inquires. But the difficulty is that the majority of respondents (persons who are required to answer the questions) does not care to fill the questionnaires in and to return them to the investigators. Sometimes, the questionnaires are returned incomplete and full of errors. In spite of these drawbacks, the method is considered as the standard method for routine business and administrative inquires. The answers to the questionnaires are very often recorded by trained enumerators to overcome the difficulties these days. It is important to note that the questions should be few, brief, very simple, easy for all respondents to answer, clearly worded and not offensive to certain respondents.

iv) **Collection through Enumerators.** Under this method, the information is gathered by employing trained enumerators who assist the informants in making the entries in the schedules or questionnaires correctly. This method gives the most reliable information if the enumerator is well trained, experienced and tactful. It is considered the best method when a large scale governmental inquiry is to be conducted. This method cannot be adopted by private individual or institution as its cost would be prohibitive to them.

v) **Collection through Local Sources.** In this method, there is no formal collection of data but the agents or local correspondents are directed to collect and to send the required information, using their own judgment as to the best way of obtaining it. This method is cheap and expeditious, but gives only the estimates.

vi) **Computer interviews.** Respondents enter data directly into a computer in response to questions presented on the monitor.

Experimental Research

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

ii) **Field experiments.** Manipulation of the independent variable(s) in a natural situation.

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

Secondary Research

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

External Secondary Data

i) **Official,** *e.g.* the publications of the Statistical Division, Ministry of Finance, the Federal and Provincial Bureaus of Statistics, Ministries of Food, Agriculture, Industry, Labour, etc.

ii) **Semi-Official,** *e.g.* State Bank of Pakistan, Railway Board, Central Cotton Committee, Boards of Economic Inquiry, District Councils, Municipalities, etc.

iii) Publications of Trade Association, Chambers of Commerce, etc.

iv) Technical and Trade journals and newspapers.

v) Research organisation such as universities, and other institutions.

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

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

A few examples of how and where Statistics is used is as under:

i) In experimental science, the experiments generate data, it must be collected and analyzed.

ii) In Government, many types of statistical data are collected all the time. This data can be used for various types of planning and also to inform the general public.

iii) In education, Statistics are used to describe the results and standards of education.

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

a) Statistics can prove anything.

b) 'There are three types of lies – lies, damned lies, and Statistics'

Statistics can also be misused in many ways such as using Not Representative Samples, Small Sample Size, Ambiguous Averages and dispersions, Detached facts, Implied Connections, Wrong and Misleading Graphs, Wrong use of Statistical techniques, Serious violation of assumptions behind the Statistical techniques and Faculty Surveys etc.

EXERCISES

OBJECTIVE

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

i) All numerical data are not Statistics.

ii) A Statistic is a summary measure computed for a population.

iii) A Parameter is a summary measure computed for a sample.

iv) Descriptive Statistics are used to make projections or estimates about the population.

v) Inferential Statistics is the study and description of data.

vi) A sample is typically a very large collection of individuals or objects of our interest.

vii) The thickness of the glass is an example of attribute data.

viii) The number of students in a class is an example of continuous data.

ix) The make of a car is an example of discrete data.

x) The main objective of Statistics is to collect a sample, analyze it and make inferences about the unknown characteristics of the population from which the sample has been drawn.

SUBJECTIVE

1.1 Explain what is meant by Statistics. Give the important uses and limitations of statistics.

1.2 Define Statistics. Discuss, giving examples, the importance of the study of statistics and show how it can help the extension of scientific knowledge. (P.U., B.A. (Hons.), 1960)

1.3 a) What do you understand by the term Statistics? Give its chief characteristics.

b) Give a brief account of the importance of statistics in different fields. (P.U., B.A./B.Sc. 1971)

1.4 Comment upon the following statement of Sir Ronald Fisher: "Statistics may be regarded as (i) the study of populations, (ii) the study of variation and (iii) the study of the methods of reduction of data". (P.U., B.A. (Hons.), Part-I, 1964)

1.5 Comment on the statement given below:

"Statistics is concerned with understanding the real world through the information that we derive from classification and measurement. Its distinctive characteristic is that it deals with variability and uncertainty which is everywhere".

1.6 a) Define Statistics and explain its characteristics.

b) What are the uses of Statistics? (P.U., B.A. (Hons. in Econ.), 1960)

1.7 Explain the difference between the following?

a) Statistics and Statistic.

b) Population and Sample.

c) Descriptive and Inferential Statistics.

d) Quantitative and Qualitative Variables.

e) Discrete and Continuous Variables.

f) Biased and Unbiased errors.

g) Primary data and Secondary data.

h) Nominal and ordinal scale, Interval and ratio scale.

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

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

b) Classify the following variables as discrete or continuous:

i) The number of students attending a class.

ii) The amount of milk produced by a cow.

iii) The number of heads in the toss of 6 coins.

iv) The yearly income of a College Professor.

v) The age of a shopkeeper.

vi) The weight of a college student.

vii) The number of petals on a flower.

viii) The life times of television tubes produced by a company.

ix) Temperature recorded every half hour at a weather bureau.

x) The number of shares sold each day in the stock market.

1.10 Classify each variable as qualitative or quantitative

i) Colour of eyes.

ii) Number of computers sold in the last month.

iii) Marital status of faculty members.

iv) Student's weight.

v) Lifetime of car batteries.

vi) Number of burgers sold by a fast food shop.

vii) Brand of cars.

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

i) Weights of cars

ii) Rankings of squash players.

iii) Temperature of the city.

iv) Salaries of the top five executives in bank.

v) Marital status.

vi) Ages of students.

vii) Ratings of five players (Poor, Fair, Good, Excellent)

viii) IQ of student.

ix) Rating of movies

x) Weights of suitcases on plane.

1.12 Round off the following continuous data to four significant digits each.

(i) 32.21705, (ii) 937.05002, (iii) 0.003599499,
(iv) 1.003599499, (v) 0.07000455, (vi) 22.2500001.

1.13 a) Distinguish between Primary and Secondary data, giving examples of each.

b) Describe the methods which can be used in the collection of statistical data, stating the advantages and disadvantages of each method.

c) Enumerate the main sources of errors in Statistics and give their effects.

(P.U., B.A./B.Sc., 1982-S)

1.14 What methods would you employ in the collection of statistical data when the field of inquiry is (i) small, (ii) fairly large and (iii) very large, if you are to pay due regard to accuracy, labour, cost and time?

1.15 What are the different methods employed in the collection of data for statistical enquiries? In what type of inquiry should each one of them be used? (P.U., B.A./B.Sc., 1961)

1.16 Select a newspaper or a magazine article that involves a statistical study and answer the following questions.

a) Is this study descriptive or inferential? Explain your answer.

b) What variables are used in the study?

c) What level of measurements was used to obtain data?

d) Is population defined in the article? If not, how could it be defined?

e) How the data might have been collected?

f) Do you agree with the conclusions given in the article?

◆◆◆◆◆◆◆◆◆◆◆

CHAPTER 2

PRESENTATION OF DATA

PRESENTATION OF DATA

2.1 INTRODUCTION

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

2.2 CLASSIFICATION

The term *classification* is defined as the process of dividing a set of observations or objects into classes or groups in such a way that (i) observations or objects in the same class or group are similar, (ii) observations or objects in each class or group are dissimilar to observations or objects in other class or group. Classification is thus the sorting of data into homogeneous classes or groups according to their being alike or not. When the data are sorted according to one criterion only, it is called a *simple classification* or a *one-way classification*. Classification is called a *two-way classification* when the data are sorted according to two criteria. A manifold classification or cross-classification is made according to several criteria.

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

2.2.1 Aims of Classification. The main aims of classification are:

i) to reduce the large sets of data to an easily understood summary;

ii) to display the points of similarity and dissimilarity;

iii) to save mental strain by eliminating unnecessary details;

iv) to reflect the important aspects of the data; and

v) to prepare the ground for comparison and inference.

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

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

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

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

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

2.3 TABULATION

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

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

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

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

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

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

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

i) Titles should be brief in the form of phrases. Complete sentences are unnecessary.

ii) Abbreviations should not be used.

iii) Main titles should be in capitals throughout. Sub-titles, if any, should be in lower case letters with major words capitalized and should indicate clearly what the table describes.

iv) The different parts of a title should be separated by commas but no full-stop at the end.

v) Words in titles should not be hyphenated except when really necessary.

vi) If a title necessitates the use of two or more lines, an *inverted pyramid* arrangement of the lines should be used.

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

i) The heading should be clear but concise.

ii) They should be arranged in such a way that the most important characteristic is placed in the first column. The column of totals is usually placed at extreme right, but some people prefer the totals on the left.

iii) Only the first word in each column caption should be capitalized. No full-stop should be put at the end.

iv) Abbreviations, when clear, may be used.

v) Main caption should be centred over the column it is to span.

vi) Extra lines should be used to avoid crowding in caption box.

vii) Whenever possible, caption width should be made roughly proportional to the size of numbers to be inserted.

c) **Row Captions and Stub.** The heading or title for a row, is called the *Row Caption* and the section containing the row captions is known as *Stub*. The necessary points in this respect are given below:

i) The principles for column captions apply to row captions in stub.

ii) If the stub is long and has several levels of classification, the major classification should be capitalized to separate the table into parts.

iii) Whenever the figures have more than four or five significant digits, the digits should be grouped in threes or fours. For example, one should write 23 178 327, not 23178327.

iv) In long tables, some space should be left after every five or ten rows.

v) Totals should usually be placed at the bottom, but some prefer to place them at the top.

vi) Items in the stub should be arranged so as to facilitate easy reading.

vii) Every stub should have an appropriate heading describing its contents. This heading should be centred in the upper left box of the table.

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

Prefatory notes give additional specifications of the data indicative of items included or excluded for all data of the table, statements of the box, etc. They are placed between the title and the boxhead. The wording should be in lower case alphabet. *Footnotes* are used to clarify anything in the table by giving a fuller description, by drawing attention to incompleteness or by stating any special circumstances affecting the data. The footnotes should be specific in nature. They are placed immediately below the bottom line of the table, above the source. Footnote symbols should be placed as follows:

i) If they refer to an entire column or a set of columns, place them at the end of the appropriate caption.

ii) If they refer to an entire row or a set of rows, place them at the end of the appropriate stub title.

iii) If they refer to a single cell in the table, place beside the cell entry in the body of the table.

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

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

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

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

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

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

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

i) A table should be simple. A complex table if possible, may be broken into relatively simple tables.

ii) Units of measurements and nature of the data should be specified in title,, captions, etc. in parentheses.

iii) Percentages should be clearly indicated as 'per cent of total' etc. and their total should be shown as 100.0.

iv) If the figures entered in the table are rounded off, this should be indicated in the prefatory note or in the stub or caption.

v) Zeros need not be entered.

vi) Minus signs are a part of the table and precede the number.

vii) The relationship of the parts to the whole should be shown by thin or heavy rulings.

viii) The item or items to be emphasized should be placed in the most prominent position of the table.

The general sketch of a table is given below:

← ..... TITLE ..... →

← Prefatory notes

Boxhead →	← Units	COLUMN CAPTIONS					
↑ STUB ↓	.....	.....	B O	D Y		.....	.....

Footnotes .....

Source notes.:...

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

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

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

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

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

We first compute the relevant figures as below:

Rural population = Total population – Urban population

= 271,076 – 46,740 = 224,336

Female population = Total population – Male population

= 271,076 – 139,699 = 131,377

Rural male population = District male population – Urban males

= 139,699 – 23,083 = 116,616

Similarly, Urban females = 46,740 – 23,083 = 23,657

Rural females = 224,336 – 116,616 = 107,720

Married population of District = 271,076 – 112,352 = 158,724

Rural unmarried population = 112,352 – 21,072 = 91,280

Rural unmarried males = 91,280 – 36,864 = 54,416

Urban unmarried females = 21,072 – 12,149 = 8,923 etc.

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

Title:

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

Boxhead (Captions)

Stub:

Source:

Areas	Both Genders			Male			Female		
	Total	Married	Un-married	Total	Married	Un-married	Total	Married	Un-married
District	271,076	158,724	112,352	139,699	73,134	66,565	131,377	85,590	45,787
Urban	46,740	25,668	21,072	23,083	10,934	12,149	23,657	14,734	8,923
Rural	224,336	133,056	91,280	116,616	54,200	54,416	107,720	70,856	36,864

2.4 FREQUENCY DISTRIBUTION

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

2.4.1 Class-limits. The *class-limits* are defined as the numbers or the values of the variables which describe the classes; the smaller number is the *lower* class limit and the larger number is the *upper* class limit. Class-limits should be well defined and there should be no overlapping. In other words, the limits should be *inclusive, i.e.* the values corresponding exactly to the lower limit or the upper limit be included in that class. The class-limits are therefore selected in such a way that they have the same number of significant places as the recorded values. Suppose the data are recorded to the nearest integers. Then an appropriate method for defining the class limits without overlapping, for example, may be 10 – 14, 15 – 19, 20 – 24, etc. The class limits may be defined as 10.0 – 14.9, 15.0 – 19.9, 20.0 – 24.9, etc. when the data are recorded to nearest tenth of an integer. Sometimes, a class has either no lower class limit or no upper class-limit. Such a class is called an *open-end class*. The open-end classes, if possible, should be avoided as they are a hindrance in performing certain calculations. A class indicated as 10 – 15 will include 10 but not 15, *i.e.* $10 \leq X < 15$.

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

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

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

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

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

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

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

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

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

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

ii) **Determine the range of variation in the data.** *i.e.* the difference between the largest and the smallest values in the data.

iii) **Divide the range of variation by the number of classes** to determine the approximate width or size of the equal class-interval. In case of fractional results, the next higher whole number is usually taken as the size or width of class-interval. If equal class-intervals are inconvenient or may be undesirable, then classes of unequal size are used. But in practice, intervals that are multiple of 5 or 10, are commonly used as people can understand them more readily.

iv) **Decide where to locate the class-limit** of the lowest class and then the lower class boundary. The lowest class usually starts with the smallest data value or a number less than it. It is better if it is a multiple of class-interval. Find the upper class boundary by adding the width of the class-interval to the lower class-boundary and write down the upper class limits too. The open-end classes, *i.e.* classes with the lowermost or uppermost class boundary unknown, should be avoided if possible.

v) **Determine the remaining class-limits** and class boundaries by adding the class-interval repeatedly. The lowest class should be placed at the top and the rest should follow according to size. In some cases, the highest class is placed at the top.

vi) **Distribute the data into the appropriate classes.** This is best done by using a "Tally-Column" where values are tabulated against appropriate classes by merely making short bars or tally marks to represent them. It is customary for convenience in counting to place the first four bars vertically and the fifth one diagonally and to leave a space. The number of tallies is then written in the frequency column. The tally column is usually omitted in the final presentation of the frequency distribution. But in case of small number of values, the actual values should be shown against each class to mitigate chances of error.

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

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

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

106	107	76	82	109	107	115	93	187	95	123	125
111	92	86	70	126	68	130	129	139	119	115	128
100	186	84	99	113	204	111	141	136	123	90	115
98	110	78	185	162	178	140	152	173	146	158	194
148	90	107	181	131	75	184	104	110	80	118	82

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

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

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

i) **By listing the actual values:**

FREQUENCY DISTRIBUTION OF WEIGHTS OF 60 APPLES

Weight	Entries	Frequency
65 – 84	76, 82, 70, 68, 84, 78, 75, 80, 82	9
85 – 104	93, 95, 92, 86, 100, 99, 90, 98, 90, 104	10
105 – 124	106, 107, 109, 107, 115, 123, 111, 119, 115, 113, 111, 123, 115, 110, 107, 110, 118	17
125 – 144	125, 126, 130, 129, 139, 128, 141, 136, 140, 131	10
145 – 164	162, 152, 146, 158, 148	5
165 – 184	178, 173, 181, 184	4
185 – 204	187, 186, 204, 185, 194	5
Total		**60**

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

ii) **By using a Tally-Column:**

FREQUENCY DISTRIBUTION OF WEIGHTS OF 60 APPLES

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

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

7.5	8.2	6.2	8.9	7.8	5.4	9.4	9.9	10.9	10.8	7.4
9.7	11.6	12.6	5.0	10.2	9.2	12.0	9.9	7.3	7.3	8.4
10.3	10.1	10.0	11.1	6.5	12.5	7.8	6.5	8.7	9.3	12.4
10.4	9.1	9.7	9.3	6.2	10.3	6.6	7.4	8.6	7.7	9.4
7.7	12.8	8.7	5.5	8.6	9.6	11.9	10.4	7.8	7.6	12.1
4.6	14.0	8.1	11.4	10.6	11.6	10.4	8.1	4.6	6.6	12.8
6.8	7.1	6.6	8.8	8.8	10.7	10.8	6.0	7.9	7.3	9.3
9.3	8.9	10.1	3.9	6.0	6.9	9.0	8.8	9.4	11.4	10.9

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

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

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

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

The following table shows the required frequency distribution:

FREQUENCY DISTRIBUTION OF MEAN DEATH RATES

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

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

41.78	29.32	31.47	35.35	32.82	39.42
61.65	28.31	44.63	22.78	44.44	48.12
81.71	33.47	50.35	29.19	51.26	50.32
26.84	18.95	48.19	43.72	43.89	47.15
60.20	44.43	41.17	37.50	22.35	29.17

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

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

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

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

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

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

Arrange the values as a frequency distribution.

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

FREQUENCY DISTRIBUTION OF ASSISTANTS IN 50 RETAIL ESTABLISHMENTS

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

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

2.4.6 Cumulative Frequency Distribution. The total frequency of a variable from its one end to a certain value (usually upper class boundary in grouped data), called the *base*, is known as the *cumulative frequency, less than or more than* the base of the variable. A table that shows the cumulative frequencies, is called a *cumulative frequency distribution.* The cumulative frequency of the last class is the sum of all frequencies in the distribution. If the cumulation process is from the lowest value to the highest, it is referred to as "*a less than*" type cumulative frequency distribution. For example, let us consider a frequency distribution having k classes, each of width h. Let us denote the midpoint of the *ith* class by x_i with frequency f_i such that $\sum_{i=1}^{k} f_i = n$. Now the lower class-boundary of the first group is $x_1 - h/2$ and the upper class boundaries are $x_1 + h/2$, $(i = 1, 2, ..., k)$. The cumulative frequency distribution is then obtained by adding each successive frequency to the cumulative total of frequencies for the preceding classes as shown below:

Class-boundary	Cumulative Frequency
less than $x_1 - h/2$	0
less than $x_1 + h/2$	f_1
less than $x_2 + h/2$	$f_1 + f_2$
less than $x_3 + h/2$	$f_1 + f_2 + f_3$
.	.
.	.
.	.
less than $x_k + h/2$	$\sum f_i = n$

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

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

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

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

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

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

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

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

2.5 STEM-AND-LEAF DISPLAY

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

leading digit	trailing digits
2	43
stem	leaf

OR

leading digit	trailing digits
24	3
stem	leaf

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

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

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

48, 31, 54, 37, 18, 64, 61, 43, 40, 71, 51, 12, 52, 65, 53
42, 39, 62, 74, 48, 29, 67, 30, 49, 68, 35, 57, 26, 27, 58

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

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

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

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

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

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

Example 2.8 Construct a stem-and-leaf display for the data of annual death rates given in Example 2.3.

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

Stem	Leaf
3	9
4	6 6
5	0 4 5
6	0 0 2 2 5 5 6 6 6 8 9
7	1 3 3 3 4 4 5 6 7 7 8 8 8 9
8	1 1 2 4 6 6 7 7 8 8 8 9 9
9	0 1 2 3 3 3 3 4 4 4 6 7 7 9 9
10	0 1 1 2 3 3 4 4 6 6 7 8 8 9 9
11	1 4 4 6 6 9
12	0 1 4 5 6 8 8
14	0

2.6 GRAPHICAL REPRESENTATION

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

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

2.7 DIAGRAMS

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

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

i) *Linear or One-Dimensional Diagrams.* They consist of Simple Bars, Multiple Bars and Component Bar charts. Here the values are represented only by one dimension, generally the length of the bar.

ii) *Areal or Two-Dimensional Diagrams.* They consist of Rectangles, Sub-divided Rectangles and Squares, the areas of which are proportional to the values of the given quantities. This device is used to represent data having moderately large variations.

iii) *Cubic or Three-Dimensional Diagrams.* They are in the form of Cubes and cylinders, whose volumes are proportional to the values they represent. These diagrams are used when the variation among the values of the data to be portrayed is so large that even the square roots of the values concerned fail to reduce the variation appreciably.

iv) *Pie-Diagrams.* They are in the form of Circles and Sectors. Here the areas of circles or sectors are in proportion to the values they represent or compare.

v) *Pictograms.* They consist of pictures or small symbolic figures representing the statistical data. A pictogram is an effective way of visual comparisons. For example, we can compare the armed strength of various countries by drawing pictures of the number of soldiers, where each pictorial soldier may denote, say, 1,000 soldiers. In a similar way, the production of wheat can be compared by means of the pictures of wheat bags of a specified size. It is essential to repeat the pictures a number of times to represent the differences in magnitudes.

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

i) An appropriate scale consistent with the size of paper available and the size of the data to be represented, should be chosen and indicated either at the side or at the bottom of the diagram. This scale *must* start at zero.

ii) A diagram like a table, must have a title, which should be brief and self-explanatory. A key, footnote or source will also be necessary.

iii) A diagram should be shaded, coloured or cross-hatched to show the different parts, if any.

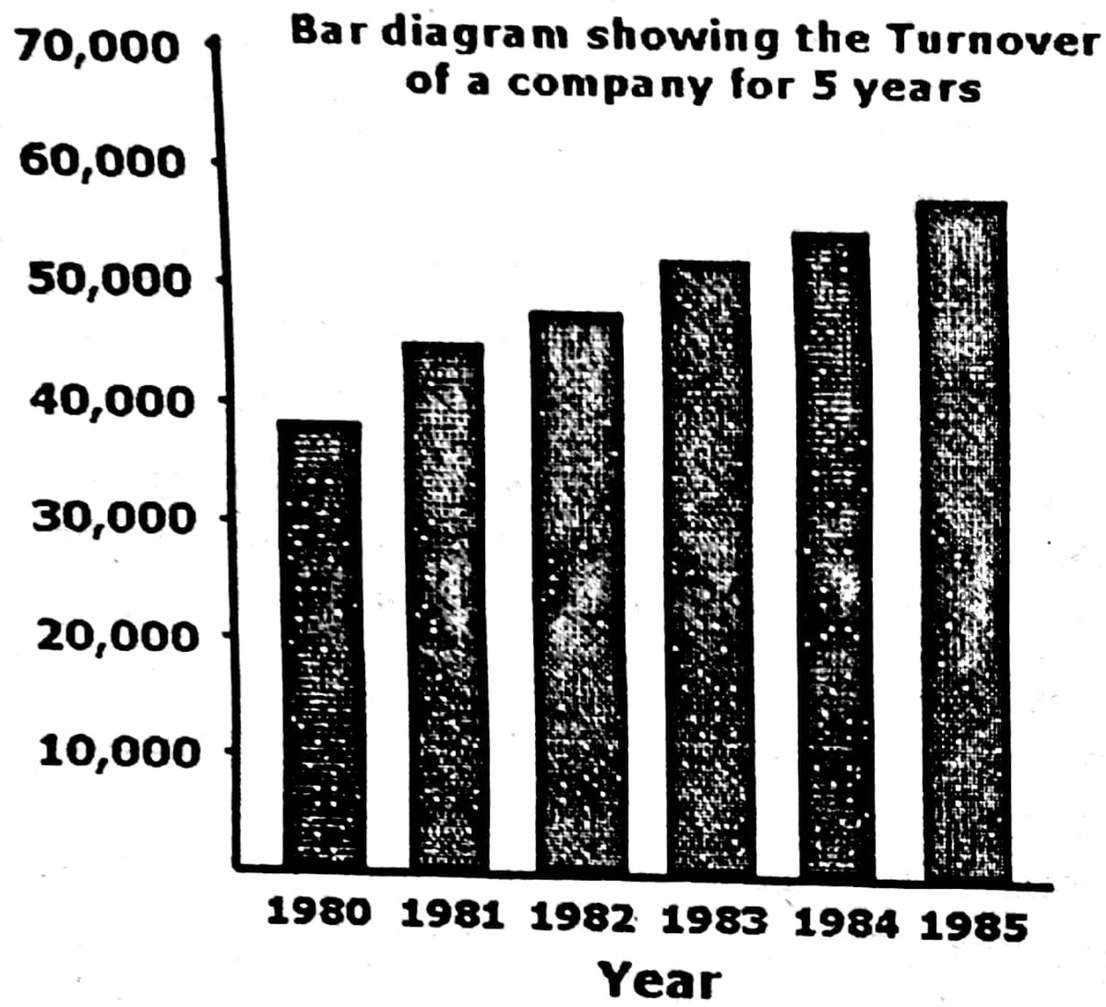
iv) Lettering should be shown horizontally.

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

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

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

The bar chart is drawn below:



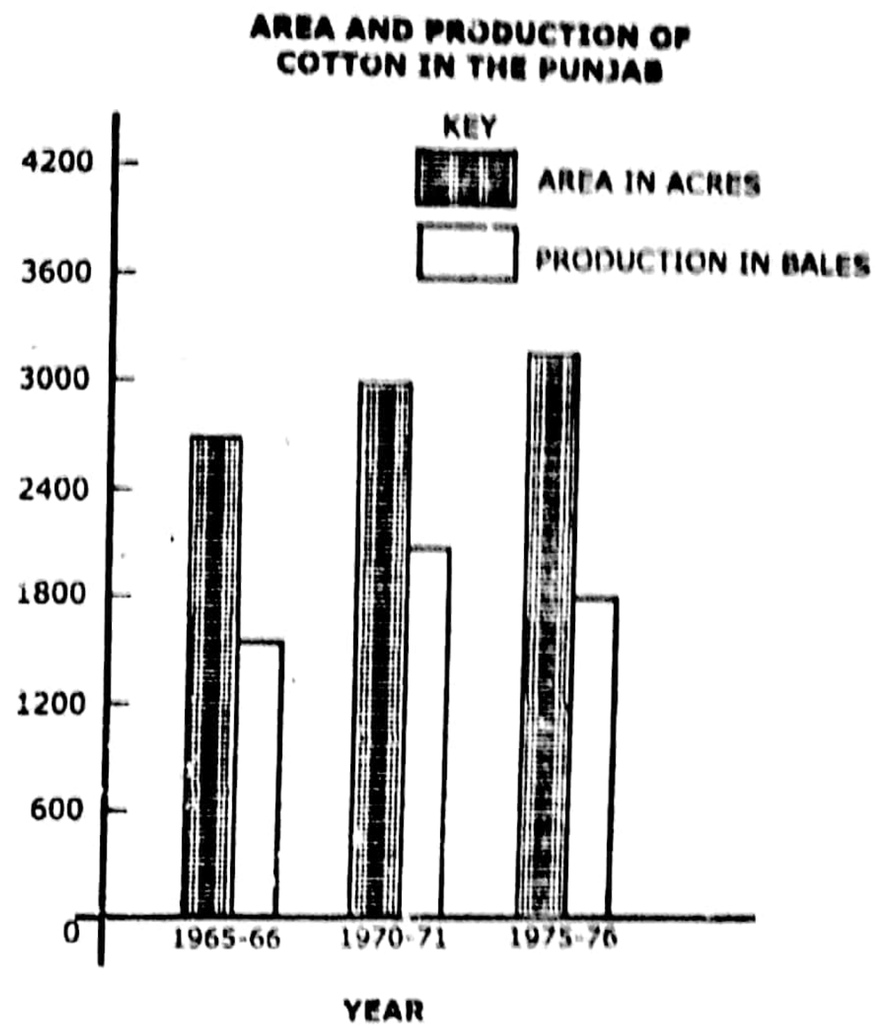
2.7.2 Multiple Bar Chart. A *multiple bar chart* shows two or more characteristics corresponding to the values of a common variable in the form of grouped bars, whose lengths are proportional to the values of the characteristics, and each of which is shaded or coloured differently to aid identification.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

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

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

(Source: Statistical Wing, Agriculture Deptt. Lahore)

The multiple bar charts are drawn below:



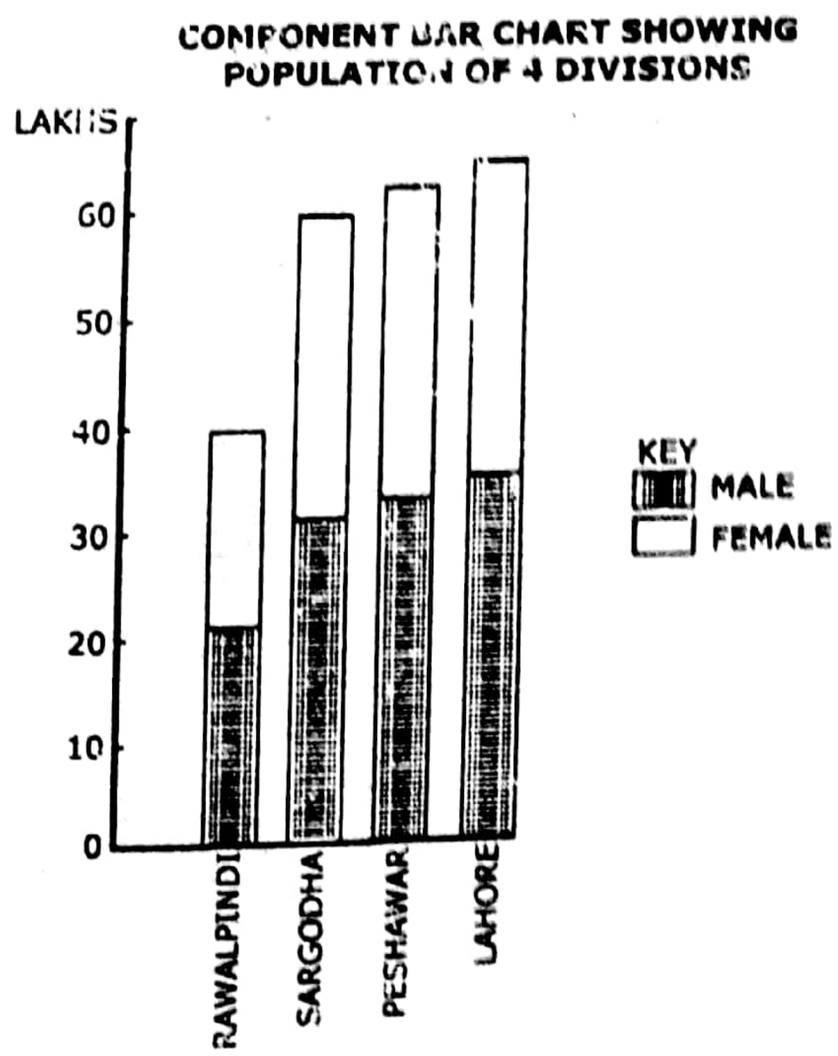
2.7.3 Component Bar Chart. A *component bar chart* is an effective technique in which each bar is divided into two or more sections, proportional in size to the component parts of a total being displayed by each bar. The various component parts shown as sections of the bar, are shaded or coloured differently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

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

(Population in Lakhs)

Division	Both Genders	Male	Female
Peshawar	64	33	31
Rawalpindi	40	21	19
Sargodha	60	32	28
Lahore	65	35	30

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



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

i) change each component into the percentage of the corresponding total,

ii) draw one rectangle for each total, taking equal lengths (100 units) and breadths proportional to the totals,

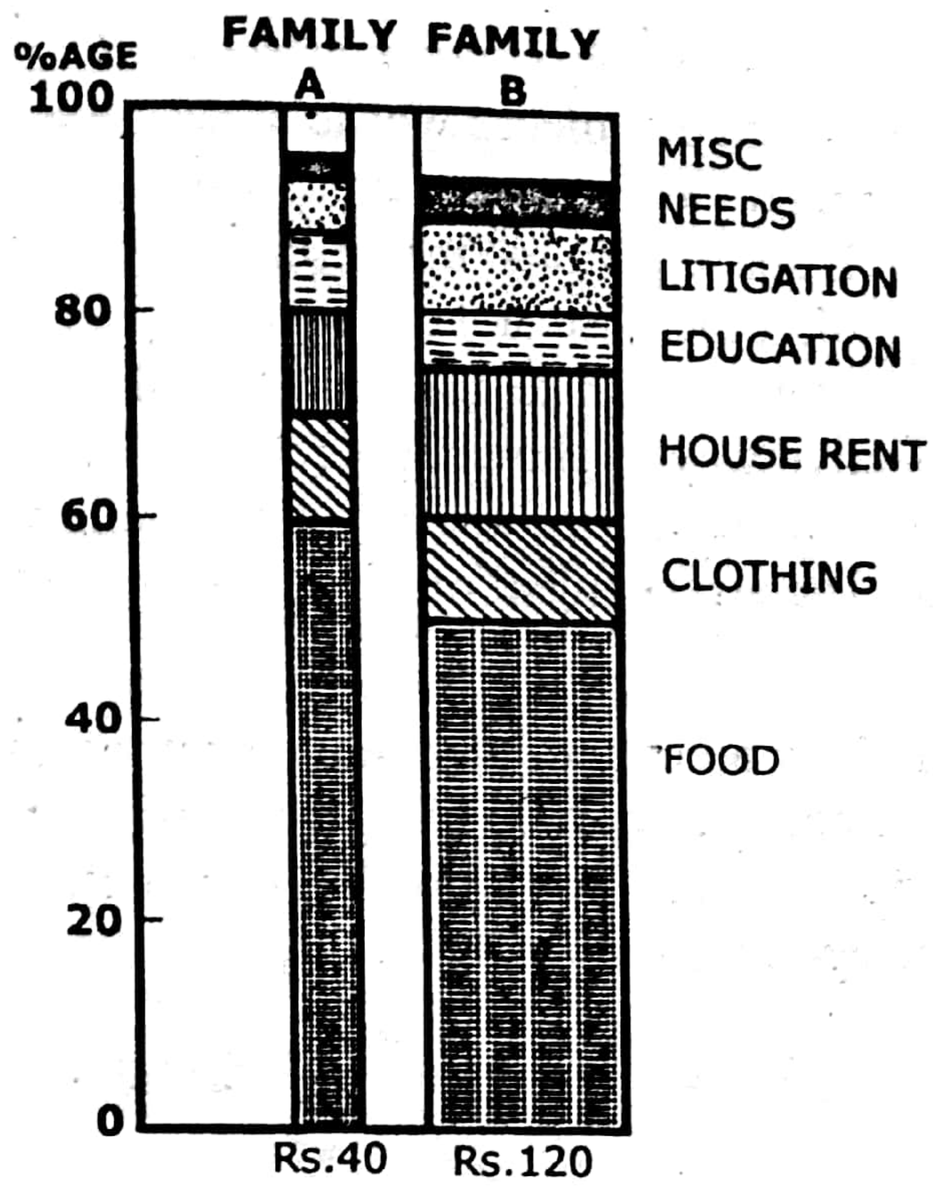
iii) divide every rectangle so drawn into parts equal in number to the number of components. Each part shaded or coloured will represent percentage size of one component.

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

Items of Expenditure	*Family A*	*Family B*
Food	24	60
Clothing	4	14
House Rent	4	16
Education	3	6
Litigation	2	10
Conventional Needs	1	6
Miscellaneous	2	8
Total	40	120

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

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

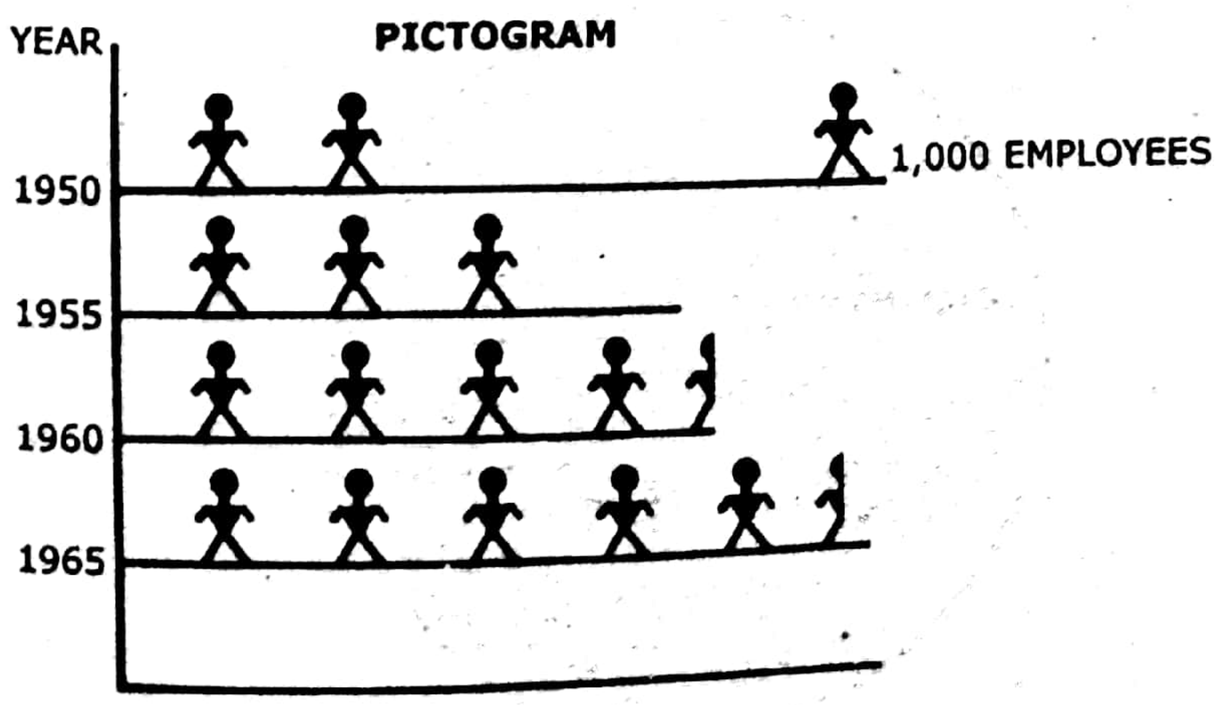


2.7.5 Pictograms. A *pictogram* is a popular device for portraying the statistical data by means of pictures or small symbols. It is said that *a picture is worth ten thousand words.* It is customary to represent a unit value of the data by a standard symbol or a picture and the whole quantity by an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

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

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

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



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

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

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

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

Example 2.14 Represent the total expenditure and expenditures on various items of a family by a pie diagram.

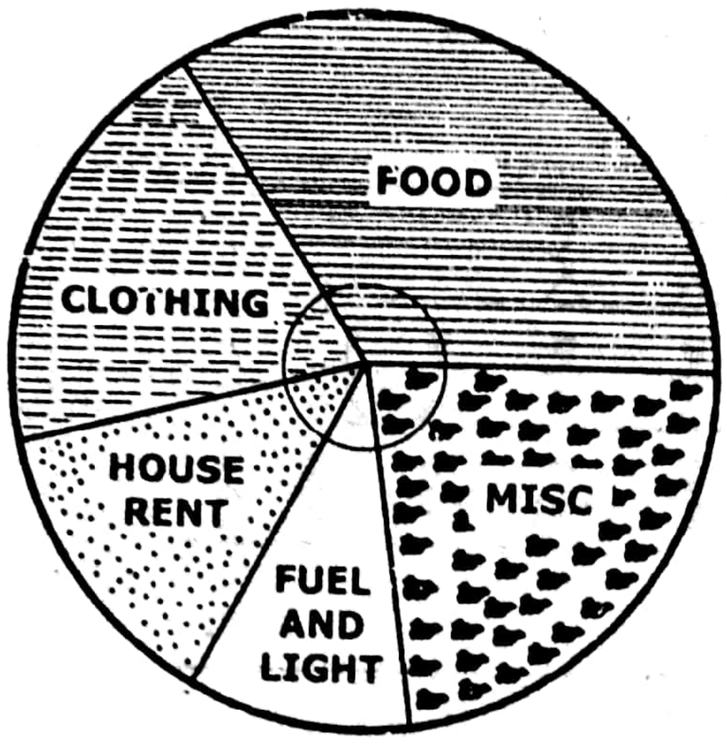
Items:	Food	Clothing	House Rent	Fuel and Light	Misc.
Expenditure: (in Rs.)	50	30	20	15	35

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

Items	Expenditure (in Rs.)	Angles of the Sectors (in Degrees)
Food	50	
Clothing	30	
House Rent	20	
Fuel and Light	15	120
Miscellaneous	35	72
Total	**150**	**360**

"The pie diagram consisting of a circle divided into five sectors defined by angles 120°, 72°, 48°, 36° and 84°, is drawn below:"

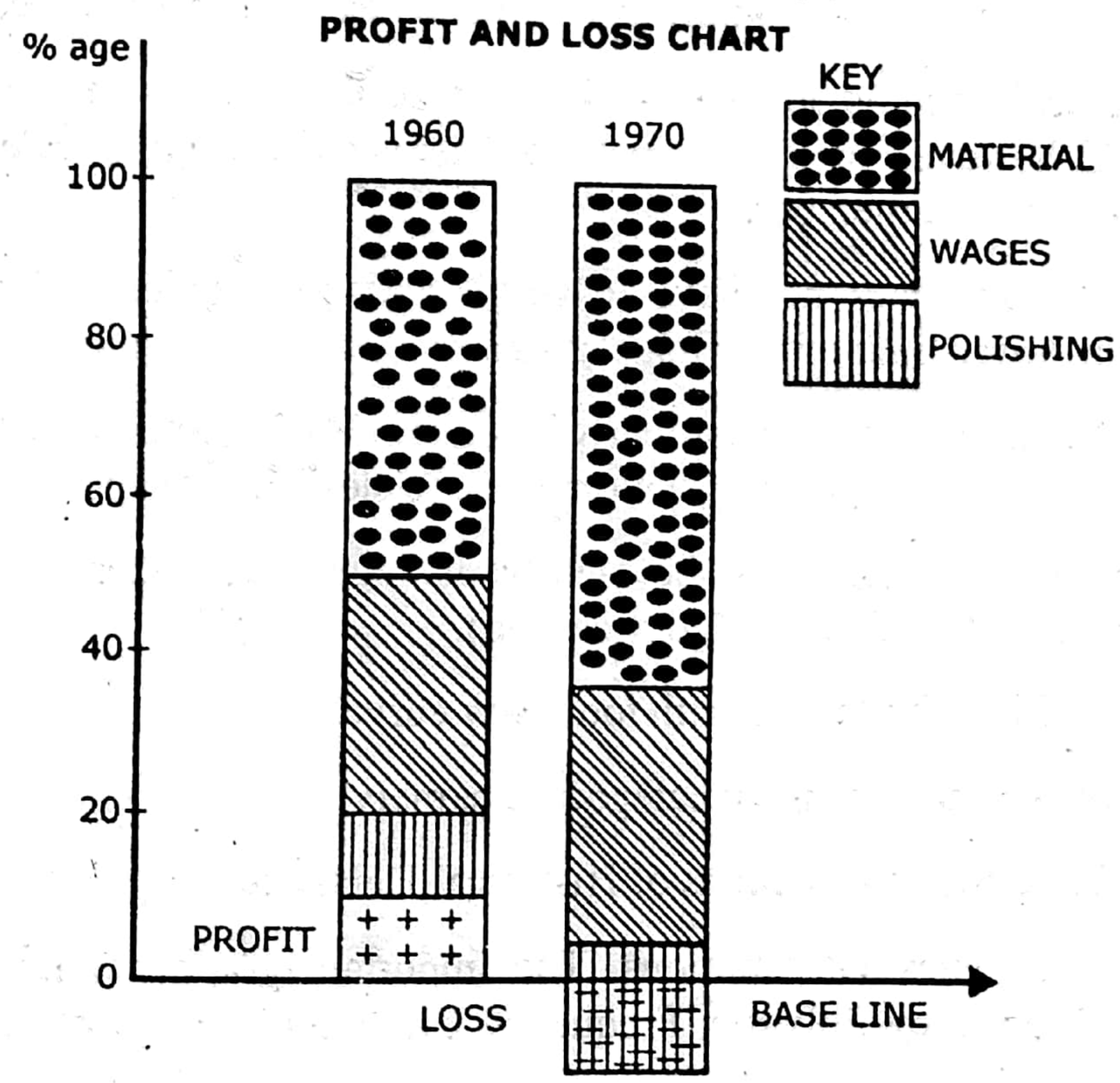
PIE DIAGRAM



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

COST, PROCEEDS, PROFIT OR LOSS PER CHAIR

Particulars	1960	1970
i) Materials	Rs. 10	Rs. 16
ii) Wages	6	8
iii) Polishing, etc.	2	4
Total cost	18	28
Proceeds	20	25
Profit (+) or loss (–)	+2	-3



A pie chart may also be used for this purpose.

2.8 GRAPHS

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

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

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

i) A scale and the form of representation is to be selected in such a way that the true impression of the data to be represented is given by the graph.

ii) Every graph must have a clear and comprehensive title at top. Where necessary, sub-titles should be added.

iii) The source of the data must be given. A key and footnotes should be provided when necessary.

iv) The independent variable should always be placed on the horizontal axis.

v) The vertical scale should always begin with zero, otherwise the graph will give a false impression. If, however, the first item of the data is quite large, a scale-break should be shown between zero and next member.

vi) The horizontal axis does not have to begin with zero unless of course, the independent variable or the lower limit of the first class interval is zero.

vii) The axes of the graph should be properly labelled. Labels should clearly state both the variable and the units, *e.g.* "Distance" and "Kilometer". "Sales" and "Rupees", etc.

viii) Curves if more than one, must be clearly distinguished either by different colours or by differentiated lines (solid, dashed, dot-dashed).

ix) The graph should not be loaded with too many curves.

Graphs can be divided into two main categories, namely:

a) Graphs of Time-Series or Graphs of Historical Data, and

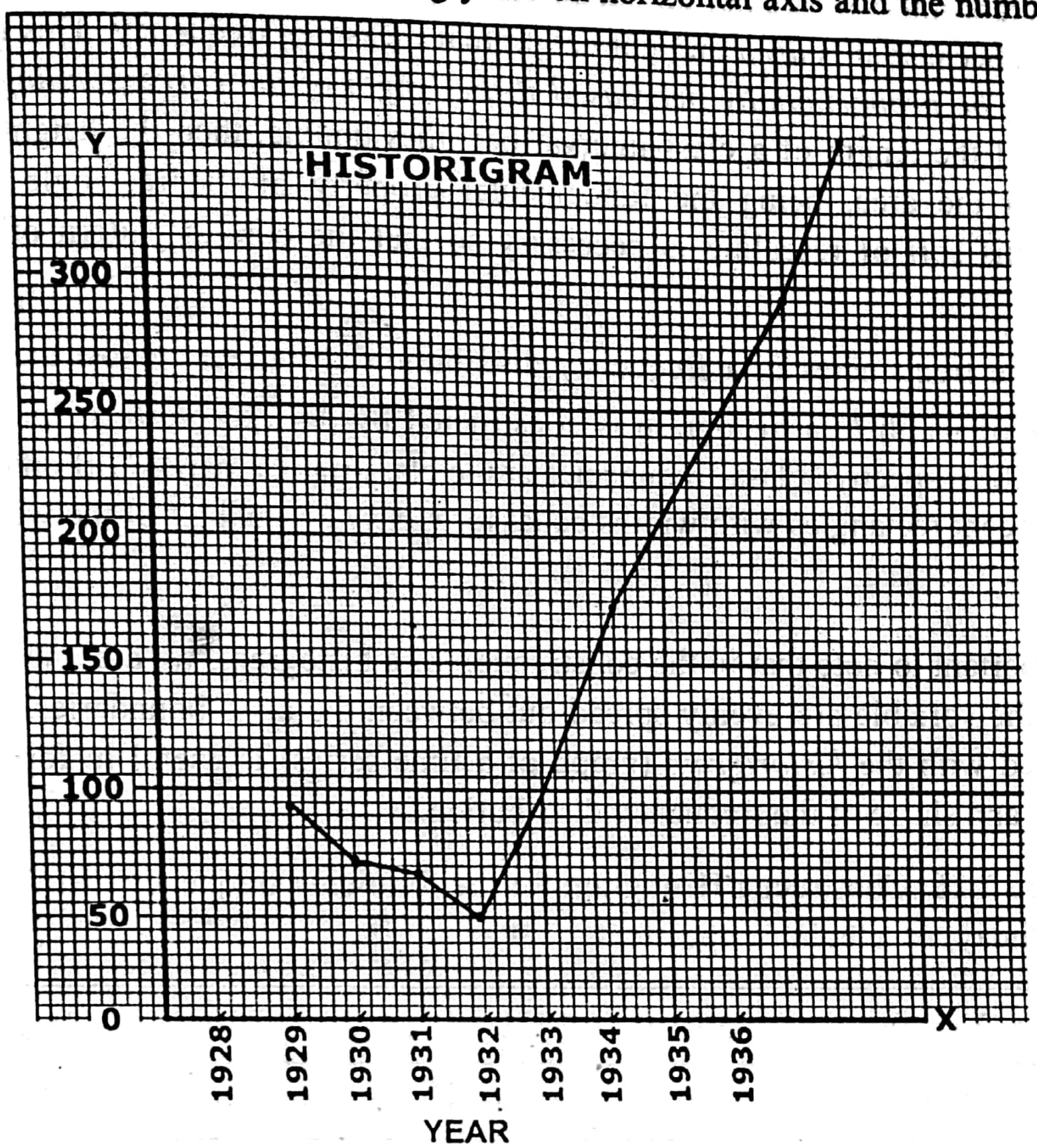
b) Graphs of Frequency Distributions. The important graphs of frequency distributions are Histogram, Frequency Polygon, Frequency Curve and the Cumulative Frequency Curve or Ogive.

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

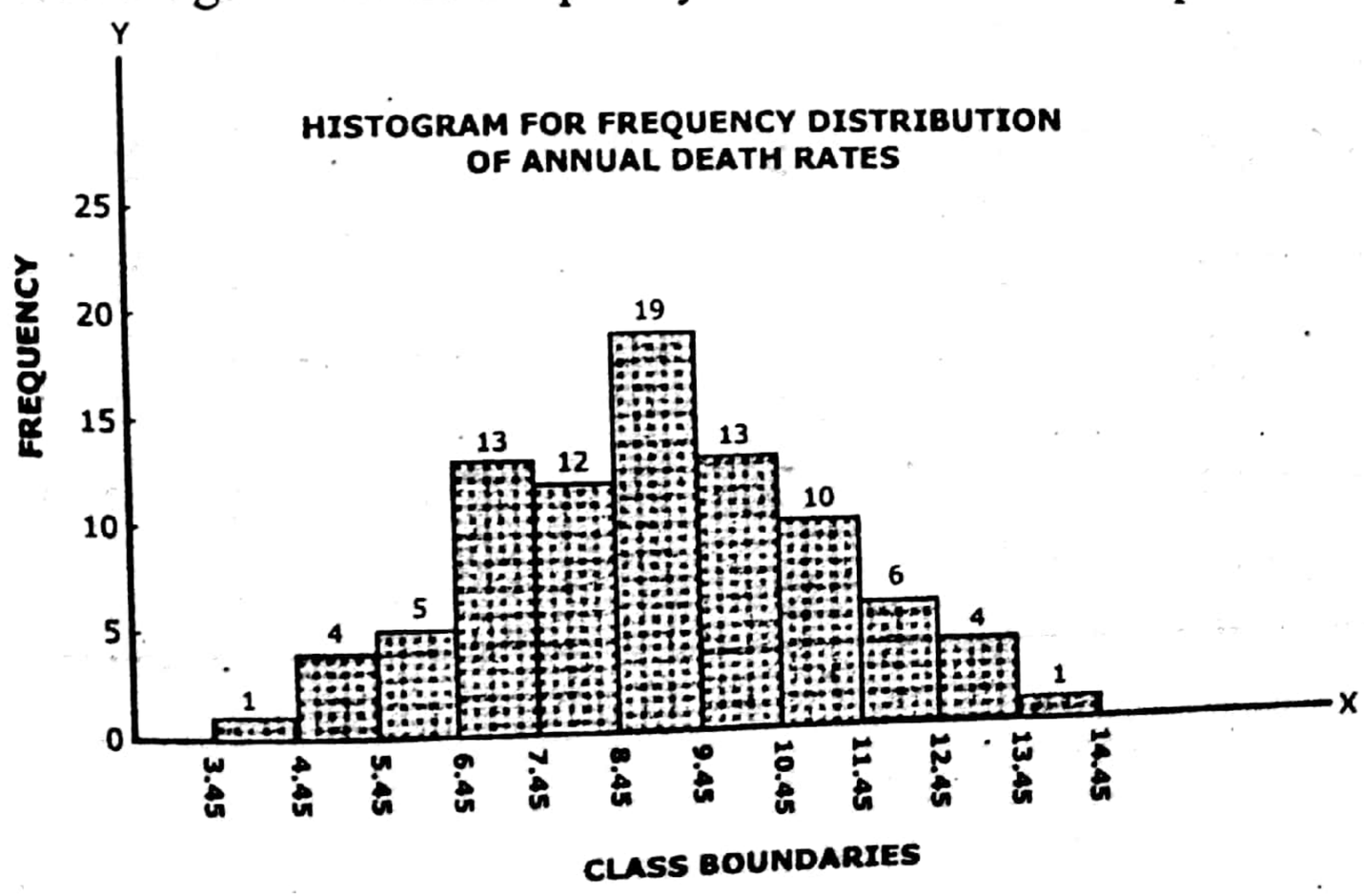
Example 2.15 The following table gives the number of cars produced in Germany during the years 1929-1936. Draw a suitable graphs, *i.e.* Historigram of the series.

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

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



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



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

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

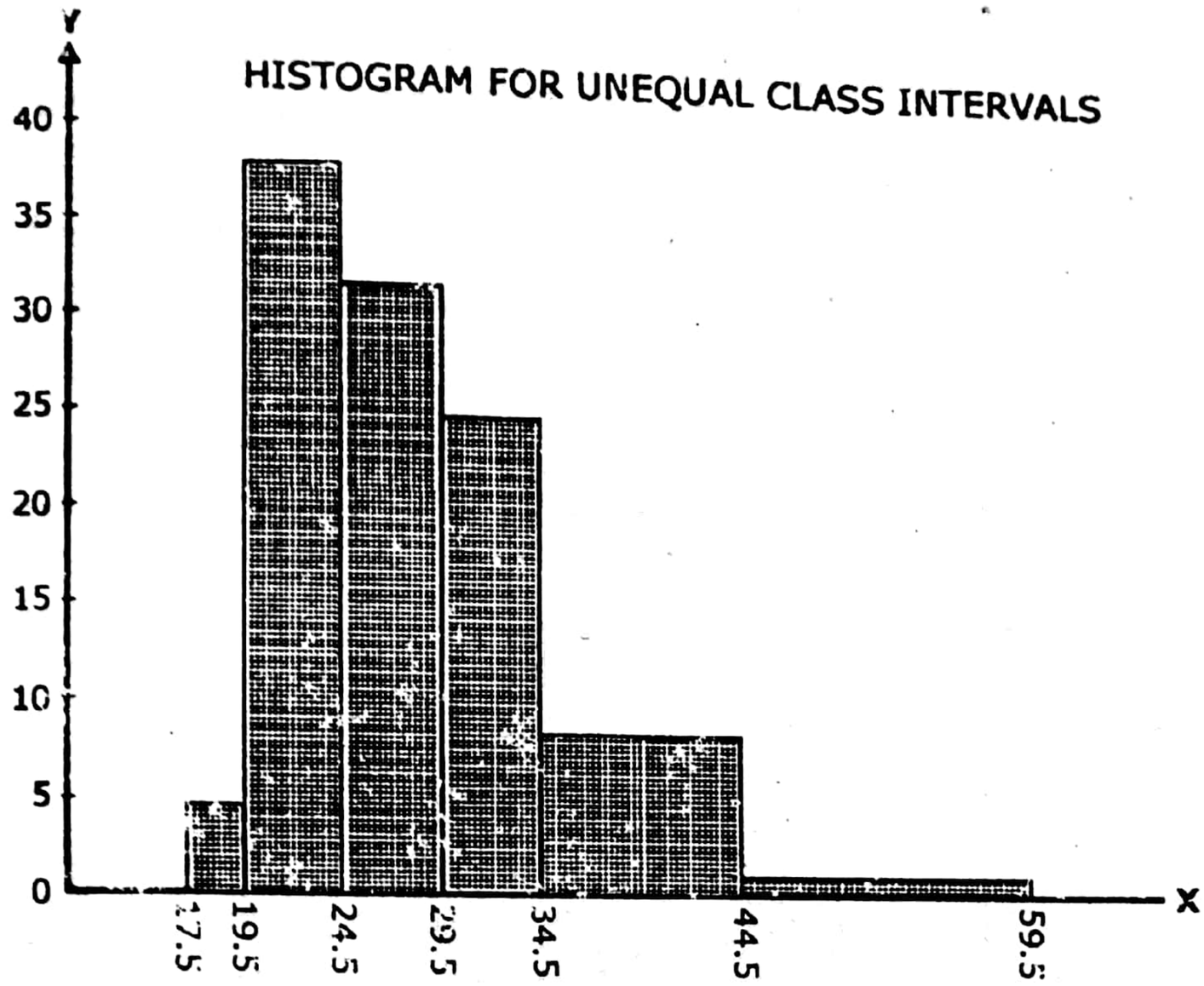
It is important to note that in the construction of a histogram, we assume that within any one class, the values of the variable are evenly spread out between the class-boundaries. A histogram which must not be confused with the *historigram* (graph of a time series) is useful in forming a rough idea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

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

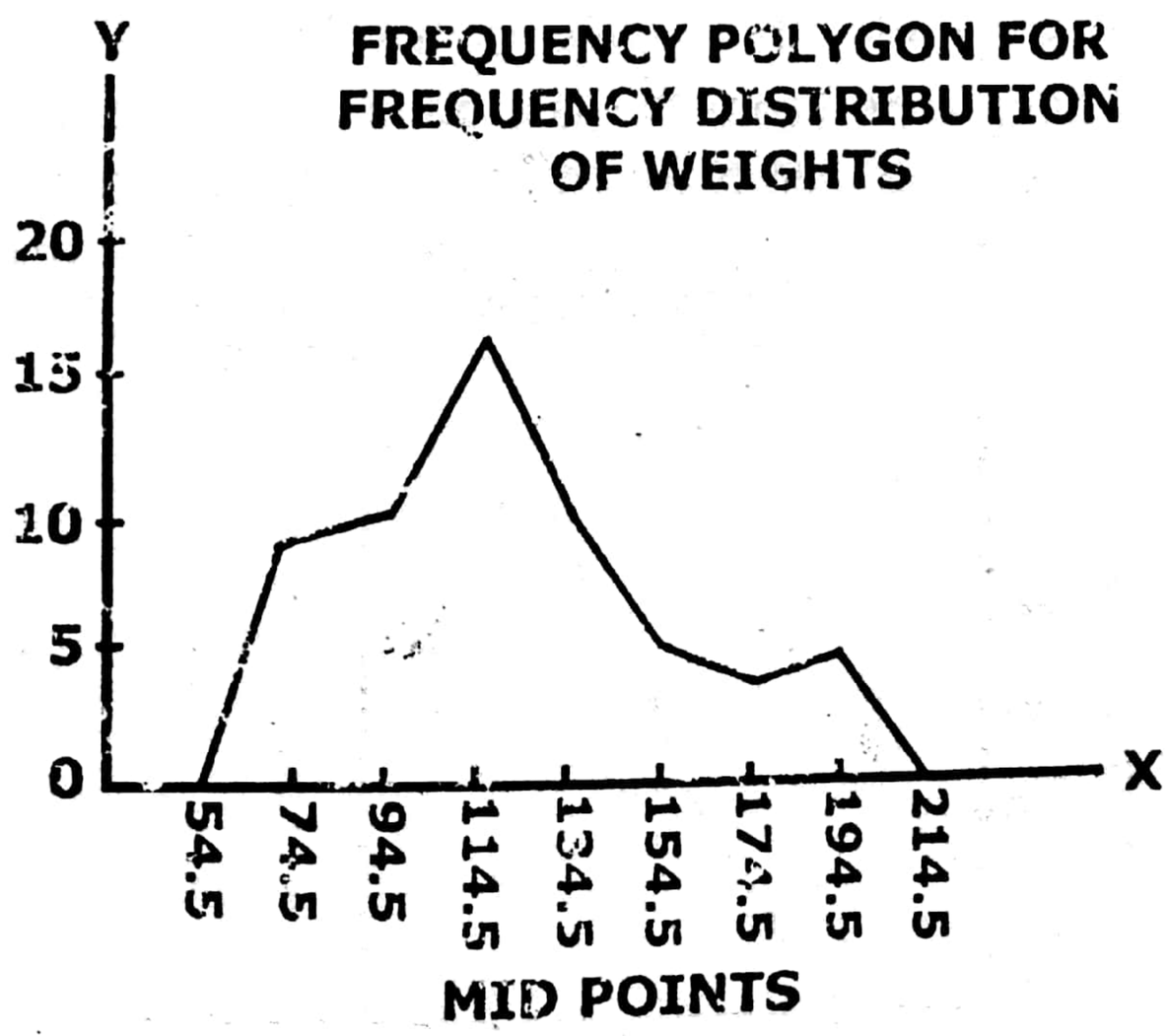
Age (Years)	18–19	20–24	25–29	30–34	35–44	45–59
No. of Operators	9	188	160	123	84	15

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

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

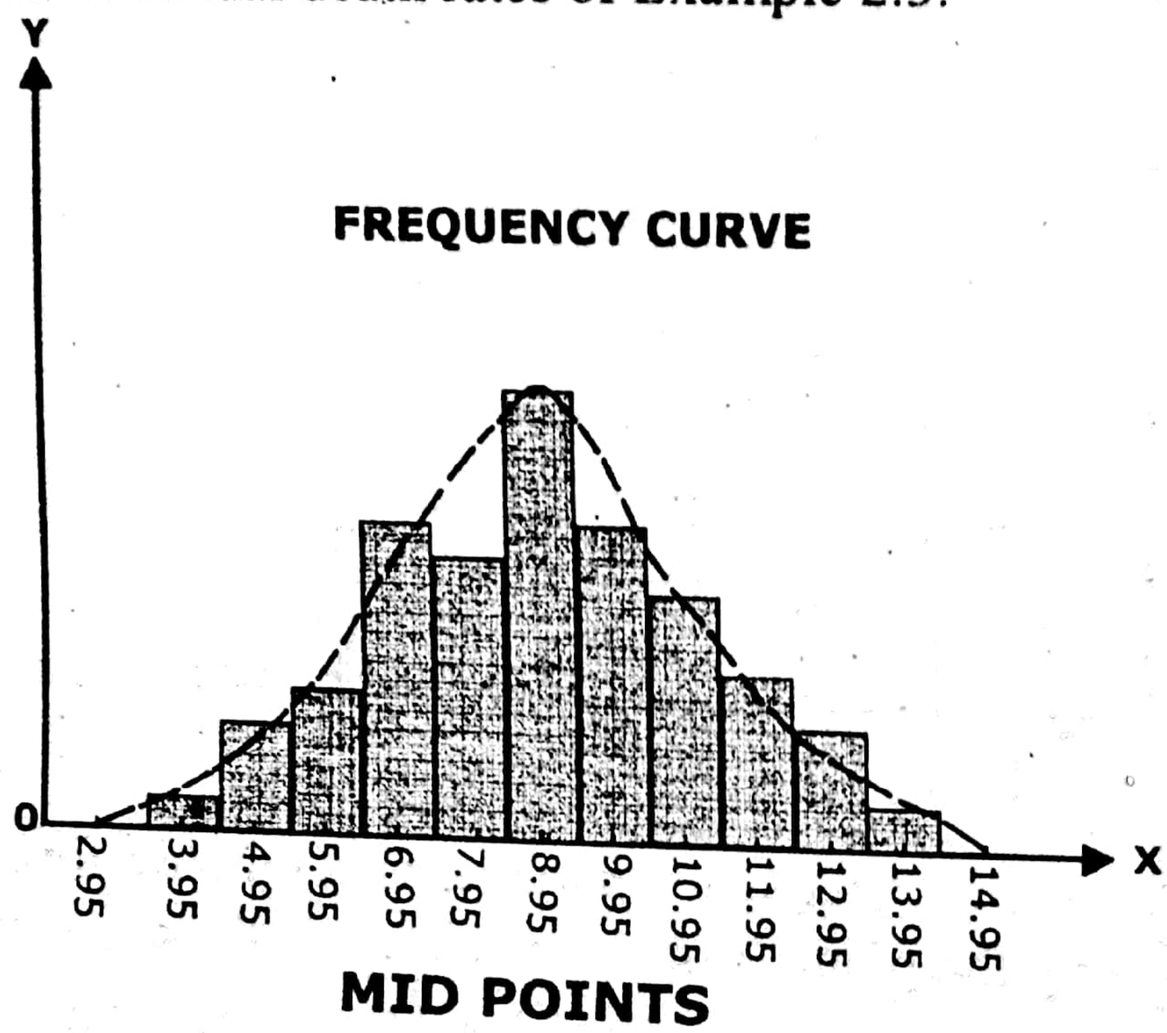


2.8.3 Frequency Polygon. A *frequency polygon* is a graphic form of a frequency distribution, which is constructed by plotting the points (x_i, f_i) where x_i is the class-mark of the ith class and f_i is the corresponding frequency, and then connecting them by straight line segments provided the class-intervals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

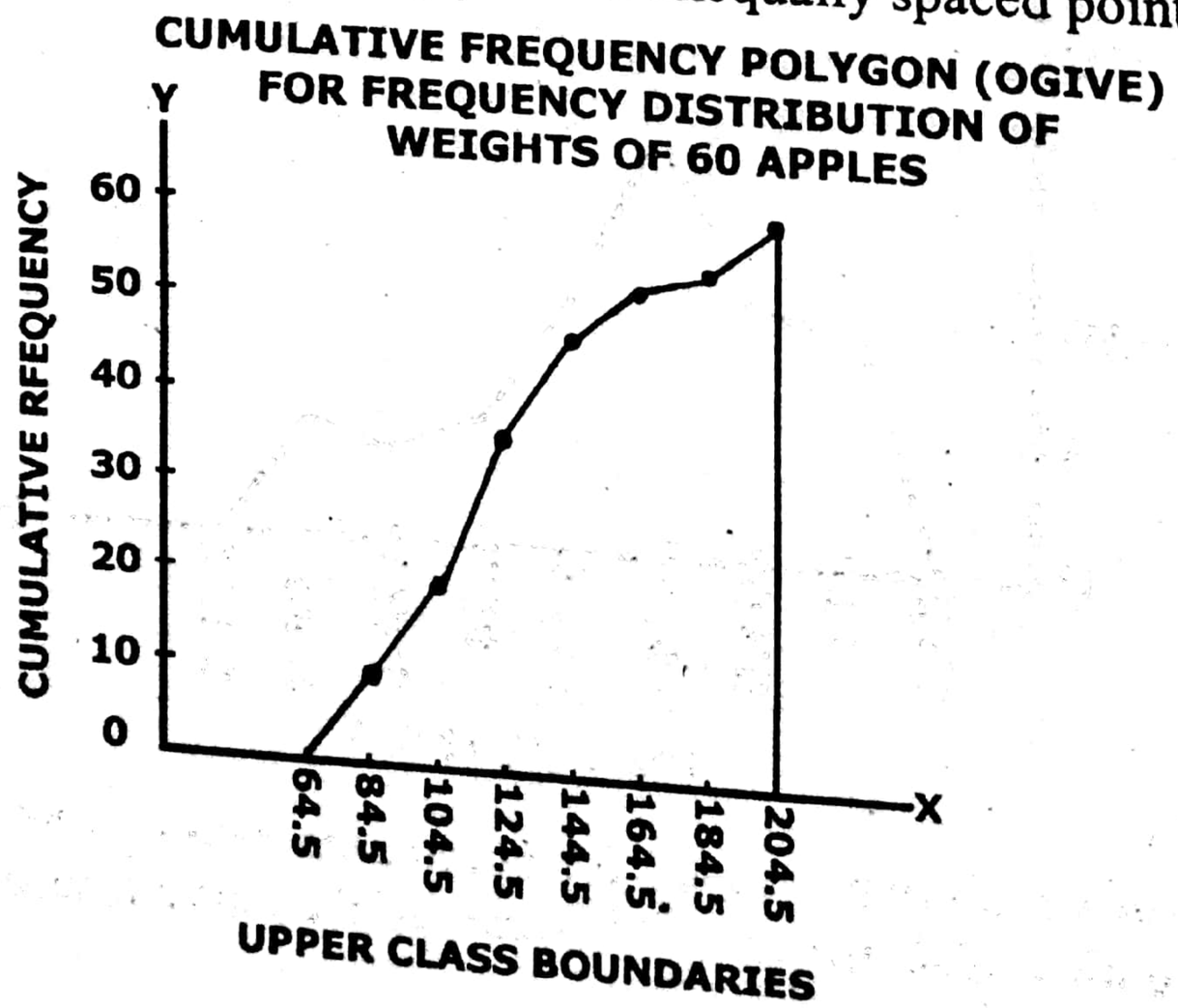


A frequency polygon which can be used for comparing two or more data sets, gives roughly the position of the *mode*, some idea of *skewness* and *kurtosis* of the curve (these terms are defined later).

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



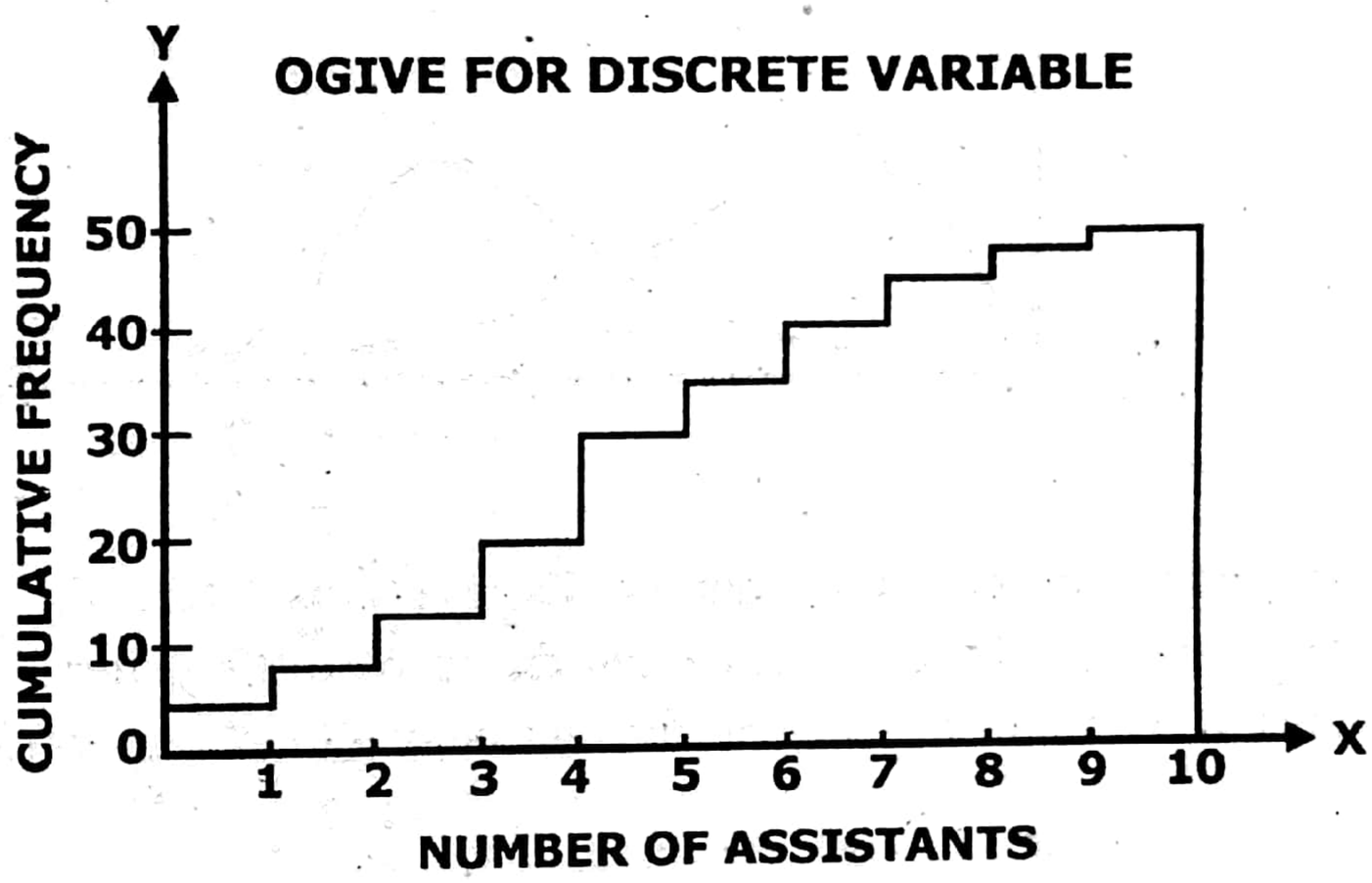
2.8.5 Cumulative Frequency Polygon or Ogive. A *cumulative frequency polygon*, popularly known as *Ogive* (rhymes with "alive" and pronounced *o' jiv*) is a graph obtained by plotting the cumulated frequencies of a distribution against the upper or lower class boundaries depending upon whether the cumulation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



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

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

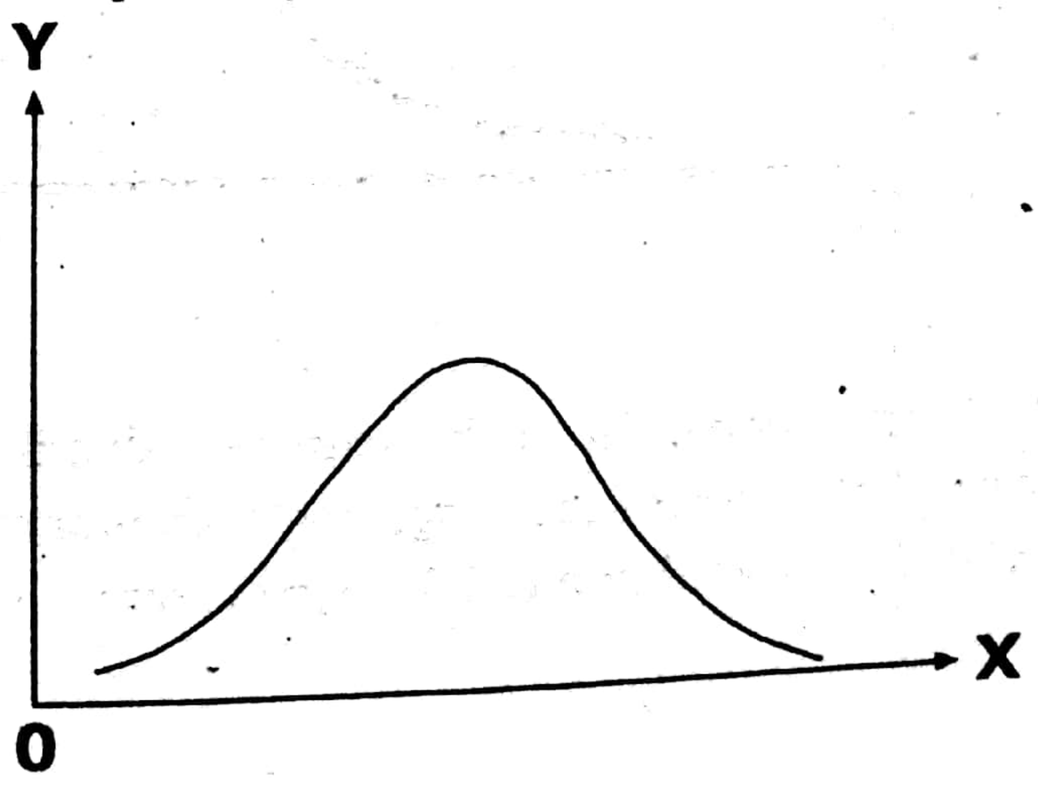
2.8.6 Ogive for a Discrete Variable. When a variable X is discrete, its cumulative frequency polygon consists of horizontal line segments between any two successive values and has a jump of height f_i at each value of x_i. In other words, the cumulative distribution increases only in jumps and is constant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



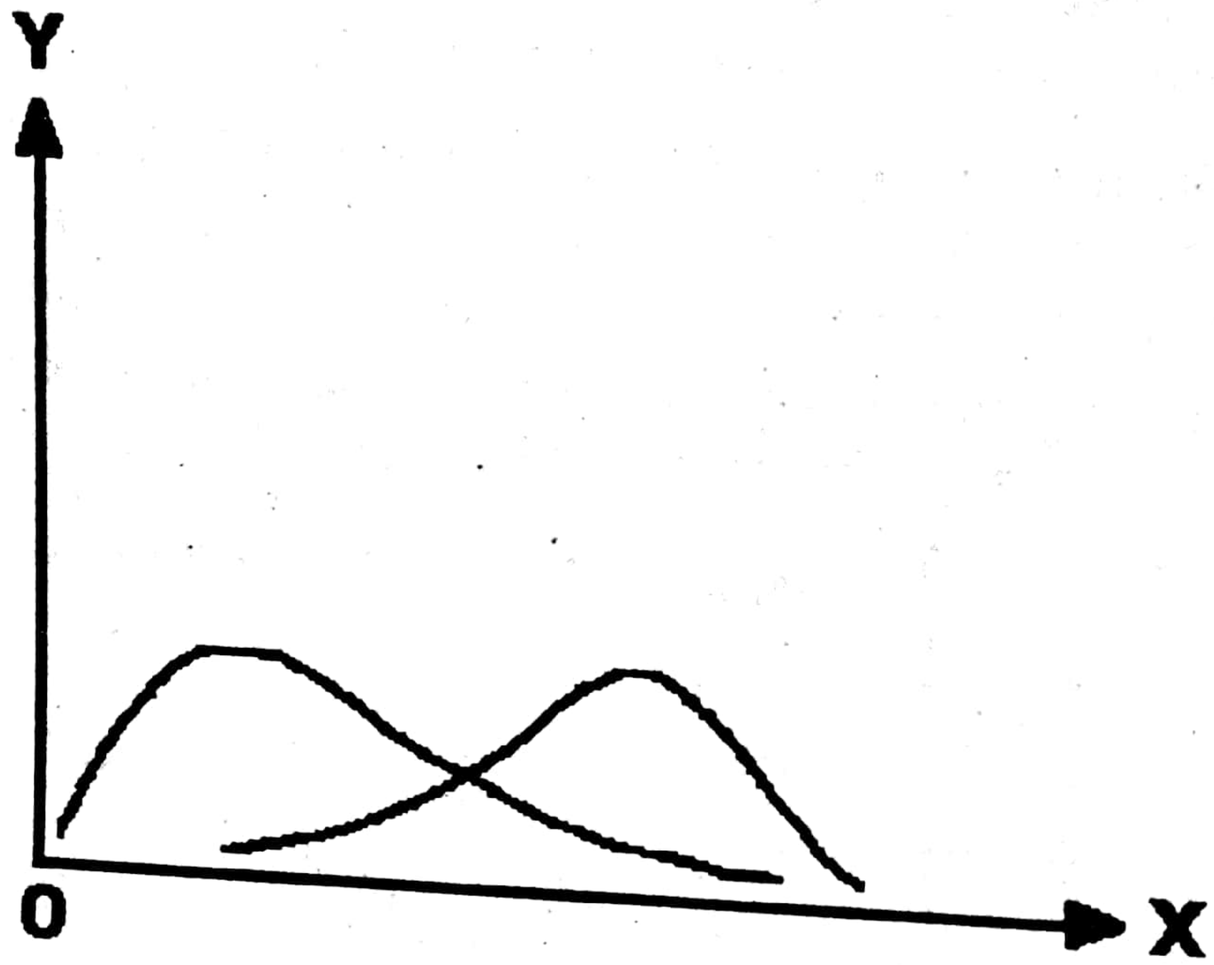
This graph shows that the cumulative frequency polygon is stepped. Such a function is called a *step-function.*

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

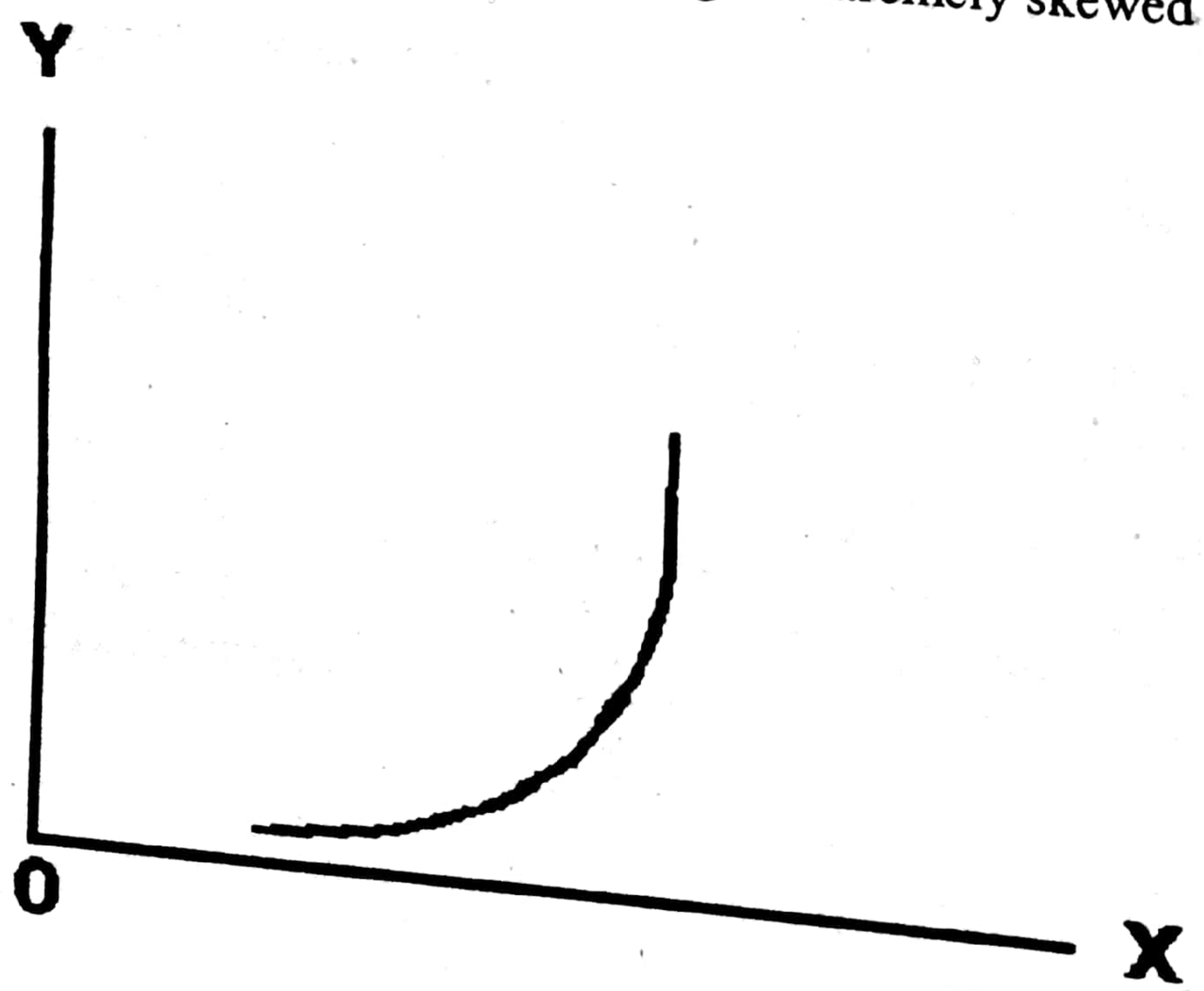
i) **The Symmetrical Distributions.** A frequency distribution or curve is said to be symmetrical if values equidistant from a central maximum have the same frequencies, i.e. the curve can be folded along the central maximum in such a way that the two halves of the curve coincide. The *Normal curve*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



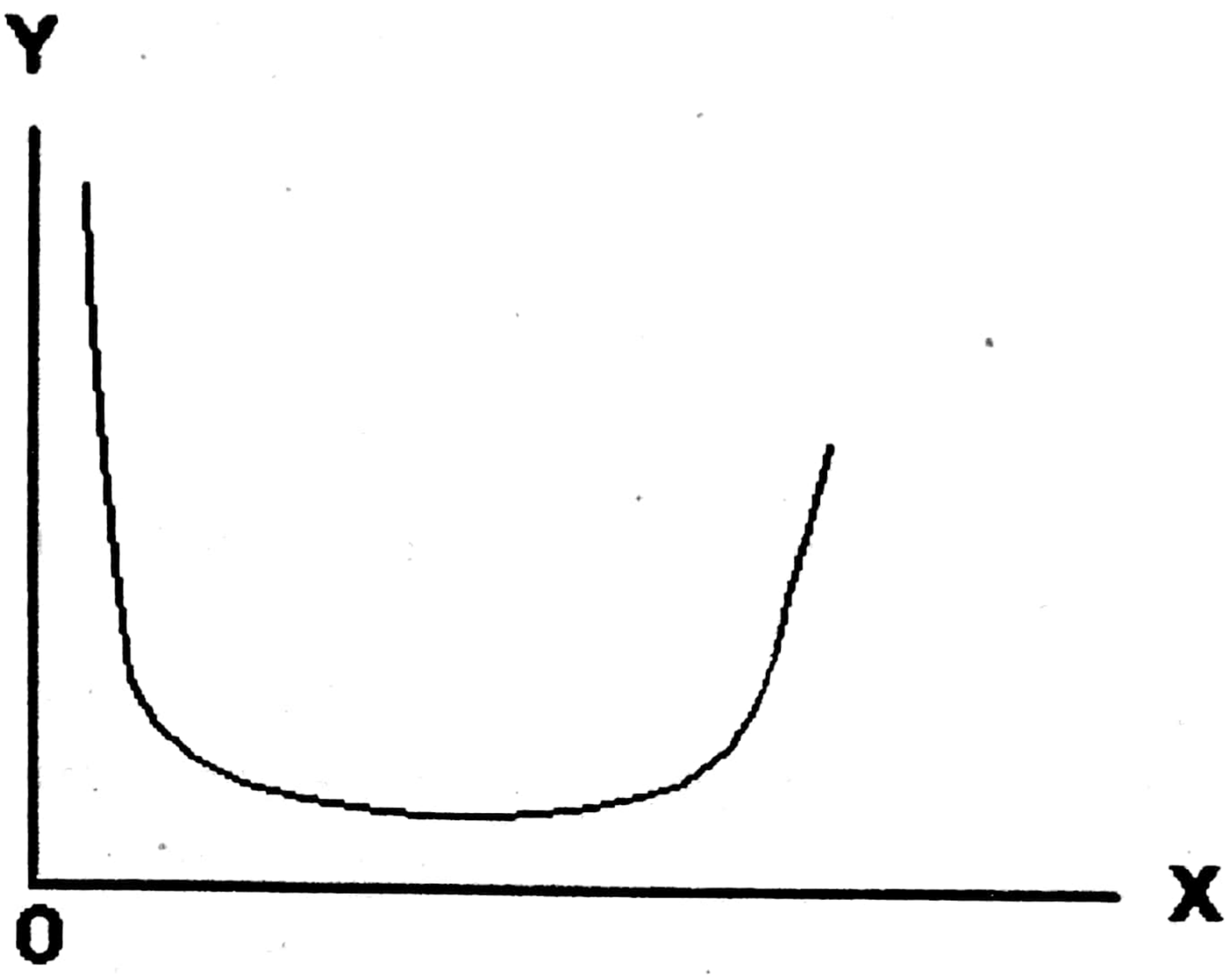
ii) **The Moderately Skewed or Asymmetrical Distributions.** A frequency distribution or curve is said to be *skewed* when it departs from symmetry. Here the frequencies tend to pile up at one end or the order end of the distribution or curve. This is the most common pattern encountered in practice.



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



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



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

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

i) the relative rates of change are to be compared;

ii) visual comparisons are to be made between two or more series which differ widely in magnitude: and

iii) the data are to be examined to see whether they are characterized by a constant rate of change.

A *ratio chart* possesses the following characteristics:

i) There is no zero line on the logarithmic scale as the logarithm of zero is minus infinity.

ii) A geometric progression when plotted on semi-logarithmic paper, forms a straight line, as the logarithms of a geometric progression form an arithmetic progression.

iii) The slope of the logarithmic scale variable indicates the rate at which the variable is changing (i.e. increasing or decreasing).

iv) In case of two or more curves, the curve having the steepest slope, has the largest rate of change.

v) Equal slopes (in case of parallel curves) indicate equal rates of change.

EXERCISES

OBJECTIVE

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

i) The term cross-sectional data refers to data that may change overtime.

ii) The frequency distribution represents data in a condensed form.

iii) The data presented in an array does not allow us to locate the largest and smallest values in the data set.

iv) The classes in any frequency distribution are generally not mutually exclusive.

v) For nominally or ordinary scaled data the frequency distribution cannot be constructed.

vi) Frequency distribution of continuous data may be represented diagrammatically.

vii) Frequency distribution can be presented graphically by using both histogram and historigram.

viii) Time series data can be tabulated using frequency distributions.

ix) Simple bar diagram is used for two-dimensional comparisons.

x) The width of a bar in histogram represents the frequency rather than the value of a variable.

xi) Class marks are the lower limits of each class.

xii) The lower class limit is the middle possible data value for a class.

xiii) The sum of relative frequencies in a relative frequency distribution should always equal 100.

xiv) A pie chart can be used to display quantitative data.

xv) A shape of the frequency distribution and the relative frequency distribution always will be different.

b) MULTIPLE CHOICE QUESTIONS.

i) Which of following is not an example of condensed data?

a) frequency distribution b) data array c) histogram d) polygon

ii) In the construction of a frequency distribution the steps are to:

a) decide the number of classes

b) arranging the data in ascending / descending order

c) locate the smallest and largest values in a data set

d) all of above

iii) The number of classes in a frequency distribution generally should be

a) less than five
b) more than twenty
c) between five and twenty
d) between ten and twenty

iv) As the number of observations and classes increase, the shape of the frequency polygon:

a) remains same
b) tends to smooth
c) become more erratic
d) none of them

v) A cumulative frequency distribution is graphically represented by:

a) frequency curve
b) frequency polygon
c) pie chart
d) ogive

vi) A relative frequency distribution presents frequencies in terms of:

a) whole numbers
b) percentages
c) fractions
d) all of above

vii) A diagram that presents properties that look like slices of a pizza is known as:

a) a bar diagram
b) a component bar diagram
c) a histogram
d) a pie diagram

viii) Observed data organized into tabular form is called:

a) a bar chart
b) a pie chart
c) a frequency polygon
d) a frequency distribution

ix) The number of occurrences of a data value is called:

a) the frequency
b) the cumulative frequency
c) the relative frequency
d) all of above

x) In the following stem-and-leaf diagram:

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

The number that occurred the most is

a) 2
b) 55
c) 42
d) 5

SUBJECTIVE

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

2.2 Define the terms "Classification" and "Tabulation". Outline the main steps in tabulation. What do you mean by *captions, stubs, title* and *prefatory notes*? (P.U., B.A./B.Sc. 1983)

2.3 What is a statistical table? What are different types of tables? Explain the different parts of a table and the main points to be kept in mind in their construction. (P.U., B.A./B.Sc. 1970)

2.4 Represent the data given in the following paragraph in the form of a table, so as to bring out clearly all the facts, indicating the source and bearing suitable title:

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

2.5 a) Write short notes on:

Class-frequency, Class-Interval, Class limits, Class Marks, Size of Class-Interval and Sturges' rule.

b) Determine class boundaries, class limits and class marks for the first and last classes in respect of the following:

i) Weights of 300 entering freshmen ranged from 98 to 226 pounds, correct to the nearest pound.

ii) The thickness of 460 washers ranged from 0.421 to 0.563 inches.

c) A sample consists of 34 observations, each recorded as correct to the nearest integer, ranging in value from 201 to 337. If it is decided to use seven classes of width 20 units and to begin the first one at 199.5, find the class boundaries, limits and marks of the seven classes.

(I.U. M.A. Econ., 1992)

2.6 a) What is meant by a frequency distribution? Describe briefly the main steps in the preparation of a frequency table from raw data.

(P.U., B.A./B.Sc. 1974)

b) Prepare a frequency table for the price data given below, taking 5 units as the width of class-interval.

100	96	92	88	86	84	82	80	78	91
87	83	79	77	75	73	71	69	58	56
73	50	57	55	53	51	48	46	63	59
55	51	49	47	45	43	41	58	54	50
56	44	42	40	38	36	46	53	50	43

2.7 a) Why are frequency distributions constructed? What are the rules to be observed in making a frequency distribution from ungrouped data?

b) A record was made of the number of absences per day from a factory over 35 days with the following results:

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

i) On how many days were there *fewer than* 4 *people* absent?

ii) On how many days were there *at least* 4 *people* absent?

iii) What is the total number of absences over the whole 35 days?

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

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

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

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

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

91	124	109	129	141	107	86	76	118	111
99	99	114	100	88	108	87	101	101	71
63	121	122	111	119	77	127	61	133	68
77	177	110	95	96	96	86	106	119	79
81	127	86	153	79	129	151	89	143	147
90	142	105	94	125	96	99	138	145	113
129	87	113	110	144	91	106	104	97	115
100	117	73	134	108	102	123	106	119	104
101	120	112	138	140	103	96	136	78	83
75	100	113	114	109	116	109	116	104	128

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

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

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

2.11 The following figures give the number of children born to 50 women:

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

Construct an ungrouped frequency distribution of these data.

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

"To forgive an injury is often considered to be a sign of weakness; it is really a sign of strength. It is easy to allow oneself to be carried away by resentment and hate into an act of vengeance; but it takes a strong character to restrain those natural passions. The man who forgives an injury proves himself to be the superior of the man who wronged him, and puts the wrong-doer to shame. Forgiveness may even turn a foe into a friend. So mercy is the noblest form of revenge."

2.13 The weights of 50 football players are listed below:

193	240	217	283	268	212	251	263	275	208
230	288	259	225	252	236	243	247	280	234
250	236	277	218	245	268	231	269	224	259
258	231	255	228	202	245	246	271	249	255
265	235	243	219	255	245	238	257	254	284

Make a stem-and leaf display for the data and convert it to a frequency table with 10 classes, beginning with 190.

2.14 Make a stem-and-leaf table for the following data. Using 8.0 as the lower limit of the first class and with a width of 1 unit, convert it to a frequency distribution.

9.0	10.2	11.3	12.1	10.7	13.8	10.8
11.6	13.6	16.4	11.0	15.8	9.3	13.7
11.7	11.0	8.0	12.0	11.5	9.7	11.6
10.1	14.1	10.0	9.9	13.4	15.7	11.5
12.3	9.8	13.0	9.1	8.3	12.9	14.0
10.5	13.2	10.5	10.6	12.5	15.1	12.8
10.4	11.2	9.3	11.7	17.7	13.9	16.9
13.4	11.8	16.8	14.2	11.8	9.6	11.9
8.7	14.7	10.9	17.9	11.5	14.7	15.9
11.8	10.6	12.6	12.6	15.7	14.9	9.9

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

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

(P.U., B.A./B.Sc. 1974)

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

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

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

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

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

Country	Birth Rate	Death Rate
India	33	24
Japan	32	19
Germany	16	10
Egypt	44	24
Australia	20	9
New Zealand	18	8
France	21	16
Russia	38	16

2.21 Represent the following data by rectangular diagram showing percentage of Income spent by two families on different items of expenditure.

Family-budgets of two families

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

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

Items of Expenditure	Family A (Rs.)	Family B (Rs.)	Family C (Rs.)
Food Articles	43	87	120
Clothing	18	17	25
Recreation	3	10	12
Education	5	9	15
Rent	10	21	17
Miscellaneous	6	15	11

2.23 Represent the following data by means of a pictogram:

(a)

Industry	No. of Employees (000)
Marine	96
Forest	187
Mineral	290
Farm	635

(b)

Year	Production of Vans
1982	2040
1983	2996
1984	4319
1985	6324

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

Item	Expenditure in Rs.
Food	190
Clothing	64
Rent	100
Medical care	46
Other items	80

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

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

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

Wheat	106	lakh acres
Gram	30	lakh acres
Barley	15	lakh acres
Pulses	10	lakh acres
Fodder	25	lakh acres
Other crops	14	lakh acres

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

(P.U., B.A./B.Sc., 1968)

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

Cost per ton disposed commercially

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

2.27 a) Describe the Graphical Methods used in Statistics, giving their respective advantages and disadvantages.

(P.U., M.A. Econ. 1981)

b) Draw up a list of rules for the construction of graphs.

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

2.28 State the general rules which should be borne in mind in the construction of graphs. Draw a suitable graph of the following time series.

Year	Gross Profit (Rs.)	Expenses (Rs.)
1931	7,9000	2,700
1932	5,550	1,700
1933	4,800	1,500
1934	4,500	1,150
1935	6,550	4,300
1936	9,100	5,800
1937	8,500	3,700
1938	7,300	3,900
1939	6,500	4,800
1940	6,200	3,650

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

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

2.30 a) Explain with the help of diagrams the difference between a frequency polygon, a histogram and an ogive. (P.U., B.A. (Part-I), 1961)

b) Construct (i) a Histogram, (ii) a Relative frequency polygon and (iii) an Ogive for the following frequency distribution of the heights of 100 male students at Islamia University, Bahawalpur.

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

(I.U. M.A., Econ. 1989)

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

Earnings	180–184	185–189	190–194	195–199	200–204	205–209	210–214	215–219
Workers	10	24	30	36	40	29	23	8

State how you would construct the histogram if the class-intervals were unequal in size.

(Engg. University, B.Sc. Final, 1977)

2.32 a) Define the statistical term *Histogram*.

b) Explain the method of constructing histograms when the class intervals are unequal.

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

No. of certificate held	No. of members
1 – 50	10
51 – 100	15
101 – 150	30
151 – 200	40
201 – 300	120
301 – 400	100
401 – 500	85

Construct a histogram of the distribution of savings certificates. (P.U., B. Com., 1961)

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

Degree of Cloudiness:	10	9	8	7	6	5	4	3	2	1	0
Frequency:	580	150	196	75	55	40	45	68	75	130	220

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

2.34 a) Draw a histogram illustrating the following data:

Age nearest birth day	20–24	25–29	30–39	40–44	45–49	50–54	55–64
Number of men	1	2	26	22	20	15	14

(P.U., B.A./B.Sc. 1973, 75, 78; M.A. Econ. 1980, P.C.S., 1971)

b) Given the following frequency table of the heights in centimeters of 1,000 students, draw its histogram showing the class marks:

x	155–157, 158–160, etc.														
f	4	8	26	53	89	146	188	181	125	92	60	22	4	1	1

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

2.35 a) What is a cumulative frequency curve? How does it differ from the ordinary curve? (P.U., M.A. Econ. 1985)

b) Describe the common types of frequency curves. Indicates their shapes.

2.36 Construct an Ogive from the following table:

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

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

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

Time taken (minutes)	< 5,	< 10,	< 15,	< 20,	< 25,	< 30,	< 35,	< 40,	< 45
Cum. Frequency (F)	28	45	81	143	280	349	374	395	400

a) Draw a cumulative frequency curve and estimate how many pupils took less than 18 minutes.

b) 6% of the pupils took x minutes or longer. Find x.

c) Take equal class-intervals of 0–, 5–, 10–, etc., construct a frequency distribution and draw a histogram.

2.38 Describe in detail the method of drawing Ratio Charts and explain their uses in economic statistics.

(P.U., M.A. Econ., 1967)

2.39 Explain what is meant by a ratio chart and discuss its advantages over the natural scale diagram. Describe and illustrate two practical applications of a ratio chart.

2.40 Toss five coins together and note the number of heads (X). Do this 64 times and count the number of times that $x = 0, 1, 2, 3, 4, 5$. Construct a frequency polygon and an Ogive to represent these results.

2.41 a) Distinguish between primary and secondary data. Describe briefly the methods of collecting primary data.

b) Find the missing entries in the following frequency distribution table?

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

(P.U., B.A./B.Sc., 2008)

2.42 The following data give the annual earnings (rounded to thousands of dollars) of 35 households.

30	5	29	7	20	31	28	10
13	24	25	7	22	15	15	22
78	79	99	30	80	80	75	63
24	35	25	90	35	33	70	63
35	17	9					

i) Prepare a stem-and-leaf display for these data.

ii) Condense the stem-and-leaf display by grouping the stems as 0–2, 3–5 and 6–9.

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

◆◆◆◆◆◆◆◆◆◆

CHAPTER 3

MEASURES OF CENTRAL TENDENCY OR AVERAGES

MEASURES OF CENTRAL TENDENCY OR AVERAGES

3.1 INTRODUCTION

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

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

3.2 CRITERIA OF A SATISFACTORY AVERAGE

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

i) rigorously defined,

ii) based on all the observations made,

iii) simple to understand and easy to interpret,

iv) quickly and easily calculated,

v) amenable to mathematical treatment,

vi) relatively stable in repeated sampling experiments, and

vii) not unduly influenced by abnormally large or small observations.

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

3.3 TYPES OF AVERAGES

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

3.4 THE ARITHMETIC MEAN

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

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

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

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

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

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

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

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

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

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

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

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

Calculate the arithmetic mean.

The mean is given by

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{45 + 32 + 37 + 46 + 39 + 36 + 41 + 48 + 36}{9}$$

$$= \frac{360}{9} = 40 \text{ marks}$$

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

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

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_n}$$

$$= \frac{\sum x_i w_i}{\sum w_i} \qquad (i = 1, 2, \ldots, n)$$

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

Example 3.2 Calculate the weighted mean from the following data:

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

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

We calculate the weighted mean as below:

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

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

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

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

The sum of the deviations $= \sum (x_i - \bar{x}), \qquad (i = 1, 2, \ldots, n)$

$$= \sum x_i - n\bar{x} \qquad (\because \bar{x} \text{ is constant})$$

$$= \sum x_i - \sum x_i = 0 \qquad (\because \bar{x} = \sum x / n)$$

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

Now $\sum (x_i - a)^2 = \sum (x_i - \bar{x} + \bar{x} - a)^2$

$$= \sum [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - a) + (\bar{x} - a)^2]$$

$$= \sum (x_i - \bar{x})^2 + 2(\bar{x} - a)\sum (x_i - \bar{x}) + n(\bar{x} - a)^2$$

$$= \sum (x_i - \bar{x})^2 + n(\bar{x} - a)^2 \quad [\because \sum (x_i - \bar{x}) = 0]$$

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

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

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

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

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots + n_k}$$

$$= \frac{\sum n_i \bar{x}_i}{n}, \qquad (i = 1, 2, \ldots, k)$$

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

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

Now summing over all values of i, we obtain

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

Dividing both sides by n, we get

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

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

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

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

Section	Number of boys	Mean height
A	40	62"
B	37	58"
C	43	61"

Find the overall mean height of 120 boys.

Here $n_1 = 40;\ n_2 = 37;\ n_3 = 43,$ and

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

The mean height of the combined class is given as

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$$

$$= \frac{(40 \times 62) + (37 \times 58) + (43 \times 61)}{40 + 37 + 43} = \frac{7249}{120} = 60.4''$$

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

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_kx_k}{f_1 + f_2 + \ldots + f_k}$$

$$= \frac{\Sigma f_i x_i}{n}, \qquad (i = 1, 2, \ldots, k)$$

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

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

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

The calculations are outlined below:

Weight (grams)	Sum of actual observations	f_i	Mean Weight of each class $(\bar{x}_i)$	$f_i\bar{x}_i$	Midpoints (x_i)	$f_i x_i$
65 – 84	695	9	77.2	694.8	74.5	670.5
85 – 104	947	10	94.7	947.0	94.5	945.0
105 – 124	1919	17	112.9	1919.3	114.5	1946.5
125 – 144	1325	10	132.5	1325.0	134.5	1345.0
145 – 164	766	5	153.2	766.0	154.5	722.5
165 – 184	716	4	179.0	716.0	174.5	698.0
185 – 204	956	5	191.2	956.0	194.5	972.5
Total	7324	60	--	7324.1	--	7350.0

Calculation based on Ungrouped data

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

$$\therefore \qquad \bar{x} = \frac{\sum x_i}{n}, \qquad (i = 1, 2, \ldots, 60)$$

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

This is the exact mean of the given data.

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

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

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

Calculation based on Grouped data

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

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

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

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

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

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

Now
$$\bar{x} = \frac{1}{n}\Sigma x_i = \frac{1}{n}\Sigma(a + hu_i)$$
$$= \frac{na}{n} + \frac{h\Sigma u_i}{n} = a + h\bar{u}$$

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

Example 3.5 Given the following frequency distribution of weights, calculate the mean weight by the Short Method.

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

To calculate the mean weight, let us take $u_i = \dfrac{x_i - 114.5}{20}$, where a = 114.5 is the midpoint corresponding to the largest frequency and h = 20 is the width of the uniform class interval. The necessary calculations are shown in the table below:

Weight (grams)	Midpoints (x_i)	f_i	u_i	f_iu_i
65 – 84	74.5	9	–2	–18
85 – 104	94.5	10	–1	–10
105 – 124	114.5	17	0	–28
125 – 144	134.5	10	1	10
145 – 164	154.5	5	2	10
165 – 184	174.5	4	3	12
185 – 204	194.5	5	4	20
Total	--	60	--	+52 / 24

Thus
$$\bar{x} = a + h\bar{u}, \quad \text{where } \bar{u} = \frac{\Sigma f_i u_i}{n}$$
$$= 114.5 + \frac{(24)(20)}{60} = 114.5 + 8.0 = 122.5 \text{ grams.}$$

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

Death Rate	Frequency
3.5 – 4.4	1
4.5 – 5.4	4
5.5 – 6.4	5
6.5 – 7.4	13
7.5 – 8.4	12
8.5 – 9.4	19
9.5 – 10.4	13
10.5 – 11.4	10
11.5 – 12.4	6
12.5 – 13.4	4
13.5 – 14.4	1
Total	**88**

The necessary calculations are given below:

Death Rate	Midpoints (x_i)	f_i	$u_i (= x_i - 8.95)$	$f_i u_i$
3.5 – 4.4	3.95	1	–5	–5
4.5 – 5.4	4.95	4	–4	–16
5.5 – 6.4	5.95	5	–3	–15
6.5 – 7.4	6.95	13	–2	–26
7.5 – 8.4	7.95	12	–1	–12
8.5 – 9.4	8.95	19	0	–74
9.5 – 10.4	9.95	13	1	13
10.5 – 11.4	10.95	10	2	20
11.5 – 12.4	11.95	6	3	18
12.5 – 13.4	12.95	4	4	16
13.5 – 14.4	13.95	1	5	5
Total	--	**88**	--	+72 –2

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

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

3.5 THE GEOMETRIC MEAN

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

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

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

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

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

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

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

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

$$G = [x_1^{f_1} . x_2^{f_2} \ldots x_k^{f_k}]^{1/n},$$

In terms of logarithms, the formula becomes

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

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

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

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

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

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

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

The geometric mean, G, is calculated as

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

Taking logs, we have

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

$$= \frac{1}{9}[1.65321 + 1.50515 + 1.56820 + 1.66276 + 1.59106 + 1.55630 + 1.61278 + 1.68124 + 1.55630]$$

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

Hence $\quad G = \text{anti} - \log(1.59856) = 39.68$

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

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

The computation of the geometric mean is shown below:

Weight (grams)	x_i	f_i	$\log x_i$	$f_i \log x_i$
65 – 84	74.5	9	1.8722	16.8498
85 – 104	94.5	10	1.9754	19.7540
105 – 124	114.5	17	2.0589	35.0013
125 – 144	134.5	10	2.1287	21.2870
145 – 164	154.5	5	2.1889	10.9445
165 – 184	174.5	4	2.2418	8.9672
185 – 204	194.5	5	2.2889	11.4445
Σ	--	60	--	124.2483

$$\therefore \quad \log G = \frac{1}{n}\Sigma f_i \log x_i = \frac{124.2483}{60} = 2.0708$$

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

3.6 THE HARMONIC MEAN

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

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

$$= \frac{n}{\Sigma \frac{1}{x_i}}, \qquad \text{where } x \neq 0$$

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

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

$$H = \text{Reciprocal of } \frac{\frac{1}{15}+\frac{1}{20}+\frac{1}{25}}{3}$$

$$= \frac{3}{0.06667+0.5000+0.04000}$$

$$= \frac{3}{0.15667} = 19.15\, km/hr \text{ approximately.}$$

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

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

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

$$\text{H} = \text{Reciprocal of } \frac{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\ldots+\frac{f_k}{x_k}}{f_1+f_2+\ldots+f_k}$$

$$= \frac{n}{\Sigma f_i \frac{1}{x_i}}$$

Similarly, the weighted harmonic mean is defined as

$$H_w = \frac{w_1+w_2+\ldots+w_n}{w_1\left(\frac{1}{x_1}\right)+w_2\left(\frac{1}{x_2}\right)+\ldots+w_n\left(\frac{1}{x_n}\right)}$$

$$= \frac{\Sigma w_i}{\Sigma w_1\left(\frac{1}{x_i}\right)}$$

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

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

We calculate the harmonic mean as below:

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

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

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

x_i	3.95	4.95	5.95	6.95	7.95	8.95	9.95	10.95	11.95	12.95	13.95
f_i	1	4	5	13	12	19	13	10	6	4	1

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

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

x_i	f_i	$\log x_i$	$f_i \log x_i$	$\frac{1}{x_i}$	$f_i\left(\frac{1}{x_i}\right)$
3.95	1	0.59660	0.59660	0.25316	0.25316
4.95	4	0.69461	2.77844	0.20202	0.80808
5.95	5	0.77452	3.87260	0.16807	0.84035
6.95	13	0.84198	10.94574	0.14388	1.87044
7.95	12	0.90037	10.80444	0.12579	1.50948
8.95	19	0.95182	18.08458	0.11173	2.12287
9.95	13	0.99782	12.97166	0.10050	1.30650
10.95	10	1.03945	10.39450	0.09132	0.91320
11.95	6	1.07740	6.46440	0.08368	0.50208
12.95	4	1.11229	4.44916	0.07722	0.30888
13.95	1	1.14459	1.14459	0.07168	0.07168
Total	88	--	82.50671	--	10.50672

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

Hence $G = \text{anti-log } (0.93758) = 8.66$, and

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

3.7 THE MEDIAN

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

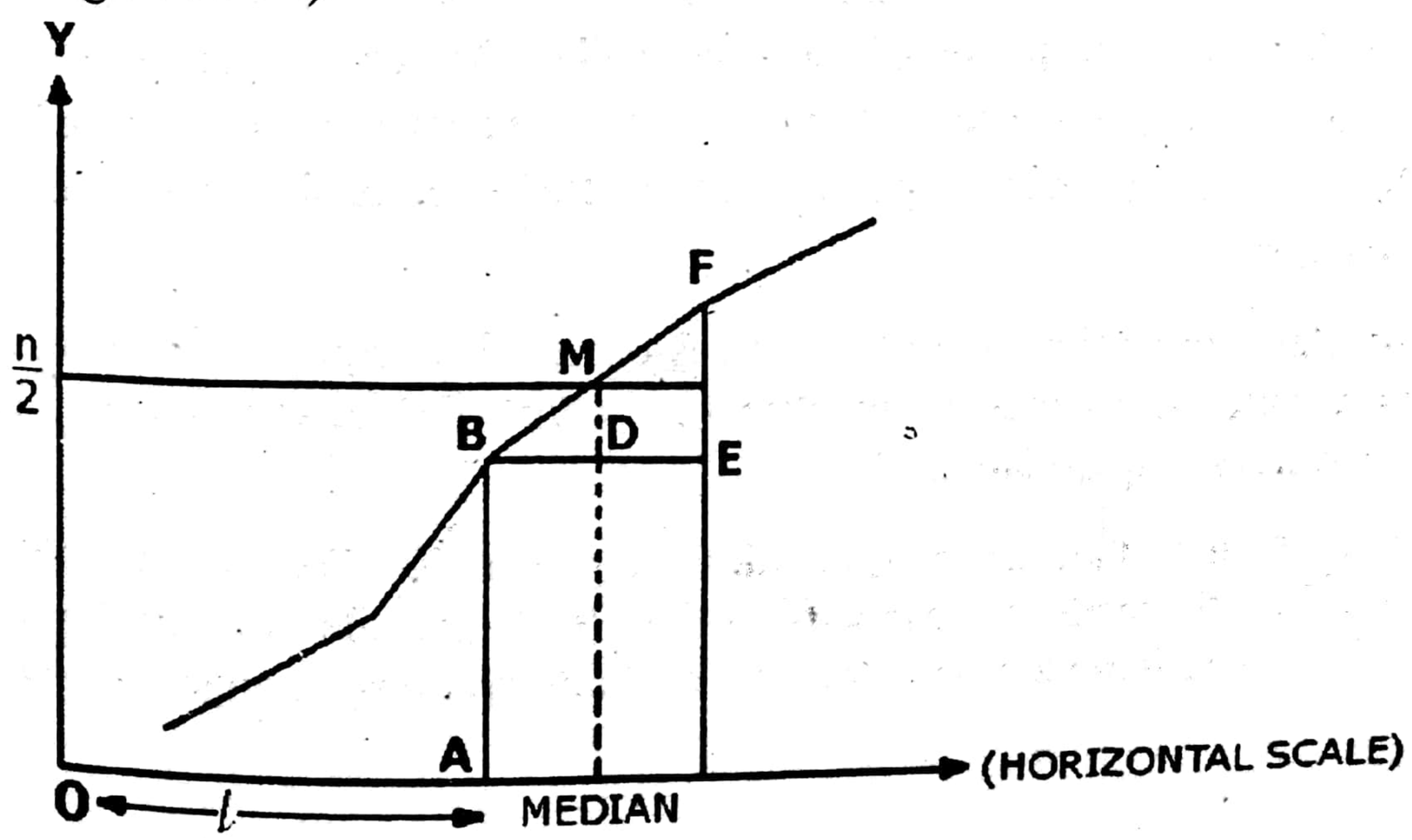
Thus the sample median of the n observations $x_1, x_2, \ldots, x_n$ when arranged in *order* from smallest to largest, is the middle value if n is odd, and the average of two middle values if n is even. Stated differently, when $\frac{n}{2}$ is not an integer, the median is $\left(\frac{n+1}{2}\right)th$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

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

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

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

Median = OA + BD (see figure below).



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

or $\quad BD = BE.\frac{DM}{EF}$

Now evidently BE is the width of the class-interval containing median and hence is equal to h.

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

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

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

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

3.7.1 Quantiles. When the number of observations is quite large, the principle according to which a distribution or an ordered data set is divided into two equal parts, may be extended to any number of divisions. The three values which divide the distribution into four equal parts, are called the *Quartiles*. These values are denoted by Q_1, Q_2 and Q_3 respectively. Q_1 is called the *first* or *lower quartile* and Q_3 is known as the *third or upper quartile*. In other words, the quartiles Q_1, Q_2 and Q_3 are the values at or below which lie respectively, the lowest 25, 50 and 75 per cent of the data. Similarly, the nine values which divide the distribution into ten equal parts, are called *Deciles* and are denoted by $D_1, D_2, \ldots, D_9$; while the ninety nine values dividing the data into one hundred equal parts, are called *Percentiles* and are denoted by $P_1, P_2, \ldots, P_{99}$. The second quartile or the fifth decile or the fiftieth percentile is obviously identical with the median.

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

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

i) When $\frac{jn}{100}$ is not an integer, the *jth* percentile is given as

P_j = Observation with ordinal number $[\frac{jn}{100}]+1$; and

ii) When $\frac{jn}{100}$ is an integer, the *jth* percentile is

P_j = Average of two observation with ordinal numbers

$$\left(\frac{jn}{100}\right) \text{ and } \left(\frac{jn}{100}\right)+1,$$

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

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

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

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

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

Find the median and the quartiles.

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

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

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

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

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

= 39 marks

Now Q_1 = Marks obtained by $\left([\frac{n}{4}]+1\right)th$ student as $\frac{n}{4}$ is not an integer.

= Marks obtained by $\left([\frac{9}{4}]+1\right)th$, *i.e. 3rd* student.

= 36 marks.

Similarly,

$$Q_3 = \text{Marks obtained by } \left([\frac{3n}{4}]+1\right)th \text{ student}$$

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

$$= 45 \text{ marks.}$$

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

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

Find the median number of assistants. Also calculate the quartiles and the *7th* decile.

This is an example of ungrouped frequency distribution with unit class interval. To locate the median, the quartiles and the *7th* decile for such a distribution, we cumulate the frequencies as shown in the table.

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

Since $\frac{n}{2}$, *i.e.* $\frac{50}{2}$ is an integer, therefore the median is the average number of assistants in $\left(\frac{n}{2}\right)th$ and $\left(\frac{n}{2}+1\right)th$, *i.e., 25th* and *26th* retail establishments. Looking at the cumulative frequency row, we find that these two values corresponding to the same value of x, *i.e.* 4.

Hence *median* number of assistants = 4.

For Q_1, we see that $\frac{n}{4}$, *i.e.* $\frac{50}{4}$ is not an integer, therefore

$$Q_1 = \text{No. of assistants in } \left([\frac{50}{4}]+1\right)th \text{ establishment.}$$

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

$$= 2 \text{ assistants.}$$

Similarly,

$$Q_3 = \text{No. of assistants in } \left([\frac{3 \times 50}{4}]+1\right)th \text{ establishment as } \frac{3n}{4} \text{ is also not an integer.}$$

$$= \text{No. of assistants in } 38th \text{ establishment.}$$

$$= 6 \text{ assistants.}$$

Again D_7 = Average number of assistants $\left(\frac{7 \times 50}{10}\right)th$ and $\left(\frac{7 \times 50}{10}+1\right)th$ establishment as $\frac{7n}{10}$ is an integer.

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

= 5 assistants (since both values correspond to 5)

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

Marks	30–39	40–49	50–59	60–69	70–79	80–89	90–99
Number of students	8	87	190	304	211	85	20

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

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

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

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

= Marks obtained by $\frac{905}{2}$, *i.e.* $452.5th$ student which corresponds to marks in the class 59.5 – 69.5. Therefore

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

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

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

And Q_1 = Marks of $\left(\frac{n}{4}\right)th$ student

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

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C\right) = 49.5 + \frac{10}{190}(226.25 - 95)$$

$$= 49.5 + 6.9 = 56.4 = 56 \text{ marks}$$

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

= Marks of $\frac{3 \times 905}{4}$, *i.e.* 678.75*th* student which lies in the class 69. 5 – 79.5

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

$$= 69.5 + \frac{10}{211}(678.75 - 589) = 69.5 + 4.2 = 73.7 = 74 \text{ marks}$$

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

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

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

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

3.8 THE MODE

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

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

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

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

where l = lower class boundary of the modal class,

f_m = frequency of the modal class,

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

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

h = width of class interval.

The mode can also be calculated by the following formula:

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

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

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

Example 3.14 Calculate the mode for the distribution of examination marks given in Example 3.13.

The class that caries the highest frequency is 59.5 – 69.5, which is thus the modal class.

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

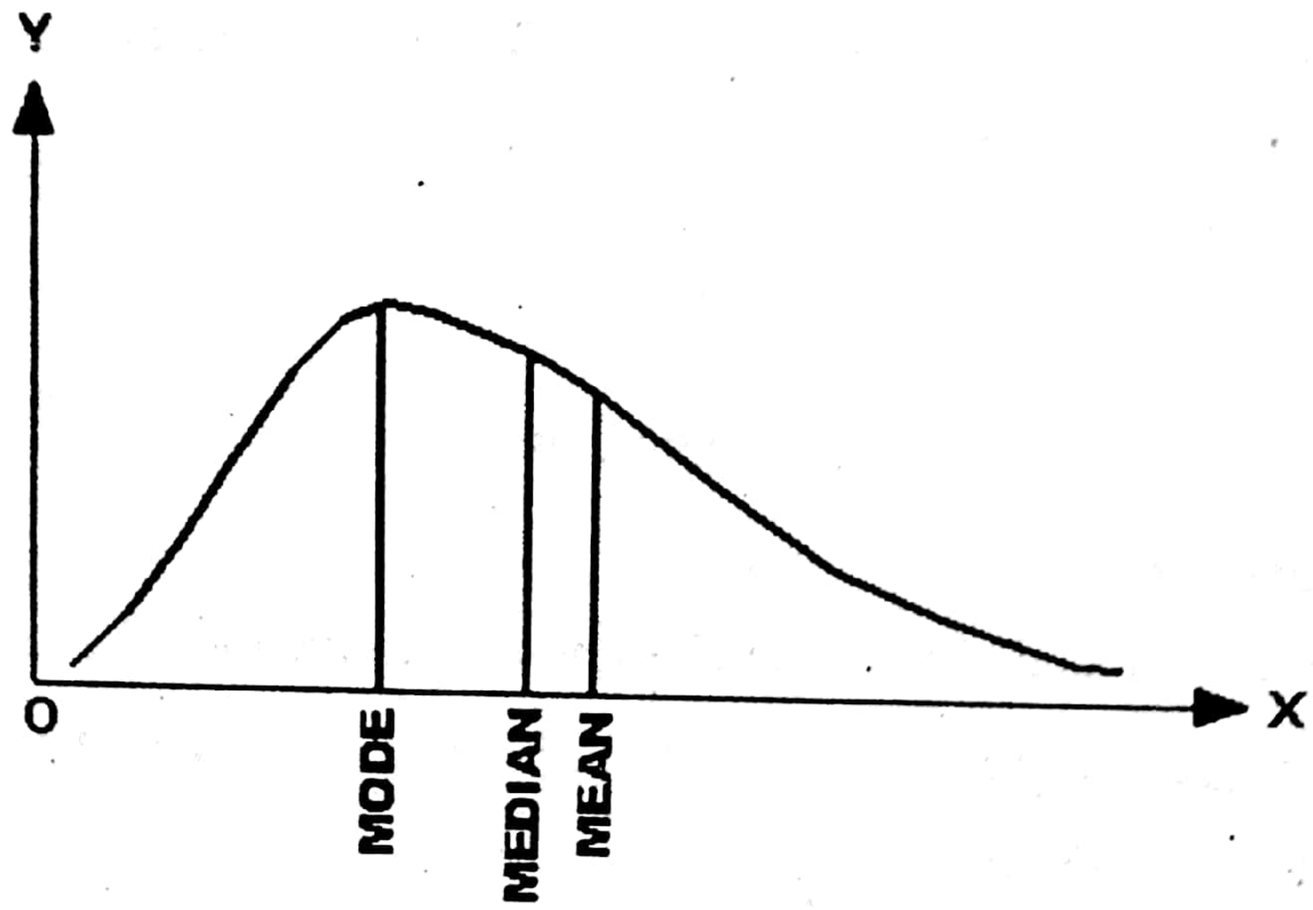
$$\text{Hence Mode} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h,$$

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

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

3.9 EMPIRICAL RELATION BETWEEN MEAN, MEDIAN AND MODE

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



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

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

or

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

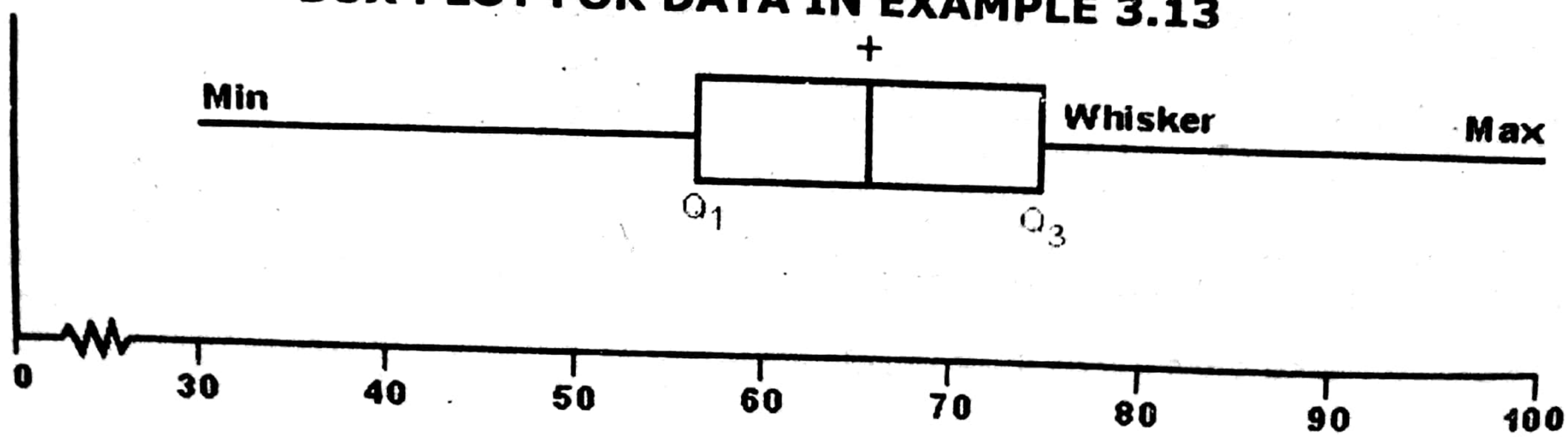
This empirical relation does not hold in case of a J-shaped or an extremely skewed distribution.

3.10 THE BOX PLOTS

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

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

BOX PLOT FOR DATA IN EXAMPLE 3.13



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

3.11 RELATIVE MERITS AND DEMERITS OF VARIOUS AVERAGES

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

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

i) It is rigorously defined by a mathematical formula.

ii) It is based on all the observations in the data.

iii) It is easy to calculate and simple to comprehend.

iv) It is determined for almost every kind of data.

v) It is a relatively stable statistic with the fluctuations of sampling. That is why it is universally used.

vi) It is amenable to mathematical treatment.

The disadvantages of the mean are:

i) It is greatly affected by extreme values in the data.

ii) It gives sometimes fallacious conclusions.

iii) In a highly skewed distribution, the mean is not an appropriate measure of average.

iv) If the grouped data have "open-end" classes, mean cannot be calculated without assuming the limits.

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

i) It is rigorously defined by a mathematical formula.

ii) It is based on all observed values.

iii) It is amenable to mathematical treatment in certain cases.

iv) It gives equal weightage to all the observations.

v) It is not much affected by sampling variability.

vi) It is an appropriate type of average to be used in case rates of change or ratios are to be averaged.

Its disadvantages are:

i) It is neither easy to calculate nor to understand.

ii) It vanishes if any observation is zero.

iii) In case of negative values, it cannot be computed at all.

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

i) It is rigorously defined by a mathematical formula.

ii) It is based on all the observations in the data.

iii) It is amenable to mathematical treatment.

iv) It is not much affected by sampling variability.

v) It is an appropriate type for averaging rates and ratios.

The disadvantages of the harmonic mean are:

i) It is not readily understood.

ii) It cannot be calculated, if any one of the observations is zero.

iii) It gives too much weightage to the smaller observations.

3.11.4 The Median. The advantages of the median are:

i) It is easily calculated and understood.

ii) It is located even when the values are not capable of quantitative measurement.

iii) It is not affected by extreme values. It can be computed even when a frequency distribution involves "open-end" classes like those of income and prices.

iv) In a highly skewed distribution, median is an appropriate average to use.

The median has the following disadvantages:

i) It is not rigorously defined.

ii) It is not capable of lending itself to further statistical treatment.

iii) It necessitates the arrangement of data into an array which can be tedious and time consuming for a large body of data.

3.11.5 The Mode. The advantages of the mode are:

i) It is simply defined and easily calculated. In many cases, it is extremely easy to locate the mode.

ii) It is not affected by abnormally large or small observations.

iii) It can be determined for both the quantitative and the qualitative data.

The disadvantages of the mode are:

i) It is not rigorously defined.

ii) It is often indeterminate and indefinite.

iii) It is not based on all the observations made.

iv) It is not capable of lending itself to further statistical treatment.

v) When the distribution consists of a small number of values, the mode may not exist.

EXERCISES

OBJECTIVE

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

i) A measure of central tendency is a quantitative value that tends to locate in some sense the middle of a set of data.

ii) The mean of a sample always divides the data into two equal halves – half larger and half smaller.

iii) For any distribution, the sum of the deviations from the mean equals zero.

iv) The arithmetic mean is not affected by the extreme values.

v) A distribution always has exactly one median score, but it can have more than one mode score.

vi) The value that occurs most frequently is known as median.

vii) The mean may not exist for some sets of data.

viii) Another name for the median is the score at the third quartile.

ix) The Harmonic mean, the geometric mean and the arithmetic mean are equal only if all the numbers $x_1, x_2, \ldots, x_n$, are identical.

x) The Harmonic mean is the Nth root of the product of the numbers.

xi) The first quartile is also referred to as the 25th percentile.

xii) If a distribution of scores is symmetric, then the median and the mode will be the same.

xiii) If a distribution is positively skewed, then the mean is smaller than the median.

xiv) If a distribution is skewed to left, generally mean > median > mode.

xv) If the mean, median, and mode of a distribution are 5, 6, and 8, respectively, the distribution is positively skewed.

b) MULTIPLE CHOICE QUESTIONS.

i) Half the observations are always larger than the

a) Mean b) Total c) Median d) Mode

ii) The value that occurs most often in a set of data is called the

a) Mean b) Mode c) Geometric mean d) Harmonic mean

iii) In case of an open-end class,

a) A median cannot be computed.

b) The arithmetic mean and the median will always be exactly equal.

c) A mean cannot be computed.

d) The distribution is always positively skewed.

iv) Which of the following is a true statement about the median?

a) It is always one of the data values.

b) It is influenced by extreme values.

c) Fifty percent of the observations are larger than the median.

d) It is the middle value of the data values.

v) Which of the following is not a characteristic of the arithmetic mean?

a) It is influenced by extreme values.

b) The sum of the observations from the mean is zero.

c) Fifty percent of the observations are larger than the mean.

d) The sum of the squared directions from mean is always minimum.

vi) Find the mean of the following sample of distances of stars from the earth:
18.2, 56.9, 24.6, 13.5

a) $\bar{X} = 28.30$ b) $\bar{X} = 43.40$ c) $\mu = 28.30$ d) $\mu = 43.40$

vii) In a positively skewed distribution, the mean is always

a) Smaller than the median b) Equal to the median

c) Larger then the median d) Equal to the mode

viii) The median is larger than the arithmetic mean when

a) The distribution is positively skewed.

b) The distribution is negatively skewed.

c) The data is organized into a frequency distribution.

d) The distribution is symmetrical.

ix) The geometric mean of the numbers 2, 4 and 8 is

a) 3.67 b) 4 c) 3.43 d) 5

x) Which of the following statement is not true for Harmonic Mean?

a) Harmonic mean is smaller than the mean.

b) It is based on all the values.

c) It is an appropriate type for averaging rates and ratios.

d) It gives equal weightage to all the values.

SUBJECTIVE

3.1 What is a statistical average? Name the important types of averages. Discuss the advantage and disadvantages of each average. (P.U., B.A. (Hons.) 1960)

3.2 What is a measure of "central tendency"? What is the purpose served by it? What are its desirable qualities?

3.3 What are the principal criteria for a satisfactory average? State giving reasons the circumstances in which it would be preferable to use (i) the mean, (ii) the median (iii) the mode, (iv) geometric mean and (v) harmonic mean.

3.4 What criteria do you apply to judge the merits of an average? Discuss the merits and demerits of the different averages in common use with special reference to these criteria.

3.5 In what circumstances would you consider the Arithmetic mean, the Geometric mean and the Harmonic mean respectively, the most suitable statistic to describe the central tendency of distributions? (P.U., B.A./B.Sc. 1989)

3.6 What are the different measures of central tendency? Describe the manner of computation of any *three* of them with suitable illustrations. (P.U., M.A. Econ. 1967)

3.7 Define weighted average and explain how it differs from simple mean. Give the method of its computation and discuss the use of weighted mean in Statistics. (P.U., M.A. Econ. 1974)

3.8 What is the median? What are its advantages and disadvantages? Give reasons why the statistician usually prefers the arithmetic mean to the median. (P.U., M.A. Econ. 1981)

3.9 Define, and explain how to compute, the following quantities for a grouped distribution: $\bar{x}, Q_1, Q_3, D_7$ and Mode.

3.10 Define the arithmetic mean, the mode and the median. Discuss the relationship of these three measures of location in a skewed distribution. State the chief advantages of the arithmetic mean as a form of average.

3.11 Define Mean, Median and Mode. What are their advantages and limitations in the analysis of data? Give various methods of calculating Arithmetic mean, with illustrations. (P.U., B.A./B.Sc. 1958, 1960)

3.12 a) Define Mean, Median and Mode. Give an empirical relation between them. Does this relation give correct value for the mode?

b) Criticise the following statements:

i) An average does not reveal all the information about the data.

ii) The median is described as *the value of the average* rather than the *average value*.

3.13 Comment on the following statements:

i) The depth of a river at four different points is 2, 7, 5, 6 feet respectively. The average depth is 5 feet. Therefore all the people with heights above 5 feet can cross it.

ii) The average marks of one class of students are 30. Therefore every student is hopeless.

iii) The average income of a king and his household servants is £20,000 per month, therefore all the household servants must be fabulously paid.

iv) On an average, the number of accidents occurring in the middle of the road are 5 per thousand. The number of deaths at other places is 30 per thousand. Therefore, it is safer to walk in the middle of the road.

v) In a country, 2,000 vaccinated persons died. Therefore vaccination is useless.

3.14 Define the mean, the median and the mode of a frequency distribution.
It is commonly true that the median lies between the other two measures and is approximately twice as far from the mode as from the mean. State with reasons, whether you expect this relationship to hold, and which of the three statistics is likely to be the most useful single statistic for each of the following distributions:

i) The annual earnings of employed males in Pakistan.

ii) The percentage of sky, to the nearest 10 per cent, covered by cloud at Karachi at mid-day.

iii) The exact length of rods cut to a standard size by machine.

(P.C.S., 1971)

3.15 a) Define arithmetic mean and describe its properties.

b) If the arithmetic mean of n numbers $x_1, x_2, \ldots, x_n$ is M and A is any arbitrary number, then show that

$$\Sigma(x_i - A)^2 = \Sigma(x_i - M)^2 + n(M - A)^2$$ (P.U., B.A./B.Sc. 1977, 82)

c) A distribution consists of three components with frequencies 3, 4 and 5, and having means 2, 5.5 and 10. Find the mean of the combined distribution. (P.U., B.A./B.Sc. 1977)

3.16 a) State the properties of the arithmetic mean.

b) Show that $\Sigma(x_i - a)^2 = \Sigma(x_i - \bar{x})^2 + n(\bar{x} - a)^2$. In other words, show that the sum of squares about $a = \bar{x}$ is smallest.

c) A distribution consists of 3 components with frequencies 45, 40 and 65, having their means 2, 2.5 and 2 respectively. Show that the mean of the combined distribution is 2.13 approximately.

3.17 a) The number of cars crossing a certain bridge in a big city in 10 intervals of five minutes each was recorded as follows:

25, 15, 18, 30, 20, 20, 12, 9, 16, 15

Calculate (i) the arithmetic mean, (ii) the median and (iii) the geometric mean.

b) Explain why the mean calculated for a set of ungrouped data might differ from the mean if the same data were grouped into a frequency distribution.

3.18 a) Define Arithmetic mean, Geometric mean and Harmonic mean; and prove that for any two positive numbers a and b,

A.M. $\geq$ G.M. $\geq$ H.M.

(P.U., B.A./B.Sc. 1991)

b) The monthly incomes of ten families in rupees in a certain locality are given below:

Family:	A	B	C	D	E	F	G	H	I	J
Income (Rs.):	85,	70,	10,	75,	500,	8,	42,	250,	40,	36.

Calculate the arithmetic mean, the geometric mean and the harmonic mean of the above incomes. Which one of the above three averages represents the above figures best?

3.19 Calculate the arithmetic mean, the geometric mean and the harmonic mean of the annual incomes of fifteen families as given below:

Rs. 60, 80, 90, 96, 120, 150, 200, 360, 480, 520, 1060, 1200, 1450, 2500, 7200.

3.20 a) In a company having 80 employees, 60 earn Rs.3.00 per hour and 20 earn Rs.2.00 per hour. (i) Determine the mean earnings per hour. (ii) Do you consider this mean hourly wage to be typical?

(P.U., B.A./B.Sc. 1980-S)

b) An examination candidate's percentages are: English, 73; French, 82; Mathematics, 57; Science, 62; History, 60. Find the candidate's weighted mean if weights of 4, 3, 3, 1, 1 respectively are allotted to the subjects.

3.21 Find (i) the simple average of prices in column 2 and (ii) the weighted average, using the quantities in column 3 as weights, and explain the difference between the two results.

(1) Piece goods	(2) Price per metre (Rs.)	(3) Quantity (millions metres)
Unbleached	8.37	286
Bleached	9.50	255
Printed flags	9.16	64
Other sorts	9.84	172
Dyed in piece	13.65	165
Of dyed yarn	11.95	80

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

139	126	114	100	88	62	77	99	103	108
144	129	148	63	69	148	132	118	142	116
123	104	95	80	85	106	123	133	140	134

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

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

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

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

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

3.24 Find the arithmetic and geometric means of the series 1, 2, 4, 8, 16, ..., 2^n. Find also the harmonic mean.

(P.U. D. St. 1960)

3.25 Find (i) arithmetic mean, (ii) geometric mean, and (iii) harmonic mean of the series 1, 3, 9, 27, 81, ..., 3^n.

(P.U., B.A./B.Sc. 1973, 82)

3.26 a) Define Geometric mean and describe its advantages and disadvantage.

b) Given two sets, each of n positive values, $x_{11}, x_{12}, \ldots, x_{1n}; x_{21}, x_{22}, \ldots, x_{2n}$; prove that the geometric mean of the ratios of corresponding values in the two sets is equal to the ratio of the geometric means of the two sets. (P.U., B.A./B.Sc. 1987)

Hint. Let a ratio be defined as $X = \dfrac{X_1}{X_2}$.

Then $\log X = \log X_1 - \log X_2$

Sum for all pairs of X_1's and X_2's.

Hence $G = \dfrac{G_1}{G_2}$.

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

3.28 a) Define Harmonic mean. How does it differ from arithmetic mean? What are its advantages and disadvantages?

b) A man travels from A to B at average speed of 30 miles per hour and returns from B to A along the same route at an average speed of 60 miles per hour. Find the average speed of the entire journey. (P.U., B.A./B.Sc. 1972)

c) Find out the average speed of person who rides the first mile at the rate of 8 miles an hour, the next mile at the rate of 7.5 miles an hour and the third mile at the rate of 5.5 miles an hour. (P.U., B.A./B.Sc. 1981)

3.29 a) A bus traveling 200 kilometers has 10 stages at equal intervals. The speed of the bus in the various stages was observed to be 10, 15, 20, 25, 20, 30, 40, 50, 30, 40 kilometers per hour. Find the average speed at which the bust travels.

b) Find our the average rate of (i) motion in the case of a person who rides the first mile at the rate of 10 miles an hour, the next mile at the rate of 8 miles per hour, and the third mile at the rate of 6 miles per hour; (ii) increase in population, which in the first decade has increased 20%, in the next 25% and in the third 44%. (P.U., B.A. (Part-I) 1962-S)

3.30 Find the geometric mean and the harmonic mean of the following frequency distribution:

Weekly Income (Rs.)	35–39	40–44	45–49	50–54	55–59	60–64	65–69
No. of workers	15	13	17	29	11	10	5

(P.U., B.A. (Hons. in Econ.) 1966)

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

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

(P.U., M.A. Econ. 1970-S)

3.32 Find the mean or median, whichever you think more suitable, in each of the following:

i) Salaries of 5 men in an industrial concern:
Rs.950, Rs.2100, Rs.1500, Rs.100, Rs.10,000.

ii) Heights of 6 boys: 64", 65", 65", 66", 66", 67".

iii) Handicaps of four golfers: 4, 18, 18, 20.

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

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

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

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

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

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

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

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

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

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

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

Age of head (yrs)	Number of households
under 25	44
25 and under 30	79
30 and under 40	152
40 and under 50	122
50 and under 60	141
60 and under 65	100
65 and under 70	58
70 and under 75	32
75 and under 85	28

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

Heights (inches)	57.5–,	60.0–,	62.5–,	65.0–,	67.5–,	70.0–,	72.5–
Number	6	26	190	281	412	127	38

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

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

Class	Number	Class	Number	Class	Number
100–104	4	124–129	298	150–154	260
105–109	14	130–134	380	155–159	128
110–114	60	135–139	450	160–164	66
115–119	138	140–144	500	165–169	28
120–124	236	145–149	430	170–174	12

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

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

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

Calculate the Mean and the Median ages of the distribution.

3.40 a) Describe the merits and demerits of mean and median.

b) Calculate the median, the upper and lower quartiles from the following data: Also draw a box plot.

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

(P.U., B.A./B.Sc. 1968)

3.41 The following distribution shows Kilowatt-Hours of Electricity used in one month by 75 residential consumers in a certain locality of Lahore.

Consumption in kilowatt hours	5–24	25–44	45–64	65–84	85–104	105–124	125–144	145–164
No. of consumers	4	6	14	22	14	5	7	3

Estimate the mean, the median and the two quartiles.

3.42 The yields of grain (x *lb*) from 500 small plots are grouped in classes with a common class-interval (0.2 lb.) in the table below, the values of x given being the midvalues of the classes. Show that the mean of the distribution is 3.95 lb.; the median is 3.95 lb.; and quartiles are 3.63 lb. and 4.28 lb.

x	f
2.8	4
3.0	15
3.2	20
3.4	47
3.6	63
3.8	78
4.0	88
4.2	69
4.4	59
4.6	35
4.8	10
5.0	8
5.2	4
Total	500

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

Weight (milligrams)	Number of Seeds	Weight (milligrams)	Number of Seeds
10 – 24.9	16	85 – 99.9	655
25 – 39.9	68	100 – 114.9	803
40 – 54.9	204	115 – 129.9	294
55 – 69.9	233	130 – 144.9	21
70 – 84.9	240	145 – 159.9	4

a) Find the average weight, the median weight and the most common weight (mode) of the seeds.

b) Find the first and third quartiles. Find the third decile and the 45*th* percentile.

c) Explain your answers as you would to a person who had never studied statistics.

3.44 In a group of 500 wage-earners, the weekly wages of 4% were under Rs.60 and those of 15% were under Rs.62.50. 15% of the workers earned Rs.95 and over, and 5% of them got Rs.100 and over.

The median and quartile wages were Rs.82.25, Rs.72.75 and Rs.90.50; the fourth and sixth decile wages were Rs.78.75 and Rs.85.25 respectively.

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

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

3.45 a) Describe the advantages and disadvantages of the mean, the median and the mode. Explain the empirical relation between them.

b) The weight of the 40 male students at a university are given in the following frequency table:

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

Calculate the mean, median and the mode. (P.U., B.A./B.Sc. 1969)

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

Max. loads (Short tons)	9.8–10.2	10.3–10.7	10.8–11.2	11.3–11.7	11.8–12.2	12.3–12.7
No. of cables	7	12	17	14	6	4

Determine the mean, the median and the mode.

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

Daily wages (Rs.)	22	24	26	28	30	32	34	36	38	40	42	44
Number of employees	3	13	43	102	175	220	204	139	69	25	6	1

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

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

b) Write down the empirical relation between mean, median, and mode for unimodal distributions of moderate asymmetry. Illustrate graphically the relative positions of the mean, median and mode for frequency curves which are skewed to the right and to the left. (P.U., B.A./B.Sc. 1972, 80-S)

c) For a certain frequency distribution, the mean was 40.5 and median 36. Find the mode approximately using the formula connecting the three. (P.U., B.A./B.Sc. Optional, 1971-S)

3.49 a) What types of averages would be suitable for the following cases? Give reasons.

i) Size of agricultural holdings.

ii) Heights of students.

iii) Marks obtained in any examination.

iv) Income of workers in a factory.

v) Per capita income in Pakistan.

vi) Comparison of intelligence.

vii) Volumes of sales of ready-made shirts, shoes and collars.

viii) Number of petals of flowers.

b) What measures of central tendency would you recommend for the following cases? Give reasons in support of your answer.

i) Symmetrical Distribution.

ii) A J-shaped Distribution.

iii) Distribution having "open-end" classes at the end of the classes.

iv) Frequency distribution of a quantitative variable.

(P.U., M.A. Econ., 1977)

3.50 a) A distribution $x_1, x_2, ..., x_r, ..., x_k$ with frequencies $f_1, f_2, ..., f_r, ..., f_k$ is transformed into the distribution $X_1, X_2, ..., X_k$, with the same corresponding frequencies by the relation $X_r = ax_r + b$, where a and b are constant. Show that the mean, mode and median of the new distribution are given in terms of those of the first distribution by the same transformation.

b) A distribution has values of the variable $x_1, x_2, ..., x_k$ with corresponding frequencies $f_1, f_2, ..., f_k$. A new distribution with the same frequencies is formed by taking $X_r = 2x_r - 3$ for values of r ($r = 1, 2, ..., k$). If the values of the mean, median and mode of the original distribution are a, b and c respectively, what are these values of the new distribution? (P.U., B.A./B.Sc. 1980)

◆◆◆◆◆◆◆◆◆◆

CHAPTER 4

MEASURES OF DISPERSION, MOMENTS AND SKEWNESS

MEASURES OF DISPERSION, MOMENTS AND SKEWNESS

4.1 INTRODUCTION

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

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

The main measures of dispersion are the following:

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

4.2 THE RANGE

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

$$R = x_m - x_0,$$

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

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

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

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

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

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

Find the range and the co-efficient of dispersion.

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

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

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

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

$$= \frac{48-32}{48+32} = \frac{16}{80} = 0.2$$

4.3 THE SEMI-INTERQUARTILE RANGE OR THE QUARTILE DEVIATION

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

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

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

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

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

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

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

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

$Q_1 = 36$ marks, $Q_3 = 45$ marks, and therefore

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

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

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

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

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

4.4 THE MEAN (OR AVERAGE) DEVIATION

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

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

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

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

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

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

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

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

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

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

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

Also calculate the co-efficient of mean deviation.

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

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

$$\therefore \quad \text{Median} = \text{Marks obtained by } \left(\left[\frac{n}{2}\right]+1\right)th \text{ student in ordered data as } \frac{n}{2} \text{ is not an integer.}$$

$$= \text{Marks obtained by } \left(\left[\frac{9}{2}\right]+1\right)th, i.e. \text{ 5th student}$$

$$= 39 \text{ marks}$$

and $$\bar{x} = \frac{\Sigma x}{n} = \frac{360}{9} = 40 \text{ marks}$$

The necessary calculations are given below:

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

$$\therefore \quad M.D. \text{ (from mean)} = \frac{\Sigma\lvert x_i - \bar{x}\rvert}{n} = \frac{40}{9} = 4.4 \text{ marks}$$

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

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

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

Example 4.4 Calculate the mean deviation of the following frequency distribution showing the weights of apples:

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

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

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

Here $\bar{x} = \dfrac{\Sigma f_i x_i}{n} = \dfrac{7350.0}{60} = 122.5$ grams

Hence $M.D. = \dfrac{\Sigma f_i \lvert x_i - \bar{x}\rvert}{n} = \dfrac{1696.0}{60} = 28.27$ grams.

4.5 THE VARIANCE AND STANDARD DEVIATION

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

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

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

The variance is also denoted by Var(X). The term *variance* was introduced in 1918 by R.A. Fisher (1890–1962).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Now $\Sigma(x_i - \mu)^2 = \Sigma(x_i^2 - 2x_i\mu + \mu^2)$

$$= \Sigma x_i^2 - 2\mu\Sigma x_i + N\mu^2$$

$$= \Sigma x_i^2 - 2N\mu^2 + N\mu^2 \qquad (\because \mu = \frac{\Sigma x_i}{N})$$

$$= \Sigma x_i^2 - N\mu^2 = \Sigma x_i^2 - \frac{(\Sigma x)^2}{N}$$

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

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

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

$$s^2 = \frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2$$

The alternative formulas for standard deviations are

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

$$s = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x_i}{N}\right)^2}.$$

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

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

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

Example 4.5 A population of N = 10 has the observations 7, 8, 10, 13, 14, 19, 20, 25, 26 and 28. Find its variance and standard deviation.

Calculations appear in the following table:

	x_i	$x_i - \mu$	$(x_i - \mu)^2$	x_i^2
	7	–10	100	49
	8	–9	81	64
	10	–7	49	100
	13	–4	16	169
	14	–3	9	196
	19	+2	4	361
	20	3	9	400
	25	8	64	625
	26	9	81	676
	28	11	121	784
Σ	170	0	534	3424

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

Therefore $$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} = \frac{534}{10} = 53.4,$$

and $$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} = \sqrt{53.4} = 7.31$$

Using the alternative method.

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

$$= \frac{3424}{10} - \left(\frac{170}{10}\right)^2 = 342.4 - 289 = 53.4$$

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

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

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

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

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

Here $$\bar{x} = \frac{\Sigma x_i}{n} = \frac{360}{9} = 40 \text{ marks}$$

Therefore $$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{232}{9} = 25.78 \text{ (marks)}^2$$

and $$S = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}} = \sqrt{25.78} = 5.08 \text{ marks}$$

Using the alternative method.

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

$$= \frac{14632}{9} - \left(\frac{360}{9}\right)^2 = 1625.78 - 1600 = 25.78 \text{ (marks)}^2$$

and $$S = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2} = \sqrt{25.78} = 5.08 \text{ marks}$$

Example 4.7 Calculate the variance and standard deviation from the data of Example 4.4.

The necessary calculations may be carried out on an *electronic calculator* as below:

x_i	f_i	$f_i x_i$	$f_i x_i^2$
74.5	9	670.5	49 952.25
94.5	10	945.0	89 302.50
114.5	17	1946.5	222 874.25
134.5	10	1345.0	180 902.50
154.5	5	772.5	119 351.25
174.5	4	698.0	121 801.00
194.5	5	972.5	189 151.25
Σ	60	7350.0	973 335.00

Thus we find $$s^2 = \frac{\Sigma fx^2}{n} - \left(\frac{\Sigma fx}{n}\right)^2$$

$$= \frac{973335}{60} - \left(\frac{7350}{60}\right)^2 = 16222.25 - 15006.25$$

$$= 1216 \text{(grams)}^2$$

and $$s = \sqrt{\frac{\Sigma fx^2}{n} - \left(\frac{\Sigma fx}{n}\right)^2} = \sqrt{1216} = 34.87 \text{ grams}$$

4.5.1 Change of Origin and Scale. The computational labour can be reduced by using the same transformation as was used for computing the arithmetic mean.

Let $u_i = \dfrac{x_i - a}{h}$. Then $x_i = a + hu$ and $\bar{x} = a + h\bar{u}$

Therefore $$\Sigma(x_i - \bar{x})^2 = \Sigma[(a + hu_i) - (a + h\bar{u})]^2$$

$$= h^2 \Sigma(u_i - \bar{u})^2$$

Further $$\Sigma(u_i - \bar{u})^2 = \Sigma u_i^2 - \frac{(\Sigma u_i)^2}{n}$$

Hence $$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{h^2}{n}[\Sigma u_i^2 - \frac{(\Sigma u_i)^2}{n}]$$

$$= h^2[\frac{\Sigma u_i^2}{n} - (\frac{\Sigma u_i}{n})^2]$$

and $$S = h\sqrt{\frac{\Sigma u^2}{n} - \left(\frac{\Sigma u}{n}\right)^2}$$

This gives us a *short method* for hand calculations.

When the data are grouped into a frequency distribution, the corresponding *short method* for hand calculations, is

$$s = h\sqrt{\frac{\Sigma f_i u_i^2}{n} - \left(\frac{\Sigma fu_i}{n}\right)^2}$$

where h is the width of the class-interval, f_i is the frequency of the *ith* class and u_i is the deviation of x_i from an assumed mean in terms of class intervals. This method is also known as the *step-deviation method* or *coding method*.

Example 4.8 Find the standard deviation by the short method from the data of Example 4.4.

Let $u_i = \dfrac{x_i - 114.5}{20}$, where $a = 114.5$, value corresponding to the highest frequency, and $h = 20$, the class-interval. Then $u_i = -2, -1, 0, 1, 2, 3, 4$. Other calculations appear below:

x_i	f_i	u_i	$f_i u_i$	$f_i u_i^2$
74.5	9	−2	−18	36
94.5	10	−1	−10	10
114.5	17	0	−28	0
134.5	10	1	10	10
154.5	5	2	10	20
174.5	4	3	12	36
194.5	5	4	20	80
Σ	60		$\dfrac{+52}{+24}$	192

Therefore $\quad s = h \times \sqrt{\dfrac{\Sigma fu^2}{n} - \left(\dfrac{\Sigma fu}{n}\right)^2}$

$$= 20 \times \sqrt{\frac{192}{60} - \left(\frac{24}{60}\right)^2} = 20 \times \sqrt{3.20 - 0.16}$$

$$= 20 \times \sqrt{3.04} = 20 \times (1.7436) = 34.87 \text{ grams.}$$

4.5.2 Interpretation of the Standard Deviation. The standard deviation (σ or s) has not a simple interpretation like the arithmetic mean (μ or $\bar{x}$) that is interpreted as the balancing point for the distribution. The standard deviation is a very important concept that serves as a basic measure of variability. A smaller value of the standard deviation indicates that most of the observations in a data set are close to the mean while a large value implies that the observations are scattered widely about the mean. However, a connection between the standard deviation and fraction of data included in intervals constructed around the mean, was discovered by the Russian mathematician P.L. Chebyshev (1821–1894). This result, generally known as *Chebyshev's rule*, is stated below:

"For any set of data, the interval $\bar{x} - ks$ to $\bar{x} + ks$, where k is any number greater than 1, contains *at least* the fraction $\left(1 - \dfrac{1}{k^2}\right)$ of the data." For example, the intervals $\bar{x} \pm 2s$ and $\bar{x} \pm 3s$ will contain respectively *at least* the fractions $\left(1 - \dfrac{1}{2^2}\right)$, *i.e.* $\dfrac{3}{4}$ and $\left(1 - \dfrac{1}{3^2}\right)$, *i.e.* $\dfrac{8}{9}$ of the data.

This rule is applied to any distribution (Population or Sample) and guarantees the inclusion of a minimum fraction of the data in the constructed interval whereas the actual fraction of the included (especially in bell-shaped distributions) will exceed $\left(1-\frac{1}{k^2}\right)$.

4.5.3 Co-efficient of Variation. The variability of two or more than two sets of data cannot be compared unless we have a relative measure of dispersion. For this purpose, Karl Pearson (1857–1936) introduced a relative measure of variation, known as the *co-efficient of variation*, abbreviated C.V. which expresses the standard deviation as a percentage of the arithmetic mean of a data set. Symbolically, it is defined as

$$\text{C.V.} = \frac{S}{\bar{x}} \times 100, \text{ for sample data,}$$

$$= \frac{\sigma}{\mu} \times 100, \text{ for population data.}$$

As the coefficient of variation is a pure number without units, it is therefore used to compare the variation in two or more data sets or distributions that are measured in different units, e.g. one may be measured in hours and the other in kilograms or rupees. A large value of C.V. indicates that the variability is great and a small value of C.V. indicates less variability.

The coefficient of variation is also used to compare the performance of two candidates or of two players given their scores in various papers or games, the smaller the coefficient of variation the more consistent is the performance of the candidates or players. Thus it is used as a criterion for the consistent performance of the candidates or the players. It should be noted that this co-efficient is unreliable when the arithmetic mean is very small.

Example 4.9 Using the co-efficient of variation, determine whether or not there is greater variation among the prices of certain similar commodities given, than among the life in hours under test.

Prince in Rupees:	8,	13,	18,	23,	30
Life in hours:	130,	150,	180,	250,	345

We have to compute the mean and the standard deviation for each set so that the corresponding coefficient of variation can be obtained. The necessary arithmetic is shown below:

Price in Rupees (X)		Life in hours (Y)	
X	X^2	Y	Y^2
8	64	130	16900
13	169	150	22500
18	324	180	32400
23	529	250	62500
30	900	345	119025
92	1986	1055	253325

Price of Commodities

$$\bar{X} = \text{Rs.}\frac{92}{5} = \text{Rs.}18.4$$

$$S_X = \sqrt{\frac{1986}{5} - \left(\frac{92}{5}\right)^2}$$

$$= \sqrt{397.2 - 338.56}$$

$$= \sqrt{58.44} = \text{Rs. } 7.66$$

$$\therefore \text{ C.V.} = \frac{7.66}{18.4} \text{x} 100 = 41.63\%$$

Life in Hours

$$\bar{Y} = \frac{1055}{5} = 211 \text{ hours}$$

$$S_Y = \sqrt{\frac{253325}{5} - \left(\frac{1055}{5}\right)^2}$$

$$= \sqrt{50665 - 44521}$$

$$= \sqrt{6144} = 78.38 \text{ hours}$$

$$\therefore \text{ C.V.} = \frac{78.38}{211} \text{x} 100 = 37.15\%$$

We see that the co-efficient of variation for the prices of commodities (X) is larger than that for the life in hours (Y). Hence the prices of certain similar commodities are showing greater variation than that among the life in hours under test.

Example 4.10 Goals scored by two teams A and B in a football season were as follows:

No. of goals scored in a match (x_i)	Number of matches	
	A	B
0	27	17
1	9	9
2	8	6
3	5	5
4	4	3

By calculating the co-efficient of variation in each case, find which team may be considered more consistent.

The necessary arithmetic is shown below:

No. of goals	Team A			Team B		
(x_i)	f_i	$f_i x_i$	$f_i x_i^2$	f_j	$f_j x_i$	$f_j x_i^2$
0	27	0	0	17	0	0
1	9	9	9	9	9	9
2	8	16	32	6	12	24
3	5	15	45	5	15	45
4	4	16	64	3	12	48
Total	53	56	150	40	68	126

Team A:

$$\text{Mean} = \frac{\Sigma f_i x_i}{n} = \frac{56}{53} = 1.06, \text{ and}$$

$$s = \sqrt{\frac{\Sigma f_i x_i^2}{n} - \left(\frac{\Sigma f_i x_i}{n}\right)^2}$$

$$= \sqrt{\frac{150}{53} - \left(\frac{56}{53}\right)^2} = \sqrt{1.7138} = 1.308$$

$$\therefore \qquad \text{C.V.} = \frac{s}{\bar{x}} \text{x} 100 = \frac{1.308}{1.06} \text{x} 100 = 123.4\%$$

Team B:

$$\text{Mean} = \frac{\Sigma f_j x_i}{n} = \frac{48}{40} = 1.20, \text{ and}$$

$$s = \sqrt{\frac{\Sigma f_j x_i^2}{n} - \left(\frac{\Sigma f_j x_i}{n}\right)^2}$$

$$= \sqrt{\frac{126}{40} - \left(\frac{48}{40}\right)^2} = \sqrt{1.71} = 1.308$$

$$\text{Thus} \qquad \text{C.V.} = \frac{s}{\bar{x}} \text{x} 100 = \frac{1.308}{1.20} \text{x} 100 = 109.0\%$$

We see that the co-efficient of variation for the team B is smaller than that for the team A. Hence team B is more consistent than team A.

4.5.4 Properties of Variance and Standard Deviation. The variance and standard deviation have the following useful and interesting properties:

i) The variance of a constant is equal to zero. If a is any constant, then

$$\text{Var}(a) = \frac{1}{N}\Sigma[a-a]^2 \quad (\because \text{mean of a constant is constant itself})$$

$$= 0$$

ii) The variance is independent of the origin, *i.e.* it remains unchanged when a constant is added to or subtracted from each observation of the variable X. Symbolically,

$$\text{Var}(X+a) = \text{Var}(X)$$

$$\text{Now} \quad \text{Var}(X+a) = \frac{1}{N}\Sigma[(x_i+a)-(\mu+a)]^2 \quad \left(\because \frac{\Sigma(x_i+a)}{N} = \mu + a\right)$$

$$= \frac{1}{N}\Sigma(x_i - \mu)^2 = \text{Var}(X)$$

Hence Var(X) is *invariant* to change of the origin.

iii) The variance is multiplied or divided by the square of the constant, when each observation of the variable X is either multiplied or divided by a constant.

$$\text{Var}(aX) = \frac{1}{N}\Sigma(ax_i - a\mu)^2$$

$$= a^2 \frac{\Sigma(x_i - \mu)^2}{N} = a^2 \text{Var}(X)$$

This may also be interpreted as that the variance increases by a^2 when the *scale* of X is changed by a.

iv) The variance of the sum or difference of two *independent* variables is equal to the sum of their respective variances.

If X and Y are two *independent* variables, then

$$\text{Var}(X \pm Y) = \frac{1}{\text{N}}\Sigma[(x_i \pm y_i) - (\mu_{x+y})]^2$$

$$= \frac{1}{\text{N}}\Sigma[(x_i - \mu_x) \pm (y_i - \mu_y)]^2$$

$$= \frac{1}{\text{N}}\Sigma(x_i - \mu_x)^2 + \frac{1}{N}\Sigma(y_i - \mu_y)^2 \pm \frac{2}{N}\Sigma(x_i - \mu_x)(y_i - \mu_y)$$

The quantity $\frac{1}{\text{N}}\Sigma(x_i - \mu_x)(y_i - \mu_y)$ is called the *covariance* and is denoted by Cov (X, Y). We shall show at some later stage that the covariance of two *independent* variables is equal to zero. Thus we are left with

$$\text{Var}(X \pm Y) = \frac{1}{\text{N}}\Sigma(x_i - \mu_x)^2 + \frac{1}{\text{N}}\Sigma(y_i - \mu_y)^2$$

$$= \text{Var}(X) + \text{Var}(Y).$$

v) If k subgroups of data consisting of $N_1, N_2, \ldots, N_k$ ($\Sigma N_i = N$) observations have respective means $\mu_1, \mu_2, \ldots, \mu_k$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$, then the variance σ^2 of the combined observations is given by

$$\sigma^2 = \frac{1}{N}\Sigma N_i(\sigma_i^2 + D_i^2), \quad i = 1, 2, \ldots, k$$

where $D_i = \mu_i - \mu$ and μ is the mean for all the data.

Let for the *ith* subgroup with mean μ_i, μ, the general mean, be considered as an arbitrary origin. Then the sum of squares of deviations of the observations in the *ith* subgroup from μ is given by

$$\sum_1^{N_i}(x_i-\mu)^2 = \sum_1^{N_i}(x_i-\mu_i+\mu_i-\mu)^2$$

$$= \sum_1^{N_i}(x_i-\mu_i)^2 + N_i(\mu_i-\mu)^2, \quad (\because \text{ product term vanishes})$$

$$= N_i\sigma_i^2 + N_i D_i^2$$

$$= N_i(\sigma_i^2 + D_i^2)$$

But the variance σ^2 of the combined observations is the mean of the *sum of the deviations of all observations in k subgroups from the general mean* μ. Hence summing over k-subgroups, we get

$$N\sigma^2 = \Sigma N_i(\sigma_i^2 + D_i^2)$$

It is relevant to note that all these properties are valid for standard deviation (*S.D*), which is the positive square root of variance. In other words,

i) $S.D.\ (a) = 0.$

ii) $S.D.\ (X + a) = S.D.\ (X)$

iii) $S.D.\ (aX) = |a|\ S.D.\ (X)$, as *S.D.* cannot be negative.

iv) $S.D.\ (X \pm Y) = \sqrt{\text{Var}(X) + \text{Var}(Y)}$

v) $\sigma = \sqrt{\frac{1}{N}\Sigma N_i(\sigma_i^2 + D_i^2)}$

For sample data, the corresponding results may be obtained in the same way.

Example 4.11 Let $\bar{x}_1$ and S_1^2 be the mean and variance respectively of n_1 observations, $\bar{x}_2$ and S_2^2 be the mean and variance respectively of n_2 observations. Then, if the variance of all n_1+n_2 observations is S^2, prove that

$$S^2 = \frac{n_1S_1^2 + n_2S_2^2}{n_1+n_2} + \frac{n_1n_2}{(n_1+n_2)^2}(\bar{x}_1-\bar{x}_2)^2$$

(P.U., B.A./B.Sc. 1986)

Let $\bar{x}$ denote the general mean and be regarded as an arbitrary origin for the set of n_1 observations and set of n_2 observations. Then the variance of all n_1+n_2 observations, by definition, is given by

$$S^2 = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2}(x_i-\bar{x})^2$$

$$= \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2 + \sum_{i=n_1+1}^{n_1+n_2}(x_i-\bar{x})^2\right]$$

$$= \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(x_i-\bar{x}_1+\bar{x}_1-\bar{x})^2 + \sum_{i=n_1+1}^{n_1+n_2}(x_i-\bar{x}_2+\bar{x}_2-\bar{x})^2\right]$$

$$= \frac{1}{n_1+n_2}[n_1\{S_1^2+(\bar{x}_1-\bar{x})^2\}+n_2\{S_2^2+(\bar{x}_2-\bar{x})^2\}]$$

$$= \frac{n_1S_1^2+n_2S_2^2}{n_1+n_2} + \frac{n_1(\bar{x}_1-\bar{x})^2+n_2(\bar{x}_2-\bar{x})^2}{n_1+n_2}$$

Since $\bar{x}$ is the mean of all n_1+n_2 observations, *i.e.* $\bar{x} = \dfrac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$,

therefore $\quad \bar{x}_1-\bar{x} = \bar{x}_1 - \dfrac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2} = \dfrac{n_2(\bar{x}_1-\bar{x}_2)}{n_1+n_2}$, and

$$\bar{x}_2-\bar{x} = \bar{x}_2 - \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2} = \frac{-n_1(\bar{x}_1-\bar{x}_2)}{n_1+n_2}.$$

Hence substituting and simplifying, we get

$$S^2 = \frac{n_1S_1^2+n_2S_2^2}{n_1+n_2} + \frac{n_1n_2}{(n_1+n_2)^2}(\bar{x}_1-\bar{x}_2)^2$$

4.5.5 Standardized Variables. A variable is defined to be *standardized* or in *standard units* if it is expressed in terms of deviations from its mean and divided by its standard deviation. It is denoted by Z. In symbols, this means that

$$Z_i = \frac{x_i-\mu}{\sigma}, \text{ for population data,}$$

$$z_i = \frac{x_i-\bar{x}}{S}, \text{ for sample data.}$$

This is a very important concept in advanced statistics as the mean of a standardized variable is equal to zero and its variance is equal to one. Thus

$$\bar{Z} = \frac{1}{N}\Sigma\left(\frac{x_i-\mu}{\sigma}\right) = \frac{1}{\sigma}\cdot\frac{\Sigma(x_i-\mu)}{N} = 0;$$

and $\quad \mathrm{Var}(Z) = \dfrac{1}{N}\Sigma\left[\dfrac{(x_i-\mu)}{\sigma}-0\right]^2$

$$= \frac{1}{\sigma^2}\frac{\Sigma(x_i-\mu)^2}{N} = \frac{1}{\sigma^2}.\sigma^2 = 1$$

The Z-values, being independent of the units of measurement, provide a basis for comparison between individual values, even though they belong to different distributions. That is why they are often used in psychological and education testing, where they are known as *standard scores*. The negative numbers are avoided by multiplying the Z values by 10, an arbitrary $S.D.$, and adding 50, an arbitrary mean, to them. The values so obtained are called the *standard Z scores*. Thus a standard Z score is given by the relation

$$Z = 50 + 10\left(\frac{x - \bar{x}}{S}\right)$$

4.6 TRIMMED AND WINSORIZED MEASURES

Data sets often contain *extreme* (unusually large or small) observations which may be very different from the main body of the data set and may seem to be incorrect. Such extreme observations are generally called *Wild observations* or *Outliers*. These outliers can cause problems. In the presence of outliers, the mean and the standard deviation, being affected by the extreme observations, are therefore misleading measures of central tendency and variability. The appropriate measures then may be the Median and the Interquartile Range, which are much less sensitive to extreme or wild observations. However, it is important to examine a data set for outliers and if present, should be excluded.

For this purpose, we *either* remove a certain percentage of the smallest and largest observations to get the so-called *Trimmed data* set *or* replace trimmed values by those next in magnitude to obtain what is known as *Winsorized data* set (proposed by C.P. Winsor). The mean and the standard deviation of such data sets are known as the *Trimmed Mean* and the *Trimmed Standard Deviation*, and the *Winsorized Mean* and the *Winsorized Standard Deviation* respectively.

Generally, the *Trimmed* mean is obtained from the data set after having removed all observations below the first quartile and all observations above the third quartile. The *Winsorized* mean is calculated from the modified data set obtained by replacing each observation below the first quartile with the value of the first quartile and each observation above the third quartile with the value of the third quartile. The *Trimmed* standard deviation and the *Winsorized* standard deviation are computed from the *trimmed* data set and *Winsorized* data set as usual. The *trimmed* and *Winsorized* measures have gained importance in recent years as they are not disturbed by the presence of a few wild observations and have been found almost as good as the corresponding measures in symmetric distributions with no unusual observations.

Example 4.12 Calculate the trimmed and Winsorized means and standard deviations for the data given in Example 3.11.

The data ordered from smallest to largest and the two quartiles were found to be

$$32, 36, \underset{\substack{\uparrow \\ Q_1}}{36}, 37, 39, 41, \underset{\substack{\uparrow \\ Q_3}}{45}, 46, 48$$

To find the trimmed mean and the trimmed standard deviation, we remove the two observations 32 and 36 below the first quartile and the two observations 46 and 48 above the third quartile. Thus we have five observations 36, 37, 39, 41, 45 as trimmed data set.

$$\therefore \quad \text{Trimmed mean} = \frac{36 + 37 + 39 + 41 + 45}{5} = \frac{198}{5} = 39.6, \text{ and}$$

$$\text{Trimmed } S.D. = \sqrt{\frac{(36)^2 + \ldots + (45)^2}{5} - \left(\frac{198}{5}\right)^2}$$

$$= \sqrt{1598.4 - 1568.16} = \sqrt{30.24} = 5.5$$

To find the Winsorized mean and standard deviation, we replace the two values 32, 36 below the first quartile with 36, and the two values 46, 48 above Q_3 with 45 to get the Winsorized data set as 36, 36, 36, 37, 39, 41, 45, 45, and 45. Thus

$$\text{the Winsorized mean} = \frac{\Sigma X_i}{n} = \frac{360}{9} = 40, \text{ and}$$

$$\text{the Winsorized S.D.} = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2} = \sqrt{\frac{14534}{9} - \left(\frac{360}{9}\right)^2}$$

$$= \sqrt{1614.89 - 1600} = \sqrt{14.89} = 3.86$$

4.7 MOMENTS

A *moment* designates the power to which deviations are raised before averaging them, *e.g.* the quantity $\frac{1}{N}\Sigma(x_i - \mu)$ is called the first population moment and is denoted by μ_1. Similarly, the quantity $\frac{1}{N}\Sigma(x_i - \mu)^2$ is called the second population moment and is denoted by μ_2. The corresponding sample moments are denoted by m_1 and m_2. In general, the *rth moment about the mean* is the arithmetic mean of the *rth* power of the deviations of the observations from the mean. In symbols, this means that

$$\mu_r = \frac{1}{N}\Sigma(x_i - \mu)^r, \quad \text{for population data.}$$

$$m_r = \frac{1}{n}\Sigma(x_i - \bar{x})^r, \quad \text{for sample data.}$$

These moments are also called the *central moments* or the *mean moments* and are used to describe a set of data.

In a similar way, *moments about an arbitrary origin*, say a, are defined by the relation

$$\mu'_r = \frac{1}{N}\Sigma(x_i - a)^r, \quad \text{for population data.}$$

$$m'_r = \frac{1}{n}\Sigma(x_i - a)^r, \quad \text{for sample data.}$$

Now, if we put $r = 0$, we see that

$$\mu_0 = \mu'_0 = 1 \quad \text{and} \quad m_0 = m'_0 = 1$$

For $r = 1$, we have

$$\mu_1 = \frac{1}{N}\Sigma(x_i - \mu) = \frac{\Sigma x_i}{N} - \mu = \mu - \mu = 0, \text{ and}$$

$$\mu_1' = \frac{1}{N}\Sigma(x_i - a) = \frac{\Sigma x_i}{N} - a = \mu - a.$$

The corresponding sample results are $m_1 = 0$ and $m_1' = \bar{x} - a$.

Putting $r = 2$ in the relation for mean moments, we see that

$$\mu_2 = \frac{1}{N}\Sigma(x_i - \mu)^2 = \sigma^2, \text{ which is the population variance,}$$

and $$m_2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2 = S^2, \text{ which is the sample variance,}$$

When $a = 0$, the moment $m_r' = \frac{1}{n}\Sigma x_i^r$ is called the *rth moment about zero*.

The moments about the mean or about the arbitrary origin are also called the *power moments*.

When the sample data are grouped into a frequency distribution having k classes with midpoints $x_1, x_2, \ldots, x_k$ and the corresponding frequencies $f_1, f_2, \ldots, f_k (\Sigma f_i = n)$, the *rth* sample moments are given by

$$m_r = \frac{1}{n}\Sigma f_i(x_i - \bar{x})^r, \text{ and}$$

$$m_r' = \frac{1}{n}\Sigma f_i(x_i - a)^r$$

4.7.1 Moments about the Mean in terms of Moments about the arbitrary origin, say *a*, and conversely. It is easier to calculate the moments in the first instance, about an arbitrary origin. They are then transformed to the mean-moments. This is done by using the relationships obtained as follows.

By definition, the *rth* sample moment about the mean is given by

$$m_r = \frac{1}{n}\Sigma f_i(x_i - \bar{x})^r$$

The quantity within brackets may be written as

$$(x_i - \bar{x}) = (x_i - a + a - \bar{x}) = (x_i - a) - (\bar{x} - a)$$

$$= D_i - m_1' \text{ where } D_i = (x_i - a) \text{ and } m_1' = (\bar{x} - a)$$

Thus, we have $m_r = \frac{1}{n}\Sigma f_i(D_i - m_1')^r$

By means of Binomial expansion, we have

$$m_r = \frac{1}{n}\sum f_i[D_i^r - \binom{r}{1}D_i^{r-1}m_1' + \binom{r}{2}D_i^{r-2}(m_1')^2 + \ldots(-1)^r(m_1')^r]$$

where $\binom{r}{j} = \frac{r!}{j!(r-j)!}$ and $r! = r(r-1)(r-2)\ldots 3 \times 2 \times 1$.

i.e.
$$m_r = \frac{1}{n}\sum f_i D_i^r - \binom{r}{1}\frac{1}{n}\sum f_i D_i^{r-1}m_1' + \binom{r}{2}\frac{1}{n}\sum f_i D_i^{r-2}(m_1')^2 - \ldots + (-1)^r(m_1')^r\frac{1}{n}\sum f_i$$

$$= m_r' - \binom{r}{1}m_{r-1}'m_1' + \binom{r}{2}m_{r-2}'(m_1')^2 + \ldots + (-1)^r(m_1')^r$$

Putting $r = 1, 2, 3$ and 4, we get

$$m_1 = m_1' - m_1' = 0;$$

$$m_2 = m_2' - \binom{2}{1}m_1'.m_1' + \binom{2}{2}(m_1')^2.m_0'$$

$$= m_2' - (m_1')^2;$$

$$m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3, \text{ and}$$

$$m_4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 - 3(m_1')^4$$

The corresponding results for population data are:

$$\mu_1 = 0;$$

$$\mu_2 = \mu_2' - (\mu_1')^2;$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3; \text{ and}$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

It should be noted that in each of these relations, the sum of the co-efficients of various terms on the right hand side equals zero and each term on the right is of the same dimension as the term on the left.

Conversely, the *rth* sample moment about an arbitrary origin, *a*, is given by

$$m_r' = \frac{1}{n}\sum f_i(x_i - a)^r$$

$$= \frac{1}{n}\sum f_i(x_i - \bar{x} + \bar{x} - a)^r$$

$$= \frac{1}{n}\sum f_i(d_i + m_1')^r, \text{ where } d_i = x_i - \bar{x} \text{ and } m_1' = \bar{x} - a$$

$$= \frac{1}{n}\sum f_i d_i^r + \binom{r}{1}m_1'\frac{1}{n}\sum f_i d_i^{r-1} + \binom{r}{2}(m_1')^2 \times \frac{1}{n}\sum f_i d_i^{r-2} + \ldots + (m_1')^r\frac{1}{n}\sum f_i$$

$$= m_r + \binom{r}{1}m_{r-1}(m_1') + \binom{r}{2}m_{r-2}(m_1')^2 + \ldots + (m_1')^r$$

Putting $r = 2, 3$ and 4, we get

$$m_2' = m_2 + (m_1')^2$$
$$m_3' = m_3 + 3m_2(m_1') + (m_1')^3$$
$$m_4' = m_4 + 4m_3(m_1') + 6m_2(m_1')^2 + (m_1')^4.$$

For a population of size N, the corresponding relations are

$$\mu_2' = \mu_2 + \mu_1'^2.$$
$$\mu_3' = \mu_3 + 3\mu_1'\mu_2 + \mu_1'^3;$$
$$\mu_4' = \mu_4 + 4\mu_1'\mu_3 + 6\mu_1'^2\mu_2 + \mu_1'^4.$$

4.7.2 Sheppard's Corrections. In the calculation of moments from a grouped frequency distribution, certain errors are introduced by the assumption that the frequencies associated with a class are located at the midpoint of the class interval. These errors therefore need corrections. It has been shown by W.F. Sheppard that, if the frequency distribution (i) is continuous and (ii) tails off to zero at each end, the corrected moments are as given below:

$$m_2 \text{ (corrected)} = m_2 \text{(uncorrected)} - \frac{h^2}{12};$$
$$m_3 \text{ (corrected)} = m_3 \text{ (uncorrected)};$$
$$m_4 \text{ (corrected)} = m_4 \text{(uncorrected)} - \frac{h^2}{2}.m_2 \text{ (uncorrected)} + \frac{7}{240}.h^4;$$

where h denoted the uniform class-interval and m's are the moments about the mean of the grouped frequency distribution. The important point to note here is that these corrections are not applicable to highly skewed distributions and distributions having unequal class-intervals.

4.7.3 Moment-Ratios. There are certain ratios in which both the numerators and the denominators are moments. The most common of these moment-ratios are β_1 and β_2 defined by the relations,

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\beta_2 = \frac{\mu_4}{\mu_2^2}$. They are independent of origin and units of measurement, *i.e.* they are pure members. Actually, β_1 is the square of the third population moment expressed in *standard* units and β_2 is the fourth standardized moment for a population, where a standardized variable has been defined as

$$Z = (x - \mu)/\sigma.$$

For symmetrical distributions, β_1 is equal to zero. It is, therefore, used as a measure of skewness. β_2 is used to explain the shape of the curve and is a measure of peakedness. For the normal distribution to be discussed later, $\beta_2 = 3$.

The moment-ratios (or the standardized moments) for sample data are similarly defined as

$$b_1 = \frac{(m_3)^2}{(m_2)^3} \quad \text{and} \quad b_2 = \frac{m_4}{(m_2)^2}$$

Example 4.13 Calculate the first four moments about the mean for the following set of examination marks: 45, 32, 37, 46, 39, 36, 41, 48 & 36.

For convenience, the observed values are written in an increasing sequence. The necessary calculations appear in the table below:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
32	–8	64	–512	4096
36	–4	16	–64	256
36	–4	16	–64	256
37	–3	9	–27	81
39	–1	1	–1	1
41	1	1	1	1
45	5	25	125	625
46	6	36	216	1296
48	8	64	512	4096
360	0	232	186	10708

Now $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{360}{9} = 40$ marks

Therefore $m_1 = \dfrac{\sum(x_i - \bar{x})}{n} = 0$

$$m_2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{232}{9} = 25.78 \text{ (marks)}^2$$

$$m_3 = \frac{\sum(x_i - \bar{x})^3}{n} = \frac{186}{9} = 20.67 \text{ (marks)}^3$$

$$m_4 = \frac{\sum(x_i - \bar{x})^4}{n} = \frac{10708}{9} = 1189.78 \text{ (marks)}^4$$

Example 4.14 Compute the first four moments for the following distribution of wages after applying Sheppard's corrections.

Weekly Earnings (Rupees)	5	6	7	8	9	10	11	12	13	14	15
No. of men	1	2	5	10	20	51	22	11	5	3	1

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

We first calculate moments about an arbitrary origin. The necessary calculations are shown below. The moments about $x = 10$ are obtained by dividing the column sums by n.

Earnings in Rs. (x_i)	Men f_i	D_i $(x_i - 10)$	f_iD_i	$f_iD_i^2$	$f_iD_i^3$	$f_iD_i^4$
5	1	−5	−5	25	−125	625
6	2	−4	−8	32	−128	512
7	5	−3	−15	45	−135	405
8	10	−2	−20	40	−80	160
9	20	−1	−20	20	−20	20
10	51	0	−68	0	−488	0
11	22	1	22	22	22	22
12	11	2	22	44	88	176
13	5	3	15	45	135	405
14	3	4	12	48	192	768
15	1	5	5	25	125	625
Sums	131	--	+76 / +8	346	+562 / +74	3718
Sums÷n	1	--	0.06 $= m_1'$	2.64 $= m_2'$	0.56 $= m_3'$	28.38 $= m_4'$

Moments about the mean are:

$$m_1 = 0$$

$$m_2 = m_2' - (m_1')^2 = 2.64 - (0.06)^2 = 2.64;$$

$$m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3$$

$$= 0.56 - 3(2.64)(0.06) + 2(0.06)^3 = 0.08;$$

$$m_4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 - 3(m_1')^4$$

$$= 28.38 - 4(0.56)(0.06) + 6(2.64)(0.06)^2 - 3(0.06)^4 = 28.30$$

Applying Sheppard's corrections, we have

$$m_2 \text{ (corrected)} = m_2 \text{ (uncorrected)} - \frac{h^2}{12} = 2.64 - 0.08 = 2.56,$$

$$m_3 \text{ (corrected)} = m_3 \text{ (uncorrected)} = 0.08,$$

$$m_4 \text{ (corrected)} = m_4 \text{ (uncorrected)} - \frac{h^2}{2} . m_2 \text{ (uncorrected)} + \frac{7h^4}{240}$$

$$= 28.30 - 1.32 + 0.03 = 27.01$$

4.7.4 Change of Origin and Scale. Let a and h denote the arbitrary origin and the class-interval. Then we define a new variable u as

$$u_i = \frac{x_i - a}{h}$$

so that $x_i - a = hu_i$; $\bar{x} - a = h\bar{u}$ and hence $x_i - \bar{x} = h(u_i - \bar{u})$.

Substituting these values in the *rth* sample moments, we get

$$m'_r = \frac{1}{n}\sum f_i(x_i - a)^r = h^r.\frac{1}{n}\sum f_i u_i^r;$$

and $$m_r = \frac{1}{n}\sum f_i(x_i - \bar{x})^r = h^r.\frac{1}{n}\sum f_i(u_i - \bar{u})^r.$$

This shows that the *rth* moments of the variable X are h^r times the corresponding moments of the variable u, and are independent of the origin 'a'. In other words, the *moments are not affected by a change of origin but are affected by a change of scale.*

4.7.5 Charlier Check. We have seen that the computation of the moments depends upon the sum of the products of the frequencies by the corresponding values of the variable. It is, therefore, desirable to check these computations so that arithmetic mistakes, if any, are avoided. For this purpose, L.V. Charlier, the Norwegian statistician, introduced a check known as *Charlier check*. This check actually consists in shifting the assumed origin in the *coded* form by one interval. The relations used for this purpose are given below:

$$\sum f(u+1) = \sum fu + n$$

$$\sum f(u+1)^2 = \sum fu^2 + 2\sum fu + n$$

$$\sum f(u+1)^3 = \sum fu^3 + 3\sum fu^2 + 3\sum fu + n$$

$$\sum f(u+1)^4 = \sum fu^4 + 4\sum fu^3 + 6\sum fu^2 + 4\sum fu + n$$

Example 4.15 Calculate the first four moments about the mean from the data of Example 4.4.

The necessary calculations by taking $u_i = \dfrac{x_i - 114.5}{20}$, are set out in the following table. The last column is used for *Charlier's* check and the column sums are divided by n to get m'_r.

Data		Computations					
x_i	f_i	u	fu	fu^2	fu^3	fu^4	$f(u+1)^4$
74.5	9	−2	−18	36	−72	144	9
94.5	10	−1	−10	10	−10	10	0
114.5	17	0	0	0	0	0	0
134.5	10	1	10	10	10	10	160
154.5	5	2	10	20	40	80	405
174.5	4	3	12	36	108	324	1024
194.5	5	4	20	80	320	1280	3125
Sum	60	--	24	192	396	1848	4740
Sums÷n	1	--	0.4 $= m'_1$	3.2 $= m'_2$	6.6 $= m'_3$	30.8 $= m'_4$	For charller's check

Charlier's check.

$$\sum f(u+1)^4 = \sum fu^4 + 4\sum fu^3 + 6\sum fu^2 + 4\sum fu + n$$

$$= 1848 + 4(396) + 6(192) + 4(24) + 60$$

$$= 1848 + 1584 + 1152 + 96 + 60 = 4740,$$ which is the sum in the last column.

Hence the moments about the mean and in *class-interval units* are obtained as below:

$$m_1 = 0$$

$$m_2 = m_2' - (m_1')^2$$

$$= 3.2 - (0.4)^2 = 3.04$$

$$m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3$$

$$= 6.6 - 3(3.2)(0.4) + 2(0.4)^3 = 2.89$$

$$m_4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 - 3(m_1')^4$$

$$= 30.8 - 4(6.6)(0.4) + 6(3.2)(0.4)^2 - 3(0.4)^4 = 23.24$$

To get the moments about the mean in *ordinary units*, we multiply m_2 by h^2, *i.e.* 400, m_3 by $(20)^3$ and m_4 by $(20)^4$. Thus $m_2 = 1216$, $m_3 = 23120$ and $m_4 = 3718400$.

Example 4.16 The first three moments of a distribution about the value 2 of the variable are 1, 16 and –40. Show that the mean is 3, the variance 15 and $m_3 = -86$. Also show that the first three moments about $x = 0$ are 3, 24 and 76.

Here we are given $m_1' = \frac{1}{n}\sum f(x-2) = 1$ ... (1)

$$m_2' = \frac{1}{n}\sum f(x-2)^2 = 16 \quad \text{... (2)}$$

$$m_3' = \frac{1}{n}\sum f(x-2)^3 = -40 \quad \text{... (3)}$$

We also know that $m_1' = \bar{x} - a$, so that

$$\bar{x} = m_1' + a = 1 + 2 = 3 \qquad (\because a = 2)$$

And variance, $S^2 = m_2$ (second moment about mean)

$$= m_2' - (m_1')^2 = 16 - 1 = 15, \text{ and}$$

$$m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3$$

$$= -40 - 3(16)(1) + 2(1)3 = -86.$$

To find the moment about $x = 0$, we need the values of $\frac{1}{n}\sum fx$, $\frac{1}{n}\sum fx^2$ and $\frac{1}{n}\sum fx^3$, which are obtained from relations (1), (2) and (3).

From (1), $\frac{\sum fx}{n} = 3$

From (2), $\frac{1}{n} = \sum f(x-2)^2 = 16$

or $\frac{1}{n}\sum f(x^2 - 4x + 4) = 16$

or $\frac{1}{n}\sum fx^2 - 4\frac{\sum fx}{n} + 4 = 16$

or $\frac{1}{n}\sum fx^2 = 16 - 4 + 4(3) = 24$

(3) on expansion can be written as

$$\frac{1}{n}\sum fx^3 - 6\frac{1}{n}\sum fx^2 + 12\frac{\sum fx}{n} - 8 = -40$$

or $\frac{1}{n}\sum fx^3 = -40 + 8 - 12(3) + 6(24) = 76$

Hence the moments about $x = 0$ are

$$m_1' = \frac{\sum fx}{n} = 3,$$

$$m_2' = \frac{\sum fx^2}{n} = 24, \text{ and}$$

$$m_3' = \frac{\sum fx^3}{n} = 76.$$

Example 4.17 Show that for discrete distributions, $\beta_2 > 1$.

By definition, $\beta_2 = \frac{\mu_4}{\mu_2^2}$. (P.U., D. St. 1962)

Now β_2 will be greater than one if the numerator is greater than the denominator, i.e. if $\mu_4 > \mu_2^2$, or if $\mu_4 - \mu_2^2 > 0$.

Now $\mu_4 - \mu_2^2 = \frac{1}{N}\sum f(x-\mu)^4 - \sigma^4 \quad (\because \mu_2 = \sigma^2)$

$$= \frac{1}{N}\sum f(x-\mu)^4 + \sigma^4 - 2\sigma^4$$

$$= \frac{1}{N}\sum f(x-\mu)^4 + \frac{\sigma^4 \sum f}{N} - 2\sigma^2 . \frac{\sum f(x-\mu)^2}{N}, \quad \left[\sigma^2 = \frac{\sum f(x-\mu)^2}{N}\right]$$

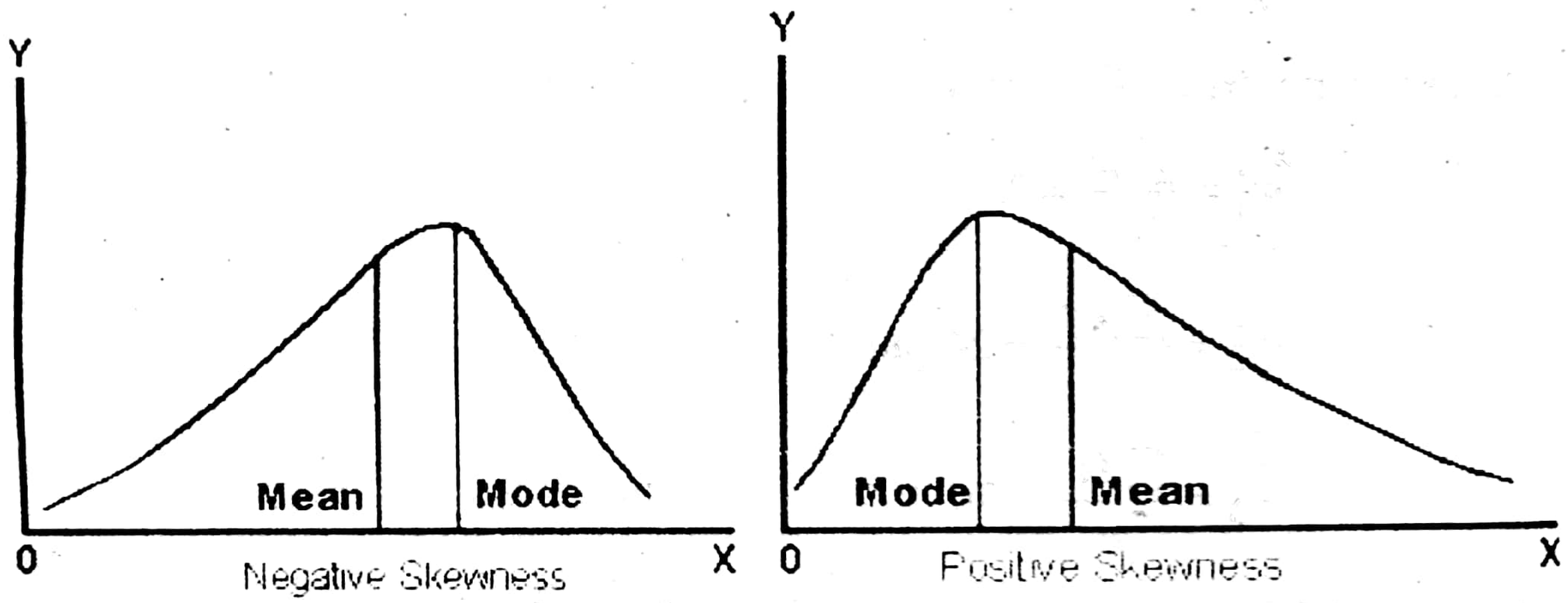
$$= \frac{1}{N}\sum f[(x-\mu)^4 + \sigma^4 - 2\sigma^2(x-\mu)^2] \quad (N\sigma^4 = \sigma^4 \sum f)$$

$$= \frac{1}{N}\sum f[(x-\mu)^2 - \sigma^2]^2 \text{ which is essentially positive.}$$

Hence $\beta_2 \geq 1$ because $\mu_4 - \mu_2^2$ is always positive.

4.8 SKEWNESS

A distribution in which the values equidistant from the mean have equal frequencies is defined to be *symmetrical* and any departure from symmetry is called *skewness*. It is important to note that in a perfectly symmetrical distribution, the mean, median and mode coincide and that the two tails of the distribution are equal in length from the mean. These values are pulled apart when the distribution departs from symmetry and consequently one tail becomes longer than the other. If the right tail is longer than the left tail, the distribution is said to have *positive skewness*. If the left tail distribution is longer than its right tail, it is said to be *negatively skewed* or to have *negative skewness*. In a positively skewed distribution, the mean is greater than the median and the median is greater than the mode, *i.e.* mean > median > mode and in a negatively skewed distribution, mode > median > mean.



The *difference* between the measures of location, being an indication of the amount of skewness or asymmetry, is used as a measure of skewness. A measure of skewness is defined in such a way that (i) the measure should be zero when the distribution is symmetric and (ii) the measure should be a pure number, *i.e.* independent of origin and units of measurements.

According, to measure the degree of skewness of a distribution or curve, Karl Pearson (1857–1936) introduced a coefficient of skewness denoted by *Sk* and defined by

$$Sk = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

We know that mode is sometimes ill-defined and is difficult to locate by simple methods. It is, therefore, replaced by its equivalent from empirical relation holding good in moderately skewed distributions. The *Pearsonian co-efficient of skewness* then becomes

$$Sk = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

The coefficient usually varies between –3(negative skewness) and +3 (positive skewness) and the sign indicates the direction of skewness. The formula satisfies both the requirements considered essential for a measure of skewness.

Arthur Lyon Bowley (1869–1957), a British statistician, has also proposed a measure of skewness that is based on the median and the two quartiles. In a symmetrical distribution, the two quartiles are equidistant from the median but in an asymmetrical distribution, this will not be the case. The *Bowley's co-efficient of skewness* is

$$Sk = \frac{Q_1 + Q_3 - 2\,\text{Median}}{Q_3 - Q_1}$$

Its values lies between 0 and ±1.

Another measure of skewness that is often used, is the third moment express in standard units (or the moment ratio) and thus is given by

$$Sk = \frac{\mu_3}{\sigma^3}, \quad \text{for population data,}$$

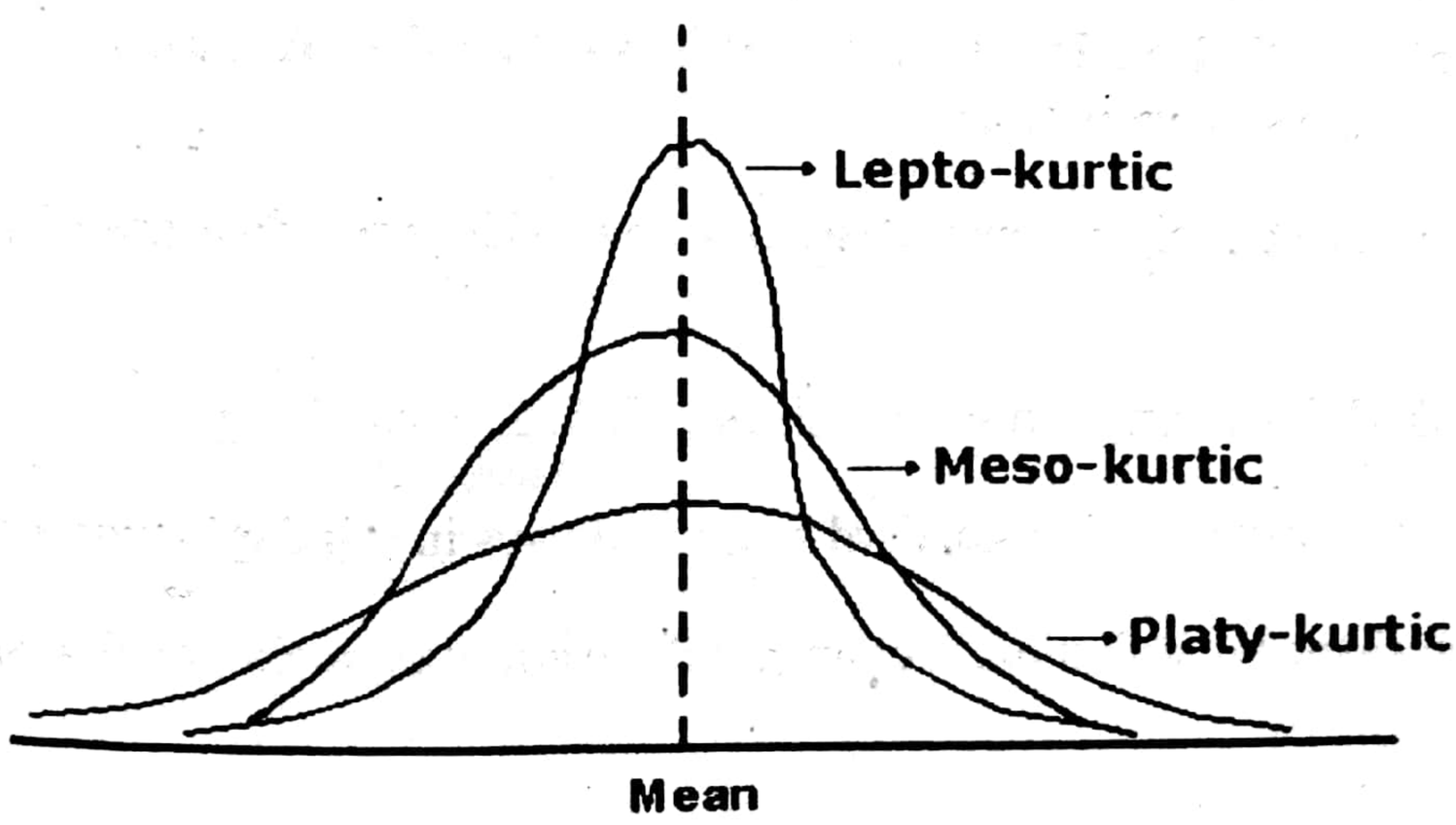
$$= \frac{m_3}{s^3}, \quad \text{for sample data,}$$

This coefficient for most distributions, will be between –2 and +2. Some statisticians denote it by α_3 or $\sqrt{\beta_1}$. If the coefficient is greater than zero, the distribution or curve is positively skewed. If $Sk < 0$, there is negative skewness. For symmetrical distributions or curves, the coefficient is zero.

4.9 KURTOSIS

Karl Pearson (1857–1936) introduce the term *Kurtosis* (literally the amount of hump) for the degree of *peakedness* or *flatness* of a unimodal frequency curve. When the values of a variable are closely bunched round the mode in such a way that the curve is *leptokurtic*. If, on the other hand, the curve is flat-topped, we say that the curve is *platykurtic*. Since the *normal curve* (to be described later) is neither vary peaked nor very flat-topped, it is taken as basis for comparison. The normal curve itself is called *mosekurtic*.

Kurtosis is usually measured by the fourth standardized moment or the moment-ratio $\beta_2 = (\mu_4 / \mu_2^2)$ whose value for a normal distribution is equal to 3. When β_2 is greater than 3, the curve is more sharply peaked and has wider tails than the normal curve and is said to be *leptokurtic*. When it is less than 3, the curve has a flatter top and relatively narrower tails than the normal curve and is said to be *platykurtic*.



The corresponding measure of kurtosis for the sample data is $b_2\left(=\frac{m_4}{m_2^2}\right)$. It should be noted that the value of b_2 for a large sample from the *normal population* is very nearly 3.

Another measure of Kurtosis not widely used, is given by

$$K = \frac{Q.D.}{P_{90} - P_{10}},$$

where Q.D. is the semi-interquartile range and P's are the *percentiles*. This is known as the *Percentile co-efficient of kurtosis*. It has been shown that K for a normal distribution is 0.263 and that it lies between 0 and 0.50.

4.10 DESCRIBING A FREQUENCY DISTRIBUTION

To describe the major characteristics of a frequency distribution, we need the calculations of the following five quantities:

i) The number of observations that describes the *size* of the data.

ii) A measure of central tendency such as the mean or median that provides information about the *centre* or *average* value.

iii) A measure of dispersion such as standard deviation that indicates the *variability* of the data.

iv) A measure of skewness that shows the *lack of symmetry* in the frequency distribution.

v) A measure of kurtosis that gives information about its *peakedness*.

It is interesting to note that all these quantities can be derived from the first four moments. For example, the first moment about $x = 0$ is the arithmetic mean, the second moment about the mean is the variance and its positive square root is the standard deviation. The third mean moment is a measure of skewness while the fourth central moment is used to measure kurtosis. Thus the first four moments play a key role in describing frequency distributions.

EXERCISES

OBJECTIVE

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

i) A measure of <u>central tendency</u> describes how widely the data are dispersed about a central value.

ii) The standard deviation for a set of values 5, 5, 5, 5, 5 and 5 is <u>5</u>.

iii) The unit of measure for the standard score is always in <u>standard deviation</u>.

iv) For a bell shaped distribution, the range will be approximately equal to <u>six</u> standard deviation.

v) The difference between the largest and the smallest observations is called the Quartile Deviation.

vi) The square of variance gives the standard deviation.

vii) The coefficient of variation is an absolute measure of dispersion.

viii) The coefficient of variation is measured in different units as the data.

ix) The Quartile Deviation is based on only two values in the series.

x) The square root of the variance of a distribution is the absolute deviation.

b) MULTIPLE CHOICE QUESTIONS

i) The main disadvantage of the range is that

a) It does not use all the observations in its calculation.

b) It can be influenced by an extreme value.

c) Both a and b are correct.

d) None of the above.

ii) Which one of the following is not a measure of dispersion?

a) Range.

b) Standard deviation.

c) Second quartile.

d) Coefficient of variation.

iii) Which of the following is not a measure of dispersion?

a) Interquartile range.

b) Difference between the values of the largest and smallest items.

c) Mean of the values of the largest and smallest items.

d) Standard deviation.

iv) The standard deviation is

a) The square of the variance.

b) Two times the standard deviation.

c) Half the variance.

d) The square root of the variance.

v) The coefficient of variation is measured in

a) The same units as the mean and the standard deviation.

b) Percent.

c) Squared units.

d) None of the above.

vi) If the original units are measured in pounds, the variance is

a) Also measured in pounds.

b) Measured in pounds squared.

c) Measured in half pounds.

d) None of the above.

vii) If the tail of a frequency distribution is in positive direction (to the right), the coefficient of skewness is

a) Zero.

b) Positive.

c) Negative.

d) None of the above.

viii) The standard deviation of a frequency distribution is 10, the mean is 250, the median is 250 and the mode is also 250. The coefficient of skewness is

a) Zero.

b) Positive.

c) Negative.

d) None of the above.

ix) Which of the following statement is true?

a) The standard deviation is less them the range.

b) The range is less than the interquartile range.

c) The arithmetic mean always exceeds the median.

d) The arithmetic man always exceeds the mode.

x) Which of the following is not a property of the standard deviation?

a) It is always negative number.

b) It is affected by extreme values in a data set.

c) It is based on all the values in the data set.

d) It is the most widely used measure of dispersion.

xi) If a distribution has zero standard deviation, then which of the following is true?

a) All observations are negative.

b) All observations are positive.

c) All observations are equal.

d) Number of positive values and negative values are equal.

xii) The empirical rule generally can be applied to

a) Bell shaped distribution.

b) Any distribution.

c) Only continuous distribution.

d) Any skewed distribution.

xiii) Symmetrical distribution will always have skewness equal to

a) Negative.

b) Positive.

c) Zero.

d) Approximately zero.

xiv) For a normal distribution the measure of Kurtosis equals to

a) Zero.

b) 3.

c) Positive number.

d) Negative number.

xv) For the given sample data set 2, 8, 10, 15, 20, 9, 18, 0, 7, 10, which is the value of coefficient of variation?

a) 70.00 percent.

b) 15.50 percent.

c) 145.00 percent.

d) 61.21 percent.

SUBJECTIVE

4.1 Explain clearly the meaning of the term Dispersion. What are the most usual methods of measuring dispersion? Indicate the advantages and disadvantages of these methods.
(P.U., B.Com. 1960; B.A. (Hons.), 1960; B.A. (Part-I), 1961)

4.2 Discuss the different measures of dispersion. Describe the method of computation of any two of them with suitable examples.
(P.U., M.A., Econ. 1969)

4.3 Describe carefully how Mean Deviation, Standard Deviation and Quartile Deviation of any given distribution are obtained. In what problems, should each be used?

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

4.4 a) What is Range and how is it calculated? What are its uses?

b) Define Quartile Deviation. Find the quartile deviation from the following data (i) graphically, (ii) using an appropriate formula.

Income per week (Rs.)	41–50	51–60	61–70	71–80	81–90	91–100	Total
No. of Earners	30	36	43	104	73	14	300

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

4.5 The members of a sports club, 60 male adults, had their weights recorded, in pounds. The weights are given below:

171	160	144	132	154	160	160	158	148	160	131	153
131	165	139	163	149	149	140	149	150	161	136	144
165	174	153	149	157	169	147	156	149	171	149	154
153	149	147	154	145	158	160	152	156	138	167	142
165	155	140	155	158	147	149	169	148	174	150	144

Construct a cumulative frequency table for these weights, using classes of width 5 lb, starting at 129.5 lb. Hence draw a cumulative frequency graph, and use this to find the median and semi-interquartile range.

Use the grouped frequency table to calculate the mean and standard deviation, and compare them with the values obtained using the original ungrouped data.

(M.A. Econ. I.U. 1990; B.Z.U. 1992)

4.6 a) Define Mean Deviation and its co-efficient. Discuss its advantages and uses.

b) Estimate the mean deviation from the arithmetic mean of the following set of examination marks.

Marks	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79
No. of students	2	3	8	24	27	40	11	5

4.7 When originally collected, the data in the following distribution had been in 5–yearly age-groups throughout the whole range 20–64. Using central values of these age-groups, the mean age had been calculated to be 44.5 years and the mean deviation without correction for grouping to be 7.15 years. Reconstruct the original table with 5-yearly age groups from the information given.

Age nearest birth day:	20–24	25–29	30–39	40–44	45–49	50–54	55–64
Number of men:	1	2	26	22	20	15	14

(P.C.S., 1971, 93; P.U., B.A./B.Sc. 1975)

Solution: Let f_1 and f_2 be the frequencies corresponding to the groups 30–34 and 55–59 in 5–yearly age groups. Then the calculations will proceed as below:

Age-group	C.V. (x_i)	f	fx	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
20–24	22	1	22	22.5	22.5
25–29	27	2	54	17.5	35.0
30–34	32	f_1	$32f_1$	12.5	$12.5f_1$
35–39	37	$(26-f_1)$	$962-37f_1$	7.5	$195.0-7.5f_1$
40–44	42	22	924	2.5	55.0
45–49	47	20	940	2.5	50.0
50–54	52	15	780	7.5	112.5
55–59	57	f_2	$57f_2$	12.5	$12.5f_2$
60–64	62	$(14-f_2)$	$868-62f_2$	17.5	$245.0-17.5f_2$
Total	--	100	$4550-5(f_1+f_2)$	--	$715+5(f_1-f_2)$

Now, Mean $=\frac{1}{n}\sum fx$,

i.e. $44.5=\frac{1}{100}[4550-5(f_1+f_2)]$

or $f_1+f_2=20$ ... (A)

and Mean Deviation $=\frac{1}{n}\sum f\lvert x-\bar{x}\rvert$

i.e. $7.15=\frac{1}{100}[715+5(f_1-f_2)]$

or $f_1-f_2=0$ ... (B)

From (A) and (B) we find that $f_1=10$ and $f_2=10$.

4.8 Find the quartile deviation and mean deviation as well as their co-efficients from the following data and comment.

Height (inches)		58	59	60	61	62	63	64	65
No. of Persons	Group A	10	18	30	42	35	28	16	8
	Group B	15	20	32	35	33	22	20	10

4.9 a) Define Variance and Standard Deviation. Describe their properties.

b) For a population of numbers 10, 8, 7, 9, 5, 12, 8, 6, 8, 2, calculate σ^2 and σ.

(P.U., M.A. Econ. 1992)

4.10 a) Prove that the variance remains unchanged when a constant is added to or subtracted from every value of the variable.

b) The scores obtained by five students on a set of examination papers are 70, 50, 60, 70, 50. These scores are changed by (i) adding 10 points to all scores, (ii) increasing all scores by 10%. What effect will these changes have on the mean and on the standard deviation?

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

4.11 a) Define the mean and standard deviation of a distribution. $\bar{x}$ is the mean and S, the standard deviation. When a provisional mean "a" is chosen, the corresponding provisional standard deviation is found to be S_1. Prove that $S_1^2 = S^2 + (\bar{x} - a)^2$. Explain briefly the advantages of this procedure in numerical work.

(P.U., B.A./B.Sc. 1978)

b) For a set of ungrouped values, the following sums are found: $n = 15$, $\Sigma x = 480$, $\Sigma x^2 = 15735$. Find the mean and the standard deviation.

(B.Z.U., B.A./B.Sc. 1976)

c) The mean and standard deviation of a sample of 20 observations were found to be 75 and 2.5 respectively. On checking the original figures, it was discovered that one observation which was actually 68, was copied down as 86. Find the correct mean and standard deviation.

4.12 a) Describe the properties of the standard deviation.

b) By multiplying each of the numbers 3, 6, 2, 1, 7, 5 by 2 and then adding 5, we obtain 11, 17, 9, 7, 19, 15. What is the relation between the standard deviations and the means for the two sets?

(P.U. B.A./B.Sc. 1972, 81)

c) A child is born to Mrs. X every year for 7 consecutive years. Compute the standard deviation of the children's age when (i) the youngest is 1 year old, and (ii) the youngest is 8 years old. Why do the answers of (i) and (ii) coincide? (P.U., B.A./B.Sc. 1980)

4.13 Show how to compute the Range and the Standard deviation of a sample of n observations, and explain briefly the meaning of these statistics, giving examples of situations in which they would be used.

Show that when $n = 2$, the two estimates are simply related.

4.14 It is often stated that in frequency distributions there exists the approximate relation: $\dfrac{\text{Mean Deviation}}{\text{Standard Deviation}} = 0.8$. Test this statement in the following distribution:

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

4.15 Calculate the mean and standard deviation for the following distribution of lengths of 200 metal bars.

x:	30	31	32	33	34	35	36	37	38	39
f:	4	8	23	35	62	44	18	4	1	1

(P.U., D.St. 1963; W.P.C.S., 1964)

4.16 Explain how Chebyshev's rule can be used to answer questions about a data set? Using this rule, find and interpret the interval $\bar{x} \pm 2s$ for question 4.15.

4.17 The following table gives the frequency distribution of expenditure on food per family per month among working class families in two localities. Find the arithmetic mean and the standard deviation of the expenditure at both places.

Range of expenditure in rupees per month	No. of Families	
	Place A	Place B
30–60	28	39
60–90	292	284
90–120	389	401
120–150	212	202
150–180	59	48
180–210	18	31
210–240	2	5

(P.U., B.A., (Part-I), 1961; M.A. Econ. 1970)

4.18 What do your understand by Variance? The wages of 1,000 employees range from Rs.4.50 to Rs.19.50. They are grouped in 15 classes with a common class interval of Rs.1, and the class frequencies from the lowest class to the highest class are: 6, 17, 35, 48, 65, 90, 131, 173, 155, 117, 75, 52, 21, 9 and 6. Find the mean wage and its standard deviation.

(P.U., D.St., 1965, P.C.S. 1986)

4.19 The breaking strength of 20 test pieces of a certain alloy is given as under:

95	103	97	130	96	73	78	95	89	68
82	79	69	67	83	108	94	87	93	117

Calculate the average breaking strength of the alloy and the standard deviation. Calculate the percentage of observations lying within the limits: mean $\pm S$; mean $\pm 2S$; mean $\pm 3S$; where S stands for the standard deviation. (P.U., B.Com., 1961, I.U., M.A. Econ. 1989)

4.20 A collar manufacturer is considering the production of a new style of collar to attract young men. The following statistics of neck circumferences are available based upon measurements of a typical group of college students.

Mid-value (inches)	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0	16.5
No. of students	4	19	30	63	66	29	18	1	1

Compute the standard deviation and use the criterion $\bar{x} \pm 3$ (standard deviation) to determine the largest and smallest size of collars he should make in order to meet the needs of practically all his customers, bearing in mind that collars are worn, on average, ¾ inches larger than neck size. (P.U., B.A./B.Sc. 1960)

4.21 The values of the arithmetic mean and standard deviation of the following frequency distribution of a continuous variable derived from the use of working origin and scale are £135.3 and £9.6 respectively. Determine the actual class-intervals.

u	–4	–3	–2	–1	0	1	2	3	Total
f	2	5	8	18	22	13	8	4	80

(P.U., D.St.; 1964; B.A./B.Sc. 1973)

4.22 In the manufacture of a certain scientific instrument great importance is attached to the life of a particular critical component. This component is obtained in bulk from two sources, A and B, and in the course of inspection, the lives of 1,000 of the components from each source are determined. The following frequency tables are obtained:

Source A		Source B	
Life (hours)	No. of components	Life (hours)	No. of components
1,000 – 1,020	40	1,030 – 1,040	339
1,020 – 1,040	96	1,040 – 1,050	136
1,040 – 1,060	364	1,050 – 1,060	25
1,060 – 1,080	372	1,060 – 1,070	20
1,080 – 1,100	85	1,070 – 1,080	130
1,100 – 1,120	43	1,080 – 1,090	350

Examine the effectiveness of the measures of dispersion with which you are familiar for comparing the dispersions of the two distributions.

4.23 Show that, for any discrete distribution, the mean deviation about the mean is not greater than the standard deviation. (P.U., B.A./B.Sc. 1981)

Solution. By definition,

Mean Deviation $= \frac{1}{n}\Sigma f_i|x_i - \bar{x}|$, and

$$s = \sqrt{\frac{1}{n}\Sigma f_i(x_i - \bar{x})^2}, \text{ where } n = \Sigma f_i$$

We are required to show that

$$\frac{1}{n}\Sigma f_i d_i^2 > \left(\frac{1}{n}\Sigma f_i d_i\right)^2, \text{ where } d_i = |x_i - \bar{x}|$$

i.e. $n[\Sigma f_i d_i^2] > (\Sigma f_i d_i)^2$

i.e. $(f_1 + f_2 + \ldots + f_k)(f_1 d_1^2 + f_2 d_2^2 + \ldots + f_k d_k^2) > (f_1 d_1 + f_2 d_2 + \ldots + f_k d_k)^2$

i.e. $\Sigma f_i^2 d_i^2 + \sum_{i \neq j} f_i f_j (d_i^2 + d_j^2) > \Sigma f_i^2 d_i^2 + 2\sum_{i \neq j} f_i f_j d_i d_j)$

i.e. $\sum_{i \neq j} f_i f_j (d_i - d_j)^2 > 0$

which is true because $(d_i - d_j)^2$ is always positive.

4.24 a) What is the co-efficient of variation? What purpose does it serve?

b) The following data have been obtained from a frequency distribution of a variable after making the substitution $X = 62 + 5u : \Sigma f = 120, \ \Sigma fu = 140, \Sigma fu^2 = 598$, calculate the co-efficient of variation, using corrected standard deviation.

(P.U., B.A./B.Sc. 1987)

4.25 a) What do you mean by absolute and relative measures of dispersion? State the uses of the co-efficient of variation in statistical analysis. (P.U., B.A./B.Sc. 1976)

b) Given below is the distribution of weekly income (to the nearest rupee) of 100 households in a locality A. Calculate the standard deviation.

Income	35–39	40–44	45–49	50–54	55–59	60–64	65–69
f	13	15	17	28	12	10	5

If 123 households in a different locality B had a mean weekly income of Rs.52.28 and a standard deviation of Rs.4.96, then compare the variability of the weekly income of two localities.

(B.Z.U., M.A. Econ. 1986; P.U., B.A./B.Sc. 1974)

4.26 a) Define Range, Mean Deviation and Standard Deviation, and compare their merits as descriptive measures of dispersion.

b) Two candidates X and Y at the B.A. (Hons.) Examination obtained the following marks in ten papers. Which of the candidate showed a more consistent performance?

Paper	I	II	III	IV	V	VI	VII	VII	IX	X
X	58	49	76	80	47	72	61	59	77	48
Y	39	38	86	72	75	69	57	49	83	66

(B.Z.U., M.A. Econ. 1986; P.U., B.A./B.Sc. 1974)

4.27 a) Calculate the corrected co-efficient of variation when mean = 67.45, variance (uncorrected) = 8.5275 and the class-interval is 3. (P.U., B.A./B.Sc. 1968)

b) The following are the scores made by two batsmen A and B in a series of innings:

A	12	15	6	73	7	19	199	36	84	29
B	47	12	76	48	4	51	37	48	13	0

Who is better as a run getter? Who is the more consistent player?

(P.U., B.A./B.Sc. 1970; B.Z.U., M.A. Econ. 1984)

4.28 a) Explain the difference between absolute dispersion and relative dispersion. Describe the properties of the standard deviation.

b) Find the co-efficient of variation from the following data using both uncorrected and corrected standard deviations.

Weight (lb)	118–126	127–135	136–144	145–153	154–162	163–171	172–180
f	3	5	9	12	5	4	2

c) A manufacturer of television tubes has two types of tubes A and B. The tubes have respective mean life-times $\bar{x}_A = 1495$ hours and $\bar{x}_B = 1895$ hours: and standard deviations $S_A = 280$ hours and $S_B = 310$ hours. Which tube has the greater (i) absolute dispersion, (ii) relative dispersion? (P.U., B.A./B.Sc. 1969, 71)

4.29 Compare the variability of expenditure of families in two towns as given below:

Expenditure (Rupees)	No. of Families	
	Town A	Town B
21 – 30	3	2
31 – 40	61	14
41 – 50	132	20
51 – 60	153	27
61 – 70	140	28
71 – 80	51	7
81 – 90	2	2

(B.Z.U., M.A. Econ. 1985)

4.30 Compare the variations in the following frequency distributions of weights of boys by computing the co-efficient of variation in each case: Also draw a box plot.

Weight (kilograms)	Classes		
	A	B	C
$20 \le x < 30$	7	5	6
$30 \le x < 40$	10	9	25
$40 \le x < 50$	20	21	24
$50 \le x < 60$	18	15	4
$60 \le x < 70$	7	6	3

4.31 If in a series of measurements, we obtain n_1 values of magnitude x_1, n_2 of magnitude x_2, and so on, and if $\bar{x}$ is the mean value of all the measurements, prove that the standard deviation is

$$\sqrt{\frac{\Sigma n_r (k - x_r)^2}{\Sigma n_r} - \delta^2}, \text{ where } \bar{x} = k + \delta.$$

(B.Z.U., B.A./B.Sc. 1986)

4.32 a) For a group of 50 boys, the mean score and standard deviation of scores on a test are 59.5 and 8.38. For a group of 40 girls, the mean and standard deviation are 54.0 and 8.23 on the same test. Find the mean and standard deviation for the combined group of 90 children.

b) A distribution consists of three components with frequencies 200, 250 and 300 having means of 25, 10 and 15, and standard deviations of 3, 4 and 5 respectively. Show that the mean of the combined distribution is 16 and its s.d. is 7.2 approximately. Find also the coefficient of variation.

4.33 a) What is meant by a standardized variable?

b) Show that for any distribution expressed in standard measures, the mean is zero and the standard deviation is one.

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

c) **The mean of scores for a group of students on a certain test was 63.7 with a standard deviation of 12.3. Find the Z score for the top student, with a score of 98 and the bottom student, with a score of 21.**

4.34 Three tests had the values in the following table for mean and standard deviations.

Test	$\bar{x}$	S
1	70	5
2	75	8
3	60	12

Student A received grades of 70, 90, 70 on the three tests, while student B received grades of 90, 70, 70. Assuming that all three tests should carry equal weight, change the grades to Z scores and average each student's scores. (P.U., B.Sc. (Hons.) Part-I, 1971)

4.35 a) Explain what is meant by the trimmed and Winsorized measures of central tendency and dispersion.

b) Calculate the trimmed and the Winsorized means and standard deviations for the following set of 15 scores:

80, 75, 42, 63, 65, 43, 78, 96, 82, 58, 79, 72, 67, 75, 68.

4.36 a) Define Moments about an arbitrary origin and about the mean. Express the moments about the mean in terms of the moments about any point and conversely.

b) Give Sheppard's corrections to moments and explain where they are used.

(P.U., B.A. (Part-I), 1961, 1962-S)

4.37 a) Prove that the second moment about an arbitrary origin equals the second moment about the mean increased by the square of the distance between the arbitrary origin and the mean. (B.Z.U., B.A./B.Sc. 1976)

b) Derive the relations which give the third and fourth moments about the mean in terms of moments about the origin.

4.38 a) What aspects or characteristics of a frequency distribution are measured by the moments? (P.U. B.A./B.Sc. 2008)

b) Prove the general formula connecting the moments about the mean with the moments about the origin

$$\mu_r = \mu'_r - r\mu'_r\,\mu'_{r-1} + \frac{r(r-1)}{2!}\mu_1'^2\,\mu'_{r-2} - \cdots$$

(P.U. B.A./B.Sc. 1985, 89, 08)

c) Explain how changes of origin and of scale affect the following:

Mean, standard deviation, moments and β_2. (P.U. B.A./B.Sc. 1985, 89)

4.39 a) The first four moments of a distribution about $x = 2$ are 1, 2.5, 5.5 and 16. Calculate the four moments about the mean and about zero. (P.U., B.A./B.Sc. 1973, 76, 08)

b) The following data have been obtained from a frequency distribution of a variable X after making the substitution $x = 10 + 5u$.

$\Sigma f = 125, \Sigma fu = -46, \Sigma fu^2 = 806, \Sigma fu^3 = -242$ and $\Sigma fu^4 = 1962$.

Calculate the Mean, Variance, b_1 and b_2 (moment-ratios). Would you consider the distribution normal? (P.U., B.A./B.Sc. 1975, 93)

4.40 Calculate the first four moments about the mean for the following data.

x	1	2	3	4	5	6	7	8	9
f	1	6	13	25	30	22	9	5	2

4.41 Calculate the first four moments and apply the test of normality to the following data which pertain to weekly earnings in rupees of 200 labourers.

Weekly wages	15	16	17	18	19	20	21	22	23	24
No. of Labourers	6	19	13	18	20	25	28	34	22	15

4.42 Compute the first four moments about the mean from the following frequency distribution. Also calculate b_1 and b_2.

Groups	2–4	4–6	6–8	8–10	10–12	12–14	14–16	16–18	18–20
Frequency	18	24	47	80	102	66	40	21	15

(P.U., D.St. 1965)

4.43 a) Define the moment-ratios β_1 and β_2. What information do they give?

b) Eight coins were tossed together 256 times and the frequency of the occurrence of 0, 1, 2, ..., 8 heads in a throw was:

Number of heads	0	1	2	3	4	5	6	7	8
Frequency	1	7	26	54	74	52	32	9	1

Calculate the values of moment-ratios b_1 and b_2.

4.44 Calculate the values of b_1 and b_2 from the following data. Use Charlier check and Sheppard's corrections.

x	10–12	12–14	14–16	16–18	18–20	20–22	22–24	24–26	26–28
f	3	30	110	218	275	222	108	32	2

4.45 a) What is meant by skewness? How would you find it in a non-symmetrical distribution? Distinguish between positive and negative skewness.

b) Define the moment-ratios β_1 and β_2, and state the purpose for which they are used. (P.U., B.A. (Part-I), 1961-S)

4.46 a) State the measures commonly employed to define Skewness and Kurtosis. What aspects of the frequency curve are measured by them?

b) What can you say of the skewness in each of the following cases?

i) The median is 49.21 while the two quartiles are 37.15 and 61.27.

ii) Mean = 1403 and Mode = 1487.

iii) The first three moments about 16 are respectively –0.35, 2.09 and –1.93. (P.U., B.A./B.Sc. 1974, 78; M.A. Econ. 1977)

4.47 Calculate skewness by (i) the Pearsonian method, and (ii) by the Bowley's formula from the following frequency distribution and interpret the result.

Ages (years)	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54
No. of Men	29	176	208	173	82	40	15	3

(P.U., M.A. Econ. 1986)

4.48 Calculate the first four moments about the mean and provide one estimate each of skewness and kurtosis of the following distribution.

Age:	25	30	35	40	45	50	55	60
Frequency:	2	8	18	27	25	16	7	2

(P.U., M.A. Econ. 1967, I.U., 1992)

4.49 The following values have been obtained from two different frequency distributions of weights (lb) having 125 and 200 observations respectively after making the substitution:

$X = 16 + 5u,\ Y = 20 + 2v,$

a) $\Sigma fu = -46,\ \Sigma fu^2 = 306,\ \Sigma fu^3 = -242,\ \Sigma fu^4 = 1962$

b) $\Sigma fv = 21,\ \Sigma fv^2 = 1265,\ \Sigma fv^3 = -627,\ \Sigma fv^4 = 14169$

Find i) which of the distributions is more consistent,

ii) which of the distributions is negatively skewed;

iii) which of the distributions is Meso-kurtic. (P.U., B.A./B.Sc. 1993)

4.50 a) The fourth mean moment of a symmetrical distribution is 243. What would be the value of the standard deviation in order that the distribution may be mesokurtic?

(P.U., B.A./B.Sc. 1981)

b) In a certain distribution, the first four moments about the point 4 are –1.5, 17, –30 and 108. Calculate b_1 and b_2 and state whether the distribution is lepto-kurtic or platykurtic. (P.U., D.St., 1962)

4.51 a) The second moments about the mean of two distributions are 9 and 16, while the fourth moments about the mean are 230 and 780 respectively. Which of the distribution is (i) leptokurtic, (ii) mesokurtic, (iii) platykurtic?

b) The second moment about the mean of a symmetrical distribution is 25. What must be the value of the fourth moment about the mean in order that the distribution be (i) leptokurtic, (ii) mesokurtic, (iii) platykurtic? (P.U., B.A./B.Sc. 1971)

4.52 Discuss the various measures or quantities by which the characteristics of frequency distributions are measured and compared. (P.C.S. 1993)

4.53 Explain briefly how averages, measures of dispersion, skewness and kurtosis are complementary to one another in understanding a frequency distributions.

4.54 The mean and standard deviation of a variable X are 60 and 8.944 respectively. Find the mean and standard deviation of a new variable if

i) All the values of X are increased by 20 points.

ii) All the values of X are increased by 25%. (P.U., B.A./B.Sc. 2008-S)

4.55 The mean, mode and standard deviation of the weekly earnings of a random sample of women workers from a locality are 3133.33, 2804.35 and 796.70 respectively.

i) Calculate Skewness of the distribution and interpret the result. Also find coefficient of variation.

ii) What will happen to the values of the mean and standard deviation if every woman has an increase of Rs.500 per week?

iii) What will the mean and standard deviation be if every woman has an increase of 10% of previous earnings?

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

◆◆◆◆◆◆◆◆◆◆