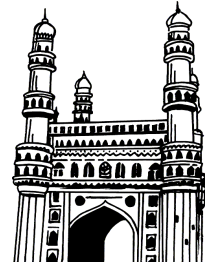


Rahul's ✓
Topper's Voice



B.B.A.

I Year II Sem

Latest 2021 Edition

BUSINESS STATISTICS

- ☞ Study Manual
- ☞ FAQ's and Important Questions
- ☞ Exercise Problems
- ☞ Short Question and Answers
- ☞ Choose the Correct Answer
- ☞ Fill in the blanks
- ☞ Solved Previous Question Papers
- ☞ Solved Model Papers

- by -

WELL EXPERIENCED LECTURER



Rahul Publications™

Hyderabad. Ph : 66550071, 9391018098

All disputes are subjects to Hyderabad Jurisdiction only

B.B.A.

I Year II Sem

BUSINESS STATISTICS

Inspite of many efforts taken to present this book without errors, some errors might have crept in. Therefore we do not take any legal responsibility for such errors and omissions. However, if they are brought to our notice, they will be corrected in the next edition.

© No part of this publications should be reporduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording and/or otherwise without the prior written permission of the publisher

Price ` 199-00

Sole Distributors :

☎ : 66550071, Cell : 9391018098

VASU BOOK CENTRE

Shop No. 3, Beside Gokul Chat, Koti, Hyderabad.

Maternity Hospital Opp. Lane, Narayan Naik Complex, Koti, Hyderabad.

Near Andhra Bank, Subway, Sultan Bazar, Koti, Hyderabad -195.

BUSINESS STATISTICS

C O N T E N T S

STUDY MANUAL

FAQ's and Important Questions	VI - IX
Unit - I	1 - 61
Unit - II	62 - 155
Unit - III	156 - 195
Unit - IV	196 - 233
Unit - V	234 - 300

SOLVED PREVIOUS QUESTION PAPERS

January - 2021	301 - 307
November - 2020	308 - 314
May - 2019	315 - 320
May - 2018	321 - 327
May - 2017	328 - 333

SOLVED MODEL PAPERS

Model Paper - I	334 - 335
Model Paper - II	336 - 337
Model Paper - III	338 - 339

SYLLABUS

UNIT - I

Statistics : Definitions – Statistical methods – Importance and Scope – Limitations – Need for Data – Principles of Measurement. Tabulation and Presentation:

Classification of Data – Data Array – Frequency Distribution – Methods of data Classification – Types of Frequency Distributions / tabulation of Data – Objectives of Tabulation – Parts and Types of Tables – Graphical Presentation – Functions of Graphs – Advantages and limitations of Graphs.

UNIT - II

Measures of Central Tendency : Introduction to Averages – Requisites for a Measure of Central Tendency, Mean - Combined mean – Weighted mean, Median – Partition values – Quartiles, Deciles and Percentiles, Relationship between Partition values–Mode– Relationship between Mean, Median and Mode.

Measures of Dispersion: Introduction – Significance and Requisites of a Measure of dispersion, Range, QD, MD and SD For Grouped and Ungrouped – Advantages and Disadvantages. Concept of Variation – Coefficient of Variation.

Skewness and Kurtosis (SK): Introduction, Measures of SK, Relative measures of SK – Advantages and Disadvantages. Moments – concepts – Calculation – Kurtosis.

UNIT - III

Index Numbers : Index Numbers - Introduction – Types – Characteristics – Construction weighted and unweighted index numbers – Price and Quantity/Volume index numbers – Tests – time reversal – Factor Reversal and Circular tests – Chain and Fixed base – Changing of base – Combining of two of more overlapping indices consumer price Index – Problems in Construction.

UNIT - IV

Probability : Concepts – Random Experiment, Sample space – Definitions of probability, Simple Problems on Probability, Addition and Multiplication theorems, conditional, Joint and Marginal Probability.

Sampling: Sampling – Reasons of Sample survey – bias in Survey, Definitions of Population, Sample, Parameter, Statistic – Principles of Sampling, Statistical Regularity, Inertia of Large Numbers, Optimization, Persistence of small numbers – Validity. Probability and non probability sampling methods – choice of sampling method, sampling distribution and Standard Error (SE).

UNIT - V

Correlation Analysis : Scatter diagram, Positive and negative correlation, limits for coefficient of correlation, Karl Pearson's coefficient of correlation, Spearman's Rank correlation.

Regression Analysis: Concept, least square fit of a linear regression, two lines of regression, properties of regression coefficients (Simple problems only)

Time Series Analysis: Components, Models of Time Series – Additive, Multiplicative and Mixed models; Trend analysis- Free hand curve, Semi averages, moving averages, Least Square methods (Simple problems only).

Contents

UNIT - I

Topic	Page No.
1.1 Statistics	1
1.1.1 Definitions	1
1.1.2 Statistical Methods	5
1.1.3 Importance of Statistics	6
1.1.4 Scope of Statistics	7
1.1.5 Limitations of Statistics	8
1.2 Need for Data	9
1.3 Principals of Measurement	10
1.4 Data Collection	11
1.5 Classification of Data	23
1.5.1 Methods of data Classification	26
1.6 Data Array	29
1.7 Frequency Distribution	31
1.7.1 Types of Frequency Distributions	33
1.8 Tabulation of Data	38
1.8.1 Objectives of Tabulation	38
1.8.2 Parts of Table	39
1.8.3 Types of Tables	41
1.9 Graphical Presentation	43
1.9.1 Functions of Graphs	45
1.9.2 Advantages and limitations of Graphs	45
1.9.3 Types of Graphs	46
1.9.4 Graphs of Frequency Distribution	47
1.9.4.1 Histogram	47
1.9.4.2 Frequency Polygon	49
1.9.4.3 Frequency Curve	51
1.9.4.4 Cumulative Frequency	52
➤ Exercise Problems	54 - 54
➤ Short Question and Answers	55 - 59
➤ Choose the Correct Answers	60 - 60
➤ Fill in the blanks	61 - 61

Topic	Page No.
UNIT - II	
2.1 Introduction to Averages	62
2.1.1 Requisites of a Measure of Central Tendency	65
2.2 Mean (or) Arithmetic Mean	65
2.2.1 Weighted Mean	77
2.2.2 Combined Mean	79
2.2.3 Geometric Mean	81
2.2.4 Harmonic Mean	86
2.3 Median	89
2.3.1 Partition Values	97
2.3.1.1 Quartiles	97
2.3.1.2 Deciles	97
2.3.1.3 Percentiles	98
2.3.1.4 Relationship between Partial Values	100
2.4 Mode	101
2.5 Relationship between Mean, Median and Mode	105
2.6 Measures of Dispersion	107
2.6.1 Introduction	107
2.6.2 Significance of Measures of Dis- persion	107
2.7 Requisites of a Measure of dispersion	109
2.7.1 Range	109
2.7.1.1 Advantages and Disadvantages of Range	110
2.7.2 Quartile Deviation	112
2.7.2.1 Advantages and Disadvantages of Quartile Deviations	112
2.7.3 Mean Deviation	118
2.7.3.1 Advantages and Disadvantages of Mean Deviation	120
2.7.4 Standard Deviation	123
2.7.4.1 Advantages and Disadvantages of Standard Deviation	125
2.7.4.2 Concept of Variation	126
2.7.5 Coefficient of Variation	126
2.8 Skewness	130
2.8.1 Introduction	130
2.8.2 Measures of Skewness	131
2.8.2.1 Relative measures of Skewness	132

Topic	Page No.
2.9 Moments	140
2.9.1 Moments About Mean	140
2.9.2 Moments about Arbitrary Point	141
2.9.3 Moments about Zero or Origin	141
2.9.4 Relationship between Central Moments and Moments about any Arbitrary point	142
2.9.5 Moments in Standard Units	142
2.9.6 Sheppard's Corrections for Moments	143
2.10 Kurtosis	144
➤ Exercise Problems	147 - 148
➤ Short Question and Answers	149 - 153
➤ Choose the Correct Answers	154 - 154
➤ Fill in the blanks	155 - 155
UNIT - III	
3.1 Index Numbers	156
3.1.1 Introduction	156
3.1.2 Characteristics of index numbers	157
3.1.3 Price and Quantity / Value index numbers	159
3.2 Types of Index Numbers	160
3.2.1 Unweighted Price Indexes Weighted Price Indexes	160
3.3 Tests of Index Numbers	169
3.3.1 Time Reversal – Factor Reversal and Circular Tests	169
3.4 Chain Index Numbers	174
3.4.1 Conversion of Chain Index of Fixed base Index	176
3.5 Changing of Base / Base Shifting of Index Numbers	177
3.6 Combining of Two of More Overlapping Indices	179
3.7 Consumer Price Index	183
3.8 Problems in Construction of Index Numbers	186
➤ Exercise Problems	188 - 189
➤ Short Question and Answers	190 - 193
➤ Choose the Correct Answers	194 - 194
➤ Fill in the blanks	195 - 195

Topic	Page No.
UNIT - IV	
4.1 Probability	196
4.1.1 Definitions of Probability	196
4.1.2 Concepts of Probability	197
4.1.2.1 Random Experiment, Sample space	197
4.2 Theorems of Probability	203
4.2.1 Addition Theorem	203
4.2.2 Multiplication Theorem	207
4.3 Statistical Independence	210
4.3.1 Conditional, Joint and Marginal Probability	210
4.4 Sampling	214
4.4.1 Reasons of Sample Survey	214
4.4.2 Bias in Survey	215
4.4.3 Definitions of Population, Sample, Parameter of Statistics	215
4.4.4 Principles of Sampling	216
4.4.4.1 Principle of Statistical Regularity	216
4.4.4.2 Principle of Inertia of Large Numbers	217
4.4.4.3 Principle of Optimization	217
4.4.4.4 Principle of Persistence of Small Numbers	217
4.4.4.5 Principles of Validity	217
4.5 Methods of Sampling	218
4.5.1 Probability and Non probability Sampling Methods	218
4.6 Choice of Sampling Method	222
4.7 Sampling Distribution	223
4.8 Standard Error (SE)	224
➤ Exercise Problems	227 - 227
➤ Short Question and Answers	228 - 230
➤ Choose the Correct Answers	231 - 232
➤ Fill in the blanks	233 - 233

Topic	Page No.
UNIT - V	
5.1 Correlation Analysis	234
5.2 Types of Correlation	236
5.2.1 Positive and Negative Correlation	236
5.2.2 Limits for Coefficient of Correlation	239
5.3 Scatter Diagram	240
5.4 Karl Pearson's Coefficient of Correlation	241
5.5 Spearman's Rank Correlation	248
5.6 Regression Analysis	259
5.6.1 Concept of Regression	259
5.6.2 Least Square Fit of a Linear Regression	261
5.6.3 Two Lines of Regression	262
5.6.4 Properties of Regression Coefficients	263
5.7 Time Series Analysis	266
5.7.1 Components of Time Series	267
5.8 Models of Time Series	269
5.8.1 Additive, Multiplicative and Mixed Models	269
5.9 Trend Analysis	270
5.9.1 Free Hand Curve	271
5.9.2 Semi Averages Method	273
5.9.3 Moving Averages Method	274
5.9.4 Least Square methods	280
➤ Exercise Problems	290 - 291
➤ Short Question and Answers	292 - 297
➤ Choose the Correct Answers	298 - 299
➤ Fill in the blanks	300 - 300

Frequently Asked & Important Questions

UNIT - I

1. State the functions of statistics ?

Ans : (May-19)

Refer Unit-I, Q.No. 2.

2. Explain the importance of statistics in economic analysis and planning.

Ans : (May-18, Imp.)

Refer Unit-I, Q.No. 5.

3. Explain the scope of statistics.

Ans : (May-18, Imp.)

Refer Unit-I, Q.No. 6.

4. What are the sources and methods of collecting data ?

Ans : (Jan.-21, Nov.-20)

Refer Unit-I, Q.No. 11.

5. Explain the various methods of collecting secondary data.

Ans : (May-19, May-18)

Refer Unit-I, Q.No. 15.

6. What are the difference between Primary and Secondary Data ?

Ans : (May-17)

Refer Unit-I, Q.No. 18.

7. Define classification. What are the objectives of classification?

Ans : (Nov.-20)

Refer Unit-I, Q.No. 19.

8. What do you mean by Graphical Presentation ? Explain the techniques of construction of graphs.

Ans : (Jan.-21)

Refer Unit-I, Q.No. 32.

9. Explain the Advantages and limitations of Graphs.

Ans : (Jan.-21)

Refer Unit-I, Q.No. 35.

UNIT - II

1. Define Arithmetic Mean. Explain the merits and demerits of Arithmetic Mean.

Ans : (Nov.-20, May-18)

Refer Unit-II, Q.No. 6.

2. Define Dispersion. Explain the objectives of Measures of Dispersion.

Ans : (Imp.)

Refer Unit-II, Q.No. 34.

3. Explain the Importance of Measures of Dispersion.

Ans : (May-18, May-17)

Refer Unit-II, Q.No. 35.

4. Define Mean Deviation. Explain how Mean Deviation is calculated for individual series.

Ans : (Imp.)

Refer Unit-II, Q.No. 45.

5. Define Standard Deviation. Explain how Standard Deviation is calculated for individual series.

Ans : (Nov.-20)

Refer Unit-II, Q.No. 48.

6. Define the concept of variation.

Ans : (May-19)

Refer Unit-II, Q.No. 52.

7. Explain briefly about Skewness.

Ans : (May-19)

Refer Unit-II, Q.No. 54.

UNIT - III

1. What do you understand by index numbers ?

Ans : (May-18)

Refer Unit-III, Q.No. 1.

2. Explain the Characteristics of Index Numbers.

Ans : (May-17)

Refer Unit-III, Q.No. 2.

3. Explain the uses of index numbers.

Ans : (Nov.-20, May-18)

Refer Unit-III, Q.No. 3.

4. Explain the limitations of index numbers

Ans : (Nov.-20, May-18)

Refer Unit-III, Q.No. 4.

5. Explain the various methods of constructing index numbers.

Ans : (May-18, May-17)

Refer Unit-III, Q.No. 6.

6. Define splicing ? What are its conditions and types ?

Ans : (May-19)

Refer Unit-III, Q.No. 17.

7. What are the various Problems in the Construction of Index Numbers?

Ans : (Imp.)

Refer Unit-III, Q.No. 21.

UNIT - IV

1. What are the various approaches of probability?

Ans : (May-17)

Refer Unit-IV, Q.No. 5.

2. Explain the principle of statistical regularity.

Ans : (May-18)

Refer Unit-IV, Q.No. 17.

3. Explain the Principle of Inertia of Large Numbers

Ans : (May-18, May-17)

Refer Unit-IV, Q.No. 18.

4. Explain the various Methods of Sampling.

Ans : (Jan.-21, May-19)

Refer Unit-IV, Q.No. 22.

5. Define Standard Error (SE). What are the uses of Standard Error?

Ans : (Jan.-21)

Refer Unit-IV, Q.No. 28.

UNIT - V

- 1. What is Karl Pearson's Coefficient of Correlation? Explain properties of Coefficient of Correlation.**

Ans : (Imp.)

Refer Unit-V, Q.No. 8.

- 2. Explain about rank correlation and its features.**

Ans : (Imp.)

Refer Unit-V, Q.No. 11

- 3. What are the differences between correlation and regression analysis ?**

Ans : (Imp.)

Refer Unit-V, Q.No. 21.

- 4. Explain the Utility of Time Series Analysis**

Ans : (Imp.)

Refer Unit-V, Q.No. 23.

- 5. What are the Components of Time Series ?**

Ans : (May-18, May-17)

Refer Unit-V, Q.No. 24.

- 6. Discuss the method of moving averages in measuring trend. What are its merits and limitations of moving average method?**

Ans : (Nov.-20)

Refer Unit-V, Q.No. 31.

- 7. Define least square method. Explain merits and demerits of least square method.**

Ans : (Imp.)

Refer Unit-V, Q.No. 32.

UNIT I

Statistics

Definitions – Statistical methods – Importance and Scope – Limitations – Need for Data – Principles of Measurement. Tabulation and Presentation:

Classification of Data – Data Array – Frequency Distribution – Methods of data Classification – Types of Frequency Distributions / tabulation of Data– Objectives of Tabulation – Parts and Types of Tables – Graphical Presentation– Functions of Graphs – Advantages and limitations of Graphs.

1.1 STATISTICS

1.1.1 Definitions

Q1. Define the term Statistics. Explain the characteristics of statistics ?

Ans : (May-18)

The word 'statistic' has been derived from the Latin word 'status' or the Italian word 'statistic' or the German word 'statisti' or the French word 'statistique' all of which mean a political state'. The 'State' collected data pertaining to population and their wealth, in order to plan new taxes and to fund wars. Hence, statistics was also known as 'science of statecraft'.

Statistics is said to be as old as recorded history. One of the earliest known statistics is the census conducted by the Emperors of Egypt in connection with construction of pyramids. Statistics have been found to be use in India even before 300 B.C. Historical evidences about the prevalence of a good system of collection of data are available in Kautilya's Arthshastra.

Definitions

1. **According to A.L. Bowley**, "Statistics are numerical facts in any department of enquiry placed in relation to each other."

The main features of the above definition are:

- (i) Statistics (or Data) implies numerical facts.
- (ii) Numerical facts or figures are related to some enquiry or investigation.

- (iii) Numerical facts should be capable of being arranged in relation to each other.

On the basis of the above features we can say that data are those numerical facts which have been expressed as a set of numerical figures related to each other and to some area of enquiry or research. We may, however, note that all the characteristics of data are not covered by the above definition.

2. **According to Yule & Kendall**, "By statistics we mean quantitative data affected to a marked extent by multiplicity of causes."

This definition covers two aspects, i.e., the data are quantitative and affected by a large number of causes.

3. **According to Webster**, "Statistics are classified facts respecting the conditions of the people in a state - especially those facts which can be stated in numbers or in tables of numbers or in any other tabular or classified arrangement."

4. **According to Achenwall**, "A collection of noteworthy facts concerning state, both historical and descriptive."

Definitions 3 and 4, given above, are not comprehensive because these confine the scope of statistics only to facts and figures related to the conditions of the people in a state. However, as we know that data are now collected on almost all the aspects of human and natural activities, it cannot be regarded as a state-craft only.

5. **According to L.R. Connor**, "Statistics are measurements, enumerations or estimates of natural or social phenomena, systematically arranged, so as to exhibit their inter-relations."

This definition also covers only some but not all characteristics of data.

6. **According to H. Secrist**, "By statistics we mean aggregate of facts affected to a marked extent by a multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."

This definition can be taken as a comprehensive definition of statistics since most of the characteristics of statistics are covered by it.

Characteristics

On the basis of the above definitions we can now state the following characteristics of statistics - as data:

1. Statistics are numerical facts

In order that any set of facts can be called as statistics or data, it must be capable of being represented numerically or quantitatively. Ordinarily, the facts can be classified into two categories:

- (a) Facts that are measurable and can be represented by numerical measurements. Measurement of heights of students in a college, income of persons in a locality, yield of wheat per acre in a certain district, etc., are examples of measurable facts,
- (b) Facts that are not measurable but we can feel the presence or absence of the characteristics. Honesty, colour of hair or eyes, beauty, intelligence, smoking habit etc., are examples of immeasurable facts. Statistics or data can be obtained in such cases also, by counting the number of individuals in different categories. For example, the population

of a country can be divided into three categories on the basis of complexion of the people such as white, wheatish or black.

2. Statistics are aggregate of facts

A single numerical figure cannot be regarded as statistics. Similarly, a set of unconnected numerical figures cannot be termed as statistics. Statistics means an aggregate or a set of numerical figures which are related to one another. The number of cars sold in a particular year cannot be regarded as statistics. On the other hand, the figures of the number of cars sold in various years of the last decade is statistics because it is an aggregate of related figures. These figures can be compared and we can know whether the sale of cars has increased, decreased or remained constant during the last decade.

It should also be noted here that different figures are comparable only if they are expressed in same units and represent the same characteristics under different situations. In the above example, if we have the number of Ambassador cars sold in 1981 and the number of Fiat cars sold in 1982, etc., then it cannot be regarded as statistics. Similarly, the figures of, say, measurement of weight of students should be expressed in the same units in order that these figures are comparable to each other.

3. Statistics are affected to a marked extent by a multiplicity of factors

Statistical data refer to measurement of facts in a complex situation, e.g., business or economic phenomena are very complex in the sense that there are a large number of factors operating simultaneously at a given point of time. Most of these factors are even difficult to identify. We know that the quantity demanded of a commodity, in a given period, depends upon its price, income of the consumer, prices of other commodities, tastes and habits of the consumer. It may be mentioned here that these factors are only the main factors but not the only factors affecting the demand of a commodity.

Similarly, the sale of a firm in a given period is affected by a large number of factors. Data collected under such conditions are called statistics or statistical data.

4. Statistics are either enumerated or estimated with reasonable standard of accuracy

This characteristic is related to the collection of data. Data are collected either by counting or by measurement of units or individuals.

For example, the number of smokers in a village is counted while height of soldiers is measured. We may note here that if the area of investigation is large or the cost of measurement is high, the statistics may also be collected by examining only a fraction of the total area of investigation.

When statistics are being obtained by measurement of units, it is necessary to maintain a reasonable degree or standard of accuracy in measurements.

The degree of accuracy needed in an investigation depends upon its nature and objectivity on the one hand and upon time and resources on the other.

For example, in weighing of gold, even milligrams may be significant whereas, for weighing of a wheat-bag, a few grams may not make much difference. Sometimes, a higher degree of accuracy is needed in order that the problem, to be investigated, gets highlighted by the data. Suppose the diameter of bolts produced by a machine is measured as 1.546 cms, 1.549 cms, 1.548 cms, etc.

If, instead, we obtain measurements only up to two places after decimal, all the measurements would be equal and as such nothing could be inferred about the working of the machine. In addition to this, the degree of accuracy also depends upon the availability of time and resources.

For any investigation, a greater degree of accuracy can be achieved by devoting more time or resources or both. As will be discussed later, in statistics, generalisations

about a large group (known as population) are often made on the basis of small group (known as sample).

It is possible to achieve this by maintaining a reasonable degree of accuracy of measurements.

Therefore, it is not necessary to always have a high degree of accuracy but whatever degree of accuracy is once decided must be uniformly maintained throughout the investigation.

5. Statistics are collected in a systematic manner and for a predetermined purpose

In order that the results obtained from statistics are free from errors, it is necessary that these should be collected in a systematic manner. Haphazardly collected figures are not desirable as they may lead to wrong conclusions. Moreover, statistics should be collected for a well-defined and specific objective, otherwise it might happen that the unnecessary statistics are collected while the necessary statistics are left out. Hence, a given set of numerical figures cannot be termed as statistics if it has been collected in a haphazard manner and without proper specification of the objective.

6. Statistics should be capable of being placed in relation to each other

This characteristic requires that the collected statistics should be comparable with reference to time or place or any other condition. In order that statistics are comparable it is essential that they are homogeneous and pertain to the same investigation. This can be achieved by collecting data in identical manner for different periods or for different places or for different conditions.

**Q2. What are the functions of statistics ?
(OR)**

State the functions of statistics ?

Ans : (May-19)

The following are the main functions of statistics:

1. Presents facts in numerical figures

The first function of statistics is to present a given problem in terms of numerical figures. We know that the numerical presentation helps in having a better understanding of the nature of a problem. Facts expressed in words are not very useful because they are often vague and are likely to be understood differently by different people. For example, the statement that a large proportion of total work force of India is engaged in agriculture, is vague and uncertain. On the other hand, the statement that 70% of the total work force is engaged in agriculture is more specific and easier to grasp. Similarly, the statement that the annual rate of inflation in a country is 10% is more convincing than the statement that prices are rising.

2. Presents complex facts in a simplified form

Generally a problem to be investigated is represented by a large mass of numerical figures which are very difficult to understand and remember. Using various statistical methods, this large mass of data can be presented in a simplified form. This simplification is achieved by the summarisation of data so that broad features of the given problem are brought into focus. Various statistical techniques such as presentation of data in the form of diagrams, graphs, frequency distributions and calculation of average, dispersion, correlation, etc., make the given data intelligible and easily understandable.

3. Studies relationship between two or more phenomena

Statistics can be used to investigate whether two or more phenomena are related. For example, the relationship between income and consumption, demand and supply, etc., can be studied by measuring correlation between relevant variables. Furthermore, a given mathematical relation can also be fitted to the given data by using the technique of regression analysis.

4. Provides techniques for the comparison of phenomena

Many a times, the purpose of undertaking a statistical analysis is to compare various phenomena by computing one or more measures like mean, variance, ratios, percentages and various types of coefficients. For example, when we compute the consumer price index for a particular group of workers, then our aim could be to compare this index with that of previous year or to compare it with the consumer price index of a similar group of workers of some other city, etc. Similarly, the inequalities of income in various countries may be computed for the sake of their comparison.

5. Enlarges individual experiences

An important function of statistics is that it enlarges human experience in the solution of various problems. In the words of A.L. Bowley, "the proper function of statistics, indeed is to enlarge individual experience." Statistics is like a master key that is used to solve problems of mankind in every field. It would not be an exaggeration to say that many fields of knowledge would have remained closed to the mankind forever but for the efficient and useful techniques and methodology of the science of statistics.

6. Helps in the formulation of policies

Statistical analysis of data is the starting point in the formulation of policies in various economic, business and government activities. For example, using statistical techniques a firm can know the tastes and preferences of the consumers and decide to make its product accordingly. Similarly, the Government policies regarding taxation, prices, investments, unemployment, imports and exports, etc., are also guided by statistical studies in the relevant areas.

7. Helps in forecasting

The success of planning by the Government or of a business depends to a large extent upon the accuracy of their forecasts. Statistics provides a scientific basis for making such

forecasts. Various techniques used for forecasting are time series analysis, regression analysis, etc.

8. Provides techniques for testing of a hypothesis

A hypothesis is a statement about some characteristics of a population (or universe). For example, the statement that average height of students of a college is 66 inches, is a hypothesis. Here students of the college constitute the population. It is possible to test the validity of this statement by the use of statistical techniques.

9. Provides techniques for making decisions under uncertainty

Many a times we face an uncertain situation where any one of the many alternatives may be adopted. For example, a person may face a situation of rain or no rain and he wants to decide whether to take his umbrella or not. Similarly, a businessman might face a situation of uncertain investment opportunities in which he can lose or gain. He may be interested in knowing whether to undertake a particular investment or not. The answer to such problems are provided by the statistical techniques of decision-making under uncertainty.

Q3. State whether statistics as a science (or) an art.

Ans :

There has been many arguments regarding the statistics as a science or art. Science is defined as systematised arrangement related to knowledge. It focuses on cause and effect relationship in terms of scientific principles or laws in order to make clear generalisations. In simple words, science provides the knowledge that is helpful in finding a way. But it does not show the direction to be chosen. On contrary to science, art refers to the ability of managing the facts in order to attain the set goals. It provides the ways of handling and presenting the data, making judgement logically and obtaining the effective and relevant results.

Earlier the natural scientists were in dilemma regarding statistics to be considered as a distinct science. In today's world, all sciences are considered as statistical.

When statistics is considered as science, it is not similar to the sciences such as chemistry, physics, zoology and so on. This is due to the reason that statistical phenomenon are influenced by multiplicity of causes that cannot be measured accurately. In short, the statistical science is not much accurate in comparison to natural sciences. Statistics is thus considered as a scientific method and its tools are widely applied in scientific studies.

"Statistics is not a body of substantive knowledge but a body of methods for obtaining knowledge".

- Wallis and Roberts

Science is knowledge whereas art is action. Thus, from this point of view, statistics can also be considered as an "art". The basic function of statistics is to apply the given method in order to extract facts, results so as to arrive at taking effective conclusion.

1.1.2 Statistical Methods

Q4. Explain various methods of statistics ?

(OR)

Elucidate the various methods of statistics ?

Ans :

Statistical methods are considered a common kit of tools for describing and analyzing data for various disciplines. In other words, statistical methods are the procedures used in collection, organization, summary, analysis, interpretation and presentation of data. They provide tools for extracting basic truths, which often lie hidden in a mass of data. The methods of statistical analysis can be divided into two groups:

i) Descriptive Statistics

In descriptive statistics an analysis is made using data gathered on a group to describe or reach conclusions about the same group. The term Descriptive Statistics is confined to the treatment of data for the purpose of

describing their characteristics. It involves techniques for summarizing data and presenting them in a usable form. It includes mean, standard deviation and graphical methods like charts and graphs.

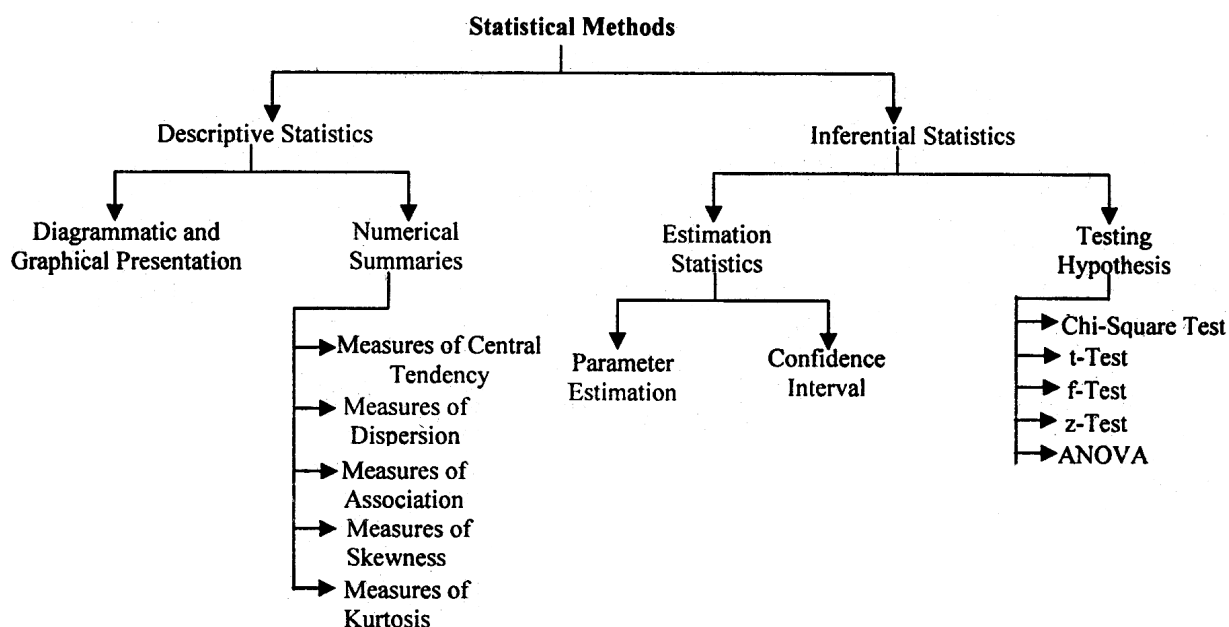


Fig : Statistical Methods

ii) Inferential (or) Inductive Statistics

Inductive statistics involves making forecasts, estimations or judgment about some larger group of data than that actually observed or about some future happening based on a study of historical data. It refers to generalizing from a sample to population, estimating unknown parameters, drawing conclusions and making decisions accordingly. They project from known to the unknown. It is used to models patterns in the data, accounting for randomness and drawing inferences about larger population, it includes techniques like ANOVA, time series analysis and hypothesis testing such as Chi-square test, t-test, z- test etc.

1.1.3 Importance of Statistics

Q5. Explain the importance of statistics in economic analysis and planning.

Ans :

(May-18, Imp.)

1. Importance to the State

We know that the subject of statistics originated for helping the ancient rulers in the assessment of their military and economic strength. Gradually its scope was enlarged to tackle other problems relating to political activities of the State. In modern era, the role of State has increased and various governments of the world also take care of the welfare of its people. Therefore, these governments require much greater information in the form of numerical figures for the fulfillment of welfare objectives in addition to the efficient running of their administration.

In a democratic form of government, various political groups are also guided by the statistical analysis regarding their popularity in the masses. Thus, it can be said that it is impossible to think about the functioning of modern state in the absence of statistics.

2. Importance in Economics

Statistics is an indispensable tool for a proper understanding of various economic problems. It also provides important guidelines for the formulation of various economic policies.

Almost every economic problem is capable of being expressed in the form of numerical figures, e.g., the output of agriculture or of industry, volume of exports and imports, prices of commodities, income of the people, distribution of land holding, etc. In each case, the data are affected by a multiplicity of factors. Further, it can be shown that the other conditions prescribed for statistical data are also satisfied. Thus, we can say that the study of various economic problems is essentially the one of a statistical nature.

3. Importance in national income accounting

The system of keeping the accounts of income and expenditure of a country is known as national income accounting. These accounts contain information on various macro-economic variables like national income, expenditure, production, savings, investments, volume of exports and imports, etc. The national income accounts of a country are very useful in having an idea about the broad features of its economy or of a particular region. The preparation of these accounts require data, regarding various variables, at the macro-level. Since such information is very difficult, if not impossible, to obtain, is often estimated by using techniques and principles of statistics.

4. Importance in planning

Planning is indispensable for achieving faster rate of growth through the best use of a nation's resources. It also requires a good deal of statistical data on various aspects of the economy. One of the aims of planning could be to achieve a specified rate of growth of the economy. Using statistical techniques, it is possible to assess the amounts of various resources available in the economy and accordingly determine whether the specified

rate of growth is sustainable or not. The statistical analysis of data regarding an economy may reveal certain areas which might require special attention, e.g., a situation of growing unemployment or a situation of rising prices during past few years. Statistical techniques and principles can also guide the Government in adopting suitable policy measures to rectify such situations. In addition to this, these techniques can be used to assess various policies of the Government in the past. Thus, it is rather impossible to think of a situation where planning and evaluation of various policies can be done without the use of statistical techniques. In view of this it is sometimes said that, "Planning without statistics is a ship without rudder and compass". Hence statistics is an important tool for the quantification of various planning policies.

5. Importance in business

With the increase in size of business of a firm and with the uncertainties of business because of cut throat competition, the need for statistical information and statistical analysis of various business situations has increased tremendously. Prior to this, when the size of business used to be smaller without much complexities, a single person, usually the owner or manager of a firm, used to take all decisions regarding its business. For example, he used to decide, from where the necessary raw materials and other factors of production were to be obtained, how much of output will be produced, where it will be sold, etc. This type of decision making was usually based on experience and expectations of this single man and as such had no scientific basis.

1.1.4 Scope of Statistics**Q6. Explain the scope of statistics.**

Ans : (May-18, Imp.)

Statistics is viewed not just as a device for collecting numerical data, but as a means of sound techniques for their handling, analysis and for drawing valid inference from them. From this perspective, the scope or subject matter to statistics can be broadly studied under 2 hands namely.

- (A) Statistical Methods and
- (B) Applied Statistics

(A) Statistical Methods

Statistical methods are the tools that are in the hands of the statistician. They include all the general principles and techniques that are commonly used in the collection, analysis and interpretation of data. These methods are applicable to all kinds of data. The stage involved in study of any kind of data are

- (i) Observation and Collection
- (ii) Organization
- (iii) Presentation
- (iv) Analysis and
- (v) Interpretation

(B) Applied Statistics

Applied Statistics deals with application of Statistical methods of specific problems or concrete forms. To illustrate, if a software services firm is experiencing attrition (loss of personnel), it may be worthwhile to investigate the reasons for the same. Special techniques can be employed to understand the underlying trends.

For example, a correlation analysis could probably indicate that increasing opportunities (growth in the industry) and attrition are highly correlated. All such techniques and the results obtained by employing such techniques form part of applied Statistics.

1.1.5 Limitations of Statistics

Q7. "The science of statistics is a most useful servant but only of great value to those who understand its proper use".

Examine the above statement and discuss the limitations of statistics.

Ans :

1. Statistics deals with numerical facts only

Broadly speaking there are two types of facts:
(a) quantitative and (b) qualitative.

Quantitative facts are capable of being represented in the form of numerical figures and therefore, are also known as numerical facts. These facts can be analysed and interpreted with the help of statistical methods. Qualitative facts, on the other hand, represent only the qualitative characteristics like honesty, intelligence, colour of eyes, beauty, etc. and statistical methods cannot be used to study these types of characteristics. Sometimes, however, it is possible to make an indirect study of such characteristics through their conversion into numerical figures. For example, we may assign a number 0 for a male and 1 for a female, etc.

2. Statistics deals only with groups and not with individuals

Statistical studies are undertaken to study the characteristics of a group rather than individuals. These studies are done to compare the general behaviour of the group at different points of time or the behaviour of different groups at a particular point of time.

For example, the economic performance of a country in a year is measured by its national income in that year and by comparing national income of various years, one can know whether performance of the country is improving or not. Further, by comparing national income of different countries, one can know its relative position vis-a-vis other countries.

3. Statistical results are true only on the average

Statistical results give the behaviour of the group on the average and these may not hold for an individual of that very group. Thus, the statement that average wages of workers of a certain factory is ` 1,500 p.m. does not necessarily mean that each worker is getting this wage. In fact, some of the workers may be getting more while others less than or equal to ` 1,500. Further, when value of a variable is estimated by using some explanatory variable, the estimated value represents the value on the average for a particular value of

the explanatory variable. In a similar way, all the laws of statistics are true only on the average.

4. **Statistical results are only approximately true**

Most of the statistical studies are based on a sample taken from the population. Under certain circumstances the estimated data are also used. Therefore, conclusions about a population based on such information are bound to be true only approximately. Further, if more observations are collected with a view to improve the accuracy of the results, these efforts are often offset by the errors of observation. In the words of Bowley, "When observations are extended, many sources of inaccuracy are found to be present, and it is very frequently impossible to remove them completely. Statistical results are, therefore, very general estimates rather than exact statements." Thus, whether statistical results are based on sample or census data, are bound to be true only approximately.

5. **Statistical methods constitute only one set of methods to study a problem**

A given problem can often be studied in many ways. Statistical methods are used to simplify the mass of data and obtain quantitative results by its analysis. However, one should not depend entirely on statistical results. These results must invariably be supplemented by the results of alternative methods of analysing the problem. It should be kept in mind that statistics is only a means and not an end.

According to D. Gregory and H. Ward, "Statistics cannot run a business or a government. Nor can the study of statistics do more than provide a few suggestions or offer a few pointers as to firm's or government's future behaviour."

6. **Statistics are liable to be misused**

Statistical data are likely to be misused to draw any type of conclusion. If the attitude of the investigator is biased towards a particular aspect of the problem, he is likely to collect

only such data which give more importance to that aspect. The conclusions drawn on the basis of such information are bound to be misleading. Suppose.

For example, the attitude of the Government is biased and it wants to compute a price index which should show a smaller rise of prices than the actual one. In such a situation, the Government might use only those price quotations that are obtained from markets having lower prices.

7. **Statistics must be used only by experts**

Statistics, being a technical subject, is very difficult for a common man to understand. Only the experts of statistics can use it correctly and derive right conclusions from the analysis. In the words of Yule and Kendall, "Statistical methods are the most dangerous tools in the hands of inexperts." Hence, this is the most important limitation of statistics.

1.2 NEED FOR DATA

Q8. **Explain the need for collecting data in statistics.**

Ans :

Statistical data are the basic material needed to make an effective decision in a particular situation. The main reasons for collecting data are as listed below:

- (i) To provide necessary inputs to a given phenomenon or situation under study.
- (ii) To measure performance in an ongoing process such as production, service, and so on.
- (iii) To enhance the quality of decision-making by enumerating alternative courses of action in a decision-making process, and selecting an appropriate one.
- (iv) To satisfy the desire to understand an unknown phenomenon.
- (v) To assist in guessing the causes and probable effects of certain characteristics in given situations.

For any statistical analysis to be useful, the collection and use of input data is extremely important. One can collect an enormous amount of data on a subject of interest in a compact and usable form from the internet. However, the reliability of such data is always doubtful.

In order to design an experiment or conduct a survey one must understand the different types of data and their measurement levels.

Types of Data

Statistical data are the outcome of a continuous process of measuring, counting, and/or observing. These may pertain to several aspects of a phenomenon (or a problem) which are measurable, quantifiable, countable, or classifiable. While conducting a survey or making a study, an investigator develops a method to ask several questions to deal with the variety of characteristics of the given population or universe. These characteristics which one intends to investigate and analyses are termed as *variables*. The data, which are the observed outcomes of these variables, may vary from response to response. Consumer behaviour (attitude), profit/loss to a company, job satisfaction, drinking and/or smoking habits, leadership ability, class affiliation or status are examples of a variable.

Table summarizes the types of variables which can be studied to yield the observed outcomes in relation to the nature of data, information, and measurement.

<i>Data Type</i>	<i>Information Type</i>	<i>Measurement Type</i>
Categorical	→ Do you practice Yoga?	Yes ☐ No ☐
Numerical	→ How many books do you have in your library?	Number
	→ What is your height?	Centimetres or Inches

Table : Nature of Data, Information, and Measurement

It may be noted from Table that categorical variables are those which are not expressed in numerical terms. Sex, religion, and language are a few examples of such variables. The numerical variables are classified into two categories:

- (i) **Discrete variables** which can only take certain fixed integer values. The number of cars sold by Maruti Udyog Ltd. in 2001, or the number of employees in an organization are examples of discrete variables.
- (ii) **Continuous variables** which can take any numerical value. Measurement of height, weight, length, in centimeters/inches, grams/kilograms are a few examples of continuous variables.

1.3 PRINCIPALS OF MEASUREMENT

Q9. What is Principals of Measurement. Explain the various levels of measurement.

Ans :

Measurement is the process of systematically assigning numbers to objects and their properties to facilitate the use of mathematics in studying and describing objects and their relationships. Some types of measurement are fairly concrete: for instance, measuring a person's weight in pounds or kilograms or his height in feet and inches or in meters. Note that the particular system of measurement used is not as important as the fact that we apply a consistent set of rules: we can easily convert a weight expressed in kilograms to the equivalent weight in pounds, for instance. Although any system of units may seem arbitrary (try defending feet and inches to someone who grew up with the metric system!), as long as the system has a consistent relationship with the property being measured, we can use the results in calculations.

In general, the principles of measurement (scaling) has three characteristics:

1. Numbers are ordered. One number is less than, equal to or greater than another number.
2. Difference between numbers are ordered. The difference between any pair of numbers is greater than, less than or equal to the difference between any other pair of numbers.
3. The number series has a unique origin indicated by the number zero.

The combinations of these characteristics of order, distance and origin provide the following widely used classification of measurement scales:

Scale of Measurement	Characteristics of Measurement	Basic Empirical Operation
(i) Nominal	No order, distance or unique origin	Determination of categorical information. Numbers only identify groups which cannot be ordered
(ii) Ordinal	Order but no distance or unique origin	Determination of greater or lesser values. Numbers allow ranking but no arithmetic
(iii) Interval	Both order and distance but not unique	Determination of equality of intervals or differences. Intervals between numbers are meaningful
(iv) Ratio	Order, distance and unique origin	Determination of equality of ratios. Intervals between numbers are meaningful and also their ratio as the lowest value is a meaningful zero.

The four scales of measurement are nominal, ordinal, interval, and ratio.

- Nominal:** Categorical data and numbers that are simply used as identifiers or names represent a nominal scale of measurement. Numbers on the back of a baseball jersey and your social security number are examples of nominal data.
- Ordinal:** An ordinal scale of measurement represents an ordered series of relationships or rank order. Individuals competing in a contest may be fortunate to achieve first, second, or third place. first, second, and third place represent ordinal data.
- Interval:** A scale that represents quantity and has equal units but for which zero represents simply an additional point of measurement is an interval scale.
- Ratio:** The ratio scale of measurement is similar to the interval scale in that it also represents quantity and has equality of units. However, this scale also has an absolute zero (no numbers exist below zero).

1.4 DATA COLLECTION

Q10. Explain briefly about collection of data ?

Ans :

Data collection is the process to gather information about the relevant topic of research, which is being done by researcher. Data collection is a term used to describe a process of preparing and collecting data, for example, as part of a process improvement or similar project.

The purpose of data collection is to obtain information to keep on record, to make decisions about important issues, to pass information on to others. Primarily, data are collected to provide information regarding a specific topic.

Data collection usually takes place early in an improvement project, and is often formalised through data collection plan which often contains the following activity :

1. Pre collection activity on goals, target data, definitions and methods.
2. Collection of data
3. Presenting findings involving some form of sorting analysis and/or presentation.

While deciding about the method of data collection to be used for the study the researcher should keep in mind two types of data viz., primary data and secondary data. The primary data are those, which are collected afresh and for first time and thus happen to be original in character.

The secondary data are those which have been collected by someone else and which have already been passed through statistical process. The researcher would have to decide which sort of data he would be using for his study. The methods for collecting primary and secondary data differ since primary data are to be originally collected while in case of secondary data the nature of data collection work is merely that of compilation.

Database

Database systems have become an important tool in many applications over the past thirty years, ranging from traditional ones in business applications to more recent ones in science and technology. A database system is one of the critical components to build business applications. It provides an array of features which can be used to ensure optimal utilization of data for enhancing decision effectiveness in organizations.

The management of data in a database system is done by means of a general-purpose software package called a database management system. The database management system is the

major software component of a database system. The DBMS helps to create an environment in which end user have better access to more and better managed data than they did before the DBMS become the data management standard.

A database is an organized list of facts and information. Database usually contain text and numbers, and frequently they hold still images, sounds and video or film clips. There is a difference between a simple list and a database? A database permits its user to extract a specific group of disparate facts from within a collection of facts. The paper filling system in an office is a kind of database. However, today people are more interested in databases constructed on computers where database management programs help people design and build collections of information.

Many databases offer citations and abstracts of journal articles and books. By searching for keywords that might appear in an article or book, a user can retrieve the citation, and often an abstract, of the journal article or book.

Uses of Database in Data Collection

Database can be used in following manner in data collection :

1. Citations and Abstracts

Many databases offer citations and abstracts of journal articles and books. By searching for keywords that might appear in an article or book, a user can retrieve the citation, and often an abstract, of the journal article or book.

2. Full text searching

A full-text database incorporates large files of text such as all of the paragraphs from a journal article or all of the chapters of a book. By searching for keywords that might appear in a text, a user can call forth the citation and abstract of a journal article or book. Today, many such databases do not offer the full text for free. Those with full text usually offer the search for citation and abstract for free, but charge for delivery of the full text of an article.

3. Updating

A typical database is designed around a central set of facts. All databases permit their operators to add new information or to update old facts whenever needed.

4. Sorting

In a database stored in a computer, the order of information can be rearranged or sorted quickly. To help people retrieve and print facts, databases sort lists into reports.

5. Electronic Storage

Most academic databases can be stored electronically, either on computers serving the Internet or on CD-ROM or DVD discs providing storage of research data.

6. Backup and Recovery

The system having this responsibility ensures that recovery is possible in the case of a system crash during execution of one or more transactions.

7. Redundancy can be Reduced

In a database system, the redundancy (duplicacy) of the data is controlled, and the system is aware of the redundancy and assumes the responsibility for propagating updates (automatic change at other places in case of duplicate data).

8. Inconsistency can be avoided

If the redundancy is controlled, the inconsistency is automatically controlled. Inconsistency means the mismatch in the two entries with same data. It must be avoided to get correct information.

Q11. What are the sources and methods of collecting data ?

Ans :

(Jan.-21, Nov.-20)

To understand the multitude of choices available to a researcher for collecting the project/ study-specific information, one needs to be fully cognizant of the resources available for the study and the level of accuracy required. To appreciate the truth of this statement, one needs to examine the gamut of methods available to the researcher. The data sources could be either contextual and primary or historical and secondary in nature (Figure).

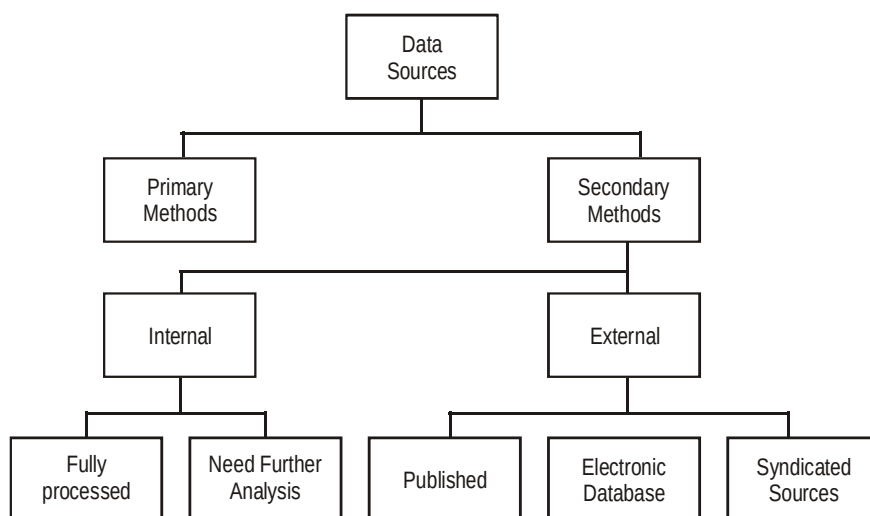


Fig. : Sources of research information

1. **Primary data** as the name suggests is original, problem- or project-specific and collected for the specific objectives and needs spelt out by the researcher. The authenticity and relevance is reasonably high. The monetary and resource implications of this are quite high and sometimes a researcher might not have the resources or the time or both to go ahead with this method. In this case, the researcher can look at alternative sources of data which are economical and authentic enough to take the study forward. These include the second category of data sources—namely the secondary data.

2. **Secondary data** as the name implies is that information which is not topical or research-specific and has been collected and compiled by some other researcher or investigative body. The said information is recorded and published in a structured format, and thus, is quicker to access and manage. Secondly, in most instances, unless it is a data product, it is not too expensive to collect. As suggested in the opening vignette, the data to track consumer preferences is readily available and the information required is readily available as a data product or as the audit information which the researcher or the organization can procure and use it for arriving at quick decisions.

In comparison to the original research-centric data, secondary data can be economically and quickly collected by the decision maker in a short span of time. Also the information collected is contextual; what is primary and original for one researcher would essentially become secondary and historical for someone else.

Q12. Explain the various methods of collecting primary data.

Ans :

Primary data is one, which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions. Primary data collection is necessary when a researcher cannot find the data needed in secondary sources.

E.g. If a researcher is inserted to know the impact of afternoon meal scheme for the school children, he has to undertake a survey and collect data on the opinion of parents and children by asking relevant questions. Such a data collected for the purpose is called primary data.

Conducting primary research is a useful skill to acquire as it can greatly supplement research in secondary sources, such as journals, magazines, or books. Primary research is an excellent skill to learn as it can be useful in a variety of settings including business, personal, and academic.

Methods of Collecting Primary Data

The various methods of collecting primary data are :

- (a) Interviews
 - Personal Interview
 - Telephone Interview
- (b) Surveys
- (c) Observations
- (d) Analysis
- (e) Questionnaires
- (f) Schedules

(a) Interviews

Interviews are non-on-one or small group question and answer sessions. Interviews will provide a lot of information from a small number of people and are useful when you want to get an expert or knowledgeable opinion on a subject. The interviews can be :

- **Personal Interview :** The investigator follows a rigid procedure and seeks answers to a set of pre-conceived questions through personal interviews. This method of collecting data is usually carried out in a structured way where output depends upon the ability of the interviewer to a large extent.
- **Telephone Interviews :** This method of collecting information involves contracting the respondents on telephone itself. This is not a very widely

used method but it plays an important role in industrial surveys in developed regions, particularly, when the survey has to be accomplished in a very limited time.

(b) Surveys

Surveys are a form of questioning that is more rigid than interviews and that involve larger groups of people. Surveys will provide a limited amount of information from a large group of people and are useful when you want to learn what a larger population thinks.

(c) Observations

Observations involve taking organized notes about occurrences in the world. Observations provide you insight about specific people, events, or locales and are useful when you want to learn more about an event without the biased viewpoint of an interview. The information obtained relates to what is currently happening and is not complicated by either the past behaviour or future intentions or attitudes or respondents.

This method is no doubt an expensive method and the information provided by this method is also very limited. As such this method is not suitable in inquiries where large samples are concerned.

(d) Analysis

Analysis involves collecting data and organizing it in some fashion based on criteria you develop. They are useful when you want to find some trend or pattern. A type of analysis would be to record commercials on three major television networks and analyze gender roles.

(e) Questionnaire

The questionnaire is an important tool for gathering primary data. Questionnaire are mailed to the respondents with a required to return after completing the same. It is most extensively used method in various economic and business surveys. Questionnaire to be used must be prepared very carefully so that it may prove to be effective in collecting the relevant information. Poorly constructed

questions can result in large errors and invalidate the research data, so significant effort should be put into the questionnaire.

(f) Schedules

Under this method the enumerators are appointed and given training. They are provided with schedules containing relevant questions. These enumerators go to respondents with these schedules. Data are collected by filling up the schedules by enumerators on the basis of replies given by respondents. Much depends upon the capability of enumerators so far as this method is concerned. Some occasional field checks on the work of the enumerators may ensure sincere work.

Q13. What are the advantages and disadvantages of primary data.

Ans :

Advantages of Primary Data

1. Reliability

The information collected for primary data is more reliable than those collected from the secondary data because this information is collected directly from the respondents.

2. Availability of a Wide Range of Techniques

There are lot of techniques that can be employed, which means that all information necessary can be obtained by using the appropriate techniques, enabling all areas of the research topic to be answered and investigating thoroughly and effectively.

3. Addresses Specific Research Issues

The organisation asking for the research has the complete control on the process and the research is streamlines as far as its objectives and scope is concerned. Researching company can be asked to concentrate their efforts to find data regarding specific market rather than concentration on mass market. Primary research is designed to collect the information the marketer wants to know, and report it in ways that benefit the marketer.

4. Greater Control

Not only does primary research enable the marketer to focus on specific issues, it also enables the marketer to have a higher level of control over how the information is collected. In this way the marketer can decide on such issues as size of project (e.g., how many responses), location of research (e.g., geographic area) and timeframe for completing the project.

5. Efficient Spending for Information

Unlike secondary research, where the marketer may spend for information that is not needed, primary data collection focuses on issues specific to the researcher and improves the chances that research funds will be spent efficiently.

6. Proprietary Information

Information collected by the marketer using primary research is their own and is generally not shared with others. Thus, information can be kept hidden from competitors, and potentially offer an "information advantage" to the company that undertook the primary research.

Disadvantages of Primary Data**1. Cost**

Compared to secondary research, primary data may be very expensive since there is a great deal of marketer involvement and the expense in preparing and carrying-out research can be high. Skilful persons are required, which is also not an economic process, it is costly.

2. Time Consuming

To be done correctly primary data collection requires the development and execution of a research plan. Going from the start-point of deciding to undertake a research project to the end-point of having results is often much longer than the time it takes to acquire secondary data.

3. Not Always Feasible

Some research projects, while potentially offering information that could prove quite valuable, are not within the reach of a marketer. Many are just too large to be carried-out by all but the largest companies, and some are not feasible at all.

4. Large Volume of Data

Since the data collected by primary methods remains in a very large amount so it becomes very complicated to handle and maintain all the data. Large volumes of data also create difficulty in the data processing.

5. Reluctancy of Respondents

In many cases, the respondents remain reluctant to give the answers of the researchers' questions. Sometimes they give such answers which create business in the research.

Q14. Explain the various ways editing of primary data.

Ans :

1. Editing for Completeness

The editor should see that each schedule and questionnaire is complete in all respects, i.e., answer to each and every question has been furnished. If some questions have not been answered and those questions are of vital importance the informants should be contacted again either personally or through correspondence. It may happen that in spite of best efforts a few questions remain unanswered. In such questions, the editor should mark "No answer" in the space provided for answers and if the questions are of vital importance then the schedule or questionnaire should be dropped.

2. Editing for Consistency

While editing the data for consistency, the editor should see that the answers to questions are not contradictory in nature. If

there are mutually contradictory answers, he should try to obtain the correct answers either by referring back the questionnaire or by contacting, wherever possible, the informant in person.

For example, if amongst others, two questions in questionnaire are: (a) Are you married? (b) State the number of children you have, and the reply to the former question is 'no' and to the latter 'three', then there is contradiction and it should be clarified.

3. Editing for Accuracy

The reliability of conclusions depends basically on the correctness of information. If the information supplied is wrong, conclusions can never be valid. It is, therefore, necessary for the editor to see that the information is accurate in all respects.

However, this is one of the most difficult tasks of the editor. If the inaccuracy is due to arithmetical errors, it can be easily detected and corrected. But if the cause of inaccuracy is faulty information supplied, it may be difficult to verify it, e.g., information relating to income, age, etc.

4. Editing for homogeneity

By homogeneity we mean the condition in which all the questions have been understood in the same sense. The editor must check all the questions for uniform interpretation. For example, as to the question of income, if some informants have given monthly income, others annual income and still others weekly income or even daily income, no comparison can be made.

Similarly, if some persons have given the basic income whereas others the total income, no comparison is possible. The editor should check up that the information supplied by the various people is homogeneous and uniform.

Q15. Explain the various methods of collecting secondary data.

Ans : (May-19, May-18)

Secondary data are those data which have been already collected and analysed by some earlier agency for its own use; and later the same data are used by a different agency.

Secondary data is data that is neither collected directly by the user nor specifically for the user, often under conditions not known to the user like Government reports. Secondary information has already been collected for some other purposes. It may be available from internal sources, or may have been collected and published by another organisation. Secondary data is cheaper and more quickly available than primary data, but needs processing before it is useful. Since the data has been collected for another purpose by somebody else, it may not be fully useful, the context could have changed or data could have been doctored.

According to W.A. Neiswanger, "A secondary source is a publication, reporting the data which have been gathered by other authorities and for which others are responsible".

The purpose of collecting secondary data is to make some changes, or to review the needs, to make solutions based on normal database management. To get some new ideas or it was also for the purpose of time saving. If researcher has short time and researcher has to complete an objective, secondary data in this regard is the best way to save time and complete task.

Tools and Techniques of Collecting Secondary Data

In most of the studies the investigator finds it impracticable to collect first-hand information on all related issues and as such he makes use of the data collected by others. There is a vast amount of published information from which statistical studies may be made and fresh statistics are constantly in a state of production. The sources of secondary data can broadly be classified under two heads :

1. Internal Secondary Data

Data that originate within the firm for which the research is being conducted are internal data. They may be adapted for the marketing research purposes. They may be formal data and informal data. Formal data are available on a regularly scheduled basis, such as monthly, quarterly or annually in a form that allows comparisons through time. Informal data reports basic marketing knowledge and are available on a non-recurring basis.

Sales analysis and invoicing are considered important sources of internal secondary data.

i) Sales Analysis

Sales analysis is an important tool of marketing research. It is the first step in the marketing research program and acts as a basis for the development of further marketing research. It reveals the current operating problems in the marketing area where the scope for marketing research can be adequately explored in smaller organisations, sales analysis is an important source of marketing information. It provides a major share of the factors for marketing research.

ii) Invoice Analysis

Company invoices have been a very useful source of information. A copy of an invoice is preserved and information from it may be punched, tabulated, processed and summarised to provide suitable information to the researcher. The invoice data may be classified according to customer, nature of product, region and area. The invoice record may be of immense use provided it has been used with precaution and scientifically.

iii) Accounting Records

The basis for accounting records concerned with sales is the sales invoice.

The usual sales invoice has a sizable amount of information on it, which generally includes name of customer, location of customer, items ordered, quantities ordered, quantities shipped, dollar extensions, back orders, discounts allowed and date. In addition, the invoice often contains information on sales territory, sales representative and warehouse of shipment. This information, when supplemented by data on costs and industry and product classification, as well as from sales calls, provides the basis for a comprehensive analysis of sales by product, customer, industry, geographic area, sales territory, and sales representative, as well as the profitability of each sales category. Unfortunately, most firms' accounting systems are designed primarily for tax reasons rather than for decision support.

2. External Secondary Data

The second form of secondary data is external sources which are generally published and are available in different forms and from different sources. Although external secondary data may be obtained from different sources, some of the sources are given here.

i) Libraries

Researchers first attend libraries to find out relevant data pertaining to research. They provide many sources where suitable data may be obtained. Public libraries, colleges and University libraries contain a large amount of business information, which provides sources of other data. Management books, theses, management journals and other publications can be consulted in these libraries. Management institutes, research institutes, banks, insurance companies, public utility companies and manufacturing units have maintained adequate libraries.

ii) Literature

A great amount of secondary data is available from literature, particularly on marketing subjects. With the development of marketing researches in different countries, new and interesting facts are coming into the picture, which are available in various publications. Consultations of this literature may provide proper guidance pertaining to publication, which can be used from time to time.

iii) Periodicals

Business periodicals published fortnightly, monthly, quarterly, semi annually and annually are often consulted by the marketing executives and researchers to plan and design their marketing research. Periodicals on economics, finance, trade, transport, industry, labour and management are being prepared by the Government as well as by the non-government agencies. Journals of the marketing association, management association, research agencies, advertising agencies and other related periodicals are becoming very common in India and abroad.

iv) Census and Registration Data

Census and registration data have become very comprehensive sources of marketing research. Previously, these were concentrated only to one population census, but it now extends too many areas. Census and registration data includes Census of Population, Census of Agriculture, Census of Cattle, Census of Trade, Census of Transport, Census of Industry, Census of Banking and Finance etc.

v) Trade Associations

Trade associations may be an excellent source of data pertaining to an industry. The trade association of one industry

may exchange data with the trade association of another industry, and within one industry a firm may exchange data with another firm with the help of trade association of the industry.

vi) Government Departments

Different government departments have different data, which are not available in libraries. But these are very useful for understanding various aspects of the economy. The researchers can utilise them for the purposes of their research. Information and data pertaining to agriculture, industry, trade, transport, banking and finance can be obtained from the respective ministries of the Government of India.

vii) Private Sources

Private sources include varied sources available in the form of books, monographs, bulletins, journals, commercial reports and so on. They are priced and publicly circulated. Some of the sources include extensive original research and some summarise the research findings of other person. Many of them are statements of facts and opinions. The All India Management Association, the Indian Marketing Association, Commerce Pvt. Ltd., Capital, the Economic Times and Financial Express etc. are important private institutions which supply suitable information and data to the public in the form of journals, books and newspapers.

viii) Commercial Data

There are several institutions and companies, which purchase and sell marketing information and data. Some of these companies are solely engaged in marketing research. They collect information and data directly from field surveys. Some such companies collect and process the secondary data and supply them to their subscribers.

ix) Financial Data

The financial data of reputed concerns are available in several magazines, newspapers, journals and in summary of statistics. The Directorate of Income Tax publishes information pertaining to taxes and income ranges. Such information and data are useful to forecast the market potential of a particular product. Private institutions such as the Economic Times of India, Commerce Private Ltd. etc. are publishing assets and investment-wise data of several large companies. The market researchers are indirectly benefited by such data and information.

x) International Organisations

International organisations such as the International Monetary Fund, the World Bank, the United Nations Organisation, the Asian Bank, the African Bank, Foreign embassies etc. publish several useful statistics, which can be used by researchers. The statistics may relate to the population problem, trade, institutions, culture, agriculture, regional festivals, superstitions, education, consumption, transportation, forestry, manufacturing and so on. There are several other publications such as the World Almanac, Thomas Register etc., which publish much useful information for marketing researchers.

xi) References and Bibliography

In every publication, the researcher can find references and a bibliography which can be very good sources of information of marketing research. The researcher can consult them for further information and data.

xii) Volumes of Statistics

There are several private and public organisations, which prepare a summary of statistics. In India, the Indian Statistical

Institute publishes the Statistical Abstract. Commerce Pvt. Ltd., the Times of India Ltd. and the Financial Express compile directories of different subjects. The Government of India publishes the economic survey of India wherein statistics relating to every field of economic activities are compiled in a suitable form.

xiii) Advertising Agencies

Advertising agencies have proved to be very useful sources of marketing research. Recently, a large number of agencies have come into the findings of the advertising researches for their clients. Advertising agencies sometimes, publish reviews, resumes and tests of marketing researches. The consumers' behaviour, consumption pattern and demographic features are generally revealed by these agencies.

xiv) Other Sources

There are several other sources of marketing researches. Individuals conduct their own researches, which may be purchased by other institutions. Marketing associations, management associations and individual business houses have been conducting marketing researches for other researches. There are a large number of researches organisations in foreign countries, which are selling their research findings to organisations requiring knowing the outcome of their researches.

Q16. What are the advantages and disadvantages of Secondary Data?

Ans :

Advantages of Secondary Data**1. Economy**

Such data is cheaper. The amount of money spent in acquiring secondary data is generally

less than that needed to obtain primary data. The various secondary data from libraries can be obtained at no cost.

2. Quickness

Most of these data are 'instant' since they already exist and merely need to be discovered. Thus, the time in collecting secondary data is largely search time and usually requires few hours or days.

3. Quality

An individual investigator cannot match the quality or size of the firms that obtain much of the existing secondary data. The information is gathered by trained personnel specialised in data collection. Also many organisations may not release their data to individual researcher but may give it to firms.

4. No Need of Measuring Instruments

When information is gathered from secondary sources, there is a problem in designing information gathering instruments as information is already collected by someone else. The only problem is to locate the appropriate source and method of recording desired information.

5. Availability

Secondary data is sometimes available even in those cases where primary investigations are not possible.

6. Bases for Comparison

Secondary data is useful in the case of exploratory researches as they provide increased understanding of the problem.

7. Useful in Exploratory Research

Secondary data act as a basis for comparison after primary data is collected.

8. Generates Feasible Alternatives

Secondary data are very useful in generating viable alternatives to solve problems. The multiplicity of data sources, research

approaches and managerial styles usually lead to a number of possibilities which should be examined by the researcher.

Disadvantages of Secondary Data

It is however difficult to find secondary data that exactly fits into the needs of some specific research investigation. The problems experienced in respect of secondary sources can be in terms of :

1. Relevance

The data may not fit into the needs of investigation. There may be difference in the units of measurement; there may be surrogated data; discrepancy of classes and data may pertain to some other period of time.

2. Accuracy

It is observed that it is rather difficult to measure the degree of approximations used in the collection of information as well as the competence of the investigator in motivating the persons to supply the desired information.

3. Existence of Obsolete Information

Information may be outdated or obsolete.

4. Non-Disclosure of Research Findings

All the findings of a research study may not be made public.

5. Seldom Catering to the Need

The available data may not suit the current purpose of research, due to incompleteness, generalities and so on.

6. Other Limitations

- i) There may be difficulties in the identification of the source.
- ii) Errors may be there in recording or transferring information from secondary sources.
- iii) The facilities or capabilities of the agency that originally collect the data might be questionable.

Q17. Explain the various ways editing of secondary data.

Ans :

Since secondary data have already been obtained it is highly desirable that a proper scrutiny of such data is made before they are used by the investigator. In fact, the user has to be extra-cautious while using secondary data. In this context, Prof. Bowley rightly points out that secondary data should not be accepted at their face value." The reason is that such data may be erroneous in many respects due to bias, inadequate size of the sample, substitution, errors of definition, arithmetical errors, etc. Even if there is no error such data may not be suitable and adequate for the purpose of the enquiry. Hence Before using such data, the investigator should consider the following aspects:

Whether the data are suitable for the Purpose of Investigation in View

Before using secondary data the investigator must ensure that the data are suitable for the purpose of the enquiry. The suitability of data can be judged in the light of the nature and scope of investigation. For example, if the object of enquiry is to study the wage levels including allowances of workers and the data relate to basic wages alone, such data would not be suitable for the immediate purpose. It may be difficult to find data which exactly fit the needs of the present project.

Quite often secondary data do not satisfy immediate needs because they have been compiled for other purpose. Even when directly pertinent to the subject under study, secondary data may be just enough off the point to make them of little or no use. The value of secondary data is frequently impaired by:

- (i) Variation in the units of measurement: consumer income, for example, may be measured by individual, family household, spending units or tax return.
- (ii) Definition of classes may be different: for example, definition of literate, educated, poor may vary from researcher to researcher.

- (iii) Variation in the date/period to which the data related: the data available may relate to a different time period and not serve the purpose of researcher. Data published to promote the interests of a particular group whether it is political, social or commercial are suspect.

Whether the Data are Adequate for the Investigation If it is found that the data are suitable for the purpose of investigation, they should be tested for adequacy. Adequacy of the data is to be judged in the light of the requirements of the survey and the geographical area covered by the available data. For example, in the illustration given above, if our object is to study the wage rates of the workers in sugar industry in India and if the available data cover only the State of U.P., it would not serve the purpose. The question of adequacy may also be considered in the light of the time period for which the data are available. For example, for studying trend of prices we may use data for the last 8-10 years but from the source known to us data may be available for the last 2-3 years only which would not serve the purpose.

Whether the Data are Reliable It is very difficult to find out whether the secondary data are reliable or not. The following tests, if applied, may be helpful to determine how far the given data are reliable:

- (i) Which specific method of data collection was used? If a source fails to give a detailed description of its method of data collection, researchers should be hesitant about using the information provided. When the methodology is described, researchers should subject it to a painstaking examination.

Data published to promote the interests of a particular group whether it is political, commercial or social are suspect.
- (ii) Was the collecting agency unbiased or did it "have an axe to grind"?
- (iii) If the enumeration was based on a sample, was the sample representative?

- (vi) Were the enumerators capable and properly trained? Incompetent or poorly trained enumerators cannot be depended upon to produce useful result.
- (v) Was there a proper check on the accuracy of field work?
- (vi) Was the editing, tabulating and analysis carefully and conscientiously done? Carelessness in either one or more of these functions can render of little value the findings of an otherwise valuable study.
- (vii) What degrees of accuracy was desired by the compiler? How far was it achieved?

Q18. What are the difference between Primary and Secondary Data ?

Ans :

(May-17)

Basis of Difference	Primary Data	Secondary Data
1) Meaning	Primary data is the one, which is collected by the investigator himself for the purpose of a specific inquiry or study.	Secondary data are those data which have been already collected and analysed by some earlier agency for its own use; and later the same data are used by a different agency.
2) Cost	The cost of obtaining primary data is typically more.	Secondary data is typically available for free or for the subscription fee to the database magazine or journal.
3) Sources	A primary data source is a publication in which the data are published by the same authority which gathered and analysed them.	A secondary data source is a publication, reporting the data which have been gathered by other authorities and for which others are responsible.
4) Methods	It is normally collected through experiments, surveys, questionnaire focus groups, interviews.	The method or way of collecting secondary data includes books, journals, census data biographies, articles and databases.
5) Reliability	The information collected for primary data is more reliable than those collected from the secondary data.	Whereas secondary data are less reliable because these information are not collected for that particular purpose.
6) Scientific Method	Primary data follows the scientific method. A hypothesis is formed, data is collected from an experiment based on the hypothesis and the hypothesis is proven correct or not.	Secondary data does not start with a hypothesis as the data is already collected. Patterns and insights are found within the secondary data and then the observation on that data is made.
7) Precaution	No extra precautions are required in primary data.	Secondary data need more care and attention.
8) Form of Data	Primary data are in the shape of raw material.	Secondary data are usually in the shape of readymade products.
9) Accuracy	Primary research is tailored specifically for the project and tends to be more accurate. Primary data is customised.	While secondary data can provide plenty of information, it is less accurate because the data collected was not collected specifically for the questions. Secondary data is not customised.
10) Example	One's own questionnaire.	Data from a magazine, Journal, etc.

1.5 CLASSIFICATION OF DATA

Q19. Define classification. What are the objectives of classification?

Ans :

(Nov-20)

Classification is the grouping of related facts into classes. Facts in one class differ from those of classification. Sorting facts on one basis of classification and then on another basis is called cross-classification. This process can be repeated as many times as there are possible bases of classification. Classification of data is a function very similar to that of sorting letters in a post-office. It is well known that letters collected in a post-office are sorted into different lots on a geographical basis, i.e., in accordance with their destinations such as Mumbai, Calcutta, Kanpur, Jaipur, etc. They are then put into separate bags, each containing letters with a common characteristic, viz., having the same destination. Classification of statistical data is comparable to the sorting operation.

The principal objectives of classifying the data are :

- To condense the mass of data in such a manner that similarities and dissimilarities can be readily apprehended. Millions of figures can thus be arranged in a few classes having common features.
- To facilitate comparison.
- To pinpoint the most significant features of the data at a glance.
- To give prominence to the important information gathered while dropping out the unnecessary elements.
- To enable a statistical treatment of the material collected.

Q20. Explain different types of classification?**Ans :****(Nov.-20)**

Broadly , the data can be classified on the base of following four criteria:

- (i) Geographical, i.e., area-wise, e.g., cities, districts, etc.
- (ii) Chronological, i.e., on the basis of time.
- (iii) Qualitative, i.e., according to some attributes.
- (iv) Quantitative, i.e., in terms of magnitudes.

(i) Geographical Classification

In this type of classification data are classified on the basis of geographical or vocational differences between the various items, like States, cities, regions, zones, areas, etc. For instance, the production of foodgrains in India may be presented State-wise in the following manner:

State-wise Estimates of Production of Foodgrains

Name of State	Total Foodgrains (Thousand tonnes)
Andhra Pradesh	1093.7
Bihar	12899.0
Haryana	11334.7
Punjab	21148.9
Uttar Pradesh	41828.6
All India	1,92,433.6

Geographical classifications are usually listed in alphabetical order for easy reference. Items may also be listed by size to emphasize the important areas as in ranking the States by population. Normally, in reference table the first approach is followed and in summary tables the second approach is followed.

(ii) Chronological Classification

When data are observed over a period of time the type of classification is known as chronological classification. For example, we may present the figures of population (or production, sales, etc.) as follows:

Population of India From 1941 To 1991

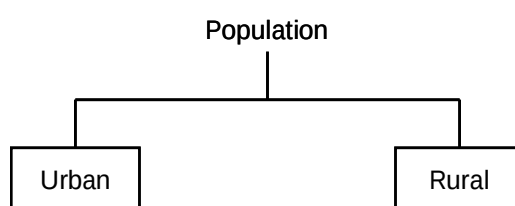
Year	Population (in crores)
1941	31.87
1971	54.82
1951	36.11
1981	68.33
1961	43.92
1991	84.63

Time series are usually listed in chronological order, normally starting with the earliest period. When the major emphasis falls on the most recent events, a reverse time order may be used.

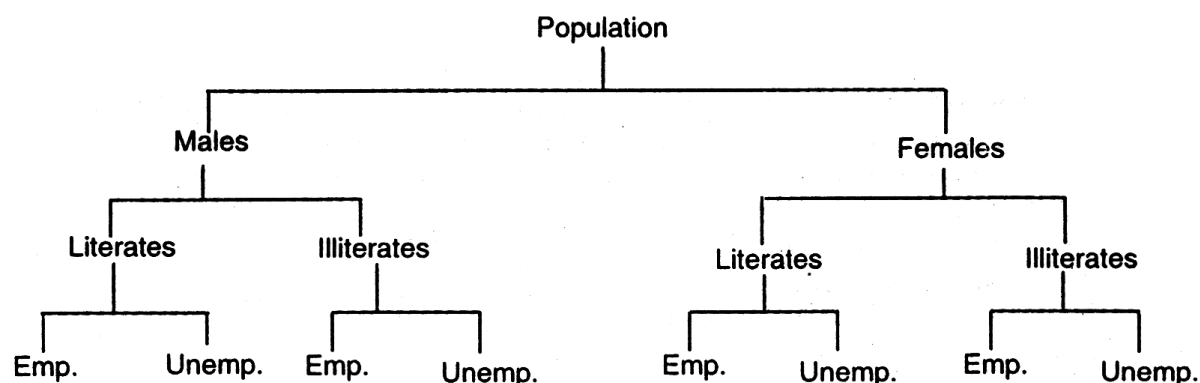
(iii) Qualitative Classification

In qualitative classification data are classified on the basis of some attribute or quality such as sex, colour of hair, literacy, religion, etc. The point to note in this type of classification is that the attribute under study cannot be measured: one can only find out whether it is present or absent in the units of the population under study.

For example, if the attribute under study is population, one can find out how many persons are living in urban area and how many in rural area. Thus when only one attribute is studied two classes are formed, one possessing the attribute and the other not possessing the attribute. This type of classification is known as simple classification. For example, the population under study may be divided into two categories as follows:



In a similar manner, we may classify population on the basis of sex, i.e., into males and females, or literacy, i.e., into literate and illiterate, and so on. The type of classification where only two classes are formed is also called two-fold or dichotomous classification. If instead of forming only two classes we further divide the data on the basis of some attribute or attributes so as to form several classes, the classification is known as manifold classification. For example, we may first divide the population into males and females on the basis of the attribute 'sex'; each of these classes may be further subdivided into 'literate' and 'illiterate' on the basis of the attribute 'literacy'. Further classification can be made on the basis of some other attribute, say, employment. An example of manifold classification is given here:



Note. Emp. indicates Employed and Unemp. indicates Unemployed.

(iv) Quantitative Classification

Quantitative classification refers to the classification of data according to some characteristics that can be measured, such as height, weight, income, sales, profits, production, etc. For example, the students of a college may be classified according to weight as follows:

Weight (in lb.)	No. of Students
90 - 100	50
100 - 110	200
110 - 120	260
120 - 130	360
130 - 140	90
140 - 150	40
Total	1,000

Such a distribution is known as empirical frequency distribution or simple frequency distribution.

In this type of classification, there are two elements, namely (i) the variable, i.e., the weight in the above example, and (ii) the frequency, i.e., the number of students in each class. There were 50 students having weight ranging from 90 to 100 lb., 200 students having weight ranging from 100 to 110 lb., and so on.

1.5.1 Methods of data Classification

Q21. What are the Methods of Data Classification ?

Ans :

There are two ways in which observations in the data set are classified on the basis of class intervals, namely

- (i) Exclusive method, and
- (ii) Inclusive method

(i) Exclusive Method

When the data are classified in such a way that the upper limit of a class interval is the lower limit of the succeeding class interval (i.e. no data point falls into more than one class interval), then it is said to be the exclusive method of classifying data. This method is illustrated in Table.

Table : Exclusive Method of Data Classification

Dividends Declared in per cent (Class Intervals)	Number of Companies (Frequencies)
0-10	5
10-20	7
20-30	15
30-40	10

Such classification ensures continuity of data because the upper limit of one class is the lower limit of succeeding class. As shown in Table, 5 companies declared dividends ranging from 0 to 10 per cent, this means a company which declared exactly 10 per cent dividend would not be included in the class 0-10 but would be included in the next class 10-20. Since this point is not always clear, therefore to avoid confusion data are displayed in a slightly different manner, as given in Table.

Table

Dividends Declared in per cent (Class Intervals)	Number of Companies (Frequencies)
0 but less than 10	5
10 but less than 20	7
20 but less than 30	15
30 but less than 40	10

(ii) Inclusive Method

When the data are classified in such a way that both lower and upper limits of a class interval are included in the interval itself, then it is said to be the inclusive method of classifying data. This method is shown in Table.

Table : Inclusive Method of Data Classification

Number of Accidents (Class Intervals)	Number of Weeks (Frequencies)
0-4	5
5-9	22
10-14	13
15-19	8
20-24	2

Remarks

1. An exclusive method should be used to classify a set of data involving continuous variables and an inclusive method should be used to classify a set of data involving discrete variables.
2. If a continuous variable is classified according to the inclusive method, then certain adjustment in the class interval is needed to obtain continuity as shown in Table.

Table

Class Intervals	Frequency
30-44	28
45-59	32
60-74	45
75-89	50
90-104	35

To ensure continuity, first calculate correction factor as:

$$x = \frac{\text{Upper limit of a class} - \text{Lower limit of the next higher class}}{2}$$

and then subtract it from the lower limits of all the classes and add it to the upper limits of all the classes.

From Table we have $x = (45 - 44) \div 2 = 0.5$. Subtract 0.5 from the lower limits of all the classes and add 0.5 to the upper limits. The adjusted classes would then be as shown in Table

Table

Class Intervals	Frequency
29.5-44.5	28
44.5-59.5	32
59.5-74.5	45
74.5-89.5	50
89.5-104.5	35

3. Class intervals should be of equal size to make meaningful comparison between classes. In a few cases, extreme values in the data set may require the inclusion of open-ended classes and this distribution is known as an open-ended distribution. Such open-ended classes do not pose any problem in data analysis as long as only a few frequencies (or values) lie in these classes. However, an open-ended distribution is not fit for further mathematical calculations because mid-value which is used to represent the class, cannot be determined for an open-ended class. An example of an open-ended distribution is given in Table

Age (Years)	Population (Millions)
Under 5	17.8
5-17	44.7
18-24	29.9
25-44	69.6
45-64	44.6
65 and above	27.4
	<u>234.0</u>

Table : Provides a tentative guide to determine an adequate Number of Classes

Number of Observations, N	Suggested Number of Classes
20	5
50	7
100	8
200	9
500	10
1000	11

Table : Guide to Determine the Number of Classes to Use

1.6 DATA ARRAY

Q22. What is Data Array ?

Ans :

A better presentation of the above raw data would be to arrange them in an ascending or descending order of magnitude which is called the 'arraying' of the data. However, this presentation (arraying), though better than the raw data does not reduce the volume of the data.

Organizing Data Using Data Array

The best way to examine a large set of numerical data is first to organize and present it in an appropriate tabular and graphical format.

Table presents the total number of overtime hours worked for 30 consecutive weeks by machinists in a machine shop. The data displayed here are in raw form, that is, the numerical observations are not arranged in any particular order or sequence.

94	89	88	89	90	94	92	88	87	85	88	93
94	93	94	93	92	88	94	90	93	84	93	84
91	93	85	91	89	95						

Table : Raw Data Pertaining to Total Time Hours Worked by Machinists

These raw data are not amenable even to a simple reading and do not highlight any characteristic/trend, such as the highest, the lowest, and the average weekly hours. Even a careful look at these data do not easily reveal any significant trend regarding the nature and pattern of variations therein. As such no meaningful inference can be drawn, unless these data are reorganized to make them more useful. For example, if we are to ascertain a value around which most of the overtime hours cluster, such a value is difficult to obtain from the raw data.

Moreover, as the number of observations gets large, it becomes more and more difficult to focus on the specific features in a set of data. Thus we need to organize the observation so that we can better understand the information that the data are revealing.

The raw data can be reorganized in a data array and frequency distribution. Such an arrangement enables us to see quickly some of the characteristics of the data we have collected.

When a raw data set is arranged in rank order, from the smallest to the largest observation or vice-versa, the ordered sequence obtained is called an *ordered array*. Table reorganizes data given in Table in the ascending order

84	84	85	85	87	88	88	88
88	89	89	89	90	90	91	91
92	92	93	93	93	93	93	93
94	94	94	94	94	95		

Table Ordered Array of Total Overtime Hours Worked by Machinists

It may be observed that an ordered array does not summarize the data in any way as the number of observations in the array remains the same. However, a few advantages of ordered arrays are as under:

Advantages of Ordered Array

The following are a few advantages of an ordered array:

- (i) It provides a quick look at the highest and lowest observations in the data within which individual values vary,
- (ii) It helps in dividing the data into various sections or parts.
- (iii) It enables us to know the degree of concentration around a particular observation.
- (iv) It helps to identify whether any values appear more than once in the array.

Disadvantages of Ordered Array

- (i) In spite of various advantages of converting a set of raw data into an ordered array, an array is a cumbersome form of presentation which is tiresome to construct. It neither summarizes nor organizes the data to present them in a more meaningful way.
- (ii) It also fails to highlight the salient characteristics of the data which may be crucial in terms of their relevance to decision-making. The above task cannot be accomplished unless the observations are appropriately condensed.
- (iii) The best way to do so is to display them into a convenient number of groupings with the number of observations falling in different groups indicated against each. Such tabular summary presentation showing the number (frequency) of observations in each of several non-overlapping classes or groups is known as frequency distribution (also referred to as grouped data).

1.7 FREQUENCY DISTRIBUTION

Q23. What is Frequency Distribution ? Explain the formation of discrete and continuous Frequency Distribution.

Ans :

Frequency Distribution

Arrangement of raw data in the form of row or column is called data array.

Frequency distribution refers to the tabular arrangement of data, when arranged into groups or categories according to conveniently established divisions of the range of the observations. In a frequency distribution, raw data is expressed in distinct groups called 'classes'. The number of observations that fall into each of these classes is known as 'frequency'. Thus, a frequency distribution has two parts, the left part represents the classes and the right part represents the frequencies.

When data is described by a continuous variable, it is called continuous data and when described using discrete variable it is called discrete data.

The following are the two examples of discrete and continuous frequency distributions.

No. of Students	No. of Colleges
100	25
120	70
140	24
160	100
180	37
200	50

Table (1) : Discrete Frequency Distribution

Age (In years)	No. of Students
10-15	12
15-20	20
20-25	14
25-30	40
30-35	18
40-45	22

Table (2) : Continuous Frequency Distribution

Construction of a Discrete Frequency Distribution

The process of preparing frequency distribution is very simple. For discrete data, place all the possible variables values in ascending order in one column and then prepare another column for 'Tally marks' (|) to count the number of marks. A block of five marks is prepared and some space is left in between the blocks. The frequency column refers to the number of "Tally marks", a particular class will contain. The construction of a discrete frequency distribution is illustrated as follows.

Consider a sample study in which 40 families were surveyed to find the number of children in each family. The data obtained are,

3	2	2	1	3	4	2	1	3	4
5	0	2	1	2	3	3	2	1	1
2	3	0	3	2	1	4	3	5	5
4	3	6	5	4	3	1	0	6	5

'Tally marks' are used to condense this data into a discrete frequency distribution as shown below.

No. of Children	No. of Families	Frequency
0	III	3
1	IIII II	7
2	IIII III	8
3	IIII IIII	10
4	IIII	5
5	IIII	5
6	II	2
Total		40

Table: Discrete Frequency Distribution

Construction of a Continuous Frequency Distribution

Before constructing the frequency distribution for continuous data, it is necessary to identify some of the important terms that are used frequently.

(i) Class Limits

The class limits represent the smallest and the largest values to be included in a class. The lowest and highest boundaries of a class, are known as the *lower* and *upper* limits of the class respectively.

For example, in a class among the values ranging between 70-79, 70 is the lower limit and 79 is the upper limit or we can say that there can be no value which is less than 70 and greater than 79.

(ii) Class Intervals

The class interval represents the width (span or size) of a class. The width may be determined by subtracting the lower limit of one class from the lower limit of the following class (alternatively successive upper limits may be used).

For example, if the two classes are 20-30 and 30-40, the width of the class interval would be the difference between the two successive lower limits i.e., $30 - 20 = 10$ or the difference between the upper limit and lower limit of the same class, i.e., $30 - 20 = 10$.

(iii) Class Midpoint

Midpoint of a class is defined as the average of the lower and upper limits of a class. The value thus obtained lies between the lower and upper class limits. For example the midpoint of a class 10-20 is,

$$\frac{10 + 20}{2} = \frac{30}{2} = 15$$

(iv) Class Frequency

The number of observations that fall into each of the class is known as the *class frequency*. For example, the class frequency for the class interval 20-30 is '10'.

Q24. What are the merits and demerits of Frequency Distribution?

Ans :

Merits of Frequency Distribution

The following are the merits of frequency distribution,

1. To obtain knowledge about the important features of the data, the data is expressed in a compact form.
2. In frequency distribution it is possible to note quickly the pattern of distribution of observations in different classes.
3. To analyze and reveal the data characteristics, statistical techniques can be used.

Demerits of Frequency Distribution

The following are the demerits of frequency distribution,

1. The identity of individual observations may be lost in the process of grouping. It becomes difficult to notice how the observations are put in each class and are distributed.
2. In different classes, there will be high clustering of observations, if the grouped observations are less than five. As a result, some important data may not be revealed.

1.7.1 Types of Frequency Distributions

Q25. Classify the different types of Frequency Distribution.

Ans :

The following are the three types of frequency distributions,

1. Cumulative frequency distribution
2. Relative frequency distribution
3. Percentage frequency distribution.

1. Cumulative Frequency Distribution

The normal frequency distribution simply tells us about the number of occurrence of a particular observation (i.e., value of the variable). However, it cannot provide the information such as “total number of observations whose value is less than or greater than a particular variable value”. In such a situation, cumulative frequency distribution can be implemented. This frequency distribution can be obtained by adding all the successive frequencies of the variable values according to some principles. These are two types of cumulative frequencies, they are,

- (i) 'Less than' cumulative frequency
- (ii) 'More than' cumulative frequency.

The less than cumulative frequency can be obtained simply by taking the first frequency as it is and then adding all the succeeding frequencies to it. On the other hand, the more than cumulative frequency can be obtained by taking the total number of frequencies as the first frequency and subtracting all the succeeding frequencies.

2. Relative Frequency Distribution

In order to evaluate data, sometimes it becomes necessary to disclose the percentage of observations which may come under each distribution class rather than disclosing the actual frequencies of class. However, for converting frequency distribution into corresponding relative frequency distribution one should use the following formula,

$$\text{Relative Frequency Distribution} = \frac{\text{Frequency Observation}}{\text{Total Number of Observations}}$$

3. Percentage Frequency Distribution

Percentage frequency distribution is calculated after calculating the relative frequency distribution. The formulae for calculating percentage frequency distribution is given below,

$$\text{Percentage Frequency Distribution} = \text{Relative Frequency} \times 100$$

(or)

$$\frac{\text{Frequency Observation}}{\text{Total Number of Observations}} \times 100$$

PROBLEMS

1. Prepare a frequency table for the following data with width of each class interval as 10. Use exclusive method of classification:

57	44	80	75	00	18	45	14	04	64
72	51	69	34	22	83	70	20	57	28
96	56	50	47	10	34	61	66	80	46
22	10	84	50	47	73	42	33	48	65
10	34	66	53	75	90	58	46	39	69

Sol :

FREQUENCY DISTRIBUTION

Marks	Tallies	Frequency
0-10	II	2
10-20	III	5
20-30	III	4
30-40	III	5
40-50	III III	8
50-60	III III	8
60-70	III II	7
70-80	III	5
80-90	IIII	4
90-100	III	2
Total		<u>50</u>

2. Prepare a frequency distribution for the following observations:

15 45 40 42 50 60 62 68 70 42
 75 75 80 81 25 26 31 32 78 45
 31 45 42 43 55 56 78 80 81 62
 60 62 58 69 70 45 50 56 72 58
 75 62 62 65 60 70 35 37 40 55

Sol :

Since the lowest value is 15 and the largest 81, we take class intervals of 10.

Frequency Distribution

Variable	Tally Bars	Frequency
15-25	I	1
25-35	III	5
35-45	III IIII	8
45-55	III I	6
55-65	III III IIII	14
65-75	III II	7
75-85	III IIII	9
Total		<u>50</u>

3. Classify the following data by taking class interval such that their mid-values are 17, 22, 27, 32, and so on.

30 42 30 54 40 48 15 17 51 42 25 41
 30 27 42 36 28 26 37 54 44 31 36 40
 36 22 30 31 19 48 16 42 32 21 22 46
 33 41 21

Sol :

Since we have to classify the data in such a manner that the mid-values are 17, 22, 27, etc. the first class should be 15-19 (mid-value = $(15 + 19)/2 = 17$, second class 20-24 etc.

FREQUENCY DISTRIBUTION

Variable	Tallies	Frequency
15-19	IIII	4
20-24	IIII	4
25-29	IIII	4
30-34	IIII III	8
35-39	IIII	4
40-44	IIII III	9
45-49	III	3
50-54	III	3
Total		<u>39</u>

4. The marks obtained by 50 students are given below:

31 13 46 31 30 45 38 42 30 9
 30 30 46 36 2 41 44 18 29 63
 44 30 19 5 44 15 7 25 12 30
 6 22 24 37 15 6 39 32 21 20
 42 31 19 14 23 28 17 53 22 21

Construct a grouped frequency distribution.

Sol :

The lowest value is 2 and largest 63. The appropriate class intervals shall be 10 because 7 classes would be formed by taking 10 as class interval.

FORMATION OF FREQUENCY DISTRIBUTION

Marks	Tallies	No. of Students
0-10	I	6
10-20		9
20-30		10
30-40		14
40-50		9
50-60	I	1
60-70	I	1
Total		50

Example

The marks of 200 students in a test were recorded and shown by the following frequency distribution.

Marks	No. of Student
10-19	7
20-29	11
30-39	20
40-49	46
50-59	57
60-69	37
70-79	15
80-89	7

Construct the Cumulative Frequency Table

Sol :

The cumulative frequency table is given below. If we keep on adding the frequency of each class starting from the frequency of the very first class, we get cumulative frequency.

Class Interval	Frequency	Cumulative (Less than)	Cumulative (More than)
10-19	7	7	200
20-29	11	$7 + 11 = 18$	$200 - 7 = 193$
30-39	20	$18 + 20 = 38$	$193 - 11 = 182$
40-49	46	$38 + 46 = 84$	$182 - 20 = 162$
50-59	57	$84 + 57 = 141$	$162 - 46 = 116$
60-69	37	$141 + 37 = 178$	$116 - 57 = 59$
70-79	15	$178 + 15 = 193$	$59 - 37 = 22$
80-89	7	$193 + 7 = 200$	$22 - 15 = 7$
	<u>200</u>		

(i) Less than Cumulative Frequency

Upper Limits	Cumulative Frequency (Less than)
Less than 19	7
Less than 29	18
Less than 39	38
Less than 49	84
Less than 59	141
Less than 69	178
Less than 79	193
Less than 89	200

(ii) More than Cumulative Frequency

Lower Limits	Cumulative Frequency (More than)
More than 10	200
More than 20	193
More than 30	182
More than 40	162
More than 50	116
More than 60	59
More than 70	22
More than 80	7

1.8 TABULATION OF DATA

Q26. What is tabulation ? Explain the significance of tabulation.

Ans :

One of the simplest and most revealing devices for summarizing data and presenting them in a meaningful fashion is the statistical table. A table is a systematic arrangement of statistical data in columns and rows. Rows are horizontal arrangements whereas columns are vertical ones. The purpose of a table is to simplify the presentation and to facilitate 'comparisons. The simplification results from the clear-cut and systematic arrangement, which enables the reader to quickly locate desired information.

The significance of tabulation will be clear from the following points:

(i) It simplifies complex data

When data are tabulated all unnecessary details and repetitions are avoided. Data are presented systematically in columns and rows. Hence, the reader gets a very clear idea of what the table presents. There is thus a considerable saving in time taken in understanding what is represented by the data and all confusion is avoided. Also a large amount of space is saved because of non-duplicating of his headings and designations; the description at the top of a column serves for all the terms beneath it.

(ii) It facilitates comparison

Tabulation facilitates comparison. Since a table is divided into various parts and for each part there are totals and sub-totals, the relationship between different parts of data can be studied much more easily with the help of a table than without it.

iii) It gives identity to the data

When the data are arranged in a table with a title and number they can be distinctly identified and can be used as a source reference in the interpretation of a problem.

iv) It reveals patterns

Tabulation reveals patterns within the figures which cannot be seen in the narrative form. It also facilitates the summation of the figures if the reader desires to check the totals.

1.8.1 Objectives of Tabulation

Q27. What are the Objectives of Tabulation ?

Ans :

- 1. Simplify Data :** The main objective of tabulation is to simplify a mass of information into its simplest form, making it easy to understand.
- 2. Provide Clarity :** Tabulation provides clarity on data collected. Many of the questions that are related to the objective of the study can be easily answered through tabulation.

3. **Facilitate Comparison :** Tabulation facilitates easy comparison. The two competing options that need to be evaluated can be placed side by side and compared.
4. **Trend Analysis :** Tabulation helps to catch the trend by more observation. Tabulated data is well placed to perform statistical calculation of trend and other features of data.
5. **Economy of Space and Time :** Presentation of data in tabular form saves space without any compromise on quality or completeness of data. Not only presentation, but usage of data also results in savings of time for the users of such data.
6. **Enables Easy Reference :** Tabulated information with appropriate title and number are easy to remember and convenient to refer at future date. Most Research papers will have a separate. Table of contents listing down the various table in which useful statistical data is presented and their page numbers.
7. **Detections of Errors and Omissions :** Particularly in case of numeric data, tabulation serves as a control. If any of the totals of rows or columns are not as they should be, it gets immediately known.
8. **Facilitate Statistical Processing :** It is only after tabulation that data becomes fit for statistical processing we can apply various statistical methods and techniques such as correlation, calculation of various measures of central tendency and dispersion etc, only after data is correctly tabulated.
9. **Clarity on Characteristics of data :** A concise tabular form clearly reveals the characteristics of data and highlights its significant characteristics.

1.8.2 Parts of Table

Q28. Explain different parts of a table.

Ans :

Title

[Head Note]:

HEADING						
Subheading 1			Subheading 2			
	Column Heading	Column Heading	Column Heading	Column Heading	Column Heading	Row Total
Row Heading	Body	Body	Body	Body	Body	"
Row Heading						"
Row Heading						"
Row Heading						"
Row Heading						"
Row Heading						"
Row total	Column Total	Column Total	Column Total	Column Total	Column Total	Grand Total

[Foot Note]:

[Source Note]:

Fig. : Format of a Statistical Table

Constructing Statistical Table

While constructing statistical table, great care should be taken. Even slight carelessness will give wrong table which results into wrong inferences. Hence, the following rules must be followed,

1. Number

Proper numbering should be given to each table at its top or bottom or centre or above the title. In case of large data, numbering should be done on left side.

2. Heading or Title

The heading of the table is the most important thing, so it must be brief, simple, clear and complete. It must be presented in such a way that the basic idea of the problem can be obtained.

3. Head Note

The head note must be placed below the title within brackets. It provides a brief description of the table contents.

4. Subheadings

It is necessary to write the subheadings of each column. The units, if any must be mentioned clearly.

5. Body

The body of the table is the most important part. It comprises of columns and rows. The number of rows and columns must be predecided and made suitable taking in view the problem under consideration.

6. Comparison

The main objective of the table is the comparative study of data. Hence, the classes should be arranged so as to make the comparison easier.

7. Totals

Space should always be provided for the total number of columns and rows and for the grand total.

8. Source

The source of collection of data should be mentioned below the table, so as to identify from where these have been taken.

9. Footnotes

Footnotes must be placed below the body of the table for providing additional information.

10. Attractive Form

The table must be elegant and artistic.

Q29. Explain the general rules for tabulation.

Ans :

- The table should suit the size of the paper usually with more rows than columns. In making a suitable layout it may be necessary to alter the original design. The alteration often consists in changing the rows to columns or the other way round. For this reason, it is desirable to make a rough draft of the table before the figures are entered in it. Space must be allowed for reference or any other matter which is to be included in the table.
- In all tables the captions and stubs should be arranged in some systematic order. It would make the table easier to read and allow more important items to be emphasized. The arrangement of items basically depends upon the type of data. However, the principal bases for arranging items are the following:
 - (a) **Alphabetical** i.e., arrangement according to alphabets. The type of arrangement is very common in general purpose or reference tables.
 - (b) **Chronological** i.e., arrangement according to time. This type of arrangement is of particular value in presenting historical data.
 - (c) **Geographical** i.e., arrangement of data in certain territorial units such as countries, cities, districts, etc.
 - (d) **Conventional** or arrangement in a customary order such as men, women and children or Hindus, Muslims, Sikhs and Christians.

- (e) Items may be arranged according to size, i.e., the numerical importance of the items, the largest items being given first and the smallest in the last. This arrangement may be reversed, if necessary.
- The point of measurement should be clearly defined and given in the table such as income in rupees or weight in pounds, etc.
 - Figures should be rounded off to avoid unnecessary details in the table and a footnote to this effect should be given. For example, the figures may be taken to the nearest rupees and paise be eliminated.
 - If certain figures are to be emphasized they should be in distinctive type or in a 'box' or circle or between thick lines.
 - The table should not be overloaded with details. If many characteristics are to be shown it is not necessary to load them all in one table; rather a number of tables should be prepared, each table complete in itself and serving a particular purpose.
 - A column entitled 'miscellaneous' should be added for data which do not fit in the classification made.
 - The arrangement of the table should be logical and items related to each other should be placed near about and, if possible, in the same group. Derivative figures such as totals, averages and percentages should be placed near the original figures. Columns and rows should be numbered for identification since reference is more easily given by quoting numbers than the title of the column.
 - Percentage and ratios should be computed and shown, if necessary. Frequently, figures in tables become more meaningful if expressed as percentages or (less often) as ratios. In constructing a table, therefore, it is important to decide whether or not it can be improved in this way. If it can, additional column should be inserted in the table and the percentages (or ratios) computed and entered. Such percentages and ratios are sometimes called derived statistics.
 - Where standard classifications have been prepared it is usually desirable to employ them, as they are superior to hastily constructed individual classification.
 - Indicate a zero quantity by a zero, and do not use zero to indicate information which is not available. If it is not available, show this fact by the letter N.A. or dash (-).
 - Abbreviations should be avoided especially in titles and headings. For example "yr" should not be used for 'year'.
 - Be explicit. The expression "etc." is bad form in a table, since the reader may not readily discover what it refers to. In fact clarity is the most important feature of tabular presentation of any kind of statistical data.

1.8.3 Types of Tables

Q30. Explain different types of Tables.

Ans :

Tables can be classified according to their purpose, stage of enquiry, nature of data or number of characteristics used. On the basis of the number of characteristics, tables may be classified as follows:

1. Simple (or) one-way table
2. Two way table
3. Manifold table

1. Simple (or) one-way Table

A simple or one-way table is the simplest table which contains data of one characteristic only. A simple table is easy to construct and simple to follow. For example, the blank table given below may be used to show the number of adults in different occupations in a locality.

The number of adults in different occupations in a locality.

Occupations	No. Of Adults
Total	

2. Two-way Table

A table, which contains data on two characteristics, is called a two way table. In such case, therefore, either stub or caption is divided into two co-ordinate parts. In the given table, as an example the caption may be further divided in respect of 'sex'. This subdivision is shown in two-way table, which now contains two characteristics namely, occupation and sex.

The number of adults in a locality in respect of occupation and sex.

Occupation	No. of Adults		Total
	Male	Female	
Total			

3. Manifold Table

Thus, more and more complex tables can be formed by including other characteristics. For example, we may further classify the caption subheadings in the above table in respect of "marital status", "religion" and "socio-economic status" etc. A table, which has more than two characteristics of data is considered as a manifold table. For instance, table shown below shows three characteristics namely, occupation, sex and marital status.

Occupation	No. of Adults						Total
	Male			Female			
	M	U	Total	M	U	Total	

Foot note: M Stands for Married and U stands for unmarried.

Manifold tables, though complex are good in practice as these enable full information to be incorporated and facilitate analysis of all related facts. Still, as a normal practice, not more than four characteristics should be represented in one table to avoid confusion. Other related tables may be formed to show the remaining characteristics.

Q31. What are the difference between Classification and Tabulation?

Ans :

Classification	Tabulation
It is the basis for tabulation	It is the basis for further analysis
It is the basis for simplification	It is the basis for presentation
Data is divided into groups and subgroups on the basis of similarities and dissimilarities.	Data is listed according to a logical sequence of related characteristics
data are separated and grouped based on a property of the data common to all values	Data is arranged into columns and rows based on characteristics /properties, or indicators.
emphasize on the presentation aspects of the data	used as a means of sorting of data for further analysis

1.9 GRAPHICAL PRESENTATION

Q32. What do you mean by Graphical Presentation ? Explain the techniques of construction of graphs.

Ans :

(Jan.-21)

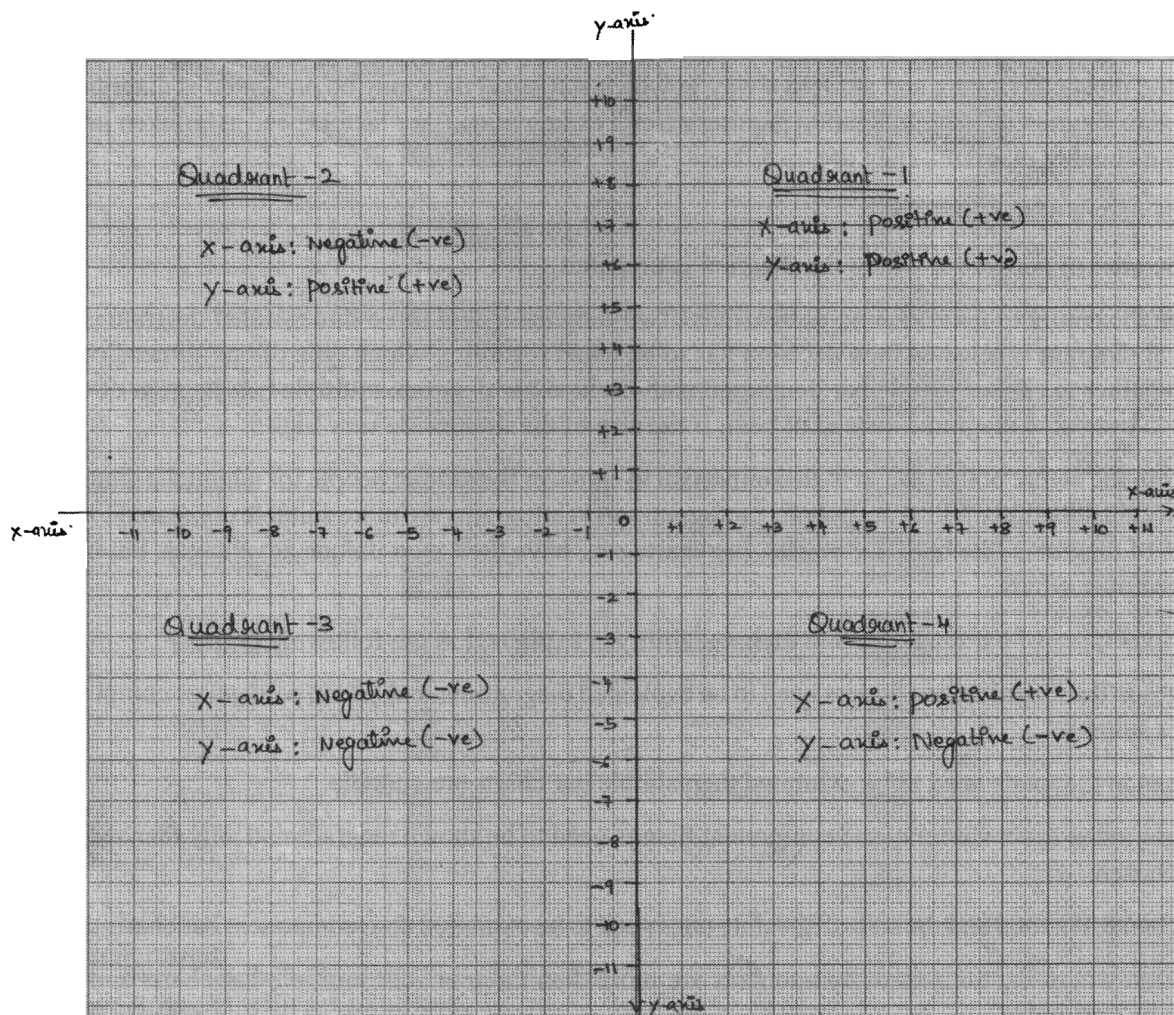
Graphical Presentation

Graphical presentation refers to the way of presenting the data with the help of graphs. Graphs play a significant role in presentation of data related to time series. Frequency distributions can also be efficiently presented using graphs.

Technique of Constructing Graphs

Graph paper is used for the construction of graphs. The graph sheet is an arrangement of horizontal and vertical intersecting lines. In any graph, firstly two simple lines are drawn intersecting each other. The intersecting point is called as origin or the 'zero' point and the two lines are called as 'coordinate axes', where the horizontal line is known as x-axis or 'abscissa' and the vertical line is known as Y-axis or 'ordinate'.

The following figure depicts a graph paper with X-axis and Y-axis and four quadrants.



In the above figure, the whole graph sheet is divided into four sections known as 'quadrants'. The point O is the origin point. In quadrant-1, both values of X-axis and Y-axis are positive. In Quadrant-2, values of X-axis is negative but values of Y-axis are positive. In quadrant-3, values of both X-axis and Y-axis are negative. In quadrant -4, X-axis values are positive and Y-axis values are negative. Usually the independent variable is shown on X-axis and dependent variable is shown on Y-axis. The scale on X-axis and Y-axis may or may not be same.

In a graph paper, the scale on X-axis and Y-axis starts with zero as origin. The curve drawn on the graph is interpreted based on the distance between the curve and the base line. In case if series of changes presented from a norm of 100% then 100% is taken as a base line. There is no fixed standard regarding the scale to be taken on X-axis and Y-axis, it completely depends upon the paper size and the given data.

The last step in the construction of graph is plotting the given data by taking into consideration the values of X and Y. The various points plotted on the graph need to be joined with a straight line.

Q33. What are the Guidelines for Construction of Graphs ?

Ans :

- Title :** A heading depicting the matter and contents of the data must be provided as title for all the graphical representations. The title must be clear, simple and precise.

2. **Organization of the Graph :** Each graph must be structurally drawn in an attractive manner. The portion of axis must be chosen and drawn neatly. We must represent independent and dependent parameters on X and Y axis respectively.
3. **Scale :** The scale selected must satisfy by all the values to be plotted on the graph.
4. **Index :** The index must be provided to show the scale of X and Y axis. The various lines drawn in the graph must be defined clearly in the index.
5. **Source of Data :** The source of data gives the information about the data retrieval and is mentioned at the bottom of the graph.
6. **False Base Line :** The vertical scale must start with zero for representing the given variable effectively, when the fluctuations are maximum and the starting values are very far from zero. Hence, the false base line is used to magnify the vertical scale and to plot the values. In a false base line, the vertical scale is broken into two parts and the values of the dependent variable from zero to the lowest value are omitted by drawing two zig-zag horizontal lines above the base line i.e., X-axis.
7. **Line Designs :** In case of representing several variables on the same graph for comparison, we may differentiate between them by using dotted lines, dash-dotted lines, broken lines, thick lines and thin lines. An index must also be provided for clearly mentioning the meanings of the lines.

1.9.1 Functions of Graphs

Q34. What are the functions of graphs ?

Ans : (Jan.-21)

- Graphic presentation of frequency distributions facilitate easy understanding of data presentation and interpretation.

- The shape of the graph offers easy answers to several questions.
- The same information can also be obtained from tabular presentation of a frequency distribution, but the same is not as effective in highlighting the essential characteristics as explicitly as is possible in the case of graphic presentation.
- The shape of the graph gives an exact idea of the variations of the distribution trends.
- Graphic presentation, therefore, serves as an easy technique for quick and effective comparison between two or more frequency distributions.
- When the graph of one frequency distribution is superimposed on the other, the points of contrast regarding the type of distribution and the pattern of variation become quite obvious.
- All these advantages necessitate a clear understanding of the various forms of graphic representation of a frequency distribution.

1.9.2 Advantages and limitations of Graphs

Q35. Explain the Advantages and limitations of Graphs.

Ans : (Jan.-21)

Advantages of Graphs

1. Graphs present information that might otherwise be lost amid the details of numerical tabulations.
2. A graph shows not only the value of a statistic but also any trend that may be present and the way in which the trend changes.
3. Graphic presentation renders complex data simple and easily understandable
4. A graph looks more attractive than a table.
5. Graphic method is the simplest method of presenting statistical data. It saves time and effort of the statistician as well as the observer.

6. Comparisons between two or more phenomena can be made easily with the help of graph.
7. No special knowledge of Mathematics is required to understand the message of data from the graph.
8. Graphic presentation is helpful in interpolation, extrapolation and forecasting.

Limitations of Graphs

Graphic presentation of data has the following limitations :

1. Actual values do not get much visibility, as there is greater attention to trend and fluctuations.
2. Complete accuracy is not possible on a graph
3. Graphs cannot be quoted to support some statement
4. Only few characteristics can be depicted on a graph
5. Selection of appropriate graph is vital to ensure that erroneous conclusions are not arrived at. This requires considerable expertise, skill and intelligence.
6. All graphic devices are not simple and straightforward. A good deal of expertise may be required to understand some of the graphs such as ratio graphs and multi-dimensional figures.

1.9.3 Types of Graphs

Q36. Explain different types of graphs.

Ans :

The graphical demonstration of statistical data in a chart is normally specified as statistical graph chart. There are many kinds of graphs and charts which are used to indicate a set of data. The data is either unremitting or separate. These graphs are very helpful to recognize the statistical data.

There are different kinds of graphical charts based on statistics as follows:

1. Line Graphs

A line graph is a diagram that shows a line joining several points, or a line that shows the relationship between the points. A line graph can be taken as xy plane, where there will be an independent variable and a dependent variable and it specifies how the two variables are related to each other and vary with respect to one another. Mostly, the independent variable is taken on the x-axis while the dependent variable on the y-axis.

2. Pie Charts

A pie chart can be taken as a circular graph which is divided into different disjoint pieces, each displaying the size of some related information. The highlight of this graph is that it represents a whole and each part represents a percentage of the whole. Hence, pie charts are best used with respect to categorical data which helps one understand what percentage each of these category constitutes. It also has a good visual treat and the percentage value of each section is instantly known.

3. Bar Graph

Bar graph is drawn on an x-y graph and it has labelled horizontal or vertical bars that show different values. The size, length and colour of the bars represent different values. Bar graph is very useful for non continuous data and it helps in comparing or contrasting the size of the different categories of the data provided.

4. Scatter Plot

A scatter plot or scatter graph is a type of graph which is drawn in Cartesian coordinate to visually represent the values for two variables for a set of data. It is a graphical representation that shows how one variable is affected by the other. The data is presented in the form of collection of points, each of which has one value of a variable positioned on the horizontal or x-axis, also called explanatory variable and the value of the other variable positioned on the vertical or y-axis, also called response variable.

5. Stem and Leaf Plot

Stem and leaf plot also called as stem plot are connected with quantitative data such that it helps in

- Displaying shapes of the distributions,
- Organize numbers and
- Set it as comprehensible as possible.

It is a descriptive technique which gives more emphases on the data provided. It concludes more about the shape of a set of data and provides better view about each of the data. The data is arranged by "place value". In Stem plots each data is taken and divided such that, they form two separate parts, a stem and a leaf. A stem is usually the first digit of the number in the data, and a leaf is the last digit of the number in the data. We write the stems in a vertical column and draw a vertical line to the right of the column. And hence we can write the leaves in the row to the right side of the corresponding stem.

1.9.4 Graphs of Frequency Distribution

Q37. Explain different graphs of frequency distribution.

Ans :

A frequency distribution can be presented graphically in any of the following ways:

- Histogram
- Frequency polygon
- Smoothed frequency curve, and
- 'Ogives' or cumulative frequency curves.

1.9.4.1 Histogram

Q38. Define Histogram. Explain with an example.

Ans :

One of the most commonly used and easily understood methods for graphical representation of frequency distribution data is called '**histogram**'. It is also known as '**column diagram**', depicting the class frequencies in a frequency distribution by vertical adjacent rectangles.

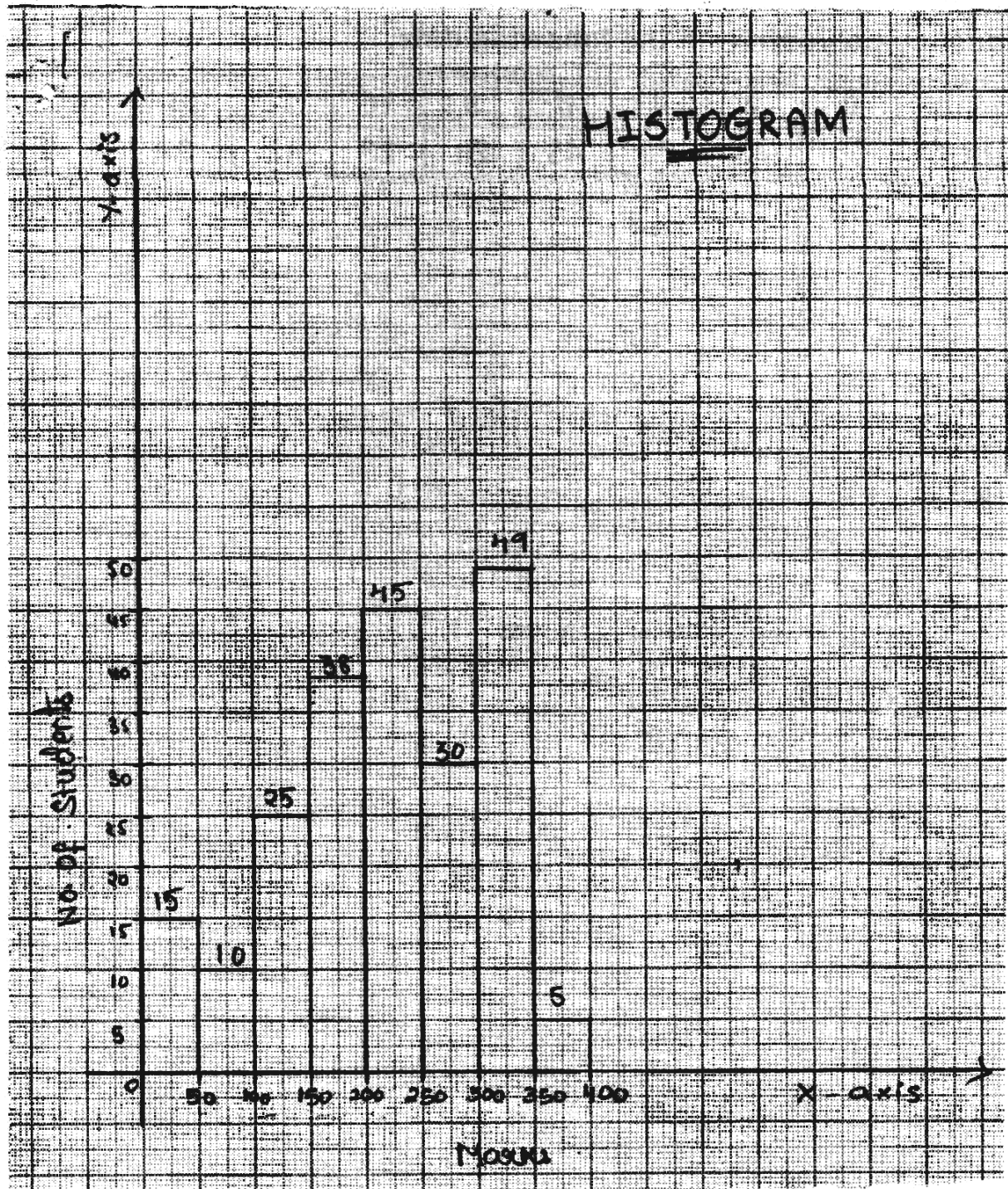
During the construction of histogram, variable is taken on the X-axis and the frequencies on the Y-axis. If the difference between the class intervals are same then distance between the rectangles on the X-axis should be same. The frequencies of each class which is equivalent to its height of rectangle can be shown on Y-axis. We get histograms by series of rectangles each with a class-interval distance as its breadth and the frequency intervals as its length. The total frequency that is distributed among the classes is the area of the histogram.

Example

Draw the histogram for the following,

Marks	No. of students
0-50	15
50-100	10
100-150	25

150-200	38
200-250	45
250-300	30
300-350	49
350-400	5



1.9.4.2 Frequency Polygon**Q39. Define Frequency Polygon. Explain with an example.****Ans :**

A frequency polygon is a graph of frequency distribution. It has more than four sides. It is particularly effective in comparing two or more frequency distributions. There are two ways in which a frequency polygon may be constructed:

- We may draw a histogram of the given data and then join by 'straight lines the mid-points of the upper horizontal side of each rectangle with the adjacent ones. The figure so formed is called frequency polygon. It is an accepted practice to close the polygon at both ends of the distribution by extending them to the base line. When this is done two hypothetical classes at each end would have to be included—each with a frequency of zero. This extension is made with the object of making the area under polygon equal to the area under the corresponding histogram. The readers are advised to follow this practice.
- Another method of constructing frequency polygon is to take the mid-points of the various class-intervals and then plot the frequency corresponding to each point and to join all these points by straight lines. The figures obtained would exactly be the same as obtained by method No. 1. The only difference is that here we have not to construct a histogram.
- By constructing a frequency polygon the value of mode can be easily ascertained. If from the apex of the polygon a perpendicular is drawn on the X-axis, we get the value of mode. Moreover, frequency polygons facilitate comparison of two or more frequency distributions on the same graph.

Frequency polygon has certain advantages over the histogram:

- The frequency polygons of several distributions may be plotted on the same axis, thereby making certain comparisons possible, whereas histograms cannot be usually employed in the same way. To compare histograms we must have a separate graph for each distribution. Because of this limitation for purposes of making a graphic comparison of frequency distributions, frequency polygons are preferred .
- The frequency polygon is simpler than its histogram counterpart.
- It sketches an outline of the data pattern more clearly.
- The polygon becomes increasingly smooth and curve-like as we increase the number of classes and the number of observations.

Example

Draw the frequency polygon from the following data,

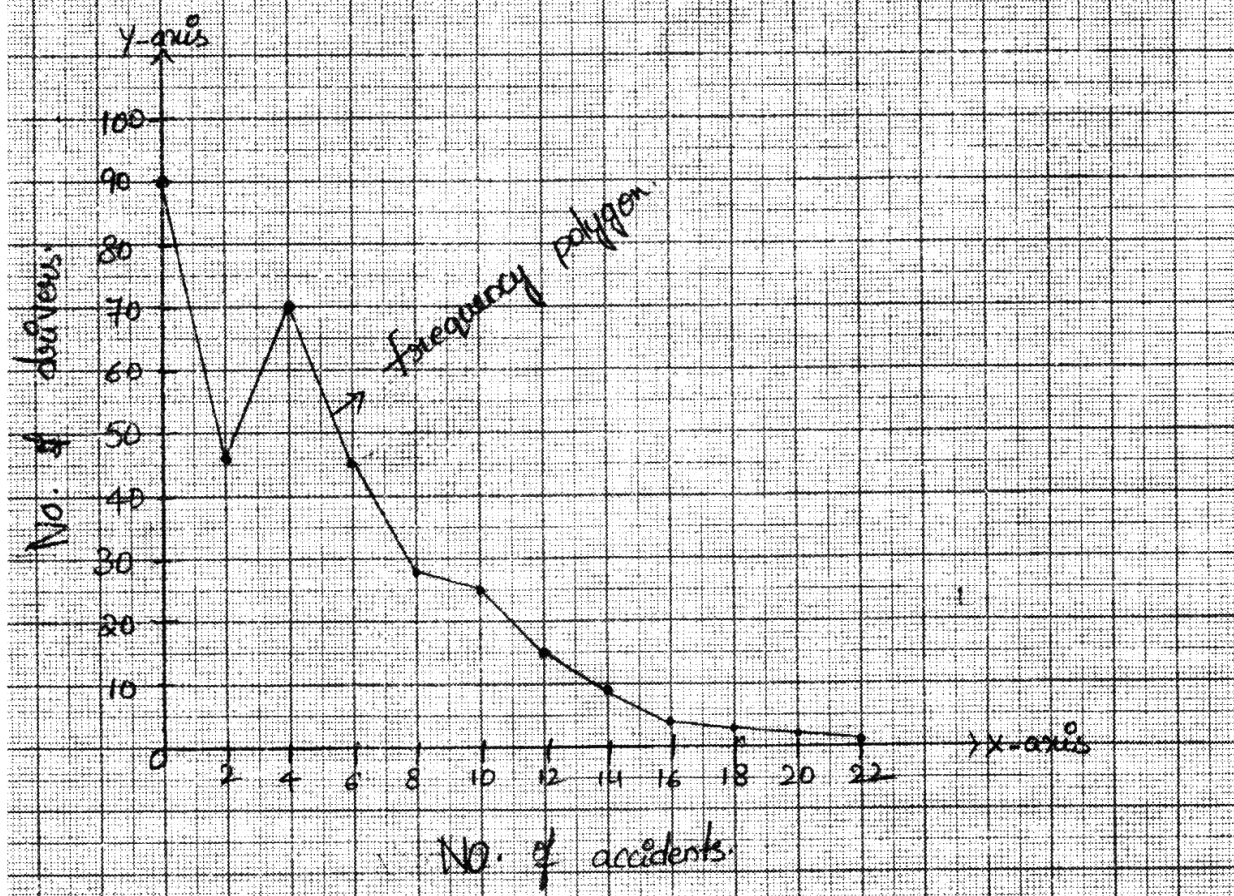
No. of Accidents	0	2	4	6	8	10	12	14	16	18	20	22
No. of Drivers	90	46	70	45	28	25	15	9	4	3	2	1

Sol :

Scale:

X-axis: 1cm = 2 accidents

Y-axis: 1cm = 10 drivers

frequency polygon of Number of Accidents.

1.9.4.3 Frequency Curve**Q40. Define Frequency Curve. Explain with an example.****Ans :**

A smoothed frequency curve can be drawn through the various points of the polygon. The curve is drawn freehand in such a manner that the area included under the curve is approximately the same as that of the polygon. The object of drawing a smoothed frequency curve is to eliminate as far as possible accidental variations that might be present in the data. While smoothing a frequency polygon the fact that it is really derived from the histogram should always be kept in mind. This would imply that the top of the curve would overtop the highest point of the polygon particularly when the magnitude of class-interval is large. The curve should look as regular as possible and sudden turns should be avoided. The extent of smoothing would, however, depend upon the nature of the data. If it is a natural phenomenon like tossing of coin, smoothing may be freely resorted to as such phenomenon normally has symmetrical curves, but if the phenomenon is social or economic the curve is generally skewed and as such smoothing cannot be carried too far.

For drawing a smoothed frequency curve it is necessary to first draw the polygon and then smooth it out. As discussed earlier, the polygon can be constructed even without first constructing histogram by plotting the frequencies at the mid-points of class-intervals. This may save some time but the smoothing of the polygon cannot be done properly without a histogram. Hence, it is desirable to proceed in a sequence, *i.e.*, first to draw a histogram then a polygon and lastly to smooth it to obtain the smoothed frequency curve. This curve should begin and end at the base line and as a general rule it may be extended to the mid-points of the class-intervals just outside the histogram. The area under the curve should represent the total number of frequencies in the entire distribution.

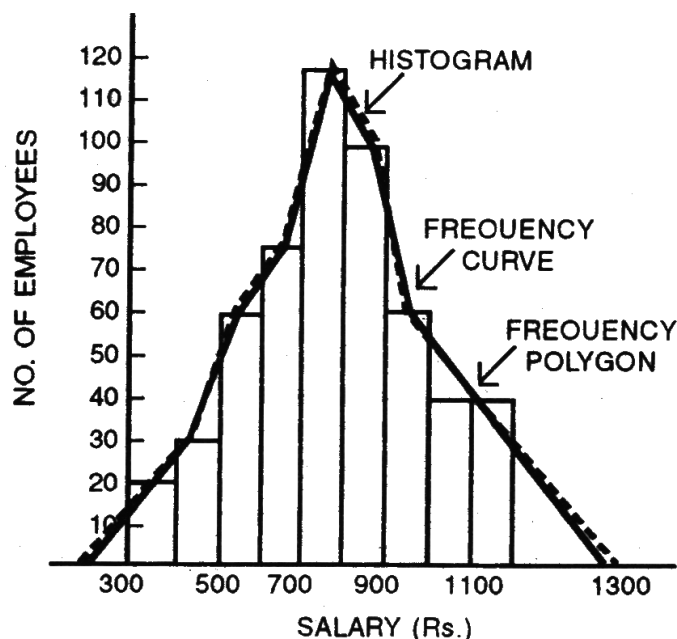
The following points should be kept in mind while smoothing a frequency curve:

- Only frequency distributions based on samples should be smoothed.
- Only continuous series should be smoothed.
- The total area under the curve should be equal to the area under the original histogram or polygon.

Illustration

Represent the following frequency distribution by means of a Histogram and superimpose thereon the corresponding frequency polygon and frequency curve:

Salary (Rs.)	No. of Employees
300-400	20
400-500	30
500-600	60
600-700	75
700-800	115
800-900	100
900-1000	60
1000-1200	40



1.9.4.4 Cumulative Frequency

Q41. Define Cumulative Frequency (or) Ogive curve. Explain with an example.

Ans :

The cumulative frequency curve also known as Ogive is a graph of cumulative frequency distribution. These cumulative frequencies are obtained by adding the frequencies.

The two techniques for constructing an Ogive are,

- (a) More than technique
 - (b) Less than technique.
- (a) 'More than' method:** In the 'more than' method we start with the lower limits of the classes and from the frequencies we subtract the frequency of each class. When these frequencies are plotted we get a declining curve.
- (b) 'Less than' method:** In the 'less than' method we start with the upper limits of the classes and go on adding the frequencies. When these frequencies are plotted we get a rising curve.

Examples

From the following table we get to know the monthly income of 750 families in a city.

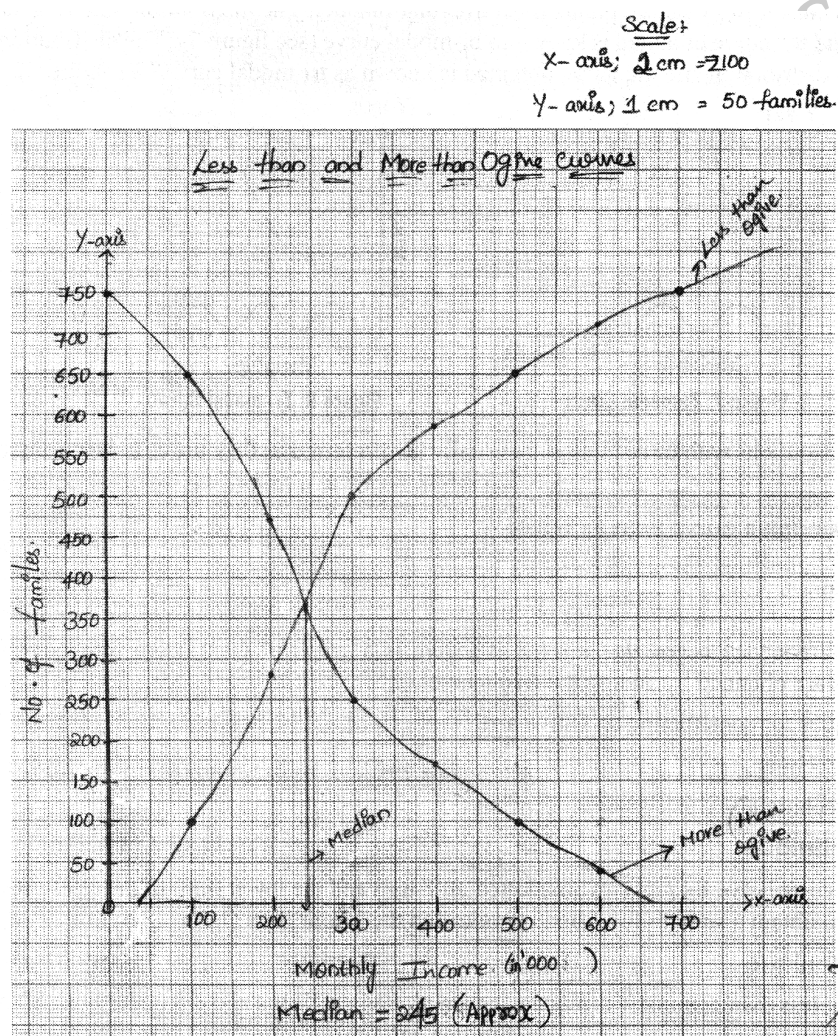
Monthly ('000) Income	Below 100	100-200	200-300	300-400	400-500	500-600	600 and above
No. of Families	100	180	220	80	70	60	40

Draw a less than and more than ogive curve, from the given data.

Sol :

Monthly Income (in '000)	No. of Families (f)	Less Than C.F.	More Than C.f.
Below 100	100	100	750
100-200	180	280	650
200 - 300	220	500	470
300 - 400	80	580	250
400- 500	70	650	170
500-600	60	710	100
600 and above	40	750	40

Note: While plotting less than C.F. on the graph, upper limits are considered and while plotting more than C.F. on the graph, lower limits are considered.



Exercise Problems

1. In a survey of 30 families in a village, the number of children per family was recorded and following data obtained.

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

Represent the data in the form of a discrete frequency distribution.

2. The class of 2C04 of Osmania University had 30 students who scored the following letter grades.

A	B	D	C	C	B	D	C	B	C	B	C	B	A	B
C	B	A	D	D	C	B	B	D	D	A	B	C	C	B

Represent the data in the form of a discrete frequency distribution.

7. Construct a frequency table (inclusive method) for the following data regarding annual profits (in thousands of rupees) of 50 firms, using Surge's formula to decide the class-intervals

28 35 61 29 36 48 57 67 69 50 48 40 47 42
 41 37 51 62 63 33 31 32 35 40 38 37 60 51
 54 56 37 46 42 38 61 59 58 44 39 57 38 44
 45 47 38 44 47 47 64

30. Represent the following data by suitable diagram :

Marks	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40
Frequency	1	8	18	36	20	10	7

35. Construct a Suitable diagram for the following datas.

Year	1999	2000	2001	2002	2003	2004
Imports (Rs. Cr.)	18	21	25	26	30	36
Exports (Rs. Cr.)	15	19	27	29	35	40

[Ans: Scale : Y Axis = 1 cm = 1 year X axis 1 cm = Rs. 5 Crores]

Short Question and Answers

1. Define the term Statistics.

Ans :

The word 'statistic' has been derived from the Latin word 'status' or the Italian word 'statistic' or the German word 'statisti' or the French word 'statistique' all of which mean a political state'. The 'State' collected data pertaining to population and their wealth, in order to plan new taxes and to fund wars. Hence, statistics was also known as 'science of statecraft'.

Statistics is said to be as old as recorded history. One of the earliest known statistics is the census conducted by the Emperors of Egypt in connection with construction of pyramids. Statistics have been found to be use in India even before 300 B.C. Historical evidences about the prevalence of a good system of collection of data are available in Kautilya's Arthshastra.

Definitions

- (i) **According to A.L. Bowley**, "Statistics are numerical facts in any department of enquiry placed in relation to each other."

The main features of the above definition are:

- (a) Statistics (or Data) implies numerical facts.
- (b) Numerical facts or figures are related to some enquiry or investigation.
- (c) Numerical facts should be capable of being arranged in relation to each other.

On the basis of the above features we can say that data are those numerical facts which have been expressed as a set of numerical figures related to each other and to some area of enquiry or research. We may, however, note that all the characteristics of data are not covered by the above definition.

- (ii) **According to Yule & Kendall**, "By statistics we mean quantitative data affected to a marked extent by multiplicity of causes."

This definition covers two aspects, i.e., the data are quantitative and affected by a large number of causes.

- (iii) **According to Webster**, "Statistics are classified facts respecting the conditions of the people in a state - especially those facts which can be stated in numbers or in tables of numbers or in any other tabular or classified arrangement."

- (iv) **According to Achenwall**, "A collection of noteworthy facts concerning state, both historical and descriptive."

Definitions 3 and 4, given above, are not comprehensive because these confine the scope of statistics only to facts and figures related to the conditions of the people in a state. However, as we know that data are now collected on almost all the aspects of human and natural activities, it cannot be regarded as a state-craft only.

- (v) **According to L.R. Connor**, "Statistics are measurements, enumerations or estimates of natural or social phenomena, systematically arranged, so as to exhibit their inter-relations."

This definition also covers only some but not all characteristics of data.

- (vi) **According to H. Secrist**, "By statistics we mean aggregate of facts affected to a marked extent by a multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."

This definition can be taken as a comprehensive definition of statistics since most of the characteristics of statistics are covered by it.

2. Scope of statistics

Ans :

Statistics is viewed not just as a device for collecting numerical data, but as a means of sound techniques for their handling, analysis and for drawing valid inference from them. From this perspective, the scope or subject matter to statistics can be broadly studied under 2 hands namely.

- (A) Statistical Methods and
- (B) Applied Statistics

(A) Statistical Methods

Statistical methods are the tools that are in the hands of the statistician. They include all the general principles and techniques that are commonly used in the collection, analysis and interpretation of data. These methods are applicable to all kinds of data. The stage involved in study of any kind of data are

- (i) Observation and Collection
- (ii) Organization
- (iii) Presentation
- (iv) Analysis and
- (v) Interpretation

(B) Applied Statistics

Applied Statistics deals with application of Statistical methods of specific problems or concrete forms. To illustrate, if a software services firm is experiencing attrition (loss of personnel), it may be worthwhile to investigate the reasons for the same. Special techniques can be employed to understand the underlying trends.

For example, a correlation analysis could probably indicate that increasing opportunities (growth in the industry) and attrition are highly correlated. All such

techniques and the results obtained by employing such techniques form part of applied Statistics.

3. What is Principals of Measurement.

Ans :

Measurement is the process of systematically assigning numbers to objects and their properties to facilitate the use of mathematics in studying and describing objects and their relationships. Some types of measurement are fairly concrete: for instance, measuring a person's weight in pounds or kilograms or his height in feet and inches or in meters. Note that the particular system of measurement used is not as important as the fact that we apply a consistent set of rules: we can easily convert a weight expressed in kilograms to the equivalent weight in pounds, for instance. Although any system of units may seem arbitrary (try defending feet and inches to someone who grew up with the metric system!), as long as the system has a consistent relationship with the property being measured, we can use the results in calculations.

4. Advantages of Primary Data

Ans :

(i) Reliability

The information collected for primary data is more reliable than those collected from the secondary data because this information is collected directly from the respondents.

(ii) Availability of a Wide Range of Techniques

There are lot of techniques that can be employed, which means that all information necessary can be obtained by using the appropriate techniques, enabling all areas of the research topic to be answered and investigating thoroughly and effectively.

(iii) Addresses Specific Research Issues

The organisation asking for the research has the complete control on the process and the research is streamlined as far as its objectives and scope is concerned. Researching company can be asked to concentrate their

efforts to find data regarding specific market rather than concentration on mass market. Primary research is designed to collect the information the marketer wants to know, and report it in ways that benefit the marketer.

(iv) Greater Control

Not only does primary research enable the marketer to focus on specific issues, it also enables the marketer to have a higher level of control over how the information is collected. In this way the marketer can decide on such issues as size of project (e.g., how many responses), location of research (e.g., geographic area) and timeframe for completing the project.

5. Disadvantages of Secondary Data

Ans :

- (i) Relevance:** The data may not fit into the needs of investigation. There may be difference in the units of measurement; there may be surrogated data; discrepancy of classes and data may pertain to some other period of time.
- (ii) Accuracy:** It is observed that it is rather difficult to measure the degree of approximations used in the collection of information as well as the competence of the investigator in motivating the persons to supply the desired information.
- (iii) Existence of Obsolete Information:** Information may be outdated or obsolete.
- (iv) Non-Disclosure of Research Findings:** All the findings of a research study may not be made public.
- (v) Seldom Catering to the Need:** The available data may not suit the current purpose of research, due to incompleteness, generalities and so on.
- (vi) Other Limitations**
 - i) There may be difficulties in the identification of the source.

- ii) Errors may be there in recording or transferring information from secondary sources.
- iii) The facilities or capabilities of the agency that originally collect the data might be questionable.

6. Define classification.

Ans :

Classification is the grouping of related facts into classes. Facts in one class differ from those of classification. Sorting facts on one basis of classification and then on another basis is called cross-classification. This process can be repeated as many times as there are possible bases of classification. Classification of data is a function very similar to that of sorting letters in a post-office. It is well known that letters collected in a post-office are sorted into different lots on a geographical basis, i.e., in accordance with their destinations such as Mumbai, Calcutta, Kanpur, Jaipur, etc. They are then put into separate bags, each containing letters with a common characteristic, viz., having the same destination. Classification of statistical data is comparable to the sorting operation.

7. What is Frequency Distribution?

Ans :

Arrangement of raw data in the form of row or column is called data array.

Frequency distribution refers to the tabular arrangement of data, when arranged into groups or categories according to conveniently established divisions of the range of the observations. In a frequency distribution, raw data is expressed in distinct groups called 'classes'. The number of observations that fall into each of these classes is known as 'frequency'. Thus, a frequency distribution has two parts, the left part represents the classes and the right part represents the frequencies.

When data is described by a continuous variable, it is called continuous data and when described using discrete variable it is called discrete data.

8. Merits and demerits of frequency distribution.

Ans :

Merits of Frequency Distribution

The following are the merits of frequency distribution,

- (i) To obtain knowledge about the important features of the data, the data is expressed in a compact form.
- (ii) In frequency distribution it is possible to note quickly the pattern of distribution of observations in different classes.
- (iii) To analyze and reveal the data characteristics, statistical techniques can be used.

Demerits of Frequency Distribution

The following are the demerits of frequency distribution,

- (i) The identity of individual observations may be lost in the process of grouping. It becomes difficult to notice how the observations are put in each class and are distributed.
- (ii) In different classes, there will be high clustering of observations, if the grouped observations are less than five. As a result, some important data may not be revealed.

9. What is tabulation ?

Ans :

One of the simplest and most revealing devices for summarizing data and presenting them in a meaningful fashion is the statistical table . A table is a systematic arrangement of statistical data in columns and rows. Rows are horizontal arrangements whereas columns are vertical ones. The purpose of a table is to simplify the presentation and to facilitate 'comparisons. The simplification results from the clear-cut and systematic arrangement, which enables the reader to quickly locate desired information.

The significance of tabulation will be clear from the following points:

- (i) **It simplifies complex data:** When data are tabulated all unnecessary details and repetitions are avoided. Data are presented systematically in columns and rows. Hence, the reader gets a very clear idea of what the table presents. There is thus a considerable saving in time taken in understanding what is represented by the data and all confusion is avoided. Also a large amount of space is saved because of non-duplicating of his headings and designations; the description at the top of a column serves for all the terms beneath it.
- (ii) **It facilitates comparison:** Tabulation facilitates comparison. Since a table is divided into various parts and for each part there are totals and sub-totals, the relationship between different parts of data can be studied much more easily with the help of a table than without it.
- (iii) **It gives identity to the data:** When the data are arranged in a table with a title and number they can be distinctly identified and can be used as a source reference in the interpretation of a problem.

10. What are the Objectives of Tabulation ?

Ans :

- (i) **Simplify Data :** The main objective of tabulation is to simplify a mass of information into its simplest form, making it easy to understand.
- (ii) **Provide Clarity :** Tabulation provides clarity on data collected, Many of the questions that are related to the objective of the study can be easily answered through tabulation
- (iii) **Facilitate Comparison :** Tabulation facilitates easy comparison. The two competing options that need to be evaluated can be placed side by side and compared.
- (iv) **Trend Analysis :** Tabulation helps to catch the trend by more observation. Tabulated data is well placed to perform statistical calculation of trend and other features of data.

- (v) **Economy of Space and Time** : Presentation of data in tabular form saves space without any compromise on quality or completeness of data. Not only presentation, but usage of data also results in savings of time for the users of such data.
-

11. What are the functions of graphs ?

Ans :

- Graphic presentation of frequency distributions facilitate easy understanding of data presentation and interpretation.
 - The shape of the graph offers easy answers to several questions.
 - The same information can also be obtained from tabular presentation of a frequency distribution, but the same is not as effective in highlighting the essential characteristics as explicitly as is possible in the case of graphic presentation.
 - The shape of the graph gives an exact idea of the variations of the distribution trends.
 - Graphic presentation, therefore, serves as an easy technique for quick and effective comparison between two or more frequency distributions.
 - When the graph of one frequency distribution is superimposed on the other, the points of contrast regarding the type of distribution and the pattern of variation become quite obvious.
-

Choose the Correct Answers

1. Sum of dots when two dice are rolled is [a]
a) a discrete variable b) a continuous variable
c) a constant d) a qualitative variable
2. The weights of students in a college/school is a [b]
a) Discrete Variable b) Continuous Variable
c) Qualitative Variable d) None of these
3. The number of accidents in a city during 2010 is [a]
a) Discrete variable b) Continuous variable
c) Qualitative variable d) Constant
4. Which of these represent qualitative data [b]
a) Height of a student
b) Liking or disliking of (500) persons of a product
c) Income of a government servant in a city
d) Yield from a wheat plot
5. Life of a T.V picture tube is a [b]
a) Discrete variable b) Continuous variable
c) Qualitative variable d) Constant
6. The "rst hand and unorganized form of data is called [c]
a) Secondary data b) Organized data
c) Primary data d) None of these
7. The data which have already been collected by some one are called [c]
a) Raw data b) Array data
c) Secondary data d) Fictitious data
8. Census reports used as a source of data is [b]
a) Primary source b) secondary source
c) Organized data d) none
9. The grouped data is also called [c]
a) Raw data b) Primary data
c) Secondary data d) Qualitative data
10. Primary data and - data are same [c]
a) Grouped b) Secondary data
c) Ungrouped d) None of these

Fill in the blanks

1. The mean of a distribution is 14 and the standard deviation is 5. What is the value of the coefficient of variation _____
2. The mean of a distribution is 23, the median is 24, and the mode is 25.5. It is most likely that this distribution is _____
3. Which of the following describe the middle part of a group of numbers _____
4. According to the empirical rule, approximately. What percent of the data should lie within _____
5. The sum of the deviations about the mean is always _____
6. The middle value of an ordered array of numbers is the _____
7. Which of the following is not a measure of central tendency _____
8. Which of the following divides a group of data into four subgroups _____
9. If the standard deviation of a population is 9, the population variance is _____
10. If a distribution is abnormally tall and peaked, then it can be said that the distribution is _____

ANSWERS

1. 35.7%
2. Negatively Skewed
3. Measure of Central Tendency
4. 95%
5. Zero
6. Median
7. Standard deviation
8. Quartiles
9. 81
10. leptokurtic

UNIT II

MEASURES OF CENTRAL TENDENCY :

Introduction to Averages – Requisites for a Measure of Central Tendency, Mean - Combined mean – Weighted mean, Median – Partition values – Quartiles, Deciles and Percentiles, Relationship between Partition values–Mode– Relationship between Mean, Median and Mode.

Measures of Dispersion:

Introduction – Significance and Requisites of a Measure of dispersion, Range, QD, MD and SDFor Grouped and Ungrouped – Advantages and Disadvantages. Concept of Variation – Coefficient of Variation.

Skewness and Kurtosis (SK):

Introduction, Measures of SK, Relative measures of SK – Advantages and Disadvantages. Moments – concepts – Calculation – Kurtosis.

2.1 INTRODUCTION TO AVERAGES

Q1. Define the term average. What are the objectives of averages.

Ans :

The average of a distribution has been defined in various ways. Some of the important definitions are:

- (i) **According to Clark and Sekkade**, “An average is an attempt to find one single figure to describe the whole of figures”.
- (ii) **According to Murray R. Spiegel**, “Average is a value which is typical or representative of a set of data”.
- (iii) **According to Croxton and Cowden**, “An average, is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of data it is sometimes called a measure of central value”.
- (iv) **According to Sipson and Kafka**, “A measure of central tendency is a typical value around which other figures congregate”.

Objectives

- 1. Represent the characteristics of the entire mass of data**

An average reduces a complex mass of data into a single typical figure to enable the user

to get a bird's eye view about the characteristics of the phenomenon under study. In the words of M.J. Mooney, “the purpose of an average is to represent a group of individual values in a simple and concise manner, so that the mind can have a quick understanding of the general size of the individuals in the group”. The objective of an average is to enable the user to make inferences about the entire group by looking at the average. It should facilitate easy expression, reference and interpretation.

- 2. To facilitate comparison**

An Average reduces a mass of complex data into one single figure. This makes it possible for the user to make comparisons between two or more sets of data and draw conclusions the characteristics of the separate sets of data. To illustrate, we can compare the average marks obtained by students from different colleges in a given examination and determine the best college. Similarly, average monthly sales of a product can help us understand if there is an impact of different seasons on the sales.

- 3. Describe Data**

Average serve the objective of describing the main characteristics of the underlying data in a simple and brief manner.

4. To quantify relationships

One of the functions of an average is to help in establishing relationships between different groups in quantitative terms. For example, stating that income of an average American is more than that of an average Indian may not mean much. It is more meaningful if the respective incomes are expressed in terms of averages.

4. To reduce the impact of extreme values

Average help in setting off the impact of extreme values. For example, if a batsman has scored 300 and 10 runs respectively in two innings of a test match, the average score will read as 155. This is far more critical when number of observations is large.

5. Basis for further Analysis

Statistical analysis covers number of measures such as standard deviation, correlation, regression, etc to name a few. All these calculations are based on the "average". Thus, averages serve the purpose of facilitating further analysis.

6. To facilitate Sampling

We have briefly learnt about Sampling and various methods of sampling in the previous chapter. The concept of Sampling is to make inferences about a Universe (population) forming a sample as the basis. This objective is achieved only with the help of averages.

7. Facilitate Decision making

A business manager is interested in numbers such as sales, production, Occupancy (in case of hotel industry), capacity utilization (particularly in case of travel industry - Air, Train, bus), etc. All these numbers are critical for decision making and policy formulation. However, none of these numbers can be used as is. To facilitate decision making, all these numbers are stated in averages.

In short, Averages are very much useful (i) for describing the distribution in concise manner (ii) For comparative study of different distributions (iii) for computing various other statistical measures and lastly (iv) to take informed decisions.

Q2. What are the characteristics of a good average?

Ans :

1. It should be rigidly defined

An average should be rigidly defined so that there is no scope for confusion or manipulation. The average value will become very unstable and non representative of the base data if it is not well defined. It is ideal to use an average that is mathematically defined by way of formula.

2. It should be easy to calculate and simple to follow

Calculation of an average should be simple to understand. It should be easy to calculate, preferably without the help of calculators. If an average is too complex to understand and calculate, its use will be very limited. It should also be capable of expression in simple numerical terms without advanced mathematical intricacies.

3. It should be based on all observations in the series

An average will be truly representative of the whole mass of data if it is computed from all the observations.

4. It should not be affected much by an extreme values

An average will be representative of the data only if it can set off extreme values against each other. A few very small or very large observations should not unduly affect the value of a good average.

5. It should be capable of further statistical processing

An average should be capable of being used in calculation of other statistical measures such as standard deviation, correlation, etc.

6. It should be capable of further algebraic treatment

An average should lend itself readily to further algebraic treatment. For example, if averages of two or more sets of data are

known, it should be possible to obtain the average of combined group.

7. It should possess sampling stability

An average should be least affected by sampling. If we take independent random samples of the same size from a given population and compute the average for each of these samples, the values so obtained from different samples should not be very different from one another.

8. It should be representative of the data

A good average should represent maximum characteristics of the underlying data. Besides the above, a good average should ideally be nearest to most items of the series, should be capable of being ascertained through graphic methods and should be popular. Lastly, an average should be meaningful.

For example, it is absurd to talk of an average of man's height and weight.

Q3. What are the limitations of averages ?

Ans :

1. Misleading Conclusions

Average is a single numerical figure representing the characteristics of a given distribution. This number is vulnerable to errors in interpretation and can lead to misleading conclusions.

For example, if the maximum and minimum temperatures of a particular city are 48° F and 2° F respectively, the average may still work out to 24° F. Looking at the average, the weather conditions might look very comfortable, but that is not really the case.

2. Choice of Average

There are different types of averages. Different types of averages are suitable for different objectives. The utility of average depends on a proper and judicious choice of the average. A wrong choice of the average might lead to erroneous results.

3. Incomplete Picture

An average does not provide the complete picture of a distribution. There may be number of distribution having the same average but differing widely in their structure and constitution. To form a complete idea about the distribution, the measures of central tendency are to be supplemented by some more measures such as dispersion, skewness and kurtosis.

4. Inadequate Representation of Data

In certain types of distribution like U- shaped (or) J-shaped distribution, an average, which is only a single point of concentration, does not adequately represent the centre series.

5. Absurd Conclusions

Sometimes, an average might throw up very absurd results that are not realistic. For example, the average size of a family might work out to a fractional number, which is not realistic.

Q4. Explain the various types of averages.

Ans :

Various measures of average can be classified into the following two categories:

(a) Mathematical Averages

The mathematical averages are as follows:

- (i) Arithmetic Mean or Mean
- (ii) Geometric Mean
- (iii) Harmonic Mean

(b) Positional Averages

The positional averages are as follows:

- (i) Median
- (ii) Mode

The above measures of central tendency will be discussed in the order of their popularity. Out of these, the Arithmetic Mean, Median and Mode, being most popular, are discussed in that order.

2.1.1 Requisites of a Measure of Central Tendency**Q5. What are the requisites of a measure of central tendency ?**

Ans : (May-17)

The following are the few requirements to be satisfied by an average or a measure of central tendency:

1. It Should be rigidly defined

The definition of an average should be clear and rigid so that there must be uniformity in its interpretation by different decision-makers or investigators. There should not be any chance for applying discretion; rather it should be defined by an algebraic formula.

2. It should be based on all the observations

To ensure that it should represent the entire data set, its value should be calculated by taking into consideration the entire data set.

3. It should be easy to understand and calculate

The value of an average should be computed by using a simple method without reducing its accuracy and other advantages.

4. It should have sampling stability

The value of an average calculated from various independent random samples of the same size from a given population should not vary much from another. The least amount of difference (if any) in the values is considered to be the sampling error.

5. It should be capable of further algebraic treatment

The nature of the average should be such that it could be used for statistical analysis of the data set. For example, it should be possible to determine the average production in a particular year by the use of average production in each month of that year.

6. It should not be unduly affected by extreme observations

The value of an average should not be unduly affected by very small or very large

observations in the given data. Otherwise the average value may not truly represent characteristics of the entire set of data.

2.2 MEAN (OR) ARITHMETIC MEAN**Q6. Define Arithmetic Mean. Explain the merits and demerits of Arithmetic Mean.**

Ans : (Nov.-20, May-18)

Arithmetic Average or Mean of a series is the figure obtained by dividing the total values of the various items by their number. In other words it is the sum of the values divided by their number. Arithmetic means is the most widely used measure of central tendency.

Merits of Arithmetic Average

1. It is rigidly defined. Hence, different interpretations by different persons are not possible.
2. It is easy to understand and easy to calculate. In most of the series it is determinate and its value is definite.
3. It takes all values into consideration. Thus, it is more representative.
4. It can be subjected to further mathematical treatment. The properties of Arithmetic mean are separately explained elsewhere in the chapter.
5. It is used in the computation of various other statistical measures.
6. It is possible to calculate arithmetic average even if some of the details of the data are lacking.

For example, it can be known even when only the number of items and their aggregate value are known, and details of various items are not available. Similarly, the aggregate value of items can be calculated if the number of items and the average are known.

7. Of all averages, arithmetic average is least affected by fluctuations of sampling. Thus, it is the most stable measure of central tendency.

8. It provides a good basis for comparison.
9. Arithmetic mean is impacted by every observation. It gives weight to all items in direct proportion to their size.

Demerits of Arithmetic Average

While Arithmetic mean satisfies most of the conditions of an ideal average, it suffers from certain drawbacks. Some of the demerits or limitations of Arithmetic Mean are listed below:

1. It cannot be determined by inspection.
2. It cannot be located graphically.
3. It cannot be used in the study of qualitative phenomena.
4. It can be significantly impacted by extreme values and may lead to erroneous conclusions. Abnormal items may considerably affect this average, particularly when the number of items is not large. Thus, it is desirable not to use arithmetic average when the distribution is unevenly spread.
5. It cannot be calculated if the distribution has open-ended classes i.e., "below 10" or "above 80".
6. It cannot be calculated even if a single observation is missing or lost or is illegible.
7. Arithmetic average can be a figure that does not exist in the series at all. Hence, it may be considered as a fictitious average.
8. It is not suitable in case of an extremely asymmetrical (skewed) Distribution. In fact, it is suitable only if the distribution is normally distributed.

Q7. Explain the uses of arithmetic mean.

Ans :

Uses of Arithmetic Mean

It is generally believed that the arithmetic mean is the best form of average for all general purposes, and should always be used unless there is a very definite reason for the choice of another form of average. However, arithmetic mean is recommended in the following situations:

1. When the frequency distribution is symmetrical.
2. When we need a stable average.
3. When other measure such as standard deviation coefficient of correlation are to be computed later.

Q8. How to calculate arithmetic mean for individual series ?

Ans :

Individual Series

The process of computing mean in case of individual observations (i.e., where frequencies are not given) is very simple. Add together the various values of the variable and divide the total by the number of items. Symbolically :

$$\bar{X}^* = \frac{X_1 + X_2 + X_3 + \dots + X_n}{N} \text{ or}$$

$$\bar{X} = \frac{\sum X}{N}$$

Here $\bar{X}$ = Arithmetic Mean, $\sum X$ = Sum of all the values of the variable X, i.e., $X_1, X_2, X_3, \dots, X_n$;
 N = Number of observations.

Steps

The formula involves two steps in calculating mean :

- i) Add together all the values of the variable X and obtain the total, i.e., $\sum X$.
- ii) Divide this total by the number of the observations, i.e., N .

Example

The following table gives the monthly income of 10 employees in an office :

Income (Rs.) 4,780 5,760 6,690 7,750 4,840 4,920 6,100 7,810 7,050
 6,950 Calculate the arithmetic mean of incomes.

Sol. :

Let income be denoted by the symbol X.

Calculation of Arithmetic Mean

Employee	Monthly Income (Rs.) X
1	4,780
2	5,760
3	6,690
4	7,750
5	4,840
6	4,920
7	6,100
8	7,810
9	7,050
10	6,950
N = 10	$\sum X = 62,650$

$$\bar{X} = \frac{\sum X}{N};$$

$$\sum X = 62,650, N = 10$$

$$\bar{X} = \frac{62,650}{10} = 6,265.$$

Hence the average income is Rs. 6,265.

Shortcut method

The arithmetic mean can be calculated by using what is known as an arbitrary origin. When deviations are taken from an arbitrary origin, the formula for calculating arithmetic mean is

$$\bar{X} = A + \frac{\sum d}{N}$$

where A is the assumed mean and d is the deviation of items from assumed mean, i.e., $d = (X - A^*)$.

Steps

- 1) Take an assumed mean.
- 2) Take the deviations of items from the assumed mean and denote these deviations by d .
- 3) Obtain the sum of these deviations,
- 4) Apply the formula : $X = A + \frac{\sum d}{N}$

Example

From Illustration 1 calculate arithmetic mean by taking 7,000 as the assumed mean.

Sol. :

Calculation of Arithmetic Mean

Employee	Income (Rs.) X	$(X-7000)$
1	4,780	-2220
3	6,690	-310
4	7,750	+750
5	4,840	-2160
6	4,920	-2080
7	6,100	-900
8	7,810	+810
9	7,050	+50
10	6,950	-50
$N = 10$		$\sum d = -7350$

$$\bar{X} = A + \frac{\sum d}{N}$$

$$A = 7000, \sum d = -7350, N = 10$$

$$\bar{X} = 7000 - \frac{7,350}{10} = 7,000 - 735 = 6,265.$$

Hence the average income is Rs. 6,265.

Q9. How to calculate arithmetic mean for discrete series ?*Ans :***Calculation of Arithmetic Mean**

In discrete series arithmetic mean may be computed by applying

- i) Direct method, ii) Short-cut method

i) Direct method

The formula for computing mean is $\bar{X} = \frac{\sum fX}{N}$

where, f = Frequency; X = The variable in question; N* = Total number of observations, i.e., $\sum f$.

Steps :

- i) Multiply the frequency of each row with the variable and obtain the total If $\sum X$.
 ii) Divide the total obtained by step (i) by the number of observations, i.e., total frequency.

Example :

From the following data of the marks obtained by 60 students of a class, calculate the arithmetic mean :

Marks	No. of Students
20	8
30	12
40	20
50	10
60	6
70	4

Sol. :

Let the marks be denoted by X and the number of students by f.

Calculation of Arithmetic Mean

Marks	No. of students	
X	f	fX
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280
	60	2,460
	N = 60	$\sum fX = 2,460$

$$\bar{X} = \frac{\sum fX}{N} = \frac{2,460}{60} = 41$$

Hence the average marks = 41.

ii) Short-cut Method. According to this method,

$$\bar{X} = A + \frac{\sum fd}{N}$$

where A = Assumed mean; d = (X – A); N - Total number of observations, i.e., $\sum f$.

Steps :

- i) Take an assumed mean.
- ii) Take the deviations of the variable X from the assumed mean and denote the deviations by *d*.
- iii) Multiply these deviations with the respective frequency and take the total $\sum fd$.
- iv) Divide the total obtained in third step by the total frequency.

Example :

Calculate arithmetic mean by the short-cut method using frequency distribution of 2.

Sol. :

Calculation of Arithmetic Mean

Marks X	No. of students f	(X – 40) d	fd
20	8	–20	–160
30	12	–10	–120
40 – A	20	0	0
50	10	+10	+100
60	6	+20	+120
70	4	+30	+120
	60		60

$$\bar{X} = A + \frac{\sum fd}{N} = 40 + \frac{60}{60} = 40 + 1 = 41$$

Q10. How to calculate arithmetic mean for continuous series ?

Ans :

Calculation of Arithmetic Mean

In continuous series, arithmetic mean may be computed by applying any of the following methods:

- i) Direct method,
- ii) Short-cut method.

i) **Direct Method.** When direct method is used

$$\bar{X} = \frac{\sum fm}{N}$$

where

m = mid-point of various classes

f = the frequency of each class

N = the total frequency.

Steps :

- i) Obtain the mid-point of each class and denote it by m.
- ii) Multiply these mid-points by the respective frequency of each class and obtain the total $\sum fm$.
- iii) Divide the total obtained in step (i) by the sum of the frequency, i.e., N.

Example :

From the following data compute arithmetic mean by direct method :

Marks:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
No. of students :	5	10	25	30	20	10

Sol. :

Calculation of Arithmetic Mean by Direct Method

Marks	Mid-point (m)	No. of Students (f)	fm
0—10	5	5	25
10—20	15	10	150
20—30	25	25	625
30—40	35	30	1,050
40—50	45	20	900
50—60	55	10	550
		N = 100	$\sum fm = 3,300$

$$\bar{X} = \frac{\sum fm}{N} = \frac{3,300}{100} = 33.$$

ii) **Short-cut Method.** When short-cut method is used, arithmetic; computed by applying the following formula :

$$\bar{X} = A + \frac{\sum fd}{N}$$

where A - assumed mean; d = deviations of mid-points from assumed mean, i.e., (m - A); N = total number of observations.

Steps :

- i) Take an assumed mean.
- ii) From the mid-point of each class deduct the assumed mean.
- iii) Multiply the respective frequencies of each class by these deviations and obtain the total Σfd .

iv) Apply the formula: $\bar{X} = A + \frac{\Sigma fd}{N}$

Calculate arithmetic mean by the short-cut method from the data of Illustration 4.

Calculation of Arithmetic Mean

Marks	Mid-points (m)	No. of Students (f)	(m - 35) d	fd
0 - 10	5	5	-30	-150
10 - 20	15	10	-20	-200
20 - 30	25	25	-10	-250
30 - 40	35 - A	30	0	0
40 - 50	45	20	+10	+200
50 - 60	55	10	+20	+200
		N = 100	$\Sigma fd = -200$	

$$\bar{X} = A + \frac{\Sigma fd}{N}$$

$$\begin{aligned}\bar{X} &= 35 + \frac{-200}{100} \\ &= 35 + (-2) = 33\end{aligned}$$

PROBLEMS ON ARITHMETIC MEAN

1. Calculate the arithmetic mean of the monthly incomes of the families in a certain locality in Delhi.

Family	A	B	C	D	E	F	G	H	I	J
Income (f)	85	70	10	75	500	8	42	250	40	36

Sol :

$$\text{Arithmetic Mean } (\bar{X}) = \frac{\Sigma X}{N}$$

$$= \frac{85 + 70 + 10 + 75 + 500 + 8 + 42 + 250 + 40 + 36}{10}$$

$$= 1116/10 = 111.6$$

$\therefore$ The arithmetic mean of the monthly incomes in the locality is Rs. 111.6.

2. Calculate arithmetic mean of the following data by direct, short-cut methods :

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks (Out of 100)	50	65	37	29	92	43	81	36	52	45

Sol :

Calculation of Arithmetic mean by

i) Direct Method

$$\text{Arithmetic mean } (\bar{X}) = \frac{\sum X}{N}$$

$$= \frac{50 + 65 + 37 + 29 + 92 + 43 + 81 + 36 + 52 + 45}{10} = \frac{530}{10} = 53$$

 $\therefore$ Average marks of the class = 53 marks.**Short Cut Method :**

X	X - A = d
50	- 42
65	- 27
37	- 55
29	- 63
92 A	0
43	- 49
81	- 11
36	- 56
52	- 40
45	- 47
	- 390

$$\begin{aligned}\bar{X} &= A + \frac{\sum d}{N} = 92 + \frac{-390}{10} \\ &= 92 + (-39) \\ &= 53\end{aligned}$$

3. "The following data related to the distance travelled by 520 villagers to buy their weekly requirements :

Miles Travelled	2	4	6	8	10	12	14	16	18	20
No. of Villagers	38	104	140	78	48	42	28	24	16	2

Sol :

$$\text{Arithmetic Mean} = \frac{\sum fx}{N}$$

Miles Travelled (X)	No. of villagers (F)	fx
2	38	76
4	104	416
6	140	840
8	78	624
10	48	480
12	42	504
14	28	392
16	24	384
18	16	288
20	2	40
N = 520		$\sum fx = 4044$

$$(\bar{X}) = \frac{\sum fx}{N} = \frac{4044}{520} = 7.7769$$

4. Find mean from the following frequency data :

Class-interval	Frequency
15 - 25	4
25 - 35	11
35 - 45	19
45 - 55	14
55 - 65	0
65 - 75	2

Sol :

Class-interval	Frequency (f)	Mid value (m)	fm
15 - 25	4	20	80
25 - 35	11	30	330
35 - 45	19	40	760
45 - 55	14	50	700
55 - 65	0	60	0
65 - 75	2	70	140
	59		2010

$$\text{Arithmetic mean } (\bar{X}) = \frac{\sum fm}{N}$$

$$= \frac{2010}{50}$$

$$(\bar{X}) = 40.2$$

5. Calculate the arithmetic average from the following data.

Value	Frequency
Less than 10	4
Less than 20	16
Less than 30	40
Less than 40	76
Less than 50	96
Less than 60	112
Less than 70	120
Less than 80	124

Sol :

Value	Mid value	Frequency	fm
0 - 10	5	4	20
10 - 20	15	12	180
20 - 30	25	24	600
30 - 40	35	36	1260
40 - 50	45	20	900
50 - 60	55	16	880
60 - 70	65	8	520
70 - 80	75	4	300
		N = 124	Σfm = 4,660

$$\text{Arithmetic mean} = \frac{\sum fm}{N} = \frac{4,660}{124}$$

$$\bar{X} = 37.58$$

6. Find the average wage of a worker from the following table :

Wages in ` above	0	10	20	30	40	50	60	70
No. of workers	670	500	420	370	300	275	250	100

Sol :

$$\text{Arithmetic mean} = \frac{\sum fm}{N}$$

Wages (Rs.)	Frequency	Simple Frequency (f)	m	fm
0 - 10	670	670 - 500 = 170	5	850
10 - 20	500	500 - 420 = 80	15	1200
20 - 30	420	420 - 370 = 50	25	1250
30 - 40	370	370 - 300 = 70	35	2450
40 - 50	300	300 - 275 = 25	45	1125
50 - 60	275	275 - 250 = 25	55	1375
60 - 70	250	250 - 100 = 150	65	9750
70 - 80	100	100 - 0 = 100	75	7500
		N = 670		25,500

$$\text{Arithmetic mean} = \frac{\sum fm}{N} = \frac{25,500}{670}$$

$$\bar{X} = 38.06$$

7. When missing frequency is to be calculated.

Find the missing frequency from the following data:

Marks	Frequency
5-10	12
10-15	16
15-20	?
20-25	14
25-30	10
30-35	8

The average mark is 16.82

Sol :

Let the frequency of 15-20 class denoted as f4.

Marks Obtained	Mid-Values	No. of Students	Deviation from assumed mean (A = 17.5)	Step deviation (d/5)	Freq x d ¹
X	m	f	d	d ¹	fd ¹
0-5	2.5	10	-15	-3	-30
5-10	7.5	12	-10	-2	-24
10-15	12.5	16	-5	-1	-16
15-20	17.5	f ₄	0	-1	0
20-25	22.5	14	+5	+1	+14
25-30	27.5	10	+10	+2	+20
30-35	32.5	8	+15	+3	+24
		70 + f ₄			Σfd ¹ = -12

$$\bar{X} = A + \left(\frac{\Sigma fd'}{N} \times C \right) = 17.54 + \frac{-12}{70 + f_4} \times 5 = 16.82, \frac{-60}{70 + f_4} = 16.82 - 17.5$$

$$\frac{-60}{70 + f_4} = -0.68, 0.68, (70 + f_4) = 60, 47.6 + 0.68 f_4 = 60, 0.68 f_4 = 12.4 f_4$$

$$= \frac{12.4}{0.68} = 18 \text{ approx.}$$

Hence, missing frequency is 18.

2.2.1 Weighted Mean

Q11. Define Weighted Arithmetic Mean.

Ans :

In the computation of simple arithmetic mean, equal importance is given to all the items. But "this may not be so in all situations. If all the items are not of equal importance, then simple arithmetic mean will not be a good representative of the given data. Hence, weighing of different items becomes necessary. The weights are assigned to different items depending upon their importance, i.e., more important items are assigned more weight. For example, to calculate the mean wage of workers of a factory, it would be wrong to compute simple arithmetic mean if there are few workers (say managers) with very high wages while majority of the workers are at low level of wages. The simple arithmetic mean, in such a situation, will give a higher value that cannot be regarded as representative wage for the group. In order that the mean wage gives a realistic picture of the distribution, the wages of managers should be given less importance in its computation. The mean calculated in this manner is called weighted arithmetic mean. The computation of weighted arithmetic is useful in many situations where different items are of unequal importance, e.g., the construction index numbers, computation of standardized death and birth rates, etc.

Formulae for weighted Arithmetic Mean

Let $X_1, X_2, \dots, X_n$ be n values with their respective weights $\omega_1, \omega_2, \dots, \omega_n$. Their weighted arithmetic mean denoted as $\bar{X}_w$ is given by,

$$(i) \quad \bar{X}_w = \frac{\sum \omega_i X_i}{\sum \omega_i} \text{ (Using direct method),}$$

$$(ii) \quad \bar{X}_w = A + \frac{\sum \omega_i d_i}{\sum \omega_i} \text{ (where } d_i = X_i - A \text{) (Using short-cut method),}$$

$$(iii) \quad \bar{X}_w = A + \frac{\sum \omega_i u_i}{\sum \omega_i} \times h \text{ (where } u_i = \frac{X_i - A}{h} \text{) (Using step-deviation method)}$$

PROBLEMS

8. A train runs 25 miles at a speed of 30 m.p.h., another 50 miles at a speed of 40 m.p.h., then due to repairs of the track travels for 6 minutes at a speed of 10 m.p.h. and finally covers the remaining distance of 24 miles at a speed of 24 m.p.h. What is the average speed in miles per hour?

Sol :

Time taken in covering 25 miles at a speed of 30 m.p.h. = 50 minutes.

Time taken in covering 50 miles at a speed of 40 m.p.h. = 75 minutes.

Distance covered in 6 minutes at a speed of 10 m.p.h. = 1 mile.

Time taken in covering 24 miles at a speed of 24 m.p.h. = 60 minutes.

Therefore, taking the time taken as weights we have the weighted mean as

Speed in m.p.h.	Time taken	
X	W	WX
30	50	1,500
40	75	3,000
10	6	60
24	60	1,440
		$\Sigma WX = 6,000$

$$\therefore \text{Average speed} = \frac{6,000}{191} = 31.41 \text{ m.p.h.}$$

9. From the following results of two colleges A and B, find out which of the two is better:

Examination	College A		College B	
	Appeared	Passed	Appeared	Passed
M.Sc.	60	40	200	160
M.A.	100	60	240	200
B.Sc.	200	150	200	140
B.A.	120	75	160	100

Sol :

Performance of the two colleges can be compared by taking weighted arithmetic mean of the pass percentage in various classes. The calculation of weighted arithmetic mean is shown in the following table.

Class	College A				College B			
	Appeared w_A	Passed	Pass Percentage X_A	$w_A X_A$	Appeared w_B	Passed	Pass Percentage X_B	$w_B X_B$
M.Sc.	60	40	66.67	4000.2	200	160	80.00	16000.0
M.A.	100	60	60.00	6000.0	240	200	83.33	19999.2
B.Sc.	200	150	75.00	15000.0	200	140	70.00	14000.0
B.A.	120	75	62.50	7500.0	160	100	62.50	10000.0
Total	480	325		32500.2	800	600		59999.2

$$\bar{X}_w \text{ for College A} = \frac{\sum w_A X_A}{\sum w_A} = \frac{32500.2}{480} = 67.71\%$$

$$\bar{X}_w \text{ for College B} = \frac{\sum w_B X_B}{\sum w_B} = \frac{59999.2}{800} = 75\%$$

2.2.2 Combined Mean

Q12. Discuss about combined mean.

Ans :

It is the arithmetic averages and the number of items in two or more related groups are known as the combined or the composite mean of the entire group can be obtained by,

$$\text{Combined Mean } (\bar{X}) = \left[\frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2} \right]$$

With the help of combined arithmetic mean, we can determine the over all mean of the combined data without going back of the original data.

PROBLEMS

10. The mean wage of 60 male workers in a factory is Rs. 40 and the mean wage of 40 female workers in the same factory is Rs. 35. Find the combined mean wage of 100 workers in the factory.

Sol :

$$\bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

$$N_1 = 60, \bar{X}_1 = 40, \quad N_2 = 40, \quad \bar{X}_2 = 35$$

$$\bar{X}_{12} = \frac{60 \times 40 + 40 \times 35}{60 + 40} = \frac{2400 + 1400}{100}$$

$$\begin{aligned} \bar{X}_{12} &= \frac{3800}{100} \\ &= 38 \end{aligned}$$

11. The mean height of 25 male workers in a factory is 61 cms, and the mean height of 35 female workers in the same factory is 58 cms. Find the combined mean height of 60 workers in the factory.

Sol :

$$\bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

$$N_1 = 25, \bar{X}_1 = 61, \quad N_2 = 35, \quad \bar{X}_2 = 58$$

$$\bar{X}_{12} = \frac{25 \times 61 + 35 \times 58}{25 + 35} = \frac{1525 + 2030}{60}$$

$$\bar{X}_{12} = \frac{25 \times 61 + 35 \times 58}{25 + 35} = \frac{1525 + 2030}{60}$$

$$= \frac{3555}{60} = 59.25$$

12. There are three sections in B.Com. First year in a certain college. The number of students in each section and the average marks obtained by them in Statistics paper are as follows. Find the average marks obtained by the students of three sections taken together.

Section	Average Marks	No. of Students
A	60	60
B	55	50
C	75	55

Sol :

Combined Mean

$$\bar{X}_{123} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3}{N_1 + N_2 + N_3}$$

$$N_1 = 60, N_2 = 50, N_3 = 55,$$

$$\bar{X}_1 = 60, \bar{X}_2 = 55, \bar{X}_3 = 75$$

$$\bar{X}_{123} = \frac{(60 \times 60) + (55 \times 50) + (75 \times 55)}{60 + 50 + 55} = \frac{3600 + 2750 + 4125}{165} = 63.48$$

- 13. The mean age of a combined group of men and women is 30 years. If the mean age of group of men is 32 and that of the group of women is 27, find out the percentage of men and women in the group.**

Sol :

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$\bar{X}_{12} = 30, \bar{X}_1 = 32, \bar{X}_2 = 27 \quad N_1 \text{ and } N_2 \text{ are not given}$$

Let the percentage of Men be N_1 . Percentage of women will be $100 - N_1$. So that $N_1 + N_2 = 100$.

$$= \frac{(N_1 \times 32) + (100 - N_1)27}{100} = 30$$

$$= 32 N_1 + 2700 - 27 N_1 = 3000$$

$$= 5 N_1 = 300$$

$$N_1 = 60, 100 - N_1 = N_2 \quad 100 - 60 = 40$$

Hence the percentage of men and women in the group is 60 and 40 respectively.

2.2.3 Geometric Mean

Q13. Define Geometric Mean. Explain how GM is calculated for individual series.

Ans :

Geometric Mean

Geometric mean is defined as the N^{th} root of the product of N items or values.

When the number of observations is more than 2 the calculation of n^{th} root becomes difficult. In order to facilitate these calculations the logarithms are used.

In case of 2 or 3 observations the Geometric mean is calculated as follows :

$$\begin{aligned} \text{G.M} &= \sqrt[n]{x_1 \times x_2 \times x_3 \dots x_n} \\ &= (x_1 \times x_2 \times \dots x_n)^{1/n} \end{aligned}$$

Whereas, when the number of observations are more the geometric mean is calculated as follows,

$$\log G.M = \frac{\log x_1 + \log x_2 + \dots + \log x_n}{N}$$

(or)

$$\log G.M = \left(\frac{\sum \log x}{N} \right)$$

$$\therefore G.M = \text{Antilog} \left(\frac{\sum \log x}{N} \right)$$

Q14. Explain how GM is calculated for discrete and continuous series.

Ans :

GM Calculation for Individual Observations :

The steps involved in calculating geometric mean for individual observation are :

1. Logarithms of variable 'x' are taken i.e., $\log x$.
2. Calculate the sum of $\log x$ i.e., $\sum \log x$.
3. Divide $\sum \log x$ by the number of observations [i.e., N].
4. Take the antilog of the value obtained in step 3.

$$G.M = \text{Antilog} \left[\frac{\sum \log x}{N} \right]$$

GM Calculation for Discrete Series :

The steps involved in calculating geometric mean for discrete series are,

1. Take the logarithms of the variable 'x'.
2. Multiply each $\log x$ value with their corresponding frequencies.
3. Calculate the sum of it i.e., $\sum f \log x$.
4. Divide $\sum f \log x$ by the sum of frequency.
5. Take the antilog for the value obtained in step 4.

$$G.M = A.L \left[\frac{\sum f \log x}{N} \right]$$

GM Calculation for Continuous Series :

The steps involved in calculating geometric mean for continuous series are,

1. Take the mid-points of the classes and denote them as 'm'.
2. Find out the logarithms for the values of midpoint and denote them as $\log x$.

3. Multiply $\log x$ with their corresponding frequencies for each class and obtain the sum of it [i.e., $\Sigma f \log x$].
4. Divide $\Sigma f \log x$ by the sum of frequency.
5. Find out the antilog for the value obtained in step 4.

$$G.M = \text{Antilog} \left[\frac{\Sigma f \log m}{N} \right]$$

PROBLEMS

- 14. Find the Geometric Mean from the following data**

2200, 220, 22, 0.22, 0.022

Sol :

X	2200	220	22	0.22	0.022	Total
$\log x$	3.2424	2.3424	1.3424	-0.6576	-1.6576	$\Sigma \log x = 4.712$

We have $n = 5$, $\Sigma \log x = 4.712$

$$\begin{aligned}
 \therefore \text{Geometric Mean (GM)} &= \text{Antilog} \left[\frac{\Sigma \log x}{n} \right] \\
 &= \text{Antilog} \left[\frac{4.712}{5} \right] \\
 &= \text{Antilog} [0.9424] \\
 &= 8.7579
 \end{aligned}$$

- 15. Calculate geometric mean of the following data :**

1, 7, 29, 92, 115 and 375

Sol :

Calculation of Geometric Mean

X	1	7	29	92	115	375	$\Sigma \log X$
$\log X$	0.0000	0.8451	1.4624	1.9638	2.0607	2.5740	8.9060

$$GM = \text{Antilog} \left[\frac{\Sigma \log X}{n} \right] = \text{Antilog} \left[\frac{8.9060}{6} \right] = 30.50$$

- 16. Find out geometric mean from the following data,**

X	10	20	30	40	50	60
f	12	15	25	10	6	2

Sol :

Calculation of Geometric Mean

x	f	logx	f logx
10	12	1	12
20	15	1.30	19.5
30	25	1.48	37
40	10	1.60	16
50	6	1.70	10.2
60	2	1.78	3.56
	N=70		Σf logx=98.26

$$\therefore \text{Geometric Mean} = \text{Antilog} \left(\frac{\Sigma f \log x}{N} \right)$$

$$= \text{Antilog} \left(\frac{98.26}{70} \right)$$

$$= \text{Antilog} (1.404)$$

$$\text{G.M} = 25.35$$

17. Calculate geometric mean of the following distribution :

X :	5	10	15	20	25	30
f :	13	18	50	40	10	6

Sol :

Calculation of GM

X	f	log X	f log X
5	13	0.6990	9.0870
10	18	1.0000	18.0000
15	50	1.1761	58.8050
20	40	1.3010	52.0400
25	10	1.3979	13.9790
30	6	1.4771	8.8626
Total	137		160.7736

$$\therefore \text{GM} = \text{Antilog} \left[\frac{160.7736}{137} \right] = \text{Antilog } 1.1735 = 14.91$$

18. Calculate Geometric Mean (GM) for the following data,

Marks	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
No. of Students	8	12	20	10	7	3

Sol :

Calculation of Mid-Values and Table Values

Marks	Mid Points (m)	No. of Students (f)	Log m	f log m
10 - 20	15	8	1.1761	9.4088
20 - 30	25	12	1.3979	16.7748
30 - 40	35	20	1.5441	30.8820
40 - 50	45	10	1.6532	16.5320
50 - 60	55	7	1.7404	12.1828
60 - 70	65	3	1.8129	5.4387

$$\text{Geometric mean} = \text{Anti log} \left[\frac{\sum f \log m}{N} \right]$$

$$= \text{Anti log} \left[\frac{91.2191}{60} \right]$$

$$= \text{Anti log} (1.5203)$$

$$\text{G.M} = 33.13599.$$

19. Calculate geometric mean of the following distribution :

Class Intervals :	5 - 15	15 - 25	25 - 35	35 - 45	45 - 55
Frequencies :	10	22	25	20	8

Sol :

Calculation of GM

Class	f	Mid Value(X)	log(X)	flogx
5 - 15	10	10	1.0000	10.0000
15 - 25	22	20	1.3010	28.6227
25 - 35	25	30	1.4771	36.9280
35 - 45	20	40	1.6020	32.0412
45 - 55	8	50	1.6990	13.5918
Total	85			121.1837

$$\text{GM} = \text{Antilog} \frac{121.1837}{85} = \text{Antilog } 1.4257 = 26.65$$

Q15. State the merits and demerits of geometric mean.*Ans :***Merits :**

1. It is a rigidly defined average.
2. It is based on all the observations.
3. It is capable of mathematical treatment. If any two out of the three values, i.e.,
 - (i) Product of observations,
 - (ii) GM of observations and
 - (iii) Number of observations, are known, the third can be calculated.
4. In contrast to AM, it is less affected by extreme observations.
5. It gives more weights to smaller observations and vice-versa.

Demerits

1. It is not very easy to calculate and hence not very popular.
2. Like AM, it may be a value which does not exist in the set of given observations.
3. It cannot be calculated if any observation is zero or negative.

2.2.4 Harmonic Mean**Q16. Define Harmonic Mean. Explain how HM is calculated for individual series.***Ans :***Harmonic Mean (H.M)**

“The reciprocal of the arithmetic mean of the reciprocal of individual observations” is known as Harmonic mean. It is represented as follows,

$$H.M = \frac{N}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} \dots \frac{1}{x_n} \right)}$$

When the number of observations are large then Harmonic mean is stated as follows,

$$H.M = \frac{N}{\Sigma(1/x)}$$

Simple H.M / Individual Observations : Steps

The steps involved in calculating harmonic mean for individual observation are

1. Take the reciprocal of each value of variable x [i.e., $\frac{1}{x}$]
2. Find out the sum of $\frac{1}{x}$ [i.e., $\Sigma \left[\frac{1}{x} \right]$]

3. Divide the number of observations [N] by $\Sigma \frac{1}{x}$.

$$H.M = \frac{N}{\Sigma \left[\frac{1}{x} \right]}$$

Q17. Explain how HM is calculated for discrete and continuous series.

Ans :

(a) Simple H.M / Discrete Series:

The steps involved in calculating harmonic mean for discrete series are as follows,

1. Find the reciprocal of each value of variable x .
2. Multiply reciprocal value of x with their corresponding frequencies and obtain the sum as $\Sigma \left[f \times \frac{1}{x} \right]$
3. Divide the sum of frequency [N] by $\Sigma \left[f \times \frac{1}{x} \right]$

The formula to calculate Harmonic mean for discrete series is as follows,

$$H.M = \frac{N}{\Sigma \left[f \times \frac{1}{x} \right]} = \frac{N}{\left(\frac{f}{x} \right)}$$

(b) Simple H.M / Continuous Series: Steps

The steps involved in calculating harmonic mean for continuous series are as follows :

1. Take the mid-points of various class intervals and denote them as m .
2. Take the reciprocal of each value of the variable " x ".
3. Multiply reciprocal of $\left(\text{i.e., } \frac{1}{m} \right)$ with their corresponding frequencies and obtain them sum $\Sigma \left[f \times \frac{1}{m} \right]$.
4. Divide the sum of frequency [N] by $\Sigma \left[f \times \frac{1}{m} \right]$

The formula to calculate Harmonic mean for continuous series is as follows,

$$H.M = \frac{N}{\Sigma \left[f \times \frac{1}{m} \right]} = \frac{N}{\Sigma \left(\frac{f}{x} \right)}$$

PROBLEMS

20. Calculate harmonic mean of the following data :

X :	10	11	12	13	14
f :	5	8	10	9	6

Sol :

Calculation of Harmonic Mean

X	10	11	12	13	14	Total
Frequency(f)	5	8	10	9	6	38
$f \times \frac{1}{X}$	0.5000	0.7273	0.8333	0.6923	0.4286	3.1815

$$\therefore HM = \frac{38}{3.1815} = 11.94$$

21. Find the harmonic mean of the following distribution :

Class Intervals	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	11	21	35	30	22	18

Sol :

Calculation of Harmonic Mean

Class Intervals	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	Total
Frequency (f)	5	8	11	21	35	30	22	18	150
Mid-values (X)	5	15	25	35	45	55	65	75	
$\frac{f}{X}$	1.0000	0.5333	0.4400	0.6000	0.7778	0.5455	0.3385	0.2400	4.4751

$$\therefore HM = \frac{150}{4.4751} = 33.52$$

Q18. State the merits and demerits of Harmonic Mean.

Ans :

Merits

1. It is a rigidly defined average.
2. It is based on all the observations.
3. It gives less weight to large items and vice-versa.
4. It is capable of further mathematical treatment.
5. It is suitable in computing average rate under certain conditions.

Demerits

1. It is not easy to compute and is difficult to understand.
2. It may not be an actual item of the given observations.
3. It cannot be calculated if one or more observations are equal to zero.
4. It may not be representative of the data if small observations are given correspondingly small weights.

2.3 MEDIAN**Q19. Define median. What are the characteristics of median?***Ans :*

If a group of N observations is arranged in ascending or descending order of magnitude, then the middle value is called median of these observations and is denoted by M or Me. That is,

$$M = \frac{N+1}{2} \text{th observation.}$$

Definition

According to Croxton and Cowden, "The median is that value which divides a series so that one half or more of the items are equal to or less than it and one half or more of the items are equal to or greater than it."

Characteristics of Median

1. Unlike the arithmetic mean, the median can be computed from open-ended distributions. This is because it is located in the median class-interval, which would not be an open-ended class.
2. The median can also be determined graphically whereas the arithmetic mean cannot be ascertained in this manner.
3. As it is not influenced by the extreme values, it is preferred in case of a distribution having extreme values.
4. In case of the qualitative data where the items are not counted or measured but are scored or ranked, it is the most appropriate measure of location.

Q20. What are the advantages and disadvantages of median?*Ans :***Advantages of Median**

- (i) The median, unlike the mean, is unaffected by the extreme values of the variable.
- (ii) It is easy to calculate and simple to understand, particularly in a series of individual observations and a discrete series.
- (iii) It is capable of further algebraic treatment. It is used in calculating mean deviation.
- (iv) It can be located by inspection, after arranging the data in order of magnitude.
- (v) It can be determined graphically.
- (vi) Median can be calculated in case of open-end classes.
- (vii) Median is defined rigidly.

Disadvantages of Median

- (i) For calculation, it is necessary to arrange the data, other averages do not need any such arrangement.
- (ii) It is amenable to algebraic treatment in a limited sense. Median cannot be used to calculate the combined median of two or more groups, like mean.
- (iii) Median is affected more by sampling fluctuations than the mean.
- (iv) It is not based on all observations. So it is a positional average.

Q21. How median is calculated for individual series ?*Ans :***Steps :**

- i) Arrange the data in ascending or descending order of magnitude. (Both arrangements would give the same answer.)

- ii) In a group composed of an **odd** number of values such as 7, add 1 to the total number of values and divide by 2. Thus, 7 + 1 would be 8 which divided by 2 gives 4—the number of the value starting at either end of the numerically arranged groups will be the median value.

In the form of formula :

$$\text{med.} = \text{Size of } \frac{N+1}{2}^{\text{th}} \text{ item}$$

- iii) In a group composed an **even** number of values, then median = $\frac{N}{2} + 1^{\text{th}}$ item

$$(\text{or}) \frac{\text{Two middle values}}{2}$$

Example

From the following data of the wages of 7 workers, compute the median wage :

Wages (in Rs.) 4100 4150 6080 7120 5200 6160 7400

Sol. :

Sl.No.	Calculation of Median Wages arranged in ascending order
1	4100
2	4150
3	5200
4	6080
5	6160
6	7120
7	7400

$$\text{Median} = \text{Size of } \frac{N+1}{2}^{\text{th}} \text{ item}$$

$$= \frac{7+1}{2}$$

$$= 4^{\text{th}} \text{ item}$$

$$= \text{Rs. 6080.}$$

Q22. How median is calculated for discrete series ?

Ans :

Steps :

- Arrange the data in ascending or descending order of magnitude,
- Find out the cumulative frequencies.

- iii) Apply the formula: Median = Size of $\left(\frac{N+1}{2}\right)^{\text{th}}$ item
- iv) Now look at the cumulative frequency column and find that total which is either equal to Size of $\frac{N+1}{2}$ or next higher to that and determine the value of the variable corresponding to it. That gives the value of median.

Example

From the following data, find the value of median :

Income (Rs.)	5,000	5,500	6,800	8,000	8,500	7,800
No. of persons	24	26	20	20	2	30

Calculation of Median

Income arranged in ascending order	No. of person (f)	CF
5,000	24	24
5,500	26	50
6,800	20	70
7,800	30	100
8,000	20	120
8,500	2	122

$$\text{Median} = \text{Size of } \frac{N+1}{2}^{\text{th}} \text{ item} = \frac{122+1}{2} = \text{Size of 61.5th item.}$$

Size of 61.5th item = Rs. 6,800. Hence the median income is Rs. 6,800.

Q23. How median is calculated for continuous series ?

Ans :

1. Arrange the given data in ascending order
2. Calculate cumulative frequency
3. Find N_1 as $N_1 = N/2$ for Median
4. In C.f find the value equal to or just greater than N_1
5. Find f and L just one step below the C.f

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

- L = Lower limit of the median class,
 C.f = Cumulative frequency of the class preceding the median class or sum of the frequencies of all classes lower than the median class.
 f = Simple frequency of the median class
 i = The class interval of the median class.

Example

From the data given below compute median of the average deposit balance of savings for the branch during march 1988.

Average deposit balance (₹)	No. of deposits
Less than 100	26
100-200	68
200-300	145
300-400	242
400-500	188
500-600	65
600-700	16

Sol :

Average deposit balance	No. of deposit f	Cf
0 - 100	26	26
100 - 200	68	94
200 - 300	145	239 Cf
L 300 - 400	242 f	481
400 - 500	188	669
500 - 600	65	734
600 - 700	16	750

- i) Median class interval = Size of $\frac{N}{2}$ th item

$$\text{Median class interval} = \text{Size of } \frac{750}{2} = 375^{\text{th}} \text{ item}$$

∴ Median class interval is 300 - 400

$$\text{Median} = L_1 + \left[\frac{\frac{N}{2} - Cf}{f} \right] \times i$$

$$\text{Median} = 300 + \left[\frac{375 - 239}{242} \right] \times 100$$

$$\text{Median} = 300 + \left[\frac{136}{242} \right] \times 100$$

$$\text{Median} = 300 + 0.5619 \times 100$$

$$\text{Median} = 300 + 56.198$$

$$\text{Median} = 356.198 \text{ (or) } 356.20.$$

PROBLEMS ON MEDIAN

22. Calculate Median from the following.

61, 62, 63, 61, 63, 64, 64, 60, 65, 63, 64, 65, 66, 64

Sol :

i) Procedure

Step 1

Arrange the data in ascending order for the given i.e.,

60, 61, 61, 62, 63, 63, 63, 64, 64, 64, 65, 65, 65, 66

Step 2

Apply the formula for :

$$\begin{aligned} \text{i) Median} &= \text{Size of } \frac{N+1}{2} \text{th item} \\ &= \frac{14+1}{2} = \text{Size of 7.5th item} \end{aligned}$$

when 7.5 is equalled to 7th and 8th items of the data. Hence 7th = 63, 8th item = 64

$$\text{median} = \frac{63 + 64}{2} = 63.5$$

23. Calculate median from the following.

22, 26, 14, 30, 18, 11, 35, 41, 12, 32

Sol :

Step I : Arrange the data in ascending order i.e.,

11, 12, 14, 18, 22, 26, 30, 32, 35, 41

Step II : Calculation of given ascending to formulae

$$\text{i) Median} = \text{Size of } \frac{N+1}{2} \text{th item}$$

$$= \frac{10+1}{2} = \frac{11}{2} = \text{Size of 5.5th item}$$

$$\text{Median} = \frac{\text{5th item} + \text{6th item}}{2} = \frac{22 + 26}{2} = \frac{48}{2} = 24$$

24. From the following data, find the value of median.

Income (₹.)	200	250	130	270	300	230
No. of Persons	34	36	26	30	16	40

Sol :

Procedure :

Step I : Arrange the data in ascending order.

Income (Rs.) (x)	No. of persons (N)	Cumulative frequency
130	26	26
200	34	60
230	40	100
250	36	136
270	30	166
300	16	182
	182	

Step II : Apply the formulae for determination of media.

$$\text{Median} = \text{Size of } \frac{N+1}{2} \text{th item}$$

$$\text{Median} = \text{Size of } \frac{182+1}{2} \text{th item}$$

$$\text{Median} = \text{Size of } \frac{183}{2} = 91.5 \text{ item}$$

Median Size of 91.5 is representing in 100 at cumulative frequency which is representing in corresponding (x) column is 230 in income level.

∴ Median = 230.

25. Calculate Median, from the following frequency distribution.

Size	10	20	30	40	50	60	70	80
Frequency	7	12	17	29	31	21	18	5

Sol :

Step I : Arrange the data in ascending order.

Size) (X)	Frequency	Cumulative Frequency (CF)
10	7	7
20	12	19
30	17	36
40	29	65
50	31	96
60	21	117
70	18	135
80	5	140
	140	

Step II : Apply the formulae according to the following :

i) **Median** = Size of $\frac{N+1}{2}$ th item

Median = Size of $\frac{140+1}{2}$ th item

Median = Size of $\frac{141}{2} = 70.5$ item

Median = Size of 70.5th item is representing 96 in cumulative frequency which representing in corresponding x column 50 in size (z)

$\therefore$ Median = 50

26. For the following data, compute Median

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	15	100	170	120	40

Sol :

Marks	Frequency	Cumulative Frequency (CF)
0 - 10	15	15
10 - 20	100	115 Cf
L 20 - 30	170 f	285
30 - 40	120	405
40 - 50	40	445

Median class interval = Size of $\frac{N}{2}$ th item

$$\text{Median class interval} = \text{Size of} = \frac{445}{2} = 222.50 \text{ item}$$

∴ 222.50th item consisting in Cf 285 which is representing in corresponding class interval is 20-30.

$$\begin{aligned} \text{Median} &= 20 + \left[\frac{\frac{N}{2} - Cf}{f} \right] \times i \\ &= 20 + \left(\frac{222.50 - 115}{170} \right) \times 10 \end{aligned}$$

$$\text{Median} = 20 + 0.632 \times 10$$

$$\text{Median} = 20 + 6.32$$

$$\text{Median} = 26.32$$

27. Following is the distribution of marks obtained by 50 students in Mercantile law

Marks more than	0	10	20	30	40	50
No. of Students	50	46	40	20	10	3

Calculate the Median marks.

Sol :

Marks	No. of students given (Cf)	Simple Frequency	Cumulative frequency (Cf)
0 - 10	50	50 - 46 = 4	4
10 - 20	46	46 - 40 = 6	10 (Cf)
20 - 30	40	40 - 20 = 20 (f)	30
30 - 40	20	20 - 10 = 10	40
40 - 50	10	10 - 3 = 7	47
50 - 60	3	3 - 0 = 3	50
		50	

i) Median = Size of $\frac{N}{2}$ th item

$$\text{Median} = \frac{50}{2} = \text{Size of } 25^{\text{th}} \text{ item}$$

∴ Class interval is 20 - 30

$$\text{Median} = L_1 + \left[\frac{\frac{N}{2} - Cf}{f} \right] \times i$$

$$\text{Median} = 20 + \left[\frac{25 - 10}{20} \right] \times 10$$

$$\text{Median} = 20 + \left[\frac{15}{20} \right] \times 10$$

$$\text{Median} = 20 + 0.75 (10)$$

$$\text{Median} = 20 + 7.5$$

$$\text{Median} = 27.5$$

2.3.1 Partition Values

Q24. Define the term Partition Value.

Ans :

Median of a distribution divides it into two equal parts. It is also possible to divide a distribution into more than two equal parts. The values that divide a distribution into more than two equal parts are commonly known as partition values or fraction.

2.3.1.1 Quartiles

Q25. Discuss briefly about Quartiles. Explain the computation of quartiles for various series.

Ans :

The values of a variable that divide a distribution into four equal parts are called quartiles. Since three values are needed to divide a distribution into four parts, there are three quartiles, viz. Q_1 , Q_2 and Q_3 , known as the first, second and the third quartile respectively.

For a discrete distribution, the first quartile (Q_1) is defined as that value of the variate such that at least 25% of the observations are less than or equal to it and at least 75% of the observations are greater than or equal to it.

For a continuous or grouped frequency distribution, Q_1 is that value of the variate such that the area under the histogram to the left of the ordinate at Q_1 is 25% and the area to its right is 75%. The formula for the computation of Q_1 can

be written by making suitable changes in the formula of median.

Formula

Lower Quartile

In Individual Series and Discrete Series

$$Q_1 = \text{Size of } \frac{N+1}{4} \text{ Item}$$

In Continuous Series

$$Q_1 = \text{Size of } \frac{N}{4} \text{ Item}$$

Upper Quartile

In individual and discrete series

$$Q_3 = \text{Size of } \frac{3(N+1)}{4} \text{ item}$$

In Continuous Series

$$Q_3 = \text{Size of } \frac{3N}{4} \text{ item}$$

2.3.1.2 Deciles

Q26. Discuss briefly about deciles. Explain the computation of deciles for various series.

Ans :

Deciles divide a distribution into 10 equal parts and there are, in all, 9 deciles denoted as D_1 , D_2 , ..., D_9 respectively.

For a discrete distribution, the i th decile D_i is that value of the variate such that at least $(10i)\%$ of the observations are less than or equal to it and at least $(100 - 10i)\%$ of the observations are greater than or equal to it ($i = 1, 2, \dots, 9$).

For a continuous or grouped frequency distribution, D_i is that value of the variate such that the area under the histogram to the left of the ordinate at D_i is $(10i)\%$ and the area to its right is $(100 - 10i)\%$.

Formula

In individual and discrete series

$$D_8 = \text{Size of } \frac{8(N+1)}{10} \text{ item}$$

In Continuous Series

$$D_8 = \text{Size of } \frac{8N^{\text{th}}}{10} \text{ item}$$

2.3.1.3 Percentiles

Q27. Discuss briefly about percentiles. Explain the computation of percentiles for various series.

Ans :

Percentiles divide a distribution into 100 equal parts and there are, in all 99 percentiles denoted as $P_1, P_2, \dots, P_{25}, \dots, P_{40}, \dots, P_{60}, \dots, P_{99}$ respectively.

For a discrete distribution, the k th percentile P_k is that value of the variate such that at least $k\%$ of the observations are less than or equal to it and at least $(100 - k)\%$ of the observations are greater than or equal to it.

For a grouped frequency distribution, P_k is that value of the variate such that the area under the histogram to the left of the ordinate at P_k is $k\%$ and the area to its right is $(100 - k)\%$.

Formula**In Individual and Discrete Series**

$$P_{50} = \text{Size of } \frac{50(N+1)^{\text{th}}}{100} \text{ item}$$

In Continuous Series

$$P_{50} = \text{Size of } \frac{50N^{\text{th}}}{100} \text{ item.}$$

PROBLEM

28. Calculate the lower and upper quartiles, fifth decile and 30th percentile.

Class Group	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25
Frequency	7	18	25	30	20

Sol :

Computation of Q_1 , Q_3 , D_5 and P_{30}

Class Group	f	c.f	
0 - 5	7	7 CF	
L 5 - 10	18 f	25	(Q_1)
10 - 15	25	50 CF	
L 15 - 20	30 f	80	(Q_3)
20 - 25	20	100	
	N = 100		

$$Q_1 = \text{Size of } \frac{N^{\text{th}}}{4} \text{ item}$$

$$= \frac{100}{4} = 25^{\text{th}} \text{ item}$$

Q₁ Lies in the Class 5–10

$$Q_1 = L + \frac{N/4 - CF}{f} \times i$$

Where

L = Lower limit i.e., 5

$$\frac{N}{4} = 25$$

f = Frequency i.e., 18

p.c.f = Preceding cumulative frequency i.e., 7

i = Class interval i.e., 5

$$Q_1 = 5 + \frac{25 - 7}{18} \times 5$$

$$Q_1 = 5 + \frac{18}{18} \times 5$$

$$Q_1 = 5 + 1 \times 5 = 10$$

$$Q_1 = 10$$

$$Q_3 = \text{Size of } \frac{3N^{\text{th}}}{4} \text{ item}$$

$$= \frac{3(100)}{4} = \frac{300}{4} = 75^{\text{th}} \text{ item}$$

Q₃ Lies in the class 15 – 20

$$Q_3 = L + \frac{3\frac{N}{4} - CF}{f} \times i$$

Where,

$$L = 15, \frac{3N}{4} = 75, \text{ p.c.f} = 50, f = 30,$$

$$i = 5$$

$$Q_3 = 15 + \frac{75 - 50}{30} \times 5$$

$$= 15 + \frac{25}{30} \times 5$$

$$= 15 + 0.833 \times 5$$

$$= 15 + 4.16 = 19.167$$

$$Q_3 = 19.167$$

5th Decile

$$D_5 = \text{Size of } \frac{5N^{\text{th}}}{10} \text{ item}$$

$$= \frac{5(100)}{10} = \frac{500}{10} = 50^{\text{th}} \text{ item}$$

D₅ = Lies in the class 10 – 15.

$$D_5 = L + \frac{\frac{5N}{10} \text{ p.c.f}}{f} \times i$$

Where,

$$L = 10, \frac{5N}{10} = 50, f = 25,$$

$$\text{p.c.f} = 25, i = 5$$

$$D_5 = 10 + \frac{50 - 25}{25} \times 5$$

$$= 10 + \frac{25}{25} \times 5$$

$$= 10 + 1 \times 5 = 15$$

$$D_5 = 15$$

30th Percentile

$$P_{10} = \text{Size of } \frac{30N^{\text{th}}}{100} \text{ item}$$

$$= \frac{30(100)}{100} = \frac{3000}{100} = 30^{\text{th}} \text{ item}$$

P₃₀ lies in the class 10 – 15

$$P_{30} = L + \frac{\frac{30N}{100} - \text{p.c.f.}}{f} \times i$$

Where,

$$L = 10, \frac{30(N)}{100} = 30, f = 25$$

$$\text{p.c.f.} = 25, i = 5$$

$$P_{30} = 10 + \frac{30 - 25}{25} \times 5$$

$$= 10 + \frac{5}{25} \times 5$$

$$= 10 + 1 = 11$$

$$P_{30} = 11$$

2.3.1.4 Relationship between Partial Values

Q28. What is the relationship between partial values ?

Ans :

The measures of central tendency which are used for dividing the data into several equal parts are called partition values.

In this section, we shall discuss data analysis by dividing it into four, ten, and hundred parts of equal size and the corresponding partition values are called quartiles, deciles, and percentiles. All these values can be determined in the same way as median. The only difference is in their location.

Quartiles The values of observations in a data set, when arranged in an ordered sequence, can be divided into four equal parts, or quarters, using three quartiles namely Q_1 , Q_2 , and Q_3 . The first quartile Q_1 divides a distribution in such a way that 25 per cent ($=n/4$) of observations have a value less than Q_1 and 75 per cent ($= 3n/4$) have a value more than Q_1 i.e. Q_1 is the median of the ordered values that are below the median.

The second quartile Q_2 has the same number of observations above and below it. It is therefore same as median value.

The quartile Q_3 divides the data set in such a way that 75 per cent of the observations have a value less than Q_3 and 25 per cent have a value more than Q_3 , i.e. Q_3 is the median of the order values that are above the median.

The generalized formula for calculating quartiles in case of grouped data is :

$$Q_i = l + \left\{ \frac{i(n/4) - cf}{f} \right\} \times h ; i = 1, 2, 3$$

where cf = cumulative frequency prior to the quartile class interval

l = lower limit of the quartile class interval

f = frequency of the quartile class interval

h = width of the class interval

Deciles The values of observations in a data set when arranged in an ordered sequence can be divided into ten equal parts, using nine deciles, D_1 ($i = 1, 2, \dots, 9$). The generalized formula for calculating deciles in case of grouped data is :

$$D_i = l + \left\{ \frac{i(n/10) - cf}{f} \right\} \times h ; i = 1, 2, \dots, 9$$

where the symbols have their usual meaning and interpretation.

Percentiles

The values of observations in a data when arranged in an ordered sequence can be divided into hundred equal parts using ninety nine percentiles, P_i ($i = 1, 2, \dots, 99$). In general, the i th percentile is a number that has $i\%$ of the data values at or below it and $(100 - i)\%$ of the data values at or above it. The lower quartile (Q_1), median and upper quartile (Q_3) are also the 25th percentile, 50th percentile and 75th percentile, respectively. For example, if you are told that you scored 90th percentile in a test (like the CAT), it indicates that 90% of the scores were at or below your score, while 10% were at or above your score. The generalized formula for calculating percentiles in case of grouped data is :

$$P_i = l + \left\{ \frac{i(n/100) - cf}{f} \right\} \times h; i = 1, 2, \dots, 99$$

where the symbols have their usual meaning and interpretation.

2.4 MODE

Q29. Define Mode. Explain how Mode is calculated for individual series.

Ans :

Mode is that value of the variable which occurs maximum number of times in a distribution and around which other items are densely distributed.

In the words of Croxton and Cowden, "The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as the most typical of a series of values."

Further, according to A.M. Tuttle, "Mode is the value which has the greatest frequency density in its immediate neighbourhood."

Determination of Mode in Individual Series

The steps involved in calculating mode for individual series are as follows :

1. First arrange the data in ascending or descending order.
2. Check which value is repeating maximum number of times. The value repeating maximum number of times is considered as the value of mode.

Q30. Explain how Mode is calculated for discrete series.

Ans :

Determination of Mode in Discrete Series

In discrete series the value of mode can be determined in two ways :

- (i) Inspection
- (ii) Grouping and analysis table.

(i) Inspection

- (a) Maximum value or highest value is selected from frequency column.
- (b) The value corresponding to highest frequency value is considered as mode.

(ii) Grouping and analysis table.

If there exists an error of maximum frequency, frequency has small value preceding or succeeding it and items are highly focused on any one side then under such case grouping and analysis table are prepared.

Grouping Table

It consists of six columns as follows,

- In column 1-highest frequency is selected and highlighted with a circle or a tick (S) marks.
- In column 2-frequency values are grouped in two's
- In column 3-Ignore first frequency and group the remaining in two's.
- In column 4-frequencies are grouped in three's
- In column 5-Ignore first frequency and group the remaining frequency values in three's.
- In column 6-Ignore first two frequencies and group remaining frequencies in three's

After preparing the grouping table the maximum value in each column should be marked with a circle.

Analysis Table

- Analysis table is prepared by taking the column number on left side and probable values of mode on right side of the table.
- The values of variable corresponding to the highest frequencies are taken at the top of the analysis table and probable values of mode are determined.
- Finally, the maximum marked variable is considered as 'Mode'.

Q31. Explain how Mode is calculated for continuous series.**Ans :**

The steps involving in calculating mode in continuous series are as follows :

1. In continuous series the class of modal class can be determined in two ways,
 - (i) By Inspection or
 - (ii) By preparing grouping and analysis table
2. Value of mode is ascertained by using the following formula,

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

where

L = Lower limit of model class

f_1 = Frequency of model class

$\Delta_1 = |f_1 - f_0|$. f_0 is a preceeding value of f_1

$\Delta_2 = |f_1 - f_2|$. f_2 is a succeeding value of f_1

i = Class interval.

PROBLEMS**29. Find the mode from the following observation,**

Marks [x] : 5, 29, 25, 10, 29, 29, 20, 18, 15, 30.

Sol :**Calculation of Mode**

Marks (Arranged in ascending order)	5	10	15	18	20	25	29	30
No. of Times occurs	1	1	1	1	1	1	3	1

Since, 29 is repeated maximum number of times i.e., 3 times.

$\therefore$ Mode = 29

PROBLEMS ON MODE**30. Compute the modal value for the following observations and give your “ reasons**

10, 12, 13, 10, 18, 16, 15, 10, 11, 17, 10, 16, 11, 10, 12, 18, 19, 20, 15, 9

Sol :

Step I : Arrange the data in ascending order,

Observations	No.of times repeated or appeared
9	1
10	5
11	2
12	2
13	1
15	2
16	2
17	1
18	2
19	1
20	1

Among of all the above number 10 is repeated at maximum i.e. 5 times than the other numbers.
Hence mode = 10.

31. Find the mode of the following data.

0, 1, 6, 7, 2, 3, 7, 6, 6, 2, 6, 0, 5, 6, 0

Sol :

Observations	No.of times repeated
0	3
1	1
2	2
3	1
5	1
6	5
7	2

Among of all the above number 6 is repeated at maximum times i.e., 5 times.

Hence mode = 6.

32.

Size of Garment	27	28	29	30	31	32
No. of persons	10	20	40	65	50	15

Sol :

- By inspection the highest value is selected from frequency column i.e., (65)
 - The value of the variable corresponding to 65 is 30
- ∴ Mode = 30

33. Calculate mode for the following data.

Class Intervals	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Frequency	4	13	21	44	33	22	7

Sol :

Class Intervals	Frequency
0 - 10	4
10 - 20	13
20 - 30	21
30 - 40	44
40 - 50	33
50 - 60	22
60 - 70	7

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

Mode lies in the class interval 30-40 since the highest frequency 44 exists there

∴ L = Lower limit = 30

$$\Delta_1 = 44 - 32 = 12$$

$$\Delta_2 = 44 - 33 = 11$$

i = Class interval = 10

$$\therefore \text{Mode} = 30 + \frac{12}{12+11} \times 10$$

$$= 30 + \frac{12}{23} \times 10$$

$$= 30 + 5.22$$

$$\text{Mode} = 35.22$$

Q32. State the merits and demerits of mode.

Ans :

Merits

1. It is easy to understand and easy to calculate. In many cases it can be located just by inspection.
2. It can be located in situations where the variable is not measurable but categorization or ranking of observations is possible.
3. Like mean or median, it is not affected by extreme observations. It can be calculated even if these extreme observations are not known.

4. It can be determined even if the distribution has open end classes.
5. It can be located even when the class intervals are of unequal width provided that the width of modal and that of its preceding and following classes are equal.
6. It is a value around which there is more concentration of observations and hence the best representative of the data.

Demerits

1. It is not based on all the observations.
2. It is not capable of further mathematical treatment.
3. In certain cases mode is not rigidly defined and hence, the important requisite of a good measure of central tendency is not satisfied.
4. It is much affected by the fluctuations of sampling.
5. It is not easy to calculate unless the number of observations are sufficiently large and reveal a marked tendency of concentration around a particular value.
6. It is not suitable when different items of the data are of unequal importance.
7. It is an unstable average because, mode of a distribution, depends upon the choice of width of class intervals.

2.5 RELATIONSHIP BETWEEN MEAN, MEDIAN AND MODE

Q33. Explain the Relationship between Mean, Median and Mode.

Ans :

There exists a relationship between the mean, median and mode. They are related to another in the following ways,

1. In symmetrical distribution, the value of mean, median and mode coincides with each other i.e., is
2. In asymmetrical (non-symmetrical or skewed) distribution, the value of median lies in between the value of mean and mode.
3. Prof. Karl Pearson has expressed the relationship between mean, median and mode as follows,

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

It means that median is more close to mean when compared to the mode.

4. In asymmetrical distribution if any two values are given out of three then the third value can be ascertained by using formula,

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

A slight difference may be between the value ascertained by using formula $\text{mode} = 3 \text{ Median} - 2 \text{ Mean}$ and the value ascertained by using exact formula.

5. In positively skewed distribution, the value of mean would be higher than the value of median and the value of median would be higher than the value of mode.

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

6. In case of negatively skewed distribution, the value of mean would be lesser than the value of median and the value of median would be lesser than the value of mode.

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

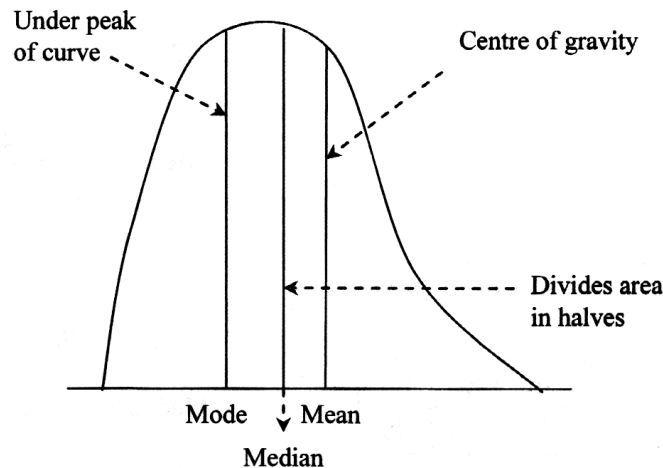


Fig : Relationship between Arithmetic Mean, Median & Mode

PROBLEMS

34. (a) The mean and median of a moderately skewed distribution are 42.2 and 41.9 respectively. Find mode of the distribution.
- (b) For a moderately skewed distribution, the median price of men's shoes is ₹ 380 and modal price is ₹ 350. Calculate mean price of shoes.

Sol :

- (a) Here, mode will be determined by the use of empirical formula.

$$\bar{X} - M_0 = 3(\bar{X} - M_d) \quad (\text{or}) \quad M_0 = 3M_d - 2\bar{X}$$

It is given that $\bar{X} = 42.2$ and $M_d = 41.9$

$$\begin{aligned} \therefore M_0 &= 3 \times 41.9 - 2 \times 42.2 \\ &= 125.7 - 84.4 = 41.3 \end{aligned}$$

- (b) Using the empirical relation, we can write

$$\bar{X} = \frac{3M_d - M_0}{2}$$

It is given that $M_d = ₹ 380$ and $M_0 = ₹ 350$

$$\therefore \bar{X} = \frac{3 \times 380 - 350}{2} = ₹ 395$$

2.6 MEASURES OF DISPERSION

2.6.1 Introduction

Q34. Define Dispersion. Explain the objectives of Measures of Dispersion.

Ans : (Imp.)

The concept of dispersion is related to the extent of scatter or variability in observations. The variability, in an observation, is often measured as its deviation from a central value. A suitable average of all such deviations is called the measure of dispersion. Since most of the measures of dispersion are based on the average of deviations of observations from an average, they are also known as the averages of second order. As opposed to this, the measures of central tendency are known as the averages of first order. Some important definitions of dispersion are given below:

Definitions

1. **According to A.L. Bowley** "Dispersion is the. measure of variation of the items."
2. **According to L.R. Connor** "Dispersion is the measure of extent to which individual items vary."
3. **According to Simpson and Kafka** "The measure of the scatteredness of the mass of figures in a series about an average is called the measure of variation or dispersion."
4. **According to Spiegel** "The degree to which numerical data tend to spread about an average value is called variation or dispersion of the data."

Objectives Of Measuring Dispersion

The main objectives of measuring dispersion of a distribution are:

1. To test reliability of an average

A measure of dispersion can be used to test the reliability of an average. A low value of dispersion implies that there is greater degree of homogeneity among various items and, consequently, their average can be taken as more reliable or representative of the distribution.

2. To compare the extent of variability in two or more distributions

The extent of variability in two or more distributions can be compared by computing their respective dispersions. A distribution having lower value of dispersion is said to be more uniform or consistent.

3. To facilitate the computations of other statistical measures

Measures of dispersions are used in computations of various important statistical measures like correlation, regression, test statistics, confidence intervals, control limits, etc.

4. To serve as the basis for control of variations

The main objective of computing a measure of dispersion is to know whether the given observations are uniform (or) not.

This knowledge may be utilized in many ways. In the words of Spurr and Bonini, "In matters of health, variations in body temperature, pulse beat and blood pressure are basic guides to diagnosis. Prescribed treatment is designed to control their variations. In industrial production, efficient operation requires control of quality variations, the causes of which are sought through inspection and quality control programmes". The extent of inequalities of income and wealth in any society may help in the selection of an appropriate policy to control their variations.

2.6.2 Significance of Measures of Dispersion

Q35. Explain the Importance of Measures of Dispersion.

Ans : (May-18, May-17)

1. To use other statistical methods

After getting value of dispersion we can proceed to other techniques such as to locate correlation or lines of regression (Regression Analysis)

2. To compare variability

We are in the general habit of comparison, may it be income, weight, height or temperature. To achieve the required degree of result one tries to improve to achieve the required result.

3. To test reliability of average

If the total of differences between the central value and the given value is smaller, the uniformity is less i.e. it means that this sum should be minimum for the reliability of the average.

4. To establish trends in time series

In time series we remove, cyclical, seasonal or random fluctuations, which we form after studying central values.

5. To control the undesired variability

According to Spurr and Bonini "In industrial production, efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programs".

Thus it is clear from above definition that we have to find the cause for variation, when we compare the degree of variability between two series.

Q36. What are the properties of dispersion ?

Ans :

A good measure of dispersion should possess, as far as possible, the following properties :

- i) It should be simple to understand.
- ii) It should be easy to compute.
- iii) It should be rigidly defined.
- iv) It should be based on each and every item of the distribution.
- v) It should be amenable to further algebraic treatment.
- vi) It should have sampling stability.
- vii) It should not be unduly affected by extreme items.

Q37. What are the two types of measures of dispersion?

OR

Write in detail about relative measures of dispersion.

Ans:

Measures of dispersion are of two types.

1. Absolute measures of dispersion
2. Relative measures of dispersion.

1. Absolute Measures of Dispersion

In case of absolute measures of dispersion, the statistical units are taken in the same manner as how it is given in the actual data. Like-kilograms, rupees, tonnes and so on. These statistical values are then utilized to compare the variations between the two distributions. The statistical units expressed in the two distributions and the average is the two distributions and the average size of the distributions should be same for the purpose of comparison.

The various absolute measures of dispersion are,

- (i) Range
- (ii) Quartile Deviation (QD)
- (iii) Mean Deviation (MD)
- (iv) Standard Deviation (SD).

2. Relative Measures of Dispersion

If the average size of the distributions and statistical units expressed in the distributions differ from each other under such a situation 'Relative Measures of Dispersion' is used. For example, one distribution is expressed in quintals of rice and another distribution is expressed in kilograms of wheat or average size of one distribution is higher than the average size of another distribution (salaries paid to managers and wages paid to workers). Thus, relative measures of dispersion helps to compare variability of two or more distribution of distinct units. Relative measures of dispersion is also known as 'Coefficient of Dispersion'.

The various relative measures of dispersion are,

- (i) Coefficient of Variation (C.V)
- (ii) Coefficient of Quartile Deviation (CQD)
- (iii) Coefficient of Mean Deviation (CMD)
- (iv) Coefficient of Range (CR)

Q38. Compare and contrast absolute measure of dispersion and relative measures of dispersion.

Ans :

S.No.	Absolute Measure of Dispersion	Relative Measure of Dispersion
1.	Absolute measure is dependent on the unit of the variable	1. Relative measure is not dependent on the unit of the variable.
2.	It is not taken into consideration for comparing two or more distributions	2. It is taken into consideration while comparing two or more distributions
3.	These measures are easy to calculate and understand.	3. These measures are complicated in calculating and understanding.

2.7 REQUISITES OF A MEASURE OF DISPERSION

2.7.1 Range

Q39. What is Range ? Explain.

Ans :

Range is the simplest method of studying dispersion. It is defined as the difference between the value of the smallest item and the value of the largest item included in the distribution. Symbolically,

$$\text{Range} = L - S$$

where

L = Largest item, and

S = Smallest item.

The relative measure corresponding to range, called the coefficient of range, is obtained by applying the following formula :

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

If the averages of the two distributions are about the same, a comparison of the range indicates that the distribution with the smaller range has less dispersion, and the average of that distribution is more typical of the group.

2.7.1.1 Advantages and Disadvantages of Range**Q40. What are the merits and Demerits of Range ?**

Ans :

The merits and limitations of Range can be enumerated here.

Merits

- (i) Amongst all the methods of studying dispersion range is the simplest to understand and the easiest to compute.
 - It is one of those measures which are rigidity defined.
- (ii) It takes minimum time to calculate the value of range. Hence, if one is interested in getting a quick rather than a very accurate picture of variability one may compute range.
 - It gives us the total picture of the problem even with a single glance.
 - It is used to check the quality of a product for quality control. Range plays an important role in preparing R-charts, thus quality is maintained.

Demerits

- 1. Range is not based on all the terms. Only extreme items reflect its size. Hence range cannot be completely representative of the data as all other middle values are ignored.
- 2. Due to above reason range is not a reliable measure of dispersion.
- 3. Range does not change even the least even if all the terms and variables are changed.
- 4. Range is too much affected by fluctuation of sampling. Range changes from sample to sample. As the size of sample increases range increases and vice versa.
- 5. It does not tell us anything about the variability of other data.
- 6. For open-end intervals, range is indeterminate because lower and upper limits of first and last interval are not given.

Q41. What are the uses of range ?

Ans :

(i) Quality Control

The object of quality control is to keep a check on the quality of the product without 100% inspection. When statistical methods of quality control are used, control charts are prepared and in preparing these charts range plays a very important role. The idea basically is that if the range - the difference between the largest and smallest mass produced items-increases beyond a certain point, the production machinery should be examined to find out why the items produced have not followed their usual more consistent pattern

(ii) Fluctuations in the Share Prices

Range is useful in studying the variations in the prices of stocks and shares and other commodities that are sensitive to price changes from one period to another.

(iii) Weather Forecasts

The meteorological department does make use of the range in determining, say, the difference between the minimum temperature and the maximum temperature. This information is of great concern to the general public because they know as to within what limits the temperature is likely to vary on a particular day.

(iv) Everyday Life

The range is a most commonly used measure of dispersion in everyday life. Questions of the form "what is the minimum and maximum temperature on a particular day"? "What is the difference between the wages earned by workers of a particular factory" ? How much one spends on petrol in his car/scooter in a month"- are all usually answered in the form of range. Answers to questions such as these are usually given in the form of 'Between such and such]. Regardless of the crudity of expression the answer is still a range.

PROBLEMS ON RANGE

- 35. The following are the wages of 10 workers of a factory. Find the range of variation and also compute the coefficient of range.**

275, 200, 370, 240, 100, 290, 400, 500, 180, 350

Sol :

Step I

Arrange the data in ascending order i.e.

100, 180, 200, 240, 275, 290, 350, 370, 400, 500

Range = L – S

where as = L = 500, S = 100

Range ($\bar{R}$) = 500 – 100 = 400

Coefficient of range = $\frac{L - S}{L + S} = \frac{500 - 100}{500 + 100}$

Coefficient of range = $\frac{400}{600} = 0.667$.

- 36. The following are the marks of 100 students of a class. Find the range of variation of marks and also compute coefficient of range.**

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of students	8	16	26	20	12	10	8

Sol :

From the given data

Highest (L) marks is 80 and

Lowest (S) marks is 10. Hence

Range ($\bar{R}$) = L – S

($\bar{R}$) = 80 – 10 = 70

Coefficient of Range = $\frac{L - S}{L + S}$

Coefficient of Range = $\frac{80 - 10}{80 + 10}$

Coefficient of Range = $\frac{70}{90}$

Coefficient of Range = 0.77.

2.7.2 Quartile Deviation

Q42. What is quartile deviations ?

Ans :

The range as a measure of dispersion discussed above has certain limitations. It is based on two extreme items and it fails to take account of the scatter within the range. From this there is reason to believe that if the dispersion of the extreme items is discarded, the limited range thus established might be more instructive.

For this purpose there has been developed a measure called the interquartile range, the range which includes the middle 50 per cent of the distribution. That is, one quarter of the observations at the lower end, another quarter of the observations at the upper end of the distribution are excluded in computing the interquartile range. In other words, interquartile range represents the difference between the third quartile and the first quartile.

$$\text{Inter quartile range} = Q_3 - Q_1$$

Very often the interquartile range is reduced to the form of the Semi-interquartile range or quartile deviation by dividing it by 2.

$$\text{Quartile Deviation or Q.D.} = \frac{Q_3 - Q_1}{2}$$

Coefficients of Quartile Deviation

Quartile deviation is an absolute measure of dispersion. Its relative measure is the coefficient of Quartile deviation.

Coefficients of Quartile Deviation

$$= \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

It is used to compare the degree of variation in different series.

2.7.2.1 Advantages and Disadvantages of Quartile Deviations

Q43. Explain merits and demerits of quartile deviations.

Ans :

Merits of Quartile Deviation

1. It is the simple to calculate and very easy to understand.
2. It is not impacted by extreme values
3. It can be computed for open ended distributions and for data containing unequal classes

Demerits of Quartile Deviation

1. It is impacted by sample size and composition of the sample.

2. It is not amenable to algebraic or statistical treatment
3. It does not tell us anything about the spread of the various data items across the measure of central tendency.

Q44. How to calculate quartile deviations?*Ans :***Individual Series**

Low Quartile (Q_1) = Size of $\frac{N+1}{4}$ th item

N = No. of observations

Upper Quartile (Q_3) = Size of $\frac{3(N+1)}{4}$ th item.

Discrete Series

i) Q_1 = Size of $\frac{N+1}{4}$ th item

ii) Q_3 = Size of $\frac{3(N+1)}{4}$ th item

iii) Quartile Deviation (Q.D) = $\frac{Q_3 - Q_1}{2}$

Continuous Series

$$(Q_1) = \frac{L_1 + \frac{N}{4} - F}{f} \times c$$

$$(Q_3) = \frac{L_3 + \frac{3N}{4} - F}{f} \times c$$

PROBLEMS ON

37. Find out the value of quartile deviation and its coefficient for the following data :

Roll No	1	2	3	4	5	6	7
Marks	25	33	45	17	35	20	55

*Sol :***Calculation of Quartile Deviation**

Marks arranged in ascending order 17 20 25 33 35 45 55

$$Q_1 = \text{Size of } \left[\frac{N+1}{4} \right]^{\text{th}} \text{ Item}$$

N = No. of observations.

$$= \text{Size of } \left[\frac{7+1}{4} \right]^{\text{th}} \text{ Item} = 2^{\text{nd}} \text{ Item} = 20$$

$$Q_3 = \text{Size of } 3 \left[\frac{N+1}{4} \right]^{\text{th}} \text{ Item}$$

$$= \text{Size of } 3 \left[\frac{7+1}{4} \right]^{\text{th}} \text{ Item} = 6^{\text{th}} \text{ Item} = 45$$

$$\text{Quartile Deviation (Q.D)} = \frac{Q_3 - Q_1}{2} = \frac{45 - 20}{2} = \frac{25}{2} = 12.5$$

$$\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{45 - 20}{45 + 20} = \frac{25}{65} = 0.38.$$

38. Compute Quartile Deviation and its coefficient for the following data.

Height (in inches)	58	59	60	61	62	63	64	65	66
No. of Students	15	20	32	35	33	22	20	10	8

Sol :

Calculation of Quartile Deviation

Height	f	C.f	
58	15	15	
59	20	35	
60	32	67	(Q ₁)
61	35	102	
62	33	135	
63	22	157	(Q ₃)
64	20	177	
65	10	187	
66	8	195	
	N = 195		

$$Q_1 = \text{Size of } \left[\frac{N+1}{4} \right]^{\text{th}} \text{ Item}$$

N = No. of observations.

$$= \text{Size of } \left[\frac{195+1}{4} \right]^{th} \text{ Item} = 49^{th} \text{ Item} = 60$$

$$\therefore Q_1 = 60$$

$$Q_3 = \text{Size of } 3 \left[\frac{195+1}{4} \right]^{th} \text{ Item}$$

$$= \text{Size of } 3 \left[\frac{196}{4} \right]^{th} \text{ Item} = 147^{th} \text{ Item} = 63$$

$$\therefore Q_3 = 63$$

$$\text{Quartile Deviation (Q.D)} = \frac{Q_3 - Q_1}{2} = \frac{63 - 60}{2} = 1.5$$

$$\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{63 - 60}{63 + 60} = \frac{3}{123} = 0.024$$

39. Calculate the Quartile Deviation and Coefficient of Quartile Deviation of the following distribution.

Marks (X)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No. of Students (f)	11	18	25	28	30	33	22	15	12	10

Sol :

	X	f	C.f
	0-10	11	11
	10-20	18	29 F
(Q ₁)	L ₁ 20-30	25 f	54
	30-40	28	82
	40-50	30	112
	50-60	33	145 F
(Q ₃)	L ₃ 60-70	22 f	167
	70-80	15	182
	80-90	12	194
	90-100	10	204

Calculation of Q₁

$$Q_1 = \left(\frac{N}{4} \right)^{th} \text{ Item}$$

$$= \left(\frac{204}{4} \right)^{th} \text{ Item} = 51^{th} \text{ Item}$$

Size of 51^{th} Item lies in cf 54. So Q_1 class interval = 20–30

$$Q_1 = \frac{L_1 + \frac{N}{4} - F}{f} \times c$$

$$= \frac{20 + 51 - 29}{25} \times 10 = 20 + \frac{220}{25} = 28.8$$

Calculation of Q_3

$$= Q_3 = 3 \left(\frac{N}{4} \right)^{th}$$

$$= 3 \left(\frac{204}{4} \right)^{th} \text{ Item} = 3(51)^{th} \text{ Item} = 153^{th} \text{ Item}$$

Size of 153^{th} Item lies in CF 167. So Q_3 class interval = 60 – 70

$$Q_3 = \frac{L_3 + \frac{3N}{4} - F}{f} \times c$$

$$= 60 + \frac{153 - 145}{22} \times 10$$

$$= 60 + \frac{80}{22} = 63.64$$

Calculation of Quartile Deviation

$$\text{Quartile Deviation (Q.D)} = \frac{Q_3 - Q_1}{2} = \frac{63.64 - 28.8}{2} = 17.4$$

$$\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{63.64 - 28.8}{63.64 + 28.8} = \frac{34.84}{92.44} = 0.37$$

40. Compute the values of Q_1 , Q_3 , Quartile Deviation and Coefficient of Quartile Deviation of the following distribution.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No.of Students	5	25	40	70	90	40	20	10

Sol :

	Marks	f	C.f
	0-10	5	5
	10-20	25	30
	20-30	40	70 F
(Q ₁) L ₁	30-40	70 f	140 F
(Q ₃) L ₃	40-50	90 f	230
	50-60	40	270
	60-70	20	290
	70-80	10	300

$$Q_1 = \text{Size of } \left(\frac{N}{4}\right)^{\text{th}} \text{ Item}$$

$$= \text{Size of } \left(\frac{300}{4}\right)^{\text{th}} \text{ Item} = 75^{\text{th}} \text{ Item}$$

Q₁ lies in the class 30 – 40.

$$Q_1 = \frac{L_1 + \frac{N}{4} - F}{f}$$

$$= 30 + \frac{75 - 70}{70} \times 10$$

$$= 30 + \frac{50}{10} = 30.71$$

$$Q_3 = 3\left(\frac{N}{4}\right)$$

$$Q_3 = \text{Size of } \frac{3 \times 300}{4} = 225^{\text{th}} \text{ Item}$$

Q₃ lies in the class 40 - 50.

$$Q_3 = \frac{L_3 + \frac{3N}{4} - F}{f} \times C$$

$$40 + \frac{225 - 140}{90} \times 10$$

$$40 + \frac{850}{90} = 49.44$$

$$\text{Thus } Q_1 = 30.71$$

$$Q_3 = 49.44$$

Quartile Deviation (Q.D)

$$= \frac{Q_3 - Q_1}{2} = \frac{49.44 - 30.71}{2} = 9.37$$

Coefficient of Q.D

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{49.44 - 30.71}{49.44 + 30.71} = \frac{18.73}{80.15} = 0.23$$

2.7.3 Mean Deviation

Q45. Define Mean Deviation. Explain how Mean Deviation is calculated for individual series.

Ans :

(Imp.)

Mean Deviation/Average Deviation

Mean deviation is usually referred to as "Average Deviation". It is defined as the average difference between the items of a distribution and the median or mean of that series.

$$\text{M.D} = \frac{1}{n} \sum |X - A|$$

It is used in computing average deviation and thus it is called as mean deviation.

$$\text{Coefficient of M.D} = \frac{\text{M.D}}{\text{Mean}}$$

Merits

Following are the merits of mean deviation,

1. It considers each and every item of the data.
2. Mean deviation is not greatly affected by the extreme items.
3. It takes deviations from the central value thus making the computation easy for various distributions and comparisons.

Steps Involved in Calculation of Mean Deviation for Individual Observation

Following are the steps to be followed while calculating mean deviation of the individual observations,

1. In the first step, median of the series is calculated.
2. In the second step, take the deviations of items from median by ignoring $\pm$ signs and denote them as $|D|$.

3. In the third step, sum of these deviations is ascertained/obtained i.e., $\sum |D|$.
4. In the last step, the sum of deviations i.e., $\sum |D|$ is divided by the number of observations.

In case of normal distribution, the mean $\pm$ mean deviation indicates the range which includes 57.7 percent of components in the series. If the distribution is adequately skewed then it might be possible to have 57.5 percent of the items within this range. When average deviation is small, then the distribution is said to be highly uniform as majority of the cases focuses on small range around the mean.

The relative measure related to the mean deviation is known as the 'coefficient of mean deviation'. It is ascertained by dividing mean deviation by the specific average which is used in calculating mean deviation. Hence, if median is used to calculate mean deviation then coefficient of mean deviation would be calculated by dividing mean deviation by median.

$\text{Coefficient of M.D} = \frac{\text{M.D.}}{\text{Median}}$

In case mean is used to calculate mean deviation then coefficient of mean deviation would be calculated by dividing mean deviation by mean.

Q46. Explain how Mean Deviation is calculated for discrete and continuous series.

Ans :

a) Steps for Calculating Mean Deviation for Discrete Series

Following are the steps for calculating mean deviation for discrete series,

1. Compute the value of median
2. Ascertain the deviations as $d = X - A$
Where,
 X = Variable
 A = Median.
Ignore negative signs in the deviations column and denote deviations by $|D|$.
3. Multiply values of deviations with their respective frequencies and determine the sum of $\sum f|D|$
4. Divide the total obtained in step 3 by N (N = The number of observations).
The value so obtained is nothing but the value of mean deviation.

b) Steps for Calculating Mean Deviation for Continuous Series

Following are the steps involved in calculation of mean deviation for continuous series,

1. Compute the value of median.
2. Ascertain the mid-point of various classes and denote mid-points column as M .
3. Ascertain the deviation as $d = M - A$
Where,
 M = Mid-points A = Median.
Ignore negative signs in the deviations column and denote them by $|D|$.
4. Multiply values of deviations with their respective frequencies and determine the sum of $\sum f|D|$.
5. Divide the total obtained in step 4 by N
Where, N = The number of observations.

2.7.3.1 Advantages and Disadvantages of Mean Deviation**Q47. State the merits and demerits of Mean Deviation.***Ans :***Merits**

1. It is easy to understand and easy to compute.
2. It is based on all the observations.
3. It is less affected by extreme observations vis-a-vis range or standard deviation (to be discussed in the next section).
4. It is not much affected by fluctuations of sampling.

Demerits

1. It is not capable of further mathematical treatment. Since mean deviation is the arithmetic mean of absolute values of deviations, it is not very convenient to be algebraically manipulated. This necessitates a search for a measure of dispersion which is capable of being subjected to further mathematical treatment.
2. It is not well-defined measure of dispersion since deviations can be taken from any measure of central tendency.

PROBLEMS**41. Calculate the mean deviation and its coefficient from the following data.**

Income of Members (₹)
1,000
1,500
2,500
3,000
3,500
4,000

Sol :

Income of Members (x)	Deviation from Median $ D = x - A (2500)$
1,000	1,500
1,500	1,000
2,000	500
2,500	0
3,000	500
3,500	1,000
4,000	1,500
N = 7	$\Sigma D = 6,000$

$$\text{Median} = \text{Size of } \frac{N+1^{\text{th}}}{2} \text{ item} = \frac{7+1}{2} = \frac{8}{2} = 4^{\text{th}} \text{ item.}$$

$$\text{Mean deviation} = \frac{\sum |D|}{N}$$

Where,

$|D|$ = Deviation from median ignoring signs.

$$\text{M.D} = \frac{6000}{7} = 857.14$$

$$\therefore \text{Coefficient of M.D} = \frac{857.14}{2500} = 0.343$$

42. Calculate mean deviation from the following series,

X	15	11	12	13	14
F	3	12	18	2	6

Sol :

Step 1

Computation of Median $|D| = X - A$

X	F	c.f	 D 	f D
15	3	3	3	9
11	12	15	1	12
12	18	33	0	0
13	2	35	1	2
14	6	41	2	12
N = 41				$\Sigma f D = 35$

$$\text{Median} = \text{Size of } \frac{N+1}{2} \text{th item}$$

$$= \frac{41+1}{2} = \frac{42}{2} = 21^{\text{th}} \text{ item}$$

$\therefore$ Median = 12

Apply formula,

$$\text{Mean deviation} = \frac{\Sigma f|D|}{N}$$

$$\text{M.D} = \frac{35}{41} = 0.85$$

$$\boxed{\text{M.D} = 0.85}$$

43. Calculate the value of co-efficient of mean deviation (from median) of the following data :

Marks	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80	80 - 90
No. of Students	2	6	12	18	25	20	10	7

Sol :

Calculation of Median and Mean Deviation

Marks	f	c.f	Mid-value (M)	M-A = D M - 54.8	F M-A = f D
10-20	2	2	15	39.8	79.6
20-30	6	8	25	29.8	178.8
30-40	12	20	35	19.8	237.6
40-50	18	38	45	9.8	176.4
50-60	25	63	55	0.2	5
60-70	20	83	65	10.2	204
70-80	10	93	75	20.2	202
80-90	7	100	85	30.2	211.4
	N=10				Σf D = 1294.8

$$\frac{N}{2} = \frac{100}{2} = 50$$

C.F is just greater than 50 is 63

∴ Median lies in the class 50 – 60

$$\text{Median} = L + \left[\frac{\frac{N}{2} - \text{p.c.f}}{f} \right] \times i$$

$$L = 50, N/2 = 50, \text{p.c.f} = 38, i = 10, f = 25$$

$$= 50 + \left[\frac{50 - 38}{25} \right] \times 10$$

$$= 50 + \left[\frac{12}{25} \right] \times 10$$

$$= 50 + 0.48 \times 10$$

$$= 50 + 4.8$$

$$\text{Median} = 54.8$$

$$\begin{aligned}\text{Mean Deviation about Median} &= \frac{\sum f|D|}{N} \\ &= \frac{1294.8}{100} = 12.948 \\ &\cong 12.95\end{aligned}$$

$$\begin{aligned}\text{Coefficient of Mean Deviation} &= \frac{\text{Mean Deviation about Median}}{\text{Median}} \\ &= \frac{12.95}{54.8} \\ &= 0.2363\end{aligned}$$

2.7.4 Standard Deviation

Q48. Define Standard Deviation. Explain how Standard Deviation is calculated for individual series.

Ans :

(Nov.-20)

Karl Pearson introduced the concept of standard deviation in the year 1823. It is widely used as a method of dispersion and satisfies most of the properties which helps in measuring dispersion. It is also called "root mean square deviation". It is represented by small Greek letter ' σ '. It is useful in measuring the absolute dispersion and judging the representativeness of the mean.

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

For step deviation method,

Where, $d = \frac{(X - A)}{i}$ and i = Class interval

For actual mean method the formula is,

$$\sigma = \sqrt{\frac{\sum fx^2}{N} \text{ where, } X = (x - \bar{X})}$$

For assumed mean method, the formula is, $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$

Where, $d = (X - A)$

Coefficient of variation (or) C. V = $\frac{\sigma}{\bar{X}} \times 100$

Methods of Standard Deviation

There are two methods available for calculating standard deviation for individual observations.

Any one of the following methods can be used.

- (i) Actual mean method and
- (ii) Assumed mean method.

(i) Actual Mean Method

Under this method, the deviations are taken from actual mean. The following formula is used in this method,

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

$$x = [X - \bar{X}]$$

Steps

The steps involved in actual mean method are,

1. Compute the actual or arithmetic mean of the given series and denote as ' $\bar{X}$ '
2. Compute the deviations i.e., $[X - \bar{X}]$ by subtracting each variable value from actual mean. Denote these deviations as ' x '.
3. Square each of the deviations obtained in step 2. Determine the sum of squared deviation and denote them as $\sum x^2$.
4. Divide $\sum x^2$ by the total number of observation ' N '.
5. Take the square-root of the value obtained in step 4.
6. The value obtained in step-5 is the value of standard deviation and is denoted as ' σ '.

Assumed Mean Method

Under this method deviations are taken from assumed mean and the following formula is applied,

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left[\frac{\sum d}{N} \right]^2}$$

Steps

The steps involved in assumed mean method are,

1. Ascertain the deviations between values (X) and assumed mean and denote it as ' d '. Calculate the sum of these deviation and denote it as $\sum d$.
2. Square each of the deviation obtained in step 1 and obtain the total of squared deviations and denote it as $\sum d^2$.
3. Finally, substitute the values of $\sum d$, $\sum d^2$ and N the standard deviation formula to obtain the value of standard deviation.

Q49. Explain how Standard Deviation is calculated for discrete series.**Ans :**

Standard deviation for discrete series may be calculated by using any one of the following three methods,

- (i) Actual mean method
- (ii) Assumed mean method and
- (iii) Step deviation method.

(i) Actual Mean Method

Under actual mean method, deviations are taken actual mean i.e., $[X - \bar{X}]$ and denoted as x .

- (a) Square each of the deviation and multiply it with its respective frequencies Σfx^2 .
- (b) Substitute the values of Σfx^2 and N in the value of standard formula

$$\sigma = \sqrt{\frac{\Sigma fx^2}{N}} \text{ to obtain the value of standard.}$$

If the actual mean is obtained in fractions then calculating standard deviation becomes a time consuming process. Due to this reason actual mean method is not used frequently.

(ii) Assumed Mean Method

Following are the steps for calculating standard deviation under assumed mean method,

1. Take the deviations from assumed mean i.e., $[X - \bar{X}]$ and denote them as 'd'.
2. Multiply each deviation with their respective frequencies and obtain sum of it and denote it as Σfd .
3. Square each of the deviations and denote it as cf .
4. Multiply each of the squared deviations with their respective frequencies and obtain the sum of them, Σfd^2 .

Finally substitute the values of Σfd , Σfd^2 and N by using the following formula to ascertain the value of standard deviation.

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left[\frac{\Sigma fd}{N} \right]^2}$$

Q50. Explain how Standard Deviation is calculated for continuous series.**Ans :**

Similar to the discrete series, in continuous series also standard deviation can be calculated by using actual mean method or assumed mean method or step deviation method. But usually step deviation method is followed,

Formula

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left[\frac{\Sigma fd}{N} \right]^2} \times i$$

Where,

$$d = \frac{(m - A)}{i}; i = \text{Class interval}$$

Steps

Following are the steps for calculating standard deviation,

1. Determine the mid-points of various classes.
2. Calculate deviations $[X - A]$ by subtracting assumed mean (A) from each value of the mid-point $[X]$.
3. Divide the deviations by the width of class interval i.e., 'i' and denote this column by 'd'.
4. Multiply each of the deviation with their respective frequencies and obtain a total and denote it as Σfd .
5. Obtain the square of each deviation i.e., d^2
6. Multiply each squared deviation with their respective frequencies and obtain a total and denote it as Σfd^2 .

2.7.4.1 Advantages and Disadvantages of Standard Deviation**Q51. State the merits and demerits of Standard Deviation.****Ans :****Merits**

1. It is a rigidly defined measure of dispersion.
2. It is based on all the observations.

3. It is capable of being treated mathematically. For example, if standard deviations of a number of groups are known, their combined standard deviation can be computed.
4. It is not very much affected by the fluctuations of sampling and, therefore, is widely used in sampling theory and tests of significance.

Demerits

1. As compared to the quartile deviation and range, etc., it is difficult to understand and difficult to calculate.
2. It gives more importance to extreme observations.
3. Since it depends upon the units of measurement of the observations, it cannot be used to compare the dispersions of the distributions expressed in different units.

2.7.4.2 Concept of Variation

Q52. Define the concept of variation.

Ans :

(May-19)

The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1913. The concept of variance is highly important in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variation in their original series. Variance is defined as follows:

$$\text{Variance} = \frac{\sum (X - \bar{X})^2}{N}$$

Thus, variance is nothing but the square of the standard deviation

$$\text{i.e., Variance} = \sigma^2 \quad \text{or} \quad \sigma = \sqrt{\text{Variance}}$$

In a frequency distribution where deviations are taken from assumed mean variance may directly be computed as follows :

$$\text{Variance} = \left\{ \frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2 \right\} \times i^2$$

$$\text{when } d = \frac{(X - A)}{i} \text{ and } i = \text{common factor.}$$

2.7.5 Coefficient of Variation

Q53. Define Coefficient of Variation.

Ans :

Coefficient of Variation (CV) was proposed by Karl Pearson. It is used to compare the variability of two or more distributions. A distribution with greater C.V. is considered as more variable or less consistent, less uniform, less stable or less homogeneous distribution and the distribution with less C.V. is considered as less variable or more consistent, more uniform, more stable or more homogeneous distribution.

$$\text{Coefficient of variation} = \frac{\text{Standard deviation}}{\text{Arithmetic mean}} \times 100$$

(OR)

$$\text{C.V} = \frac{\sigma}{\bar{X}} \times 100$$

PROBLEMS

44. Calculate the standard deviation from the data using actual mean. 140, 160, 190, 145, 175.

Sol :

Calculation of Standard Deviation

X	x = [X - $\bar{X}$]	x²
140	- 22	484
160	- 2	4
190	28	784
145	- 17	289
175	13	169
$\Sigma X = 810$		$\Sigma x^2 = 1730$

$$\text{Arithmetic mean } (\bar{X}) = \frac{\Sigma X}{N}$$

$$= \frac{810}{5} = 162$$

$$\therefore \bar{X} = 162$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{1730}{5}} = \sqrt{346}$$

$$\sigma = 18.6$$

45. Calculate the standard deviation from data using assumed mean 140, 160, 190, 145, 175.

Sol :

Calculation of Standard Deviation

X	d = [X - 190]	d²
140	- 50	2500
160	- 30	900
190	0	0
145	- 45	2025
175	- 15	225
	$\Sigma d = - 140$	$\Sigma d^2 = 5650$

$$\therefore \sigma = \sqrt{\frac{\sum d^2}{N} - \left[\frac{\sum d}{N} \right]^2}$$

$$\therefore \sum d^2 = 5650, N = 5, \sum d = -140$$

$$\sigma = \sqrt{\frac{5650}{5} - \left[\frac{-140}{5} \right]^2}$$

$$\sigma = \sqrt{1,130 - (-28)^2}$$

$$\sigma = \sqrt{1,130 - 784}$$

$$\sigma = \sqrt{346}$$

$$\sigma = 18.6$$

46. The annual salaries of a group of employees are given in the following table.

Salaries (in `000)	45	50	55	60	65	70	75	80
Number of persons	3	5	8	7	9	7	4	7

Calculate standard deviation.

Sol :

Calculation of Standard Deviation with Assumed Mean

X	f	[X - 60] d	d ²	fd	fd ²
45	3	-15	225	-45	675
50	5	-10	100	-50	500
55	8	-5	25	-40	200
60	7	0	0	0	0
65	9	5	25	45	225
70	7	10	100	70	700
75	4	15	225	60	900
80	7	20	400	140	2800
	N = 50			Σfd = 180	Σfd² = 6000

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N} \right]^2}$$

$$= \sqrt{\frac{6000}{50} - \left[\frac{180}{50}\right]^2}$$

$$= \sqrt{120 - 12.96}$$

$$= \sqrt{107.04}$$

$$\sigma = 10.35$$

47. The following table gives the marks obtained by a group of 80 students in an examination. Calculate the variance.

Marks obtained	No. of students
10 - 14	2
14 - 18	4
18 - 22	4
22 - 26	8
26 - 30	12
30 - 34	16
34 - 38	10
38 - 42	8
42 - 46	4
46 - 50	6
50 - 54	2
54 - 58	4

Sol :

Calculation of Variance

Marks	midpoint	f	(m - 32) / 4 d	fd	fd ²
10 - 14	12	2	- 5	- 10	50
14 - 18	16	4	- 4	- 16	64
18 - 22	20	4	- 3	- 12	36
26 - 30	28	12	- 1	- 12	12
30 - 34	32	16	0	0	0
34 - 38	36	10	+ 1	+ 10	10
38 - 42	40	8	+ 2	+ 16	32
42 - 46	44	4	+ 3	+ 12	36
36 - 50	48	6	+ 4	+ 24	96
50 - 54	52	2	+ 5	+ 10	50
54 - 58	56	4	+ 6	+ 24	144
		N = 80		Σfd = 30	Σfd ² = 562

$$\text{Variance} = \left[\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2 \right] \times i^2$$

$$\sum fd^2 = 562, \sum fd = 30, N = 80, i = 4$$

Substituting the values

Variance =

$$\left\{ \frac{562}{80} - \left(\frac{30}{80} \right)^2 \right\} \times 4^2 = (7.025 - 0.141) \times 16$$

$$= 6.884 \times 16 = 110.144$$

2.8 SKEWNESS

2.8.1 Introduction

Q54. Explain briefly about Skewness.

Ans : (May-19)

Skewness refers to lack of symmetry in a frequency distribution. In order to have an idea regarding the shape of the curve, skewness is studied. By making use of the given data, the shape of the curve can be drawn. The distribution is called symmetrical or skewed, if the distribution satisfies the following conditions.

- (a) Mean, median and mode should fall at distinct points.
- (b) Quartiles should not be at equal distance from median.
- (c) The shape of the curve drawn by utilizing the given data is not symmetrical.
- (d) The curve stretched more towards one side than to the other side.

Some important definitions of skewness are as follows :

1. **According to Croxton & Cowden** "When a series is not symmetrical it is said to be asymmetrical or skewed."
2. **According to Morris Hamburg** "Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution."
3. **According to Simpson & Kafka** "Measures of skewness tell us the direction

and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness."

4. **According to Garrett** "A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other - to left or right."

The analysis of above definitions shows that the term 'skewness' refers to lack of symmetry, i.e., when a distribution is not symmetrical (or is asymmetrical) it is called a skewed distribution. Any measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with a symmetrical (or normal) distribution. If, for example, skewness is positive, the frequencies in the distribution are

Spread out over a greater range of values on the high-value end of the curve (the right-hand side) than they are on the low-value end. If the curve is normal, spread will be the same on both sides of the centre point and the mean, median and mode will all have the same value. The concept of skewness gains importance from the fact that statistical theory is often based upon the assumption of the normal distribution. A measure of skewness is, therefore necessary in order to guard against the consequences of this assumption.

The concept of skewness will be clear from the following three diagrams: a symmetrical distribution, a positively skewed distribution and a negatively skewed distribution.

Q55. Distinguish between dispersion and skewness.

Ans :

Difference between Dispersion and Skewness

1. Dispersion is concerned with measuring the amount of variation in a series. Skewness is concerned with direction of variation or the departure from symmetry.

2. Dispersion gives us the extent to which the values of a given series are scattered. Skewness explains the extent and direction in which the distribution differs from symmetry.
3. Dispersion tells us about the composition of the series. Skewness tells us about the shape of the series.
4. Dispersion helps us to ascertain the extent to which a central value is representative of the series. Skewness deals with the nature of variations on either side of the central value.
5. Measures of Dispersion are based on averages of the first order (i.e. measures of central tendency). They are averages of the second order. Measures of Skewness are based on averages of first and second order (i.e. measures of central tendency and dispersion). Measures of Skewness are not averages at all.

2.8.2 Measures of Skewness

Q56. Explain the different measures of skewness.

Ans :

Skewed distribution may be either positively skewed or negatively skewed.

1. Positively Skewed Distribution

If the longer tail of the distribution is towards the higher values or right hand side, the skewness is positive. Positive skewness occurs when mean is increased by some unusually high values thus satisfying the following properties,

- (i) $\text{Mean} > \text{Median} > \text{Mode}$
- (ii) Right tail is longer than its left tail
- (iii) $(Q_3 - \text{Median}) > (\text{Median} - Q_1)$.

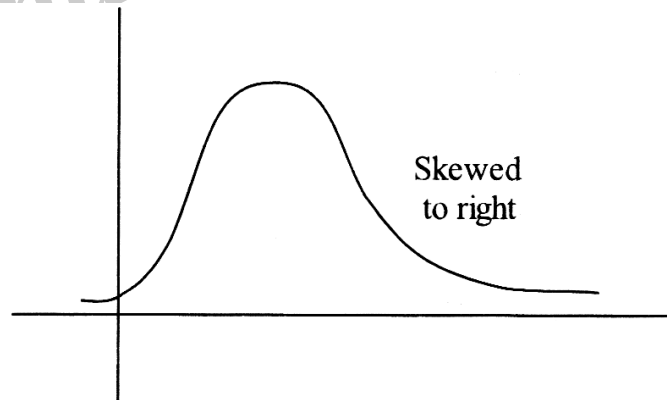


Fig: Positively Skewed Distribution

2. Negatively Skewed Distribution

If the longer tail is towards the lower value or left hand side, the skewness is negative.

Negative skewness arises when mean is decreased by some extremely low values, thus satisfying the following properties,

- (i) $\text{Mean} < \text{Median} < \text{Mode}$

- (ii) Left tail is longer than its right tail
 (iii) $(Q_3 - \text{Median}) < (\text{Median} - Q_1)$

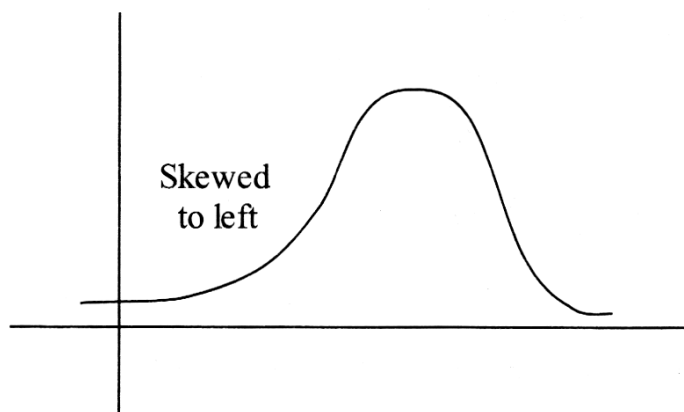


Figure: Negatively Skewed Distribution

2.8.2.1 Relative measures of Skewness

Q57. Explain the Relative measures of Skewness.

Ans :

Relative measure of skewness is also known as the coefficient of skewness. It is defined as an absolute measure of skewness expressed in relation to some measure of dispersion. The following are the coefficient of skewness,

1. Karl Pearson's coefficient of skewness
2. Bowley's coefficient of skewness
3. Kelly's coefficient of skewness

1. Karl Pearson's Coefficient of Skewness

The coefficient of skewness given by Karl Pearson can be defined as,

$$S_k(P) = \frac{\text{Mean} - \text{Mode}}{\text{S.D}}$$

If the mode is inadequately defined then an approximate relationship, $\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$ is used hence, the above formula reduces to,

$$S_k(P) = \frac{3(\text{Mean} - \text{Median})}{\text{S.D}}$$

2. Bowley's Coefficient of Skewness

The coefficient of skewness given by Bowley's is defined as,

$$S_k(B) = \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{(Q_3 - \text{Median}) + (\text{Median} - Q_1)}$$

$$S_k(B) = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

This measure is based on quartiles and is also called the quartile coefficient of skewness.

3. Kelly's Coefficient of Skewness

The following formula is proposed by Kelly for measuring the skewness on the 10th and 90th percentiles and also upon the first and ninth deciles.

$$Sk_k = \frac{P_{10} + P_{90} - 2 \text{ Median}}{P_{90} - P_{10}}$$

$$Sk = \frac{D_1 + D_9 - 2\text{Median}}{D_9 - D_1}$$

Whereas,

SK_k = Kelly's coefficient of skewness

PROBLEMS

48. Calculate Karl Pearson's co-efficient of skewness from the following data:

Age in Years	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
No. of Persons	18	20	30	22	10

Sol :

Calculation of Karl Pearson's Coefficient of Skewness

Age	Midpoint	f	$d = \frac{x - A}{10}$ (A = 35)	d ²	fd	fd ²
10 - 20	15	18	- 2	4	- 36	72
20 - 30	25	20	- 1	1	- 20	20
30 - 40	35	30	0	0	0	0
40 - 50	45	22	1	1	22	22
50 - 60	55	10	2	4	20	40
		Σf(N) = 100			Σfd = -14	Σfd² = 154

Karl Pearson's Coefficient of Skewness,

$$S_{k_p} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Calculation of Mean ($\bar{x}$)

$$\text{Mean}(\bar{x}) = A + \frac{\sum fd}{N} \times i$$

$$A = 35, \sum fd = -14, N = 100 \text{ and } i = 10$$

$$\text{Mean}(\bar{x}) = 35 + \frac{(-14)}{100} \times 10$$

$$\text{Mean}(\bar{x}) = 33.6$$

Calculation of Mode

Mode lies in the class 30 – 40

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

$$L = 30$$

$$\Delta_1 = 30 - 20 = 10$$

$$\Delta_2 = 30 - 22 = 8$$

$$i = 10$$

$$\begin{aligned} \text{Mode} &= 30 + \frac{10}{10+8} \times 10 \\ &= 30 + 5.55 \\ &= 35.55 \end{aligned}$$

Calculate of Standard Deviation (σ)

$$\begin{aligned} \sigma &= \sqrt{\left(\frac{\sum fd^2}{N}\right) - \left(\frac{\sum fd}{N}\right)^2} \times i \\ &= \sqrt{\left(\frac{154}{100}\right) - \left(\frac{-14}{100}\right)^2} \times 10 \\ &= \sqrt{1.54 - 0.0196} \times 10 \\ &= 12.33 \end{aligned}$$

Karl Pearson coefficient of skewness,

$$\begin{aligned} S_{k_p} &= \frac{\text{Mean} - \text{Mode}}{\sigma} \\ &= \frac{33.6 - 35.55}{12.33} \end{aligned}$$

$$\therefore S_{k_p} = -0.158$$

49. From the following calculate skewness based on quartiles.

Income (₹)	Below 200	200 - 400	400 - 600	600 - 800	800 - 1000	1000 - Above
No. of Persons	25	40	80	75	20	16

Sol :

Bowley's co-efficient of skewness is based on quartiles.

Calculating of Bowley's Co-efficient of Skewness

	Income	No. of Persons	Cumulative Frequencies (C.F)
	0 – 200	25	25 CF
(Q ₁)	L 200 – 400	40 f	65 CF
(Median)	L 400 – 600	80 f	145 CF
(Q ₃)	L 600 – 800	75 f	220
	800 – 1000	20	240
	1000 – 1200	16	256
		Σf = 256(N)	

Bowley's coefficient of skewness is,

$$S_{KB} = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

Calculation of Median

Median = Size of $\left[\frac{N}{2} \right]^{\text{th}}$ observation

$$N = 256$$

Median $\left[\frac{256}{2} \right]^{\text{th}}$ observation

$$= 128^{\text{th}} \text{ observation}$$

Median lies in the class 400 – 600

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

where,

$$L = 400, \frac{N}{2} = 128, CF = 65, f = 80, i = 200$$

$$\text{Median} = 400 + \frac{128 - 65}{80} \times 200$$

$$= 400 + \frac{63}{80} \times 200$$

$$= 400 + 157.5$$

$$\boxed{\text{Median} = 557.5}$$

Calculation of Quartiles Q_3 and Q_1

$$Q_3 = \text{Size of } \left[\frac{3N}{4} \right]^{\text{th}} \text{ observation}$$

$$N = 256$$

$$Q_3 = \left[\frac{3(256)}{4} \right]^{\text{th}}$$

$$= \left[\frac{768}{4} \right]^{\text{th}}$$

$$= 192^{\text{th}} \text{ observation.}$$

Q_3 lies in the class 600 - 800

$$\boxed{Q_3 = L + \frac{\frac{3N}{4} - CF}{f} \times i}$$

Where,

$$L = 600, \frac{3N}{4} = 192, CF = 145, f = 75,$$

$$i = 200$$

$$Q_3 = 600 + \frac{192 - 145}{75} \times 200$$

$$= 600 + \frac{47}{75} \times 200$$

$$= 600 + 125.33$$

$$\boxed{Q_3 = 725.33}$$

$$Q_1 = \text{Size of } \left[\frac{N}{4} \right]^{\text{th}} \text{ observation}$$

$$N = 256$$

$$Q_1 = \text{Size of } \left[\frac{256}{4} \right]^{\text{th}} \text{ observation}$$

$$= 64^{\text{th}} \text{ observation}$$

Q_1 lies in the class 200 - 400

$$\boxed{Q_1 = L + \frac{\frac{N}{4} - \text{p.c.f}}{f} \times i}$$

Where,

$$L = 200, \frac{N}{4} = 64, CF = 25, f = 40, i = 200$$

$$= 200 + \frac{64 - 25}{40} \times 200$$

$$= 200 + \frac{39}{40} \times 200$$

$$= 200 + 195$$

$$\boxed{Q_1 = 395}$$

Now substituting the values of median, Q_3 and Q_1 in S_{KB} formula

$$\boxed{S_{KB} = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}}$$

$$Q_3 = 725.33, Q_1 = 395 \text{ and Median} = 557.5$$

$$= \frac{725.33 + 395 - 2(557.5)}{725.33 - 395}$$

$$= \frac{1120.33 - 1115}{330.33} = \frac{5.33}{330.33}$$

$$S_{KB} = 0.016$$

50. Calculate Coefficient of Skewness based on Quartiles.

Profit	Below 10	10 - 20	20 - 30	30 - 40	40 - 50	Above 50
No. of companies	5	12	20	16	5	2

Sol :

Bowley's coefficient of Skewness is based on quartiles.

Calculation of Bowley's Coefficient of Skewness

Profits	No. of Companies (f)	Cumulative Frequencies (C.f)
0 - 10	5	5
10 - 20	12	17
20 - 30	20	37
30 - 40	16	53
40 - 50	5	58
50 - 60	2	60
	$\Sigma f = 60(N)$	

Calculation of Median

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

$N = 60$

Median = $\left(\frac{60}{2}\right)^{\text{th}}$ observation
 = 30^{th} observation

Median lies in the class 20 – 30

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

Where,

L = Lower limit of the median class interval

N = Number of observations

cf = Preceding cumulative frequency to the median class

f = Frequency of the median class

i = Class interval of the median class

$L = 20, N = 60, cf = 17, f = 20$ and $i = 10$

$$\begin{aligned}
 \text{Median} &= 20 + \frac{60/2 - 17}{20} \times 10 \\
 &= 20 + \frac{30 - 17}{20} \times 10 \\
 &= 20 + \frac{13}{20} \times 10 = 20 + \frac{13}{2} \\
 &= 20 + 6.5 \\
 &= 26.5
 \end{aligned}$$

$$\therefore \text{Median} = 26.5$$

Calculation of Upper Quartile (Q_3) and Lower Quartile (Q_1)

$$Q_3 = \text{Size of } \left(\frac{3N}{4} \right)^{\text{th}} \text{ observation}$$

$$N = 60$$

$$Q_3 = \text{Size of } \left(\frac{3(60)}{4} \right)^{\text{th}} \text{ observation}$$

$$= \left(\frac{180}{4} \right)^{\text{th}} \text{ observation}$$

$$= 45^{\text{th}} \text{ observation}$$

Q_3 lies in the class 30 – 40

$$Q_3 = L + \frac{\frac{3N}{4} - CF}{f} \times i$$

$$L = 30, N = 60, cf = 37, f = 16 \text{ and } i = 10$$

$$Q_3 = 30 + \frac{3(60)/4 - 37}{16} \times 10$$

$$= 30 + \frac{8}{16} \times 10 = 30 + 5 = 35$$

$$\therefore Q_3 = 35$$

$$Q_1 = \text{Size of } \left(\frac{N}{4} \right)^{\text{th}} \text{ observation}$$

$$N = 60$$

$$Q_1 = \text{Size of } \left(\frac{60}{4} \right)^{\text{th}} \text{ observation}$$

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

Q_1 lies in the class 10 – 20

$$Q_1 = L + \frac{\frac{N}{4} - CF}{f} \times i$$

$$L = 10, N = 60, cf = 5, f = 12 \text{ and } i = 10$$

$$Q_1 = 10 + \frac{60/4 - 5}{12} \times 10$$

$$= 10 + \frac{15 - 5}{12} \times 10$$

$$= 10 + \frac{10}{12} \times 10$$

$$= 10 + \frac{100}{12} = 10 + 8.33$$

$$= 18.33$$

$$\therefore Q_1 = 18.33$$

By substituting median, Q_3 and Q_1 values in S_{KB} formula, we get,

$$S_{KB} = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

$$Q_3 = 35, Q_1 = 18.33 \text{ and Median} = 26.5$$

$$S_{KB} = \frac{35 + 18.33 - 2(26.5)}{35 - 18.33}$$

$$= \frac{53.33 - 53}{16.67}$$

$$= \frac{0.33}{16.67} = + 0.0197$$

$$\therefore S_{KB} = + 0.0197$$

- 51. Calculate percentile co-efficient of skewness from the following positional measures given below :**

$$P_{90} = 110, P_{10} = 50.67, \\ \text{Median} = 75.03$$

Sol :

Kelly's Coefficient of Skewness (using percentiles)

$$S_k = \frac{P_{10} + P_{90} - 2\text{Median}}{P_{90} - P_{10}} \\ = \frac{50.67 + 110 - 2(75.03)}{110 - 50.67} \\ = \frac{160.67 - 150.06}{59.33} \\ = \frac{10.61}{59.33} = 0.179$$

Kelly's Coefficient of Skewness = 0.179

Kelly's Coefficient of Skewness = 0.179

- 52. For a distribution , Karl Pearson's Coefficient of Skewness is 0.40. Its standard deviation is 8 and mean is 30. Find the mode of the distribution.**

Sol :

Karl Pearson's Coefficient of Skewness

$$= \frac{\bar{X} - \text{Mode}}{\sigma} = SK = \frac{30 - \text{Mode}}{8} \\ = 0.40 = 30 - \text{Mode} = 3.2 \\ \text{Mode} = 26.8$$

- 53. For a distribution , the arithmetic mean is 100, standard deviation is 35 and the Karl Pearson's Coefficient of Skewness is 0.2 Find the median.**

Sol :

$$= \frac{3(\bar{X} - \text{Median})}{s}$$

$$= SK = \frac{3(100 - \text{Median})}{35} = 0.2$$

$$= 300 - 3 \text{ Median} = 7, = 3$$

$$\text{Median} = 293, \text{Median} = 97.67$$

- 54. Find the value of standard deviation of Mean = 50, Mode = 25 and Coefficient of Skewness = 0.4**

Sol :

$$= \frac{\bar{X} - \text{Mode}}{\sigma} = \text{Coefficient of Skewness}$$

$$= \frac{50 - 25}{\sigma} = 0.4 = \frac{25}{\sigma} = 0.4$$

$$= 0.4, \sigma = 25, \sigma = 62.5$$

- 55. In a frequency distribution, the coefficient of skewness based on quartiles is 0.6. If the sum of the Upper and Lower Quartiles is 100 and Median is 38, find the value of Upper Quartile.**

Sol :

$$\text{Given S.K.} = 0.6, Q_3 + Q_1 = 100,$$

$$\text{Median} = 38$$

Coefficient of Skewness

$$= \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1} = \frac{100 - 76}{Q_3 - Q_1}$$

$$= 0.6 = \frac{24}{Q_3 - Q_1} = 0.6$$

$$= \frac{24}{Q_3 - Q_1} = 0.6 = 0.6 (Q_3 - Q_1) = 24$$

$$= Q_3 - Q_1 = \frac{24}{0.6} = 40$$

$$= Q_3 + Q_1 = 100, Q_3 - Q_1 = 40,$$

by adding the two we get

$$2 Q_3 = 140, Q_3 = 70$$

2.9 MOMENTS

Q58. What do you understand by moments?

Ans :

Definition

According to R. A. Fisher, 'A quantity of data which by its mere bulk may be incapable of entering the mind is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.'

Among these 'relatively few quantities' are those which are known as moments. Two of them, the mean and the variance, have already been discussed and two other higher moments are in common use. The higher moments are basically Used to describe the characteristic of populations rather than samples.

The measures of central tendency, variability and skewness which have been discussed to describe a frequency distribution, may be classified into two groups :

- i) Percentile system, and
- ii) Moment system

The percentile system includes measures like median, quartiles, deciles, percentiles, and so on. The value of these measures represents a given proportion of frequency distribution.

The moment system includes measures like mean, average deviation, standard deviation, and so on. The value of these measures is obtained by taking the deviation of individual observations from a given origin. The term 'moment' is used in physics and refers to the measure of a force which may generate rotation.

The possibility of generating such a force depends upon (i) the amount of force needed and (ii) the distance from the origin of the point at which the force is applied. The term moment used in statistics is analogous to the term used in physics, where

- i) size of class intervals represents the 'force' and
- ii) deviation of mid-value of each class from an observation represents the distance.

While calculating moments, if deviations are taken from the actual mean, then such moments are denoted by the Greek letter μ . (μ).

2.9.1 Moments About Mean

Q59. Explain about moments about mean.

Ans :

Let $x_1, x_2, \dots, x_n$ be the n observations in a data set with mean $\bar{x}$. Then the r th moment about the actual mean of a variable both for ungrouped and grouped data is given by :

$$\text{For ungrouped data : } \mu_r = \frac{1}{n} \sum (x - \bar{x})^r ;$$

$$r = 1, 2, 3, 4.$$

$$\text{For grouped data : } \mu_r = \frac{1}{n} \sum f(x - \bar{x})^r ;$$

$$r = 1, 2, 3, 4; N = \sum f_i$$

For different values of $r = 1, 2, 3, 4$, different moments can be obtained as shown below :

$$\text{Ungrouped data : } \mu_1 = \frac{1}{n} \sum (x - \bar{x})$$

$$= 0; \mu_2 = \frac{1}{n} \sum (x - \bar{x})^2 = \sigma^2 \text{ (variance)}$$

$$\mu_3 = \frac{1}{n} \sum (x - \bar{x})^3; \mu_4 = \frac{1}{n} \sum (x - \bar{x})^4$$

$$\text{For grouped data : } \mu_1 = \frac{1}{n} \sum f(x - \bar{x});$$

$$\mu_2 = \frac{1}{n} \sum f(x - \bar{x})^2$$

$$\mu_3 = \frac{1}{n} \sum f(x - \bar{x})^3;$$

$$\mu_4 = \frac{1}{n} \sum f(x - \bar{x})^4$$

The first moment, μ_1 about origin gives the mean and is a measure of central tendency.

$$\mu_1 = \frac{1}{n} \sum (x - 0) = \frac{1}{n} \sum x \leftarrow \text{A.M.}$$

The second moment, μ_2 about the mean indicates the symmetry or asymmetry of the distribution; its value is zero for symmetrical distribution.

$$\mu_2 = \frac{1}{n} \sum (x - \bar{x})^2$$

The fourth moment, μ_4 about the mean is a measure of Kurtosis (or flatness) of the frequency curve.

$$\mu_4 = \frac{1}{n} \sum (x - \bar{x})^4 \leftarrow \text{Kurtosis}$$

2.9.2 Moments about Arbitrary Point

Q60. Explain about moments about arbitrary point.

Ans :

When actual mean is in fractions, moments are first calculated about an assumed mean, say A, and then are converted about the actual mean, as shown below :

$$\text{For grouped data : } \mu'_r = \frac{1}{n} \sum f(x - A)^r; \\ r = 1, 2, 3, 4$$

$$= \frac{1}{n} \sum f d^2 \times h^2, \text{ where}$$

$$d = \frac{x - A}{h} \text{ or } dh = x - A$$

For ungrouped data : μ'_r

$$= \frac{1}{n} \sum (x - A)^r; r = 1, 2, 3, 4$$

For $r = 1$, we have

$$\mu'_r = \frac{1}{n} \sum (x - A) = \frac{1}{n} \sum x - A = \bar{x} - A$$

2.9.3 Moments about Zero or Origin

Q61. Explain about moments about zero or origin.

Ans :

The moments about zero or origin are obtained as follows :

$$v_r = \frac{1}{n} \sum f x^r; r = 1, 2, 3, 4$$

The relationship among moments about zero and other moments is as follows :

$$v_1 = A + \mu'_1, v_2 = \mu_2 + (v_1)^2$$

$$v_2 = \mu_3 + 3v_1v_2 - 2v_1^3,$$

$$v_4 = \mu_4 + 4v_1v_3 - 6v_1^2v_2 + 3v_1^4$$

2.9.4 Relationship between Central Moments and Moments about any Arbitrary point

Q62. What is the relationship between central moments and moments about any Arbitrary point ?

Ans :

$$\begin{aligned}\mu_r &= \frac{1}{n} \Sigma (x - \bar{x})^r = \frac{1}{n} \Sigma \{x - A - (\bar{x} - A)\}^r \\ &= \frac{1}{n} \Sigma (x - A - \mu'_1)^r; \mu_1 = \bar{x} - A \\ &= \frac{1}{n} [S(x - A)^r - {}^r C_1 \mu'_1 \mu'_{r-1} \\ &\quad + {}^r C_2 (\mu'_1)^2 \Sigma (x - A)^{r-2} + \dots \\ &\quad + (-1)^r (\mu'_1)^r] \\ &= \mu'_r - {}^r C_1 \mu'_1 \mu'_{r-1} + \\ &\quad {}^r C_2 (\mu'_1)^2 \mu'_{r-2} + \dots + (-1)^r (\mu'_1)^r\end{aligned}$$

From Equation (5 - 5) for various values of r, we have

$$\begin{aligned}\mu_1 &= \mu'_1; r = 1, \mu_2 = \mu'_2 - (\mu'_1)^2; r = 2 \\ \mu_3 &= \mu'_3 - 3\mu'_1 \mu'_2 + 2(\mu'_1)^3; r = 3 \\ \mu_4 &= \mu'_4 - 4\mu'_1 \mu'_3 + 6\mu'_2 (\mu'_1)^2 - 3(\mu'_1)^3; \\ &\quad r = 4\end{aligned}$$

2.9.5 Moments in Standard Units

Q63. Explain about moments in standard units.

Ans :

When expressed in standard units, moments about the mean of the population are usually denoted by the Greek letter α (alpha). Thus

$$\alpha_r = \frac{1}{n} \Sigma f z^r = \frac{1}{n} \Sigma f \left\{ \frac{x - \mu}{\sigma} \right\}^r$$

by definite of z

$$= \frac{x - \mu}{\sigma} = \frac{1}{\sigma^r} \frac{1}{n} \Sigma f (x - \mu)^r = \frac{\mu_r}{\sigma^r}$$

Hence, $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = \frac{\mu_3}{\sigma^3}$ and $\alpha_4 =$

$\frac{\mu_4}{\sigma^4}$ for $r = 1, 2, 3$, and 4 respectively.

In the notations used by Karl Pearson, we have

$$(i) \beta_1 \text{ (Beta one)} = \alpha_3^2 = \frac{\mu_3^2}{\mu_2^2}$$

$$(ii) \beta_2 \text{ (Beta two)} = \alpha_4 = \frac{\mu_4}{\mu_2^2}$$

In the notations used by R.A. Fisher, we have

$$(i) \gamma_1 \text{ (Gamma one)}$$

$$= \alpha_3 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$

$$(ii) \gamma_2 \text{ (Gamma two)}$$

$$\begin{aligned}&= \alpha_4 - 3 \text{ or } \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 \\ &= \frac{\mu_4 - 3\mu_2^2}{\mu_2^2}\end{aligned}$$

All these coefficients are pure numbers and are independent of whatever unit the variable x may be expressed in. The values of α_3 and α_4 depend on the shape of the frequency curve and, therefore, can be used to distinguish between different shapes. Thus, α_3 or β_1 is a measure of asymmetry about the mean or skewness and $\beta_1 = 0$ for a symmetrical distribution. A curve with $\alpha_3 > 0$ is said to have positive skewness and one with $\alpha_3 < 0$, negative skewness. For most distributions, α_3 lies between -3 and 3.

The coefficient of skewness in terms of moments is given by

$$Sk = \frac{(\beta_2 + 3)\sqrt{\beta_1}}{2(5\beta_2 - 6\beta_1 - 9)}$$

If $\beta_1 = 0$ or $\beta_2 = -3$, then skewness is zero.

But $\beta_2 = \frac{\mu_4}{\mu_2^2}$ can not be zero and hence the only condition for skewness to be zero is $\beta_1 = 0$ and the coefficient has no sign.

2.9.6 Sheppard's Corrections for Moments

Q64. Explain about sheppard's corrections for moments.

Ans :

While calculating higher moments of grouped frequency distributions, it is assumed that frequencies are concentrated at the mid-values of class-intervals. However, it causes certain errors while calculating moments. W. E. Sheppard proved that, if

- the frequency curve of the distribution is continuous.
- the frequency tapers off to zero at both ends, and
- the member of classes are not too large.

then the above assumption that frequencies are concentrated at the mid-value of the class intervals is corrected using Shappard's corrections as follows:

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12}$$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2}h^2\mu_2 + \frac{7}{240}h_4$$

where h is the width of the class interval, μ_1 and μ_3 need no correction.

PROBLEM

56. The first four moments of a distribution about the origin are 1, 4, 10, and 46 respectively. Obtain the various characteristics of the distribution on the basis of the information given. Comment upon the nature of the distribution.

Sol :

In the usual notations, we have

$$A = 0, \mu'_1 = 1, \mu'_2 = 4, \mu'_3 = 10 \text{ and } \mu'_4 = 46$$

$$\bar{x} = \text{first moment about origin} = \mu'_1 = 1$$

Variance (σ^2)

$$= \mu_2 = \mu'_2 - (\mu'_1)^2 = 4 - 1 = 3$$

$$\text{S.D. } (\sigma) = \sqrt{\sigma^2} = \sqrt{3} = 1.732$$

$$\begin{aligned} \mu_3 &= \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3 \\ &= 10 - 3(4)(1) + 2(1)^3 = 0 \end{aligned}$$

Karl Pearson's coefficient of skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ or } \gamma_1 = \frac{\mu_3}{\sqrt{\beta_1}} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$

Substituting values in the above formula, we get $\gamma_1 = 0$ ($\beta_1 = 0$). This shows that the given distribution is symmetrical, and hence Mean = Median = Mode for the given distribution.

57. The first three moments of a distribution about the value 1 of the variable are 2, 25 and 80. Find the mean, standard deviation and the moment-measure of skewness.

Sol :

From the data of the problem, We have

$$\mu'_1 = 2, \mu'_2 = 25, \mu'_3 = 80 \text{ and } A = 1.$$

The moments about the arbitrary point A = 1 are calculated as follows :

$$\text{Mean, } \bar{x} = \mu'_1 + A = 2 + 1 = 3$$

$$\text{Variance, } \mu_2 = \mu'_2 - (\mu'_1)^2 = 25 - (2)^2 = 21$$

$$\begin{aligned} \mu_3 &= \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3 \\ &= 80 - 3(25) + 2(2)^2 = -54 \end{aligned}$$

Standard deviation,

$$\sigma = \sqrt{\mu_2} = \sqrt{21} = 4.58$$

Moment-measure of skewness,

$$\begin{aligned}\gamma_1 &= \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-54}{(21)^{3/2}} \\ &= \frac{-54}{96.234} \\ &= -0.561\end{aligned}$$

2.10 KURTOSIS

Q65. What is Kurtosis?

Ans :

Kurtosis in statistics can be referred to as another measure for defining the frequency distribution in addition to central tendency, dispersion and skewness. It enables us to measure the flatness or peakedness of the curve. The knowledge about kurtosis and skewness is essential for working of statistical procedures because there exist some procedures that do not support skewed data or non-mesokurtic data.

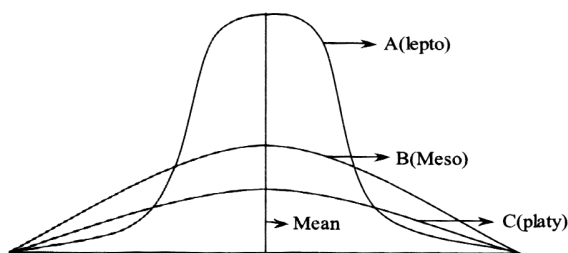
Formulae for Measuring Kurtosis

Kurtosis means convexity of a frequency curve. It gives us an idea regarding the flatness or peakedness of the curve. Kurtosis is measured by making use of the coefficients β_2 or γ_2

The formula for β_2 or γ_2 is as follows,

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$\gamma_2 = \beta_2 - 3$$



Figure

Leptokurtic (Leptokurtosis)

Curve 'A' is more sharply peaked than the normal curve 'S'. Hence, curve A is called leptokurtic curve. For leptokurtic curve $\beta_2 > 3$ and $\gamma_2 > 0$.

Mesokurtic (Mesokurtosis)

Curve 'S' is neither flat nor sharply peaked. Hence, curve 'S' is called normal curve or mesokurtic curve. For a normal curve $\beta_2 = 3$ and $\gamma_2 = 0$.

Platykurtic (Platykurtosis)

Curve 'C' is flatter than the normal curve 'S'. Hence, curve 'C' is called platykurtic. For platykurtic curve $\beta_2 = 3$ and $\gamma_2 = 0$.

PROBLEMS

58. The first four central moments of distribution are 0, 2.5, 0.7, and 18.75. Comment on the skewness and kurtosis of the distribution.

Sol :

Testing Skewness

We are given $\mu_1 = 0$, $\mu_2 = 2.5$,
 $\mu_3 = 0.7$ and $\mu_4 = 18.75$

Skewness is measured by the coefficient β_1 ,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \mu_2 = 2.5, \mu_3 = 0.7$$

$$\therefore \beta_1 = \frac{(0.7)^2}{(2.5)^3} = 0.031$$

Since $\beta_1 = 0.031$, the distribution is slightly skewed, i.e., it is not perfectly symmetrical.

Testing Kurtosis :

For testing kurtosis we compute the value of β_2 . When a distribution is normal or symmetrical, $\beta_2 = 3$. When distribution is more peaked than the normal, β_2 is more than 3 and when it is less peaked than the normal, β_2 is less than 3.

$$\beta_2 = \frac{\mu_4}{\mu_2^2}; \text{ where } \mu_4 = 18.75, \mu_2 = 2.5$$

$$\beta_2 = \frac{18.75}{(2.5)^2} = \frac{18.75}{6.25} = 3$$

59. Using moments, calculate a measure of relative skewness and a measure of relative kurtosis for the following distribution and comment on the result obtained :

Daily wages (in rupees)	No. of Workers
70 but below 90	8
90 but below 110	11
110 but below 130	18
130 but below 150	9
150 but below 170	4

Sol :

Daily Wages (Rs.)	m.p.m	f	(m-120)/10 d	fd	fd ²	fd ³	fd ⁴
70 - 90	80	8	-2	-16	32	-64	128
90 - 110	100	11	-1	-11	11	-11	11
110 - 130	120	18	0	0	0	0	0
130 - 150	140	9	+1	+9	9	+9	9
150 - 170	160	4	+2	+8	16	+32	64
		N = 50		Σfd = -10	Σfd ² = 68	Σfd ³ = -34	Σfd ⁴ = 212

$$\mu_1' = \frac{\Sigma fd}{N} \times i = \frac{-10}{50} \times 10 = -2$$

$$\mu_2' = \frac{\Sigma fd^2}{N} \times i^2 = \frac{68}{50} \times 100 = 136$$

$$\mu_3' = \frac{\Sigma fd^3}{N} \times i^3 = \frac{-34}{50} \times 1000 = -680$$

$$\mu_4' = \frac{\Sigma fd^4}{N} \times i^4 = \frac{212}{50} \times 10000 = 42400$$

Convert moments about arbitrary origin to moments about mean

$$\mu_2 = \mu_2' - (\mu_1')^2 = 136 - (-2)^2 = 132$$

$$\begin{aligned}\mu_3 &= \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 \\ &= -680 - 3(-2)(-680) + 2(-2)^3 = -680 + 816 - 16 = 120\end{aligned}$$

$$\begin{aligned}\mu_4 &= \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2\mu_2' - 3\mu_1'^4 \\ &= 42400 - 4(-2)(-680) + 6(-2)^2(136) - 3(-2)^4 \\ &= 42400 - 5440 + 3264 - 48 = 40176\end{aligned}$$

$$\beta_1 = \frac{\mu_3'^2}{\mu_2'^3} = \frac{(120)^2}{(132)^3} = \frac{14400}{2299968} = 0.006$$

$$\beta_2 = \frac{\mu_4'}{\mu_2'^2} = \frac{40176}{(132)^2} = \frac{40176}{17424} = 2.306$$

Rahul Publications

Exercises Problems

1. Compute Mean and Median from the following data

Sales (₹)	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of companies	6	10	20	25	16	12	5

[Ans: Arithmetic Mean : 34.68, Median = 34.40]

2. Given that the median sale of shops is Rs. 2,400 find out the missing frequency :

Sales (Rs.'00s)	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of Shops	5	25	?	18	7

[Ans: 25]

3. Calculate mode from the following data :

Class interval	Frequency
0 - 10	5
10 - 20	15
20 - 30	30
30 - 40	55
40 - 50	55
40 - 50	85
50 - 60	60
60 - 70	45
70 - 80	25

[Ans: 45.45]

4. Calculate Co-efficient of Quartile Deviation from the following data :

Class Interval	Frequency	Class interval	Frequency
10 - 15	4	30 - 40	10
15 - 20	12	40 - 50	8
20 - 25	16	50 - 60	6
25 - 30	22	60 - 70	4

[Ans: Coefficient of Q.D. = 0.27]

5. Find the Co-efficient of Quartile Deviation from the following Table :

Size	4 – 8	8 – 12	12 – 16	16 – 20	20 – 24	24 – 28	28 – 32	32 – 36	36 – 40
Frequency	6	10	18	30	15	12	10	6	2

[Ans: Q.D. = 5.2085, Coefficient of Q.D. = 0.264]

6. Calculate Karl Pearson's Coefficient of Skewness from the following data

Class	Frequency
70 - 80	18
60 - 70	22
50 - 60	30
40 - 50	35
30 - 40	21
20 - 30	11
10 - 20	6
0 - 10	5

[Ans: 0.0456]

Short Question and Answers

1. Define the term average.

Ans :

The average of a distribution has been defined in various ways. Some of the important definitions are:

- (i) **According to Clark and Sekkade**, "An average is an attempt to find one single figure to describe the whole of figures".
- (ii) **According to Murray R. Spiegel**, "Average is a value which is typical or representative of a set of data".
- (iii) **According to Croxton and Cowden**, "An average, is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of data it is sometimes called a measure of central value".
- (iv) **According to Sipson and Kafka**, "A measure of central tendency is a typical value around which other figures congregate".

2. Merits of Arithmetic Average.

Ans :

1. It is rigidly defined. Hence, different interpretations by different persons are not possible.
2. It is easy to understand and easy to calculate. In most of the series it is determinate and its value is definite.
3. It takes all values into consideration. Thus, it is more representative.
4. It can be subjected to further mathematical treatment. The properties of Arithmetic mean are separately explained elsewhere in the chapter.
5. It is used in the computation of various other statistical measures.
6. It is possible to calculate arithmetic average even if some of the details of the data are lacking.

For example, it can be known even when only the number of items and their aggregate value are known, and details of various items are not available. Similarly, the aggregate value of items can be calculated if the number of items and the average are known.

3. Combined mean.

Ans :

It is the arithmetic averages and the number of items in two or more related groups are known as the combined or the composite mean of the entire group can be obtained by,

$$\text{Combined Mean } (\bar{X}) = \left[\frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2} \right]$$

With the help of combined arithmetic mean, we can determine the over all mean of the combined data without going back of the original data.

4. Define Geometric Mean.

Ans :

Geometric mean is defined as the N^{th} root of the product of N items or values.

When the number of observations is more than 2 the calculation of n^{th} root becomes difficult. In order to facilitate these calculations the logarithms are used.

In case of 2 or 3 observations the Geometric mean is calculated as follows :

$$\begin{aligned} \text{G.M} &= \sqrt[n]{x_1 \times x_2 \times x_3 \times \dots \times x_n} \\ &= (x_1 \times x_2 \times \dots \times x_n)^{1/n} \end{aligned}$$

Whereas, when the number of observations are more the geometric mean is calculated as follows,

$$\log \text{G.M} = \frac{\log x_1 + \log x_2 + \dots + \log x_n}{N}$$

(or)

$$\log G.M = \left(\frac{\sum \log x}{N} \right)$$

$$\therefore G.M = \text{Antilog} \left(\frac{\sum \log x}{N} \right)$$

5. State the merits and demerits of geometric mean.

Ans :

Merits :

1. It is a rigidly defined average.
2. It is based on all the observations.
3. It is capable of mathematical treatment. If any two out of the three values, i.e.,
 - (i) Product of observations,
 - (ii) GM of observations and
 - (iii) Number of observations, are known, the third can be calculated.
4. In contrast to AM, it is less affected by extreme observations.
5. It gives more weights to smaller observations and vice-versa.

Demerits

1. It is not very easy to calculate and hence not very popular.
2. Like AM, it may be a value which does not exist in the set of given observations.
3. It cannot be calculated if any observation is zero or negative.

6. Advantages of Median.

Ans :

- (i) The median, unlike the mean, is unaffected by the extreme values of the variable.
- (ii) It is easy to calculate and simple to understand, particularly in a series of individual observations and a discrete series.
- (iii) It is capable of further algebraic treatment. It is used in calculating mean deviation.

- (iv) It can be located by inspection, after arranging the data in order of magnitude.
- (v) It can be determined graphically.
- (vi) Median can be calculated in case of open-end classes.
- (vii) Median is defined rigidly.

7. Discuss briefly about Quartiles.

Ans :

The values of a variable that divide a distribution into four equal parts are called quartiles. Since three values are needed to divide a distribution into four parts, there are three quartiles, viz. Q_1 , Q_2 and Q_3 , known as the first, second and the third quartile respectively.

For a discrete distribution, the first quartile (Q_1) is defined as that value of the variate such that at least 25% of the observations are less than or equal to it and at least 75% of the observations are greater than or equal to it.

For a continuous or grouped frequency distribution, Q_1 is that value of the variate such that the area under the histogram to the left of the ordinate at Q_1 is 25% and the area to its right is 75%. The formula for the computation of Q_1 can be written by making suitable changes in the formula of median.

8. Discuss briefly about deciles.

Ans :

Deciles divide a distribution into 10 equal parts and there are, in all, 9 deciles denoted as D_1 , D_2 , ..., D_9 respectively.

For a discrete distribution, the i th decile D_i is that value of the variate such that at least $(10i)\%$ of the observations are less than or equal to it and at least $(100 - 10i)\%$ of the observations are greater than or equal to it ($i = 1, 2, \dots, 9$).

For a continuous or grouped frequency distribution, D_i is that value of the variate such that the area under the histogram to the left of the ordinate at D_i is $(10i)\%$ and the area to its right is $(100 - 10i)\%$.

Formula

In individual and discrete series

$$D_8 = \text{Size of } \frac{8(N+1)^{\text{th}}}{10} \text{ item}$$

In Continuous Series

$$D_8 = \text{Size of } \frac{8N^{\text{th}}}{10} \text{ item}$$

9. State the merits and demerits of mode.

Ans :

Merits

1. It is easy to understand and easy to calculate. In many cases it can be located just by inspection.
2. It can be located in situations where the variable is not measurable but categorization or ranking of observations is possible.
3. Like mean or median, it is not affected by extreme observations. It can be calculated even if these extreme observations are not known.
4. It can be determined even if the distribution has open end classes.
5. It can be located even when the class intervals are of unequal width provided that the width of modal and that of its preceding and following classes are equal.
6. It is a value around which there is more concentration of observations and hence the best representative of the data.

Demerits

1. It is not based on all the observations.
2. It is not capable of further mathematical treatment.
3. In certain cases mode is not rigidly defined and hence, the important requisite of a good measure of central tendency is not satisfied.
4. It is much affected by the fluctuations of sampling.
5. It is not easy to calculate unless the number of observations are sufficiently large and

reveal a marked tendency of concentration around a particular value.

6. It is not suitable when different items of the data are of unequal importance.
7. It is an unstable average because, mode of a distribution, depends upon the choice of width of class intervals.

10. What are the properties of dispersion ?

Ans :

A good measure of dispersion should possess, as far as possible, the following properties :

- i) It should be simple to understand.
- ii) It should be easy to compute.
- iii) It should be rigidly defined.
- iv) It should be based on each and every item of the distribution.
- v) It should be amenable to further algebraic treatment.
- vi) It should have sampling stability.
- vii) It should not be unduly affected by extreme items.

11. Absolute Measures of Dispersion.

Ans :

In case of absolute measures of dispersion, the statistical units are taken in the same manner as how it is given in the actual data. Like-kilograms, rupees, tonnes and so on. These statistical values are then utilized to compare the variations between the two distributions. The statistical units expressed in the two distributions and the average is the two distributions and the average size of the distributions should be same for the purpose of comparison.

The various absolute measures of dispersion are,

- (i) Range
- (ii) Quartile Deviation (QD)
- (iii) Mean Deviation (MD)
- (iv) Standard Deviation (SD).

12. Relative Measures of Dispersion**Ans :**

If the average size of the distributions and statistical units expressed in the distributions differ from each other under such a situation 'Relative Measures of Dispersion' is used. For example, one distribution is expressed in quintals of rice and another distribution is expressed in kilograms of wheat or average size of one distribution is higher than the average size of another distribution (salaries paid to managers and wages paid to workers). Thus, relative measures of dispersion helps to compare variability of two or more distribution of distinct units. Relative measures of dispersion is also known as 'Coefficient of Dispersion'.

The various relative measures of dispersion are,

- (i) Coefficient of Variation (C.V)
- (ii) Coefficient of Quartile Deviation (CQD)
- (iii) Coefficient of Mean Deviation (CMD)
- (iv) Coefficient of Range (CR)

13. What is quartile deviations ?**Ans :**

The range as a measure of dispersion discussed above has certain limitations. It is based on two extreme items and it fails to take account of the scatter within the range. From this there is reason to believe that if the dispersion of the extreme items is discarded, the limited range thus established might be more instructive.

For this purpose there has been developed a measure called the interquartile range, the range which includes the middle 50 per cent of the distribution. That is, one quarter of the observations at the lower end, another quarter of the observations at the upper end of the distribution are excluded in computing the interquartile range. In other words, interquartile range represents the difference between the third quartile and the first quartile.

$$\text{Inter quartile range} = Q_3 - Q_1$$

Very often the interquartile range is reduced to the form of the Semi-interquartile range or quartile deviation by dividing it by 2.

$$\text{Quartile Deviation or Q.D.} = \frac{Q_3 - Q_1}{2}$$

14. State the merits and demerits of Mean Deviation.**Ans :****Merits**

1. It is easy to understand and easy to compute.
2. It is based on all the observations.
3. It is less affected by extreme observations vis-a-vis range or standard deviation (to be discussed in the next section).
4. It is not much affected by fluctuations of sampling.

Demerits

1. It is not capable of further mathematical treatment. Since mean deviation is the arithmetic mean of absolute values of deviations, it is not very convenient to be algebraically manipulated. This necessitates a search for a measure of dispersion which is capable of being subjected to further mathematical treatment.
2. It is not well-defined measure of dispersion since deviations can be taken from any measure of central tendency.

15. What is quartile deviations ?**Ans :**

Distinguish between dispersion and skewness.

Ans :**Difference between Dispersion and Skewness**

1. Dispersion is concerned with measuring the amount of variation in a series. Skewness is concerned with direction of variation or the departure from symmetry.

2. Dispersion gives us the extent to which the values of a given series are scattered. Skewness explains the extent and direction in which the distribution differs from symmetry.
3. Dispersion tells us about the composition of the series. Skewness tells us about the shape of the series.
4. Dispersion helps us to ascertain the extent to which a central value is representative of the series. Skewness deals with the nature of variations on either side of the central value.
5. Measures of Dispersion are based on averages of the first order (i.e. measures of central tendency). They are averages of the second order. Measures of Skewness are based on averages of first and second order (i.e. measures of central tendency and dispersion). Measures of Skewness are not averages at all.

16. Moments.

Ans :

Definition

According to R. A. Fisher, 'A quantity of data which by its mere bulk may be incapable of entering the mind is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.'

Among these 'relatively few quantities' are those which are known as moments. Two of them, the mean and the variance, have already been discussed and two other higher moments are in common use. The higher moments are basically Used to describe the characteristic of populations rather than samples.

The measures of central tendency, variability and skewness which have been discussed to describe a frequency distribution, may be classified into two groups :

- i) Percentile system, and
- ii) Moment system

Choose the Correct Answers

1. The sum of deviations taken from mean is : [a]
(a) Always equal to zero (b) Some times equal to zero
(c) Never equal to zero (d) Less than zero
2. The sum of the squares of the deviations of the values of a variable is least when the deviations are measured from : [d]
(a) Harmonic mean (b) Geometric mean
(c) Median (d) Arithmetic mean
3. Step deviation method or coding method is used for computation of the : [a]
(a) Arithmetic mean (b) Geometric mean
(c) Weighted mean (d) Harmonic mean
4. If the arithmetic mean of 20 values is 10, then sum of these 20 values is: [c]
(a) 10 (b) 20
(c) 200 (d) 20 + 10
5. Ten families have an average of 2 boys. How many boys do they have together? [d]
(a) 2 (b) 10
(c) 12 (d) 20
6. If the arithmetic mean of the two numbers X_1 and X_2 is 5 if $X_1 = 3$, then X_2 is : [c]
(a) 3 (b) 5
(c) 7 (d) 10
7. When the values in a series are not of equal importance, we calculate the: [c]
(a) Arithmetic mean (b) Geometric mean
(c) Weighted mean (d) Mode
8. When all the values in a series occur the equal number of times, then it is not possible to calculate the : [d]
(a) Arithmetic mean (b) Geometric mean
(c) Harmonic mean (d) Weighted mean
9. The mean for a set of data obtained by assigning each data value a weight that reflects its relative importance within the set, is called : [c]
(a) Geometric mean (b) Harmonic mean
(c) Weighted mean (d) Combined mean
10. The arithmetic mean of 10 items is 4 and the arithmetic mean of 5 items is 10. The combined arithmetic mean is : [c]
(a) 4 (b) 5
(c) 6 (d) 90

Fill in the Blanks

1. Any measure indicating the centre of a set of data, arranged in an increasing or decreasing order of magnitude, is called a measure of _____
2. Scores that differ greatly from the measures of central tendency are called _____
3. The measure of central tendency listed below is _____
4. The total of all the observations divided by the number of observations is called _____
5. While computing the arithmetic mean of a frequency distribution, the each value of a class is considered equal to _____
6. Change of origin and scale is used for calculation of the _____
7. The sample mean is a _____
8. The population mean μ is called _____
9. The arithmetic mean is highly affected by:
10. If a constant value is added to every observation of data, then arithmetic mean is obtained by _____

ANSWERS

1. Central tendency
2. Extreme scores
3. The mean
4. Arithmetic mean
5. Lower limit
6. Arithmetic mean
7. Statistic
8. Parameter
9. Extremely large values
10. Adding the constant

UNIT III

Index Numbers

Index Numbers - Introduction – Types – Characteristics – Construction weighted and unweighted index numbers – Price and Quantity/Volume index numbers– Tests – time reversal – Factor Reversal and Circular tests – Chain and Fixed base – Changing of base – Combining of two of more overlapping indices consumer price Index – Problems in Construction.

3.1 INDEX NUMBERS

3.1.1 Introduction

Q1. What do you understand by index numbers ?

Ans : (May-18)

Some prominent definitions of index numbers are given below:

According to Croxton & Cowden “Index numbers are devices for measuring differences in the magnitude of a group of related variables.”

According to Spiegel “An index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc.”

According to Patterson “In its simplest form an index number is the ratio of two index numbers expressed as a per cent. An index number is a statistical measure a measure designed to show changes in one variable or in a group of related variables over time, or with respect to geographic location, or other characteristic.”

According to Morris Hamburg “In its simplest form, an index number is nothing more than a relative number, or a ‘relative’ which expresses the relationship between two figures, where one of the figures is used as a base.”

According to Berenson and Levine “Generally speaking, index numbers measure the Size or magnitude of some object at a particular point in time as a percentage of some base or reference object in the past.”

For a proper understanding of the term index number, the following points are worth considering :

- (i) Index numbers are specialized averages. As explained in the chapter on measures of central value, an average is a single figure representing a group of figures. However, to obtain an average the items must be comparable: for example, the average weight of men, women and children of a certain locality has no meaning at all. Furthermore, the unit of measurement must be the same for all the items. Thus an average of the weight expressed in kg., lb., etc., has no meaning. However, this is not so with index numbers. Index numbers are used for purposes of comparison in situations where two or more series are expressed in different units or the series are composed of different types of items. For example, while constructing a consumer price index the various items are divided into broad heads, namely (i) Food, (ii) Clothing, (iii) Fuel and Lighting, (iv) House Rent, and (v) Miscellaneous. These items are expressed in different units : thus, under the head ‘food’ wheat and rice may be quoted per quintal, ghee per kg., etc. Similarly, cloth may be measured in terms of metres. An average of all these items expressed in different units is obtained by using the technique of index numbers.
- (ii) Index numbers measure the net change in a group of related variables. Since index numbers are essentially averages they describe in one single figure the increase or

decrease in a group of related variables under study. The group of variables may be the prices of a specified set of commodities, the volume of production in different sectors, etc. Thus, if the consumer price index of working class for Delhi has gone up to 113 in February 2000 compared to February 1999 it means that there is a net increase of 13 per cent in the prices of commodities included in the index.

- (iii) Index numbers measure the effect of changes over a period of time. Index numbers are most widely used for measuring changes over a period of time. Thus we can find out the net change in agricultural prices from the beginning of First Plan period to the end of the Eighth Plan period, i.e., from 1951 to 1996. Similarly, we can compare the agricultural production, industrial production, imports, exports, wages, etc., at two different times. However, it should be noted that index numbers not only measure changes over a period of time but also compare economic conditions of different locations, different industries, different cities or different countries.

3.1.2 Characteristics of index numbers

Q2. Explain the Characteristics of Index Numbers.

Ans : (May-17)

Characteristics of Index Numbers

The features of index numbers are as follows,

1. Measures the Change in Percentages

Index numbers measure the change in price or quantity in terms of percentages say i.e., (10%, 20%, 15%, 25% and so on). Increase or decrease in value is represented by one single figure. Like 10% increase in sales from that previous year to current year, 30% decrease in profits when compared to that of the last year and so on.

2. Specialized Averages

A single figure known as "Average" is used for representing the characteristics of the complete set of data. Average acts as a basis

for comparing different data sets with each other, when they have common unit of measurement of observations. When data sets do not have common unit of measurement, specialized averages of index numbers are used for the comparison.

3. The Measured Changes cannot be Observed Directly

Index numbers do not measure the changes directly but it studies the relative changes or variations in factors resulting to changes-like for measuring changes in Export-Imports related factors such as available raw materials, technology, competitors etc., are studied and measured.

4. Measures Changes in Relation to Time or Place

Index numbers measure change by comparing the values at different time periods or at different places like standard of living at one place is being compared with standard of living of the other place. Sales or revenue of current year is compared with that of the previous years sales or revenue.

Q3. Explain the uses of index numbers.

Ans : (Nov.-20, May-18)

Index numbers are indispensable tools of economic and business analysis. Their significance can be best appreciated by the following points:

(i) They help in framing suitable policies.

Many of the economic and business policies are guided by index numbers. For example, while deciding the increase in dearness allowance of the employees, the employers have to depend primarily upon the cost of living index. If wages and salaries are not adjusted in accordance with the cost of living, very often it leads to strikes and lock - outs which in turn cause considerable waste of resources. The index numbers provide some guideposts that one can use in making decisions.

Though index numbers are most widely used in the evaluation of business and economic conditions, there is a large number

of other fields also where index numbers are useful. For example, sociologists may speak of population indices; psychologists measure intelligence quotients which are essentially index numbers comparing a person's intelligence score with that of an average for his or her age; health authorities prepare indices to display changes in the adequacy of hospital facilities and educational research organizations have devised formulae to measure changes in the effectiveness of school systems.

(ii) They reveal trends and tendencies.

Since index numbers are most widely used for measuring changes over a period of time, the time series so formed enable us to study the general trend of the phenomenon under study. For example, by examining index number of imports for India for the last 8-10 years we can say that our imports are showing an upward tendency, i.e., they are rising year after year. Similarly, by examining the index numbers of industrial production, business activity, etc., for the last few years we can conclude about the trend of production and business activity. By examining the trend of the phenomenon under study we can draw very important conclusions as to how much change is taking place due to the effect of seasonality, cyclical forces, irregular forces, etc. Thus index numbers are highly useful in studying the general business conditions.

(iii) They are important in forecasting future economic activity.

Index numbers are useful not only in studying the past and present workings of our economy, but they are also important in forecasting future economic activity. Index numbers then are often used in time series analysis, the historical study of long-term trend, seasonal variations and business cycle development, so that business leader may keep pace with changing economic and business conditions and have better information available for decision-making purposes.

(iv) Index numbers are very useful in de-flating.

Index numbers are highly useful in deflating, i.e., they are used to adjust the original data for price changes, or to adjust wages for cost of living changes and thus transform nominal wages into real wages. Moreover, nominal income can be transformed into real income and nominal sales into real sales through appropriate index numbers.

Q4. Explain the limitations of index numbers

Ans : **(Nov.-20, May-18)**

Index Numbers are of great use in studying trends, events, pertaining to a defined phenomenon. However, they suffer from certain limitations. There are

1. Approximate Representation

Index Numbers are based on sample data. Hence, they are only approximate indicators. They may not fully reflect the changes in the relative level of a phenomenon.

2. Likelihood of Error

There is a possibility of error in the choice of the sample, selection of base period, data collection, and assignment of weights or the appropriate formula to be used.

3. Not Responsive

There are rapid changes in Technology, tastes, fashions, customs and consequently, the consumption pattern of various commodities. Index numbers may not be able to keep pace with such changes. Hence, they may not be able to reflect the changes in the phenomenon being studied.

4. No Universally Acceptable formula

There is no universally acceptable standard formula that can capture the changes in a phenomenon with perfection. None of the formulae seems to satisfy all the tests of consistency. Hence, some amount of formula error is present in any calculation of Index numbers.

5. Scope for manipulation

On account of wide variety of formulae index numbers can be manipulated by appropriate choice of base year, sample set, price and quantity quotations.

6. Other Limitations

Apart from the above, Index numbers cannot capture the aspects of quality, reliability service etc. There is no index that is universally applicable. An Index constructed for one purpose cannot be used for another. Index numbers being averages, they are subject to all limitations of an average. Lastly, the utility of Index Numbers is limited by the availability of adequate and accurate data.

3.1.3 Price and Quantity / Value index numbers

Q5. Define :

- (a) Price Index Numbers**
- (b) Quantity Index Numbers**
- (c) Value Index Numbers**

Ans :

Index numbers are classified into three types as follows,

- (a) Price index numbers
- (b) Quantity index numbers and
- (c) Value index numbers.

(a) Price Index Numbers

Price index method is used to compare the prices of various commodities of one year with that of the another year. It is further classified into two types,

- (i) Wholesale Price Index Numbers:** Wholesale price indices are used to study the changes that takes place in the general price level of a country.
- (ii) Retail Price Index Numbers:** It is used to study the changes that takes place in retail prices of various commodities like fruits, vegetables, rice, wheat etc.

(b) Quantity Index Numbers

Quantity index numbers measure the variation in the volume of goods manufactured or consumed or distributed in quantitative terms. It measures the change in quantitative terms like, increase (or) decrease in volume of production construction or employment.

(c) Value Index Numbers

Value index number are used to measure the changes that takes place in the total value of the variable. The total value is obtained by multiplying price (p) with quantity (q).

$$\text{Value} = \text{Price} \times \text{Quantity}$$

3.2 TYPES OF INDEX NUMBERS

3.2.1 Unweighted Price Indexes Weighted Price Indexes

Q6. Explain the various methods of constructing index numbers.

Ans :

(May-18, May-17)

The various methods of constructing index, numbers are shown in the following figure:

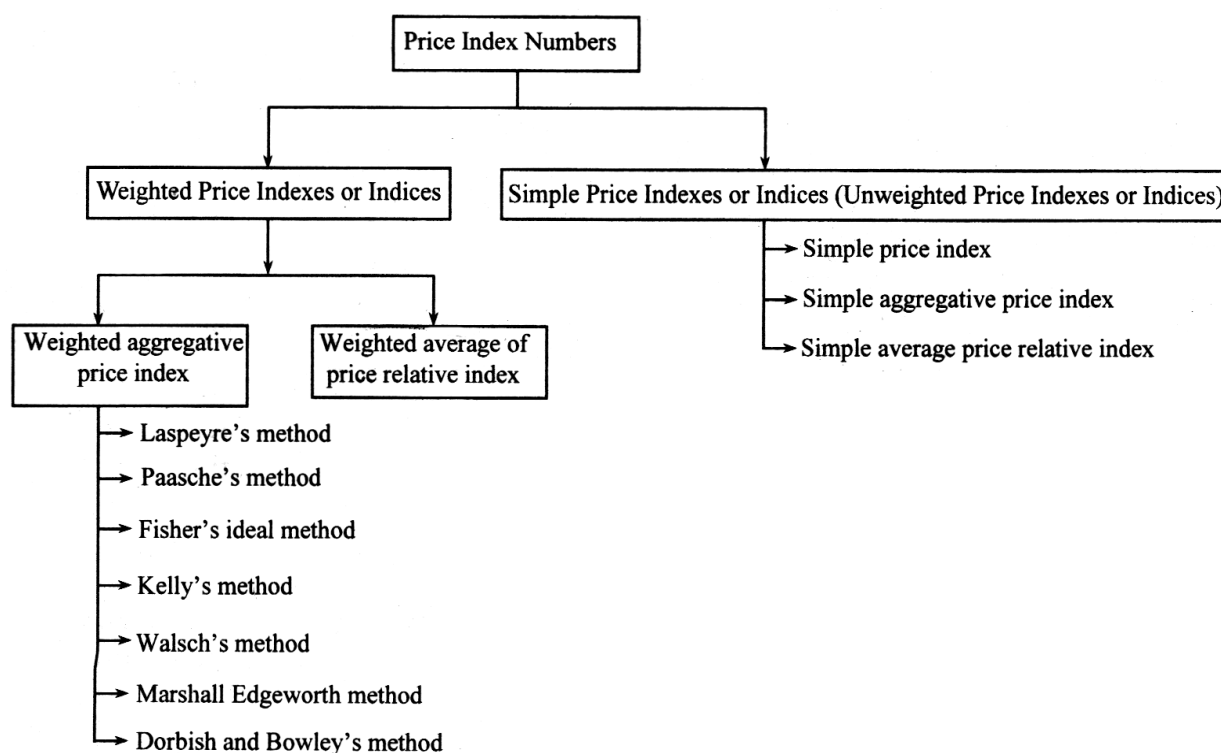


Fig. : Types and Methods of Price Index Numbers

1. Weighted Price Indexes

At the time of constructing the weighted price indexes or indices, the rational weights are allocated in an explicit manner. These rational weights show the relative significance of items or commodities which are related with the computation of an index. Quantity weights and value weights are used in this weighted indexes or indices. Weighted price indexes, or indices are further divided into two types as follows,

- (a) Weighted aggregate price index
- (b) Weighted average of price relative index.

(a) Weighted Aggregate Price Index

In a weighted aggregate price index, certain weight is assigned to each and every commodity or item of group in accordance with its significance. This helps in gathering more information and improving accuracy of the estimates. The following methods are used in weighted aggregate price index:

- (i) Laspeyre's method
- (ii) Pasche's method
- (iii) Fisher's ideal method
- (iv) Kelly's method
- (v) Walsch's method
- (vi) Marshal Edgeworth's method
- (vii) Dorbish and Bowley's method

(b) Weighted Average of Price Relative Index

In weighted average of price relative index, value of each commodity or item related with the calculation of composite index is ascertained by multiplying the price of each item with its quantity consumed. Quantity consumed is considered for computing the weighted average of price relative. The formula for weighted average of price relative index is as follows,

$$P_{01} = \frac{\sum ((p_1 \div p_0) \times 100)(p_0 q_0)}{\sum p_0 q_0} = \frac{\sum PV}{\sum V} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Where,

$V (= p_0 q_0)$ = Base prices and quantities determining values

$$P \left(= \frac{p_1}{p_0} \times 100 \right) = \text{Price relative}$$

This formula is equivalent to the formula of Laspeyre's index formula, If 'V' is taken as p_0, q_1 , then the formula would,

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Then it would be equal to Paasche's index method.

2. Unweighted Price Indexes or Indices

The unweighted indexes or indices include the following methods,

- (a) Simple price index
- (b) Aggregate price index
- (c) Average price relative index

(a) Simple/Single Price Index

Single price index is computed by dividing the current year price of the commodity with its base year price. It is a percentage ratio which represents the comparison of a particular commodity price. The general formula used for single price index is as follows,

$$\text{Single price index in period 'n'} = \frac{P_n}{P_0} \times 100$$

Where,

P_n = Price of the commodity in the n^{th} year

P_0 = Price of the commodity in the base year

(b) Simple Aggregate Price Index

In aggregate price index, the sum of current year prices of various commodities is divided with the sum of base year prices of that various commodities. The formula is given as follows,

$$\text{Aggregate price index, } P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

Where,

P_1 = Unit price of a current year prices of all commodities

P_0 = Sum of base year prices of all commodities

(c) Simple Average Price Relative Index

The method is an improvement over the aggregate price method is,

$$P_{01} = \sum \left(\frac{P_1}{P_0} \times 100 \right)$$

Where,

n = Number of commodities included in the computation of the index.

Q7. Explain Laspeyre's Method of index numbers.

Ans :

Laspeyre's Method

This method takes the quantities of the commodities in the base period as the weight of that commodity for the purpose of calculating the index numbers. The following steps may be following.

Step I

Multiply the current year price (represented by p_1) with the quantities of the base year (q_0) for each commodity.

Step 2

Add the numbers obtained in step 1. The resultant sum is represented as $\sum p_1 q_0$

Step 3

Multiply the prices of base year (represented by p_0) with the quantities of the base year for each commodity.

Step 4

Add the numbers obtained in step 3. The resultant sum is represented as $\sum p_0 q_0$

Step 5

The index number as per Laspeyre's method

$$= P_{0,1} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Q8. Explain Paasche's Method of index number.

Ans :

Paasche's Method

Paasche's Method is similar to Laspeyre's method. The only difference is in assignment of weights. As per this method quantities consumed of the commodities in the current year is taken as basis. The following steps need to be followed.

Step 1

Multiply current year's prices (p_1) with current year's quantities (q_1)

Step 2

Add the numbers obtained in step (1). The resultant sum is $\sum p_1 q_1$

Step 3

Multiply base year's Prices (p_0) with current year's quantities (q_1)

Step 4

Add the numbers obtained in step (2). The resultant sum is $\sum p_0 q_1$

Step 5

Index Number as per Paasche's method =

$$P_{0,1} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Q9. Explain Fisher's Ideal Index*Ans :***Fisher's Ideal Index :**

This is the most popular amongst all weighted aggregative index numbers. It is obtained by calculating the Geometric Mean (G.M) of Laspeyre's and Paasche's index numbers. The formula for calculating fisher's ideal index is an under.

$$P_{0.1} = \left[\sqrt{\frac{\sum p_1 q_0 \times \sum p_0 q_1}{\sum p_0 q_0 \times \sum p_1 q_1}} \right] \times 100$$

Reasons for Fisher's Index being called an Ideal Index:

We noticed that Fisher's index is called an Ideal index. The reasons for the same are as follows:

1. It gives weightage to both current consumption and base year consumption.
2. It is free from upward or downward bias.
3. It satisfies both time reversal and factor reversal tests (to be discussed later)
4. It is a Geometric mean of Laspeyre's index and Paasche's index.

Q10. Define :

- (a) Kellye's price index
- (b) Walsh price index
- (c) Dorbish Bowley price index

*Ans :***(a) Kellye's price index**

Kelly's Method emphasizes that the weights chosen for construction of an index number should be constant. While quantities are taken as weight, such quantities need not pertain to either base period or current period. The quantities may be an arithmetic or geometric mean of quantities of two or three years. They may also belong to some landmark year in recent history. The advantage of this method is that the weights need not be changed even if there is a change in base year. The formula for calculating Kelly's index is as follows.

$$P_{0.1} = \frac{\sum p_1 q}{\sum p_0 q} \times 100$$

Where q = fixed quantity weights

(b) Walsh price index

This method takes the Geometric Mean of quantities of base year and current year as weights. It is not a very popular method. The formula for calculating Walsh Index is an under.

$$P_{0.1} = \frac{\sum p_1 w}{\sum p_0 w} \times 100, \text{ Where } w = \sqrt{q_0 q_1}$$

(c) Dorbish Bowley price index

This method is a combination of Laspeyre's method and Paasche's method. It is obtained by Calculating the Arithmetic mean of Laspeyre's and Paasche's index numbers. It is also called 'L-P' formula. Thus, the index number is

$$P_{0.1} = \frac{1}{2} \left[\left(\frac{\sum p_1 q_0}{\sum p_0 q_0} \right) + \left(\frac{\sum p_1 q_1}{\sum p_0 q_1} \right) \right] \times 100$$

Q11. Define Marshall - Edge Worth's Index Method.

Ans :

Marshall-Edgeworth's Index Method

In Marshall-Edgeworth's index method, both the current year as well as base year quantities are considered as to calculate the index. The formula for Marshall-Edgeworth's Index method is as follows,

$$I_p(\text{ME}) = \frac{\sum(q_0 + q_1)P_1}{\sum(q_0 + q_1)P_0} \times 100$$

$$= \left(\frac{\sum(P_1 q_0 + \sum P_1 q_1)}{\sum q_0 q_0 + \sum P_0 q_1} \right) \times 100$$

Advantages of Marshall Edgeworth's Index Method

Following are the advantages of Marshall-Edgeworth's index method,

1. This method is easy to understand and calculate.
2. This method uses current and base year's quantities and prices.
3. This method satisfies both the unit and time reversal tests of consistency

Disadvantages of Marshall Edgeworth's Index Method

Following are the disadvantages of Marshall- Edgeworth's index method,

1. This method needs current weights while constructing index numbers.
2. This method is not satisfactory for circular and factor reversal test of consistency.

PROBLEMS ON INDEX NUMBER

1. From the following data, construct an Index number for 1994 taking 1993 as base as per simple Aggregative method

Commodities	Price in 1993 (Rs)	Price in 1994 (Rs)
A	40	60
B	60	90
C	85	125
D	25	35
E	30	40

Sol :

Construction of Price Index

Commodities	Price in 1993 in Rs. (P_0)	Price in 1994 in Rs. (P_1)
A	40	60
B	60	90
C	85	125
D	25	35
E	30	40
	$\Sigma P_0 = 240$	$\Sigma P_1 = 350$

$$P_{0,1} = \left(\frac{\Sigma p_1}{\Sigma p_0} \right) \times 100 = \frac{350}{240} \times 100 = 145.83$$

2. Following are the prices of commodities in 1970 and 1975. Calculate a price index based on price relatives using the arithmetic mean as well as geometric mean.

Year	Commodity					
	A	B	C	D	E	F
1970	45	60	20	50	85	120
1975	55	70	30	75	90	130

Sol :

(a) Arithmetic Mean

Commodity	P_0 (Price in 1970)	P_1 (Price in 1971)	Price Relative P	Log P
A	45	55	$55/45 \times 100 = 122.22$	2.0871
B	60	70	116.67	2.0668
C	20	30	150.00	2.1761
D	50	75	150.00	2.1761
E	85	90	105.88	2.0245
F	120	130	108.33	2.0346
			$\Sigma P = 753.10$	12.565.2 $= \Sigma \text{Log} p$

$$\text{Average of Price Relatives} = \frac{753.10}{6} = 125.52$$

(b) Geometric Mean

$$P_{0,1} = \text{Antilog} \left[\frac{\Sigma \text{Log} p}{N} \right] = \text{Antilog} \left(\frac{12.5652}{6} \right) = \text{Antilog} (2.0942) = 124.3$$

3. Construct index numbers of price from the following data by applying :

1. Laspeyres method
2. Paasche method
3. Fisher's ideal method

Commodity	1999		2000	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Sol :

Calculation of Various Indices

Commodity	1999		2000		$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
	Price	Qty	Price	Qty				
	p_0	q_0	p_1	q_1				
A	2	8	4	6	32	16	24	12
B	5	10	6	5	60	50	30	25
C	4	14	5	10	70	56	50	40
D	2	19	2	13	38	38	26	26
					200	160	130	103

1. **Laspeyres Method** : $P_{0,1} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$; where $\sum p_1 q_0 = 200$, $\sum p_0 q_0 = 160$

$$P_{0,1} = \frac{200}{160} \times 100 = 125$$

2. **Paasche's Method** : $P_{0,1} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$; where $\sum p_1 q_1 = 130$, $\sum p_0 q_1 = 103$

$$P_{0,1} = \frac{130}{103} \times 100 = 126.21$$

3. **Fisher's Ideal Method**: $P_{0,1} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{200}{160} \times \frac{130}{103}} \times 100$

$$= \sqrt{1.578} \times 100 = 1.256 \times 100 = 125.6$$

4. From the following data construct a price index number of the group of four commodities by using the appropriate formula :

Commodity	Base Year		Current Year	
	Price per unit	Expenditure (Rs.)	Price per unit	Expenditure (Rs.)
A	2	40	5	75
B	4	16	8	40
C	1	10	2	24
D	5	25	10	60

Sol :

Since we are given the base year and current year prices and expenditure, Fisher's ideal. Formula shall be most appropriate

Construction of Price Index

Commodity	p_0	q_0	p_1	q_1	p_1q_0	p_0q_1	p_1q_1	p_0q_1
A	2	20	5	15	100	40	75	30
B	4	4	8	5	32	16	40	20
C	1	10	2	12	20	10	24	12
D	5	5	10	6	50	25	60	30
					202	91	199	92

$$P_{0.1} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$$

$$= \sqrt{\frac{202}{91} \times \frac{199}{92}} \times 100 = 2.1912 \times 100 = 219.12$$

5. Compute Index Numbers from the following data for 2010 with 2005 as base by

(i) Lasperyes Method

(ii) Pasche's Method and

(iii) Fisher's Ideal formula.

Commodity	2005		2010	
	Price	Qty	Price	Qty
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

Sol :

Comm's	p ₀	q ₀	p ₁	q ₁	p ₁ q ₀	p ₀ q ₀	p ₁ q ₁	p ₀ q ₁
A	20	8	40	6	320	160	240	120
B	50	10	60	5	600	500	300	250
C	40	15	50	15	750	600	750	600
D	20	20	20	25	400	400	500	500
					2070	1600	1790	1470

i) Laspayer's Method

$$P_{0,1} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{2070}{1600} \times 100 = 129.375$$

ii) Paasches Method

$$P_{0,1} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{1790}{1470} \times 100 = 121.77$$

iii) Fisher's ideal formula

$$\begin{aligned}
 P_{0,1} &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 \\
 &= \sqrt{\frac{2070}{1600} \times \frac{1790}{1470}} \times 100 \\
 &= \sqrt{1.2469 \times 1.2176} \times 100 \\
 &= \sqrt{1.5182} \times 100 \\
 &= 1.2321 \times 100 \\
 &= 123.21.
 \end{aligned}$$

6. From the following data, calculate Price Index numbers by (i) Laspeyre's (ii) Paasche's and (iii) Fisher's ideal method.

Commodity	2005		2009	
	Price	Qty.	Price	Qty.
A	10	4	20	3
B	25	5	30	3
C	20	8	25	8
D	10	10	10	13

Sol :

Calculate price index

Commodity	2005		2009					
	P_0	q_0	P_1	q_1	$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$
A	10	4	20	3	80	40	60	30
B	25	5	30	3	150	125	90	75
C	20	8	25	8	200	160	200	450
D	10	10	10	13	100	100	130	100
					530	425	480	655

$$\begin{aligned} \text{i) Laspeyres } (P_{0,1}) &= \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 \\ &= \frac{530}{425} \times 100 = 124.70 \end{aligned}$$

$$\begin{aligned} \text{ii) Paasches } (P_{0,1}) &= \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100 \\ &= \frac{480}{655} \times 100 = 73.28 \end{aligned}$$

$$\begin{aligned} \text{iii) Fishers ideal } (P_{0,1}) &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \\ &= \sqrt{\frac{530}{425} \times \frac{480}{655}} \times 100 \\ &= \sqrt{1.247 \times 0.732} \times 100 \\ &= 0.9554 \times 100 = 95.54 \end{aligned}$$

3.3 TESTS OF INDEX NUMBERS

3.3.1 Time Reversal – Factor Reversal and Circular Tests

Q12. Explain the criterion for testing the consistency of good index numbers.

Ans :

1. Test

This test states that the formula of index number should be independent of the units in which the prices or quantities of various commodities (or items) are quoted. All the formulae, except the index number based on simple aggregate of prices (quantities) satisfy this test.

2. Time Reversal Test

This test was proposed by Prof. Irwin Fisher. According to Fisher 'the formula for calculating the index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as the base, or putting it another way, the index number reckoned forward should be reciprocal of the one reckoned backward.

In simple terms, given two time periods I and II, if an index number is calculated for period II taking period I as base, its value should be the reciprocal value of the index number for period I taking period II as base. The index numbers for the purpose of the test, should be in decimal form and not in percentage form. In other words, $P_{0,1}$ and $P_{1,0}$ should not be multiplied with 100. Symbolically,

$$P_{0,1} \times P_{1,0} = 1$$

Where $P_{0,1}$ = Fishers index for period II taking period I as base and

$P_{1,0}$ = Fishers index for period I taking Period II as base

Time Reversal test is satisfied by Marshall-Edge worth, Fisher, Walsh, Kelly's index numbers and also by simple Aggregative Index, simple geometric mean of price relatives and weighted average of price relatives. Laspeyre's and Paasche's index numbers do not satisfy the time reversal test.

3. Factor Reversal Test

This test was also proposed by Prof. Fisher. This test requires that the product of two index numbers, one measuring price taking quantities as base, and the other measuring quantities taking price as base, should be equal to the net increase in total value from one period to another. Let us illustrate the same with the help of an example.

If $P_{0,1}$ is price index number, $Q_{0,1}$ is quantity index number, the product of the two index numbers should be equal to the value index number $V_{0,1}$

$$P_{0,1} \times Q_{0,1} = V_{0,1}$$

Fishers index number satisfies the factor reversal test. No other method satisfies the factor reversal test.

4. Circular Test

The circular test was proposed by Weztergaard. It is an extension of the time-reversal test. It more than two time periods are considered, price index is calculated for each period with the previous year as base period. Lastly, the price index for the first year is calculated taking the last period as the base. The product of all the price index numbers should be equal to 1. Symbolically, if three time periods are considered,

$$P_{0,1} \times P_{1,2} \times P_{2,1} = 1$$

Only simple geometric mean of price relatives method and Kelly's method Satisfy the circular test.

PROBLEMS

7. Compute Fisher's Index Number and apply both factor reversal test and time reversal test from the following data :

Commodity	Base year		Current year	
	Price	Expenditure	Price	Expenditure
	Rs.	Rs.	Rs.	Rs.
A	5	25	10	60
B	1	10	2	24
C	4	16	8	40
D	2	40	5	75

Sol :

Comm's	P ₀	q ₀	P ₁	q ₁	P ₁ q ₀	P ₀ q ₁	P ₁ q ₁	P ₀ q ₁
A	5	5	10	6	50	25	60	30
B	1	10	2	12	20	10	24	12
C	4	4	8	5	32	16	40	20
D	2	20	5	15	100	40	75	30
					202	91	199	92

Fishers Ideal Index

$$\begin{aligned}
 P_{0,1} &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{202}{91} \times \frac{199}{92}} \times 100 \\
 &= \sqrt{\frac{40,198}{8372}} \times 100 = \sqrt{4.8015} \times 100 \\
 &= 2.191 \times 100 = 219.1
 \end{aligned}$$

Time Reversal Test

$$\begin{aligned}
 P_{0,1} &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \\
 P_{1,0} &= \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} \\
 P_{0,1} \times P_{1,0} &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} \\
 &= \sqrt{\frac{202}{91} \times \frac{199}{92} \times \frac{92}{199} \times \frac{91}{202}} = \sqrt{1} = 1 \\
 \therefore P_{0,1} \times P_{1,0} &= 1
 \end{aligned}$$

Factor Reversal Test

$$\begin{aligned}
 P_{0,1} \times q_{0,1} &= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_1 q_0}} \\
 &= \sqrt{\frac{202}{91} \times \frac{199}{92} \times \frac{92}{91} \times \frac{199}{202}} = \sqrt{\frac{199}{91} \times \frac{199}{91}} = \frac{199}{91} \\
 V_{0,1} &= \frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{199}{91} \\
 \therefore P_{0,1} \times Q_{0,1} &= V_{0,1}
 \end{aligned}$$

8. Calculate Fisher's price index from the following data and check whether Time Reversal Test is satisfied :

Commodity	Base Year		Current Year	
	Price (Rs.)	Qty. (Kg.)	Price (Rs.)	Qty (Kg.)
A	32	50	30	50
B	30	35	25	40
C	16	55	18	50

Sol :

Calculation of Fisher's Price Index

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_1q_1	p_0q_1	p_1q_0
A	32	50	30	50	1,600	1,500	1,600	1,500
B	30	35	25	40	1,050	1,000	1,200	875
C	16	55	18	50	880	900	800	990
					3,530	3,400	3,600	3,365

$$\text{Fisher's Ideal Index} = P_{0,1} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{3,365}{3,530} \times \frac{3,400}{3,600}} \times 100$$

$$= \sqrt{0.95 \times 0.94} \times 100$$

$$= \sqrt{0.893} \times 100$$

$$= 0.298 \times 100$$

$$= 29.8$$

$$\text{Time Reversal Test} = P_{0,1} \times P_{1,0} = 1$$

$$= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

$$= \sqrt{\frac{3,365}{3,530} \times \frac{3,400}{3,600}} \times \sqrt{\frac{3,600}{3,400} \times \frac{3,530}{3,365}}$$

$$= \sqrt{1} = 1$$

Hence Time Reversal Test is Satisfied.

9. Calculate Fisher's ideal index from the following data and prove that it satisfies both the time reversal and factor reversal tests :

Commodity	2000		2001	
	Price	Expenditure	Price	Expenditure
A	8	80	10	120
B	10	120	12	96
C	5	40	5	50
D	4	56	3	60
E	20	100	25	150

Sol :

Calculation of Fisher's Ideal Index

Commodity	2000		2001					
	p ₀	q ₀	p ₁	q ₁	p ₁ q ₀	p ₀ q ₁	p ₁ q ₁	p ₀ q ₁
A	8	10	10	12	100	80	120	96
B	10	12	12	8	144	120	96	80
C	5	8	5	10	40	40	50	50
D	4	14	3	20	42	56	60	80
E	20	5	25	6	125	100	150	120
					451	396	476	426

$$P_{0,1} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{451}{396} \times \frac{476}{426}} \times 100$$

$$\sqrt{1.2726} \times 100 = 1.128 \times 100 = 112.8$$

Time Reversal Test : Time reversal test is satisfied when $P_{0,1} \times P_{1,0} = 1$

$$P_{1,0} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

$$= \sqrt{\frac{426}{476} \times \frac{396}{451}}$$

$$P_{0,1} \times P_{1,0} = \sqrt{\frac{451}{396} \times \frac{476}{426} \times \frac{426}{476} \times \frac{396}{451}} = \sqrt{1} = 1$$

Hence, time reversal test is satisfied.

Factor Reversal Test : Factor reversal test is satisfied when :

$$P_{0,1} \times Q_{0,1} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

$$Q_{0,1} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_1 p_0}}$$

$$P_{0,1} \times Q_{0,1} = \sqrt{\frac{451}{396} \times \frac{476}{426} \times \frac{426}{396} \times \frac{476}{451}} = \frac{476}{396}$$

This is also the value of $\frac{\sum p_1 q_1}{\sum p_0 q_0}$. Hence, the above data also satisfies the Factor Reversal Test.

3.4 CHAIN INDEX NUMBERS

Q13. Define Chain Index Numbers. Explain the steps in constructing a Chain Index Numbers.

Ans :

In the fixed base method, discussed so far the base remains the same throughout the series of the index. This method, though convenient, has certain limitations. As time elapses conditions which were once important become less significant and it becomes more difficult to compare accurately present conditions with those of a remote period. New items may have to be included and old ones may have to be deleted in order to make the index more representative. In such cases it may be desirable to use the chain base method.

Steps in Constructing a Chain Index

In constructing a chain index following steps are desirable:

- (i) Express the figures for each year as percentages of the preceding year. The results so obtained are called link relatives.
- (ii) Chain together these percentages by successive multiplication to form a chain index. Chain index of any year is the average link relative of that year multiplied by chain index of previous year divided by 100. In the form of formula

Chain Index for current year

Average link relative of current year × Chain Index of previous year
100

The link relatives obtained in step (i) facilitate comparison from one year to another, i.e., between closely situated periods in which the q's are a process of chaining binary comparisons facilitate long-term comparisons.

Chain relatives differ from fixed-base relatives in computation. Chain relatives are computed from link relatives whereas fixed base relatives are computed directly from the original data. The results obtained by the two different methods should be the same, but they may differ from each other slightly due to rounding off of decimal places. Since the process of computing chain relatives is quite complicated and the results are same as the fixed base relatives obtained from the original data, chain relatives should be used when the original data are not available but the link relatives are.

PROBLEMS

10. Convert the following series of index numbers to chain base indices :

Years :	1960	1961	1962	1963	1964	1965	1966	1967
Index No. : (1960 Base)	100	110	125	133	149	139	150	165

Sol :

Year	Fixed Base Index	Link Relative	Chain Index
1960	100	—	100
1961	110	$\frac{110}{100} \times 100 = 110.00$	$\frac{110 \times 100}{110} = 110$
1962	125	$\frac{125}{110} \times 100 = 113.64$	$\frac{113.64 \times 110}{100} = 125$
1963	133	106.40	133
1964	149	112.03	149
1965	139	93.29	139
1966	150	107.91	150
1967	165	110.00	165

Note : The chain base indices will always be equal to fixed base indices for any single fixed series of index numbers.

11. Calculate the chain base index numbers chained to 1993 from the average price of following three commodities :

Commodity	1993	1994	1995	1996	1997
Wheat	4	6	8	10	12
Rice	16	20	24	30	36
Sugar	8	10	16	20	24

Sol :

Computation of Chain Indices Link Relatives based on previous year

Commodities	1993	1994	1995	1996	1997
Wheat	100	$\frac{6}{4} \times 100$ = 150	$\frac{8}{6} \times 100$ = 133.33	125	120
Rice	100	125	120	125	120
Sugar	100	125	160	125	120
Total of					
Link Relative	300	400	413.33	375	360
Average of					
Link Relatives	100	133.33	137.78	125	120
Chain Index					
(1993 = 100)	100	$\frac{133.33 \times 100}{100}$	$\frac{137.78 \times 133.33}{100}$	229.63	275.55

3.4.1 Conversion of Chain Index of Fixed base Index

Q14. Explain the Conversion of Chain Index of Fixed base Index.

Ans :

At times it may be desired to convert the chain base index numbers into fixed base index numbers. In such a case the following procedure is followed :

1. For the first year the fixed base index will be taken the same as the chain base index. However, if the index numbers are to be constructed by taking first year as the base in that case the index for the first year is taken as 100.
2. For calculating the indices for other years the following formula is used :

$$\text{Current year's F.B.I} = \frac{\text{Current year's C.B.I.} \times \text{Previous year's F.B.I.}}{100}$$

F.B.I = Fixed base index no. ; C.B.I. = Chain base index no.

PROBLEM

12. From the chain base index numbers given below prepare fixed base index numbers :

1991	1992	1993	1994	1995
80	110	120	90	140

Sol :

Computation of Fixed Base Index Numbers

Year	Chain base index number	Fixed base index numbers
1990	80	80
1991	110	$\frac{110 \times 80}{100} = 88.00$
1992	120	$\frac{120 \times 88}{100} = 105.60$
1993	90	$\frac{90 \times 105.6}{100} = 95.04$
1994	140	$\frac{140 \times 95.04}{100} = 133.06$

Q15. State the merits and demerits of chain based index numbers.

Ans :

Merits of the Chain Base Method

The merits of this method are enumerated here:

1. The chain base method has a great significance in practice because in economic and business data, we are more often concerned with making com-parisons with the previous period and not with any distant past. The link relatives obtained by chain base method serve this purpose.

2. Chain base method permits the introduction of new commodities and the detection of old ones without necessitating either the recalculation of entire series or other drastic changes. Because of this flexibility, chain index is used in many types of indices such as the consumer price index and the wholesale price index.
3. Weights can be adjusted as frequently as possible. This flexibility is of great significance in many types of index numbers.
4. Index numbers calculated by the chain base method are free to a greater extent from seasonal variations than these obtained by the other method.

Limitations of the Chain Index

- (i) The main limitation of the chain index is that while the percentages of previous year figures give accurate comparisons of year-to-year changes, the long-range comparisons of chained percentages are not strictly valid.
- (ii) When the index number user wishes to make year-to-year comparisons, as is so often done by the businessman, the percentages of the preceding year provide a flexible and useful tool.

3.5 CHANGING OF BASE / BASE SHIFTING OF INDEX NUMBERS

Q16. Define Base Shifting.

Ans :

For a variety of reasons, it frequently becomes necessary to change the reference base of an index number series from one time period to another without returning to the original raw data and recomputing the entire series. This change of reference base period is usually referred to as “**shifting the base**”. There are two important reasons for shifting the base:

1. The previous base has become too old and is almost useless for purposes of comparison. By shifting the base it is possible to state the series in terms of a more recent time period.
2. It may be desired to compare several index number series which have been computed on different base periods: particularly if the several series are to be shown on the same graph, it may be desirable for them to have the same base period. This may necessitate a shift in the base period.

When base period is to be changed, one possibility is to recompute all index numbers using the new base period. A simpler approximate method is to divide all index numbers for the various years corresponding to the old base period by the index number corresponding to the new base period, expressing the results as percentages. These results represent the new index numbers, the index number for the new base period being 100 per cent.

Mathematically speaking, this method is strictly allocable only if the index numbers satisfy the circular test.

PROBLEMS

13. The following are the index numbers of prices (1990 = 100)

Year	Index
1990	100
1991	110
1992	120
1993	200
1994	400
1995	410
1996	400
1997	380
1998	370
1999	340

Shift the base from 1990 to 1996 and recast the index numbers.

Sol :

INDEX NUMBERS WITH 1990 AS BASE (1990 = 100)

Year	Index Numbers (1990 = 100)	Index Numbers (1996 = 100)
1990	100	$\frac{100}{400} \times 100 = 25.0$
1991	110	$\frac{110}{400} \times 100 = 27.5$
1992	120	$\frac{120}{400} \times 100 = 30.0$
1993	200	$\frac{200}{400} \times 100 = 50.0$
1994	400	$\frac{400}{400} \times 100 = 100.0$
1995	410	$\frac{410}{400} \times 100 = 102.5$
1996	400	= 100.0
1997	380	$\frac{380}{400} \times 100 = 95.0$
1998	370	$\frac{370}{400} \times 100 = 92.5$
1999	340	$\frac{340}{400} \times 100 = 85.0$

The new series with 1996 as base is obtained very easily by dividing each entry of the first column by 400, i.e., the values of the index for 1996 and multiplying the ratio by 100. Thus,

$$\text{Index number for 1990} = \frac{\text{Index number for 1996}}{\text{Index number for 1990}} \times 100 = \frac{100}{400} \times 100 = 25$$

$$\text{Index number for 19901} = \frac{\text{Index number for 1991}}{\text{Index number for 1996}} \times 100 = \frac{100}{400} \times 110 = 27.5$$

In a similar manner other indices can also be obtained.

14. The following are the index numbers of wholesale prices of a commodity 'based on 1990 :

Year	Index Numbers
1993	100
1994	108
1995	120
1996	150
1997	210
1998	225
1999	240

Prepare new Index Numbers taking 1995 as base.

Sol:

Shifting the base from 1993 to 1995

Year	Index Numbers (1993 = 100)	Index Numbers (199 = 100)
1993	100	83.33
1994	108	90.00
1995	120	100
1996	150	125
1997	210	175.5
1998	225	187.5
1999	240	200.0

$$\text{Index number for 1993} = \frac{\text{Index number for 1993}}{\text{Index number for 1995}} \times 100 = \frac{100}{120} \times 100 = 83.333$$

$$\text{Index number for 1994} = \frac{108}{120} \times 100 = 90$$

3.6 COMBINING OF TWO OF MORE OVERLAPPING INDICES

Q17. Define splicing ? What are its conditions and types ?

Ans :

(May-19)

Splicing is the process of combining two or more index numbers which may convert various bases into a single series. In other words, splicing of index numbers can be defined as the procedure of converting two or more series of index numbers of different bases into a single or continuous series of index numbers.

Conditions of Splicing

Following conditions are essential for splicing the index numbers,

- Atleast two series of index numbers should be from same group.
- The base period of different indices are built on different series.
- Indices of two bases should be there in a year. For example, in 2015 there may be two index numbers with 2005 and another with 2013.

Types of Splicing

Generally, splicing is categorized into two types and they are as follows,

1. Backward splicing
2. Forward splicing.

1. Backward Splicing

Backward splicing is used for splicing a new series of indices to continue with old series of indices.

Formula

$$\text{Backward Splicing} = \frac{\text{Index A of Current Year}}{\text{Index A of Common Year}} \times 100$$

2. Forward Splicing

Forward splicing is used for splicing old series of indices to continue with new series of indices.

Formula

$$\text{Forward Splicing} = \frac{\text{Index B of Current Year} \times \text{Index A of Common Year}}{100}$$

15. The following table gives two series of index numbers with 2005 and 2010 as base year. Obtain a continuous series of index number by considering 2010 as base year.

Years	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Index No. (2005 = 200)	200	240	250	280	300	320				
Index No. (2010 = 200)						200	220	240	260	280

Sol :

$$\text{Backward Splicing} = \frac{\text{Index A of Current Year}}{\text{Index A of Common Year}} \times 100$$

Years	Index Number (2005 = 200)	Index Number (2010 = 200)	Backward Splicing
2005	200		$\frac{200}{320} \times 100 = 62.5$
2006	240		$\frac{240}{320} \times 100 = 75$
2007	250		$\frac{250}{320} \times 100 = 78.12$
2008	280		$\frac{280}{320} \times 100 = 87.50$
2009	300		$\frac{300}{320} \times 100 = 93.57$
2010	320	200	$\frac{320}{320} \times 100 = 100$
2011		220	220
2012		240	240
2013		260	260
2014		280	280

16. The following table gives two series of index numbers with 2006/2010 as base year. Obtain a combined series of index numbers with 2006 as base year.

Years	2006	2007	2008	2009	2010	2011	2012	2013	2014
Index No. (2006 = 100)	100	105	118	120	125				
Index No. (2010 = 100)					100	120	132	145	160

Sol :

$$\text{Forward Splicing} = \frac{\text{Index B of Current Year} \times \text{Index A of Common Year}}{100}$$

Years	Index No. (2006 = 100) Series A (Forward Splicing)	Index No. (2010 = 100) Series B
2006	100	
2007	105	
2008	118	
2009	120	
2010	125	100
2011	$\frac{120 \times 125}{100} = 150$	120
2012	$\frac{132 \times 125}{100} = 165$	132
2013	$\frac{145 \times 125}{100} = 181.25$	145
2014	$\frac{160 \times 125}{100} = 200$	160

Q18. Define deflating. Explain with an example.

Ans :

Deflating implies making adjustments for price changes. A rise of price level implies a fall in the value of money. Therefore, in a situation of rising prices, the workers who are getting a fixed sum in the form of wages are in fact getting less real wages. Similarly, in a situation of falling prices, the real wages of the workers are greater than their money wages. Thus, to determine the real wages, the money wages of the workers are to be adjusted for price changes by using relevant price index number.

The following formula is used for conversion of money wages into real wages.

$$\text{Real Wage} = \frac{\text{Money Wage}}{\text{Consumer Price Index}} \times 100 \quad \dots (1)$$

Another application of the process of deflating is to find the value of output at constant prices so as to facilitate the comparison of real changes in output. It may be pointed out here that the output of a given year is often valued at the current year prices. Since prices in various years are often different, the comparison of output at current year prices has no relevance.

The output at constant prices is obtained using the following formula.

$$\text{Output at Constant Prices} = \frac{\text{Output at Current Prices}}{\text{Price Index}} \times 100 \quad \dots (2)$$

Example

The following table gives the average monthly wages of a worker along with the respective consumer price index numbers for ten years.

Years Average Monthly	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Wages (₹)	500	525	560	600	630	635	700	740	800	900
Consumer Price Index	100	110	120	125	135	160	185	200	210	240

Compute his real average monthly wages in various years.

Sol :

Computation of Real Wages

Years	Average Monthly Wage	Consumer Price Index	Real Average Monthly Wage
1980	500	100	$\frac{500}{100} \times 100 = 500.00$
1981	525	110	$\frac{525}{110} \times 100 = 477.27$
1982	560	120	$\frac{560}{120} \times 100 = 466.67$
1983	600	125	$\frac{600}{125} \times 100 = 480.00$
1984	630	135	$\frac{630}{135} \times 100 = 466.67$
1985	635	160	$\frac{635}{160} \times 100 = 396.88$
1986	700	185	$\frac{700}{185} \times 100 = 378.38$
1987	740	200	$\frac{740}{200} \times 100 = 370.00$
1988	800	210	$\frac{800}{210} \times 100 = 380.95$
1989	900	240	$\frac{900}{240} \times 100 = 375.00$

3.7 CONSUMER PRICE INDEX

Q19. Discuss in detail consumer price index numbers/cost of living index numbers.

Ans :

Consumer price index numbers is also known as cost of living index number is used to measure the purchasing power of a particular class of people in relation to the changes in retail prices. In other words, it studies how price variations effect the cost of living or purchasing power of a group of people.

While constructing cost of living index number, a particular section of society is selected like [Rich, middle, poor] and a study is conducted to know how price variations effect the consumption levels of that section. Based on such information Cost of Living Index Number (CLIN) is constructed.

Steps in Construction of Consumer Price Index Numbers (CPIN)

1. Selection of Group of people

A group of people or class of people is selected to construct Cost of Living Index Numbers [CLIN]. Apart from class of people, the area (i.e., rural or urban, city or town) should be clearly specified. The group of people selected for constructing cost of living index numbers must be homogenous to a maximum extent.

2. Conducting Family Budget Enquiry

An enquiry of family budget is conducted to know how much money average family spends on the consumption of different items. These items are broadly categorized into five groups namely,

- (a) Food
- (b) Clothing
- (c) Fuel and lighting
- (d) House Rent and
- (e) Miscellaneous.

Each of the above groups is further sub-categorized into small groups,

Example

The group "food" is subdivided into cereals (like wheat, rice, pulses and so on) meat, fish, milk, fruits, vegetables and so on.

3. Price Quotations

While gathering information about retail prices a proper care should be taken as retail prices varies from place to place from shop to shop. Information about retail prices should be gathered from those local markets where selected class of people are located.

Uses of CPIN

1. Consumer Price Index Numbers (CPIN) are used in the preparation of wage contracts and wage negotiations.
2. CPIN assists the government and business organization in deciding dearness allowance [D.A] to be paid to their employees.
3. CPIN are used for deflating income and value series in National Income of the country.
4. CPIN are used to measure purchasing power of money.
5. They assist in calculating Real wage by considering the variations in money income and price level.

Q20. Explain the methods of Consumer Price Index Numbers.

Ans :

There are two methods for the compute of consumer price index numbers.

- (a) Aggregate Expenditure method
- (b) Family Budget Method

Cost of Living Index Numbers are weighted index numbers. The commodities constitute the index are given weights according to their importance. Normally, the weights are in the ratio of amounts spent on each item. There are two methods of constructing the cost of Living Index Numbers.

a) Aggregate Expenditure Method or Weighted Aggregate

Thus method is similar to Laspeyre's method. The quantities consumed in the base year are taken as weights. The formula is :

$$\text{Consumer Price Index} = \text{Cost of Living Index} = \frac{\sum p_1 q_0}{\sum p_0 q_1} \times 100$$

Since P represents price and q represents quantity, pq is the amount spent of given commodity. Thus $\sum pq$ represents total amount spent on all items. In other words it represents total expenditure. $\sum p_0 q_0$ is total expenditure incurred in the base period. $\sum p_1 q_0$ is total expenditure in the current year at base period. Thus

$$\text{Consumer Price Index} = \frac{\text{Total Expenditure Current Year at Base Year Price}}{\text{Total Expenditure in Base Year}} \times 100$$

b) Family Budget Method or Method of Weighted Relatives

The cost of living index is obtained by taking a weighted average of price relatives. The quantities consumed in the base year are taken as weights.

The formula is :

$$\text{Cost of Living Index} = \frac{\sum PV}{\sum V} \text{ Where}$$

$$P = (p_1/p_0) \times 100 \text{ for each item and } V = \text{Values Weight} = 100$$

$$\text{Thus, Cost of Living Index} = \frac{\sum \left(\frac{p_1}{p_0} \times 100 \right) \times p_0 q_0}{\sum p_0 q_0} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Thus, the cost of living index figure is one and the same, irrespective of method of construction.

PROBLEMS ON CONSUMER PRICE INDEX

17. Calculate the index number using both the Aggregate Expenditure method and Family Budget method for the year 1973 with 1960 as base year from the following data.

Commodity	Quantity in units in 1960	Price per units in 1960 (Rs)	Price per units in 1973 (Rs)
A	100	8.00	12.00
B	25	6.00	7.50
C	10	5.00	5.25
D	20	48.00	52.00
E	25	15.00	16.50
F	30	9.00	27.00

Sol :

Calculation of Consumer Price Index

Commodity	Quantity (q_0)	Price (p_0)	$p_0 q_0$ p_1	$p_0 q_0$ $= V$	$p_1 q_0$ $\times 100$	$P = (P_1/P_0)$	PV
A	100	8.00	12.00	800	1200	150	120000
B	25	6.00	7.50	150	187.50	125	18750
C	10	5.00	5.25	50	52.50	105	52.50
D	20	48.00	52.00	960	1040.00	108.33	1040000
E	25	15.00	16.50	375	412.50	110	41250
F	30	9.00	27.00	270	810.00	300	81000
				2605	3702.50		370205

(i) Aggregate Expenditure Method

$$CP_1 = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{3702.05}{2605} \times 100 = 142.13$$

(ii) Family budget Method

$$CP_1 = \frac{\sum PV}{\sum V} = \frac{370250}{2605} = 142.13$$

18. In the construction of a certain cost of living number, the following group index numbers are found. Calculate the Cost of Living Index Number by using (i) Weighted Arithmetic Mean and (ii) Weighted Geometric Mean.

Group	Index Numbers	Weights
Food	350	5
Fuel and Lighting	200	1
Clothing	240	1
House Rent	160	1
Miscellaneous	250	2

Sol :

Computation of Consumer Price Index

Group	Index No. (I)	Weights (W)	Weighted (WI)	Log I	W.log I
Food	350	5	1750	2.5441	12.7205
Fuel Lighting	200	1	200	2.3010	2.3010
Clothing	240	1	240	2.3802	2.3802
House Rent	160	1	160	2.2041	2.2041
Miscellaneous	250	2	500	2.3979	4.7958
		10	2850		24.4016

Consumer Price Index

$$(i) \quad \text{Using Arithmetic Mean} = \frac{\sum IW}{\sum W} = \frac{2850}{10} = 285$$

$$(ii) \quad \text{Using Geometric Mean} = \text{Antilog} \left[\frac{\sum W \log I}{\sum W} \right] = \text{antilog} \left[\frac{24.4016}{10} \right]$$

$$= \text{Antilog} (2.44016) = 275.55$$

3.8 PROBLEMS IN CONSTRUCTION OF INDEX NUMBERS**Q21. What are the various Problems in the Construction of Index Numbers?****Ans :****(Imp.)**

The following are some general problems that are faced in the construction of any index number:

1. Definition of the purpose

Since it is possible to construct index numbers for a number of purposes and one cannot have an all purpose index, therefore, it is very essential to define the specific purpose of its construction. For example, if we are interested in the construction of a price index number, we must have knowledge about the purpose to be served by it, i.e., what is to be measured by it; like the cost of living of workers or the change in wholesale prices, etc. In the absence of this information, it may be difficult to carry out various steps in the construction of an index number. The questions like what are the items to be included, from which markets the price quotations are to be obtained, what will be the weights of different items, etc., cannot be answered unless the purpose of the index number construction is known. Further, an index number can be of sensitive or general nature. In case of sensitive index, only those items are included whose variables (like prices in case of price index) fluctuate very often; while efforts are made to include as many items as possible when the index is of general nature. It may be pointed out that the index numbers are specialized tools and as such are more useful and efficient when properly used. The first step in this direction is a specific definition of the purpose of its construction.

2. Selection of the base period

Every index number is constructed with reference to a base period. There are two important points that must be kept in mind while selecting the base period of an index number.

- (a) The base period should correspond to a period of relative economic and political stability, i.e., it should be a normal or representative period in some way. In certain situations where identification of such a period is not possible, the average of certain periods can also be taken as base.
- (b) The base period should not be too distant from the current period. The comparison of current period with a remote base doesn't have much relevance. In the words of Morris Hamburg, "It is desirable that the base period be not too far away in time from the present. The further away we move from the base period the dimmer are our recollections of economic conditions prevailing at that time. Consequently, comparisons with these remote base periods tend to lose significance and become rather tenuous in meaning."

3. Selection of number and type of items

An index number of a particular group of items is in fact based on a sample of items taken from it. It is neither possible nor necessary to include all the items of the group in the construction of an index number. The number of items to be included depends largely upon the purpose of the index number.

There are no hard and fast rules that can be laid down with regard to the selection of the number of items, however, it must be remembered that more is the number of items the more representative will be the index number and more cumbersome will be the task of computations. Therefore, it is necessary to have some sort of balance between having a representative index and the work of computation involved in its construction.

The following points should be kept in mind in selecting the type of items:

- (a) The items should be representative of the tastes, habits and customs of the people for whom the index is to be constructed.
- (b) The selected items should be of stable quality. The standardized items should be given preference.
- (c) As far as possible, the non-tangible items like personal services, goodwill, etc., should be excluded because it is difficult to ascertain their value.

4. Collection of data

The next important step in the construction of an index number is the collection of data. For example, for the construction of price index, price quotations are to be obtained. Since the prices of commodities may vary from one market to another and in certain cases from one shop to another, it is necessary to select those markets which are representative in the sense that the group under consideration generally make purchases from these markets. The next logical step is to select an agency through which price quotations are to be obtained. The selected agency should be highly reliable and if necessary the accuracy of price quotations reported by it may also be checked by appointing some other agency or agencies. Furthermore, care should always be taken to obtain price quotations for the same quality of items.

Similar type of considerations are necessary for the collection of data for the construction of index numbers such as quantity index, value index, unemployment index, etc.

5. Selection of a suitable average

Since the index numbers are also averages, any of the five averages, viz. arithmetic mean, median, mode, geometric mean and harmonic mean can be used in its construction. However, since in most of the situations we have to average ratios of the values in current period to that in base period, geometric mean is the most suitable average in the construction of index numbers. The main difficulty of using the geometric mean is the complexities of its computations and hence, the use of arithmetic mean is more popular inspite of its being less suitable.

6. Selection of suitable Weights

According to John I. Griffin, "Weighing is designed to give component series an importance in proper relation to their real significance." The basic purpose of weighing is to enable each item to have an influence, on the index number, in proportion to its importance in the group. It is, therefore, necessary to design a system of weighing such that true importance of the items is reflected by it. The system of weighing may be either arbitrary or rational. Arbitrary or chance weighing implies that the statistician is free to assign weights to different items as he thinks fit or reasonable. Rational or logical weighing, on the other hand, implies that some criterion has been fixed for assigning weights. Two types of weights are commonly used in the construction of a price index number; (i) physical quantities and (ii) money values. These weights can be quantities (or values) produced or consumed or sold in base or current or in any other period.

Exercise Problems

1. Compute by Fishers index formula, the quantity index number from the data given below.

Commodity (`)	Base Year		Current Year	
	Price Value	Total (`)	Price (`)	Total Value (`)
A	10	100	8	96
B	16	96	14	98
C	12	36	10	40

(Ans : 120.65)

2. The following figures relate to the prices and quantities of certain commodities. Using Kelly's method, construct an Index number:

Commodity	Quantity Consumed	Price in 1980 (`)	Price in 1985 (`)
A	50	32	40
B	35	30	42
C	55	16	24
D	45	40	52
E	15	35	42

(Ans: 132.53)

3. Compute (i) Laspeyres index and (ii) Paasche's index from the data given below and test whether they satisfy (i) Time Reversal Test and (ii) Factor Reversal Tests.

Commodity	p_0	q_0	p_1	q_1
A	5	10	4	12
B	8	6	7	7
C	6	3	5	4

(Ans : (i) $P_{0.1} = 83.62$, $P_{1.0} = 119.66$ $Q_{0.1} = 120.69$, $V_{0.1} = 100.86$

(ii) $P_{0.1} = 83.57$, $P_{1.0} = 119.59$ $Q_{0.1} = 120.62$, $V_{0.1} = 100.86$)

4. Convert CBI to FBI (chained to 1996):

Year :	1996	1997	1998	1999	2000
CBI :	100	120	90	100	125

(Ans : 100,120,108,108,135)

5. From the Chain base index numbers given below, find fixed base index numbers.

Year :	1975	1976	1977	1978	1979
Chain Base Index :	80	110	120	90	140

(Ans : 80,88,105.60,95.04,133.06)

6. The following are the index numbers of wholesale prices of a commodity based on 1990.

Year	Index Numbers	Year	Index Numbers
1990	100	1994	210
1991	108	1995	225
1992	120	1996	240
1993	150		

Prepare new index numbers taking 1992 as base.

(Ans : 83.33, 90,100,125,175,187.5, 200)

Rahul Publications

Short Question & Answers

1. What do you understand by index numbers ?

Ans :

Some prominent definitions of index numbers are given below:

According to Croxton & Cowden "Index numbers are devices for measuring differences in the magnitude of a group of related variables."

According to Spiegel "An index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc."

According to Patterson "In its simplest form an index number is the ratio of two index numbers expressed as a per cent. An index number is a statistical measure a measure designed to show changes in one variable or in a group of related variables over time, or with respect to geographic location, or other characteristic."

According to Morris Hamburg "In its simplest form, an index number is nothing more than a relative number, or a 'relative' which expresses the relationship between two figures, where one of the figures is used as a base."

According to Berenson and Levine "Generally speaking, index numbers measure the Size or magnitude of some object at a particular point in time as a percentage of some base or reference object in the past."

2. Characteristics of Index Numbers

Ans :

The features of index numbers are as follows,

i) Measures the Change in Percentages

Index numbers measure the change in price or quantity in terms of percentages say i.e., (10%, 20%, 15%, 25% and so on). Increase or decrease in value is represented by one single figure. Like 10% increase in sales from that previous year to current year, 30% decrease in profits when compared to that of the last year and so on.

ii) Specialized Averages

A single figure known as "Average" is used for representing the characteristics of the complete set of data. Average acts as a basis for comparing different data sets with each other, when they have common unit of measurement of observations. When data sets do not have common unit of measurement, specialized averages of index numbers are used for the comparison.

iii) The Measured Changes cannot be Observed Directly

Index numbers do not measure the changes directly but it studies the relative changes or variations in factors resulting to changes-like for measuring changes in Export-Imports related factors such as available raw materials, technology, competitors etc., are studied and measured.

iv) Measures Changes in Relation to Time or Place

Index numbers measure change by comparing the values at different time periods or at different places like standard of living at one place is being compared with standard of living of the other place. Sales or revenue of current year is compared with that of the previous years sales or revenue.

3. Explain the limitations of index numbers

Ans :

Index Numbers are of great use in studying trends, events, pertaining to a defined phenomenon. However, they suffer from certain limitations. There are

(i) Approximate Representation

Index Numbers are based on sample data. Hence, they are only approximate indicators. They may not fully reflect the changes in the relative level of a phenomenon.

(ii) Likelihood of Error

There is a possibility of error in the choice of the sample, selection of base period, data collection, and assignment of weights or the appropriate formula to be used.

(iii) Not Responsive

There are rapid changes in Technology, tastes, fashions, customs and consequently, the consumption pattern of various commodities. Index numbers may not be able to keep pace with such changes. Hence, they may not be able to reflect the changes in the phenomenon being studied.

(iv) No Universally Acceptable formula

There is no universally acceptable standard formula that can capture the changes in a phenomenon with perfection. None of the formulae seems to satisfy all the tests of consistency. Hence, some amount of formula error is present in any calculation of Index numbers.

(v) Scope for manipulation

On account of wide variety of formulae index numbers can be manipulated by appropriate choice of base year, sample set, price and quantity quotations.

4. Explain Laspeyre's Method of index numbers.

Ans :

Laspeyre's Method

This method takes the quantities of the commodities in the base period as the weight of that commodity for the purpose of calculating the index numbers. The following steps may be following.

Step I

Multiply the current year price (represented by p_1) with the quantities of the base year (q_0) for each commodity.

Step 2

Add the numbers obtained in step 1. The resultant sum is represented as $\Sigma p_1 q_0$

Step 3

Multiply the prices of base year (represented by p_0) with the quantities of the base year for each commodity.

Step 4

Add the numbers obtained in step 3. The resultant sum is represented as $\Sigma p_0 q_0$

Step 5

The index number as per Laspeyre's method

$$= P_{0,1} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

5. Explain Paasche's Method of index number.

Ans :

Paasche's Method

Paasche's Method is similar to Laspeyre's method. The only difference is in assignment of weights. As per this method quantities consumed of the commodities in the current year is taken as basis. The following steps need to be followed.

Step 1

Multiply current year's prices (p_1) with current year's quantities (q_1)

Step 2

Add the numbers obtained in step (1). The resultant sum is $\Sigma p_1 q_1$

Step 3

Multiply base year's Prices (p_0) with current year's quantities (q_1)

Step 4

Add the numbers obtained in step (2). The resultant sum is $\Sigma p_0 q_1$

Step 5

Index Number as per Paasche's method =

$$P_{0,1} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

6. Define Chain Index Numbers.*Ans :*

In the fixed base method, discussed so far the base remains the same throughout the series of the index. This method, though convenient, has certain limitations. As time elapses conditions which were once important become less significant and it becomes more difficult to compare accurately present conditions with those of a remote period. New items may have to be included and old ones may have to be deleted in order to make the index more representative. In such cases it may be desirable to use the chain base method.

Steps in Constructing a Chain Index

In constructing a chain index following steps are desirable:

- (i) Express the figures for each year as percentages of the preceding year. The results so obtained are called link relatives.
- (ii) Chain together these percentages by successive multiplication to form a chain index. Chain index of any year is the average link relative of that year multiplied by chain index of previous year divided by 100. In the form of formula

Chain Index for current year

$\frac{\text{Average link relative of current year} \times \text{Chain Index of previous year}}{100}$

7. Define Base Shifting.*Ans :*

For a variety of reasons, it frequently becomes necessary to change the reference base of an index number series from one time period to another without returning to the original raw data and recomputing the entire series. This change of reference base period is usually referred to as **"shifting the base"**. There are two important reasons for shifting the base:

1. The previous base has become too old and is almost useless for purposes of comparison. By shifting the base it is possible to state the series in terms of a more recent time period.
2. It may be desired to compare several index number series which have been computed on different base periods: particularly if the several series are to be shown on the same graph, it may be desirable for them to have the same base period. This may necessitate a shift in the base period.

When base period is to be changed, one possibility is to recompute all index numbers using the new base period. A simpler approximate method is to divide all index numbers for the various years corresponding to the old base period by the index number corresponding to the new base period, expressing the results as percentages. These results represent the new index numbers, the index number for the new base period being 100 per cent.

8. Define deflating.*Ans :*

Deflating implies making adjustments for price changes. A rise of price level implies a fall in the value of money. Therefore, in a situation of rising prices, the workers who are getting a fixed sum in the form of wages are in fact getting less real wages. Similarly, in a situation of falling prices, the real wages of the workers are greater than their money wages. Thus, to determine the real wages, the money wages of the workers are to be adjusted for price changes by using relevant price index number.

The following formula is used for conversion of money wages into real wages.

$$\text{Real Wage} = \frac{\text{Money Wage}}{\text{Consumer Price Index}} \times 100$$

Another application of the process of deflating is to find the value of output at constant prices so as to facilitate the comparison of real changes in output. It may be pointed out here that the output of a given year is often valued at the current year prices. Since prices in various years are often different, the comparison of output at current year prices has no relevance.

The output at constant prices is obtained using the following formula.

$$\text{Output at Constant Prices} = \frac{\text{Output at Current Prices}}{\text{Price Index}} \times 100$$

9. Consumer Price Index

Ans :

Consumer price index numbers is also known as cost of living index number is used to measure the purchasing power of a particular class of people in relation to the changes in retail prices. In other words, it studies how price variations effect the cost of living or purchasing power of a group of people.

While constructing cost of living index number, a particular section of society is selected like [Rich, middle, poor] and a study is conducted to know how price variations effect the consumption levels of that section. Based on such information Cost of Living Index Number (CLIN) is constructed.

Choose the Correct Answers

1. Indices calculated by the chain base method are free from: (a)
(a) Seasonal variations (b) Errors
(c) Percentages (d) Ratios
2. The chain base indices are not suitable for: (a)
(a) Long range comparisons (b) Short range comparisons
(c) Percentages (d) Ratios
3. An index number that can serve many purposes is called: (a)
(a) General purpose index (b) Special purpose index
(c) Cost of living index (d) None of them
4. Another name of consumer's price index number is: (b)
(a) Whole-sale price index number (b) Cost of living index
(c) Sensitive (d) Composite
5. Consumer price index indicates : (c)
(a) Rise (b) Fall
(c) Both (a) and (b) (d) Neither (a) and (b)
6. Purchasing power of money can be accessed through: (c)
(a) Simple index (b) Fisher's index
(c) Consumer price index (d) Volume index
7. Cost of living at two different cities can be compared with the help of: (a)
(a) Value index (b) Consumer price index
(c) Volume index (d) Un-weighted index
8. Consumer price index numbers are obtained by: (a)
(a) Laspeyre's formula (b) Fisher ideal formula
(c) Marshall Edgeworth formula (d) Paasche's formula
9. Laspeyre's index = 110, Paasche's index = 108, then Fisher's Ideal index is equal to: (d)
(a) 110 (b) 108
(c) 100 (d) 109
10. Most commonly used index number is: (c)
(a) Volume index number (b) Value index number
(c) Price index number (d) Simple index number

Fill in the Blanks

1. An index number is called a simple index when it is computed from _____.
2. Index numbers are expressed in _____.
3. If all the values are of equal importance, the index numbers are called _____.
4. Index numbers can be used for _____.
5. Index for base period is always taken as _____.
6. When the prices of rice are to be compared, we compute _____.
7. When index number is calculated for several variables, it is called _____.
8. How many types are used for the calculation of index numbers _____.
9. In chain base method, the base period is _____.
10. Price relatives are a percentage ratio of current year price and _____.

ANSWERS

1. Single variable
2. Percentages
3. Unweighted
4. Forecasting
5. 100
6. Price index
7. Composite index
8. 2
9. Not fixed
10. Base year price

UNIT IV

Probability

Concepts – Random Experiment, Sample space – Definitions of probability, Simple Problems on Probability, Addition and Multiplication theorems, conditional, Joint and Marginal Probability.

Sampling: Sampling – Reasons of Sample survey – bias in Survey, Definitions of Population, Sample, Parameter, Statistic – Principles of Sampling, Statistical Regularity, Inertia of Large Numbers, Optimization, Persistence of small numbers – Validity. Probability and non probability sampling methods – choice of sampling method, sampling distribution and Standard Error (SE).

4.1 PROBABILITY

Q1. Explain the overview of probability.

Ans :

Introduction

The term 'probability' was coined by an Italian mathematician, 'Galileo'. He was the first man who measured probability quantitatively while dealing with the problems associated with rolling of dice in gambling.

The word 'probability' or 'chance' is the most common word used in our day-to-day life.

For example, in our daily life we use certain statements like 'Probably he may win the elections', 'It is likely that India may win the match', 'She may score above 90% in the coming examination' etc.

A probability is a quantitative measure of uncertain events. It helps in determining the chances of the occurrence of an event.

The systematic and extensive study of probability theory was made by 'B.Pascal'. Pierre de Fermat', Jacques Bernoulli' in mid-seventeenth century.

Probability is a part of our daily lives, as life is uncertain. In personal and managerial decisions, the use of probability theory helps the individual to make educated guess in matters where the outcome is uncertain. Probability helps in making decisions under uncertain conditions and predicts the outcome of an event.

4.1.1 Definitions of Probability

Q2. Discuss various Definitions of Probability.

Ans :

Definitions of Probability

Probability can be defined as the chance or 'likelihood of occurrence' of an experiment or event. Probability of any event ranges from 0 to 1. The following are the various definitions of probability.

1. Classical Probability

It is also called as 'mathematical probability'.

When there are 'n' cases which are equally likely, mutually exclusive and collectively exhaustive out of which 'm' are favorable cases to an event A, then the probability of occurrence of the event A is

given by the ratio $\frac{m}{n}$.

Symbolically

$$P(A) = \frac{\text{Number of favourable cases}}{\text{Total number of exhaustive cases}}$$

$$\therefore P(A) = \frac{m}{n}$$

$$m \leq n$$

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

2. Statistical Probability

Suppose an experiment is repeated 'n' times under essentially identical conditions.

Let an event A happens m-times then $\frac{m}{n}$ is defined as the relative frequency of A.

It is also known as 'relative frequency' probability. The limit of this relative frequency as $n \rightarrow \infty$ is defined as the probability of A.

$$\therefore P(A) = \lim_{n \rightarrow \infty} \frac{m}{n} \quad 0 \leq \frac{m}{n} \leq 1$$

3. Axiomatic Probability

In axiomatic probability three axioms or postulates are explained on the basis of which probability is calculated. They are,

- (i) Probability of an event ranges from, zero to one $\Rightarrow 0 \leq P(A) \leq 1$
- (ii) Probability of entire sample space, $P(S) = 1$
- (iii) If 'A' and 'B' are mutually exclusive events, then the probability of occurrence of either A or B is $P(A \cup B) = P(A) + P(B)$.

In general, there are three chances which can be expected for any event,

(a) Certainty

A certain event is an event whose occurrence is certain. Probability of such an event is 1.

(b) Impossibility

The occurrence of these events is not possible. Probability of such an event is 0 (zero).

(c) Uncertainty

Probability of uncertain event ranges between 0 and 1.

Thus, the probability of any event will always lie between 0 and 1.

4.1.2 Concepts of Probability

4.1.2.1 Random Experiment, Sample space

Q3. Explain the various concepts are used in probability.

(or)

What are the various terms used in probability?

Ans :

Concepts of Probability

Number of terms are used in the concepts, definitions and approaches of probability. These terms are as follows,

1. Random Experiment

Random experiment refers to a situation where erratic, nonspecific and arbitrary results can be obtained.

If an experiment is conducted repeatedly for 'n' number of times under identical conditions and if the outcome of an experiment varies in each case, then such an experiment is called as 'random experiment'.

Example

Tossing a coin, rolling a dice etc

2. Outcome

The result of a random experiment is usually referred as an outcome.

If we toss a coin, the outcome may be a head or a tail. In such a case, number of outcomes = 2.

3. Event

An event is a possible outcome of an experiment or a result of trial.

Basically there are two types of events. Simple and compound event.

(a) Simple Event

The probability of happening or non-happening of a single event is considered as a simple event.

Example

When we are selecting two black coins from a box containing 10 white and 5 black coins.

(b) Compound Event

If the joint occurrence of two or more event is considered then it is known as 'compound event' (or) 'composite event'.

Example

A box containing 5 white balls, 3 black balls and 8 red balls, when we draw 2 white balls in first draw, 2 black balls in second draw and 5 red balls in third draw.

4. Mutually Exclusive Events

When two events cannot occur simultaneously in a single trial then such events are called as mutually exclusive or incompatible events.

Example

- (a) When a single coin is tossed, either a head or a tail can turn up, both cannot come at the same time
- (b) At a single point of time, a person can be alive or dead.

5. Non-mutually Exclusive Events

When two events can occur simultaneously in a single trial then such events are said to be non-mutually exclusive events.

Example

From a pack of cards, drawing a red card and drawing a queen are the two events. These two events can occur simultaneously while drawing a red queen.

Hence, these two events are said to be non-mutually exclusive events which can occur at the same time.

6. Collectively Exhaustive Events

Collectively exhaustive events are those events whose totality contains all the potential outcomes of a random experiment.

Example

When a dice is thrown, the total possible outcomes are 1, 2, 3, 4, 5 and 6 and thus the number of exhaustive cases is 6.

7. Equally Likely Events

Events are considered as equally likely events when the probability of occurrence of all the events is equal.

These events are also called as 'equally, probable events'.

Example

If a coin is tossed, the two possible outcomes are head and tail. The probability of their occurrence is equal (i.e., $\frac{1}{2}$)

8. Independent Events

Two or more events are considered as independent events, when the outcome of one event does not influence and is not influenced by the other event.

Example

When a student has appeared in physics and chemistry examinations, his marks obtained in physics is independent of the marks obtained in chemistry.

9. Dependent Events

Dependent events are the events in which the occurrence or non-occurrence of one event in any one trial influences the probability of occurrence of other events in other trials.

Example

When a card is drawn from a pack of playing cards and is not replaced then this changes the probability of occurrence of the second card.

Here probability of occurrence of second event is dependent on the occurrence of first event.

10. Complementary Events

Two events are said to be complementary events if they are mutually exclusive and collectively exhaustive.

Example

When a coin is tossed, getting a head or a tail are mutually exclusive and collectively exhaustive.

Hence, if we get Mail then head is considered as its complementary event.

PROBLEMS

1. The following table give the details of the consumer preference for a new product to be introduced in the market.

Category	Number of consumers		
	Like	Dislike	Neutral
Male	500	250	125
Female	200	350	75

What is the probability that the consumer selected at random from the group will be,

- (i) A male who disliked the product?
 (ii) One who likes the product given that the person is a female?
 (iii) Either male or female who dislikes the product?

Sol :

Category	Number of Consumers			
	Like	Dislike	Neutral	Total
Male	500	250	125	875
Female	200	350	75	625
Total	700	600	200	1500

Total number of male and female consumers = 1500

Total number of male consumers = 875

Total number of female consumers = 625

Total number of consumers who liked the new product = 700

Total number of consumers who disliked the new product = 600

Total number of consumers who are neutral = 200

- (i) **Probability of a Male Who Disliked the Product**

Let 'A' be the event of selecting a male consumer 'D' be the event of selecting a consumer who disliked the product.

$$= \frac{P(M \cap D)}{P(C)} \text{ \{Where } P(C) \text{ - Total number of consumers\}}$$

$$\text{Probability of a male who disliked the product} = \frac{\text{Total number of males who disliked the product}}{\text{Total number of consumers}}$$

$$= \frac{250}{1500} \text{ (From table)}$$

$$= \frac{1}{6}$$

Therefore, the probability that a consumer selected at random from the group will be a male who disliked the product is $\frac{1}{6}$.

(ii) Probability of a Consumer Who Likes the Product Given that the Person is a Female

Let 'F' be the event of selecting a female consumer.

'L' be the event of selecting a consumer who likes the product.

Probability of selecting a female consumer,

$$P(F) = \frac{625}{1500}$$

Probability of selecting a female who likes the product,

$$P(F \cap L) = \frac{200}{1500}$$

$\therefore$ Probability of a consumer who likes the product given that the person is a female is

$$\begin{aligned} P(L/F) &= \frac{P(F \cap L)}{P(F)} \\ &= \frac{200}{1500} \times \frac{1500}{625} = \frac{8}{25} \end{aligned}$$

Therefore the probability that a consumer selected at random from the group will be one who likes the product given that the person is a female, $P(L/F) = \frac{8}{25}$

(iii) Probability of Either a Male or a Female Who Dislikes the Product

Probability of a male who dislike the product, $P(M \cap D) = \frac{250}{1500}$ (From table)

Probability of a female who dislike the product, $P(F \cap D) = \frac{350}{1500}$ (From table)

$\therefore$ The probability that either male or female who dislikes the product is,

$$\begin{aligned} &\frac{P(M \cap D)}{P(M)} + \frac{P(D \cap F)}{P(F)} \quad (\text{By addition theorem of probability}) \\ P(M) &= \frac{875}{1500}, P(F) = \frac{625}{1500} \\ &= \frac{\frac{250}{1500}}{\frac{875}{1500}} + \frac{\frac{350}{1500}}{\frac{625}{1500}} \\ &= \left(\frac{250}{1500} \times \frac{1500}{875} \right) + \left(\frac{350}{1500} \times \frac{1500}{625} \right) \\ &= \frac{2}{7} + \frac{14}{25} = \frac{50+98}{175} = \frac{148}{175} = 0.845 \end{aligned}$$

Therefore, the probability that a consumer selected at random from the group will be either a male or a female who dislikes the product is $= 0.845 = 84.5\%$.

Q4. Explain the importance of probability.**Ans :**

The various practical applications of the theory of probability are as follows,

- (i) The fundamental laws of statistics, viz the law of statistical regularity and law of inertia of large numbers depends mainly on theory of probability
- (ii) Probability helps in taking effective decisions under uncertain conditions
- (iii) The main objective of the theory of probability is to reduce the complication in betting and games. It helps in deciding whether it is worthwhile betting at a game, by calculating the expected value of an action with likely return on each of the outcome, given the chances of success attached to each outcome.
- (iv) The use of probability theory is increasing in economic decision making. It is very helpful in situations of risk and uncertainty.

Example

A stock broker is interested in knowing whether the price of a certain stock will increase or decrease during the next week, a sales manager is interested knowing the probability of doubling the sales of a particular product, management is interested in knowing the probability of the success of a new product etc. For making effective decisions in uncertain conditions the fundamental laws of probabilities are used.

- (v) Decision theories are based on fundamental laws of probability and expected value.
- (vi) The empirical probability concept, based on experimental tests, provides scope for the application of probability to real life situations.
- (vii) Subjective probabilities are used when it is not possible to do actual measurement. This has added a new dimension to the theory of probability.
Such subjective probabilities are revised by the application of Baye's rule.
- (viii) Probability models helps in making predictions.

(ix) The concept of probability not only helps in carrying out different types of scientific investigations but also helps in solving the day-to-day problems in life.

(x) probability tree diagrams helps in calculating probability Values of different situations.

Q5. What are the various approaches of probability?**Ans :****(May-17)**

Based on the concept of probability there are four different approaches of probability, which are explained below,

1. Classical (or) priori approach
2. Relative frequency approach
3. Subjective approach
4. Axiomatic approach.

1. Classical (or) Priori Approach

This approach assumes that all the outcomes of a random experiment are mutually exclusive, and equally likely. According to this approach in a random experiment when there are 'm' favorable cases when a favourable event A occurs and 'n' total possible cases when favorable event A does not occur. Then the probability of getting favorable cases P(A) can be calculated as,

$$P(A) = \frac{\text{Number of favourable cases}}{\text{Total number of possible outcomes}} = \frac{m}{n}$$

Classical or priori approach is given by, "James Bernoulli". He was the first person who found the uncertainty quantitatively.

According to this approach,

- (i) **Probability of an event always lies between 0 and 1.**

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

- (ii) **Sum of probability of an event and its complement is always 1.**

$$p + q = 1$$

⇒ Sum of the probability of success and failure is 1.

For example,

'p' is the probability of getting a head and

'q' is the probability of getting a tail.

Then,

$$P + q = \frac{1}{2} + \frac{1}{2} = 1$$

- (iii) **The probability of occurrence for any event has 3 chances.**

- In case of certain event, $P(E) = 1$
- In case of impossible event, $P\{E\} = 0$
- In case of uncertain event, $0 \leq P(E) \leq 1$.

Limitations

Classical theory does not holds good,

- When all the outcomes are not equally likely
- When the collectively exhaustive events of an experiment are infinite
- Classical theory does not provide answers to certain question which occurs in our daily life.

2. Relative Frequency Approach

This is given by 'VONMISES'. In some cases, the desired event may or may not occur. This approach assumes that a random experiment is repeated number of times under identical conditions and the trials are independent to each other.

This approach calculates the proportion of time known as relative frequency with which the event takes place repeatedly over an infinite number of times under identical conditions.

Therefore, the probability of occurring an event A, can be calculated as,

$$P(A) = \lim_{n \rightarrow \infty} \left\{ \frac{a}{n} \right\}$$

Where,

'a' is the number of times an event 'A' is repeated

'n' is the trials of an experiment.

As the probability of an event is ascertained by repetitive empirical observations, this probability is known as 'empirical probability'.

Limitations

- As the experiments are repeated large number of times, it takes large amount of time.
- During the experimental time, the conditions may not be always identical and homogenous.

3. Subjective Approach

This approach was initiated by "**Frank Ramsey**" in the year 1926 in his book. "The Foundation of Mathematics and Other Logical Essays".

This approach depends on the extent to which the experiences focusses on the chances of the occurrence of random event. Hence, it is also known as 'personalistic approach'.

Limitations

- As the probability is just an estimation based on one's personal beliefs it may differ from one person to another person
- We cannot calculate probability accurately as the results are not known
- There is no particular formula for calculating probability quantitatively.

4. Axiomatic Approach

This approach was given by the Russian mathematician, A.N.Kolmogorov in the year 1933 in his book "**Foundation of Probability**".

According to axiomatic approach, the probability of any event is calculated on basis of the axioms or postulates. Three axioms are taken into consideration for knowing the probability stated below,

- (i) The probability of an event, if it does not occurs is 0 (zero). For certain event it is 1 (one) and for uncertain event it always ranges from 0 to 1
- (ii) The probability of the whole sample space is always $P(S) = 1$
- (iii) When 'A' and 'B' are two mutually exclusive events, then the probability of occurrence of either A or B is equal to the sum of their individual probabilities.

$$P(A \cup B) = P(A) + P(B)$$

4.2 THEOREMS OF PROBABILITY

Q6. Explain various Probability Theorems.

Ans :

The important theorems of probability are,

1. Addition theorem
2. Multiplication theorem.

4.2.1 Addition Theorem

Q7. Explain Addition Theorem of Probability.

Ans :

Addition Theorem

Addition theorem is different for mutually exclusive events and non-mutually exclusive events.

(a) For Mutually Exclusive Events

When 'A' and 'B' are two mutually exclusive events (i.e., both cannot occur at the same time) then the probability of occurrence of A or B is equal to the sum of their individual probabilities.

$$P(A \text{ or } B) = P(A) + P(B) \\ \Rightarrow P(A \cup B) = P(A) + P(B)$$

Diagrammatically it can be represented as,

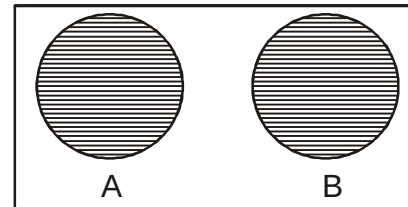


Fig. Mutually Exclusive Events

In case of 3 events A, B and C,

$$P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

(b) For Non-mutually Exclusive Events

In case of non-mutually exclusive event (i.e., if the events occur together) there is a variation in the addition theorem.

When 'A' and 'B' are non-mutually exclusive events then the probability of occurrence of A or B is the sum of their individual probability which should be deducted from the probability of A and B occurring together.

$$P(A \text{ or } B) \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Diagrammatically it can be represented as,

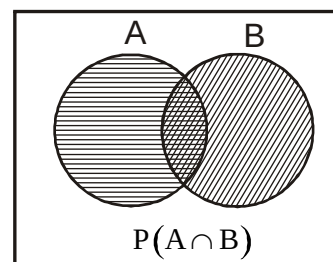


Fig. Non-mutually Exclusive Events

In case of three non-mutually exclusive events.

A, B and C the probability of occurrence of A or B or C can be calculated by the following formula.

$$P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

PROBLEMS

2. The probability that a contractor will get a plumbing contract is $\frac{3}{4}$ and the probability that he will not get electric contract is $\frac{4}{9}$. If the probability of getting at least one contract is $\frac{5}{6}$, what is the probability that he will get both the contracts?

Sol :

Let 'A' be the event that the contractor will get plumbing contract

B' be the event that the contractor will get electric contract.

The probability of getting plumbing contract = $P(A)$

The probability of not getting a plumbing contract = $P(\bar{A})$

The probability of getting electric contract = $P(B)$

The probability of not getting electric contract = $P(\bar{B})$

The probability of getting at least one contract is,

$$P(A \text{ or } B) = \frac{5}{6}$$

Given that,

$$P(A) = \frac{3}{4}; P(\bar{A}) = \frac{1}{4} (\because P(A) + P(\bar{A}) = 1)$$

$$P(\bar{B}) = \frac{4}{9}; P(B) = \frac{5}{9}$$

$\therefore$ The probability that he will get both the contract can be given by using addion theorem

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A \text{ and } B) = P(A) + P(B) - P(A \text{ or } B)$$

$$= \frac{3}{4} + \frac{5}{9} - \frac{5}{6}$$

$$= \frac{27 + 20 - 30}{36} = \frac{47 - 30}{36} = \frac{17}{36} = 0.472 = 47.2\%$$

$\therefore$ The probability that the contractor will get both the contracts = $\frac{17}{36} = 0.472 = 47.2\%$

3. A card is drawn from a standard pack of 52 cards.

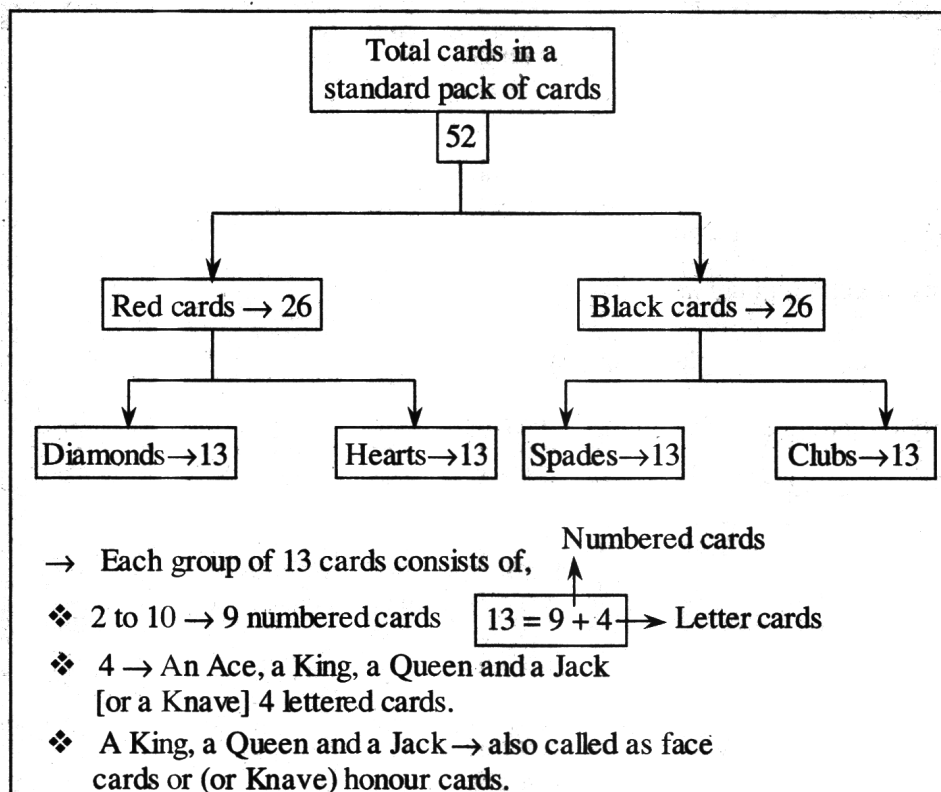
(i) What is the probability that it is a black or a red card?

(ii) What is the probability that it is an Ace or a King or a Queen?

(iii) What is the probability that it is a red Knave or a black King?

Sol :

Before solving the problem, one should know the entire picture of standard pack of cards.

**Note**

In a standard pack of 52 cards,

- There are 13 diamonds, 13 hearts, 13 spades and 13 clubs.
- There are 26 red cards and 26 black cards.
- In total, there are 36 numbered cards and 16 lettered cards.
- There are 12 face cards (honour cards) in a pack of 52 cards, i.e., 4 Kings, 4 Queens and 4 Jacks (Knives).
- There are 4 Aces in total in a standard pack of cards.

(i) Probability of Getting a Black (or) a Red Card

Let,

'A' be the event of getting a black card

'B' be the event of getting a red card.

Probability of getting a black card,

$$P(A) = \frac{26}{52} = \frac{1}{2}$$

Probability of getting a red card

$$P(B) = \frac{26}{52} = \frac{1}{2}$$

Since, the events "A" and "B" are mutually exclusive events, the probability of getting a black or red card,

$$P(A \cup B) = P(A) + P(B)$$

[By addition theorem of probability]

$$= \frac{1}{2} + \frac{1}{2} = 1$$

(ii) Probability of Getting an Ace or a King or a Queen

As we know, there are 4 Aces, 4 Kings and 4 Queens in a pack of 52 cards.

Let,

'A' be the event of getting an Ace

'B' be the event of getting a King

'C' be the event of getting a Queen.

Probability of getting an Ace, $P(A) = \frac{4}{52} = \frac{1}{13}$

Probability of getting a King, $P(B) = \frac{4}{52} = \frac{1}{13}$

Probability of getting a Queen, $P(C) = \frac{4}{52} = \frac{1}{13}$

Since A, B, C are mutually exclusive events, the probability of getting an Ace or a King or a Queen is given

by,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

[By addition theorem of probability]

$$= \frac{1}{13} + \frac{1}{13} + \frac{1}{13} = \frac{3}{13}$$

∴ The probability of getting on Ace on a King or a Queen is $\frac{3}{13}$

(iii) Probability of Getting a Red Knave or a Black King

As we know, there are 26 red cards and 26 black cards in a pack of 52 cards.

Out of 26 red cards, there are 2 Knaves.

There are 2 Kings out of 26 black cards.

Let,

'A' be the probability of getting a red Knave and

'B' be the probability of getting a black king.

Probability of getting a red Knave,

$$P(A) = \frac{2}{26} = \frac{1}{13}$$

Probability of getting a black King,

$$P(B) = \frac{2}{26} = \frac{1}{13}$$

As 'A' and 'B' are mutually exclusive events, the probability of getting a red Knave or a black King is given by,

$$P(A \cup B) = P(A) + P(B)$$

[By addition theorem of probability for mutually exclusive events]

$$\therefore P(A \cup B) = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$

∴ The probability of getting a red Knave or a black King = $\frac{2}{13}$

4. One card is drawn from a standard pack of 52. What is the probability that it is either King or a Queen?

Sol :

Let 'A' be the event of drawing a king from a standard pack of 52.

'B' be the event of drawing a queen.

P(A) = The probability of drawing a king

P(B) = Probability of drawing a queen.

As, the probability of drawing either a king or a queen has to be determined by drawing a single card, the events are said to be mutually exclusive.

For mutually exclusive events,

$$P(A \cup B) = P(A) + P(B)$$

(By addition theorem of probability)
Given data,

$$P(A) = \frac{4}{52} \text{ (As there are 4 kings in a standard pack of 52)}$$

$$P(B) = \frac{4}{52} \text{ (As there are 4 queens in a standard pack of 52)}$$

$$\therefore P(A \cup B) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

Therefore, the probability of drawing either a king or a queen = $\frac{2}{13}$

4.2.2 Multiplication Theorem

Q8. Explain Multiplication Theorem of Probability.

Ans :

Multiplication Theorem

If 'A' and 'B' are two independent events then the probability of occurrence of both the events is equal to the product of their individual probabilities.

For independent events,

$$P(A \cap B) = P(A) \cdot P(B)$$

Similarly,

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C) \text{ and so on}$$

If 'A' and 'B' are two dependent events, in such a case multiplication theorem is altered and is given as follows. For dependent events,

$$P(A \cap B) = P(A/B) \cdot P(B) \\ = P(B/A) \cdot P(A)$$

Where, $P(A/B)$ is a conditional probability of A given that B has occurred (The probability of occurrence of event A when event B has already occurred is the conditional probability of A given B).

PROBLEMS

5. A committee of 4 people is to be appointed from 3 officers of the production department, 4 officers of the purchase department, two officers of the sales department and one chartered accountant. Find the probability of forming the committee in the following manner.

- (i) There must be one from each category
- (ii) It should have at least one from the purchase department
- (iii) The chartered accountant must be in the committee.

Sol :

Total number of officers = 3 officers of the production department, 4 officers of the purchase department, 2 officers of the sales department and one chartered accountant = $3 + 4 + 2 + 1 = 10$

(i) There must be one from each category,

$$\frac{{}^3C_1 \times {}^4C_1 \times {}^2C_1 \times {}^1C_1 \times {}^1C_1}{{}^{10}C_4} = \frac{n!}{r!(n-r)!}$$

Where,

$${}^nC_n = 1, {}^nC_1 = n$$

$$\frac{24}{210} = 0.114$$

$\therefore$ The probability of forming the committee where there must be one member from each category is 0.114.

(ii) It should have atleast one from the purchase department

There are total 4 officers from purchase department.

- ∴ Probability of committee having atleast one from purchase department ($P(X > 1) = 1 - P(X < 0)$)
 $= 1 - (\text{probability of committee has no purchase officer})$

Out of 10 people if 4 members from purchase department are not considered only 6 people are left.

The probability of getting 4 people from $6 = {}^6C_4 = 15$ ways

- ∴ The probability of forming a committee without any purchase officer

$$= \frac{{}^6C_4}{{}^{10}C_4} = \frac{15}{210} = \frac{1}{14}$$

- ∴ The probability of committee having at least one from purchase department

$$= 1 - \frac{1}{14} = \frac{14-1}{14} = \frac{13}{14} = 0.93$$

(iii) **The chartered accountant must be in the committee.**

In this case, chartered accountant is must in the committee, hence, the remaining 3 members are from the total of remaining 9 people.

∴ The required probability,

$$= \frac{{}^1C_1 \times {}^9C_3}{{}^{10}C_4}$$

$$= \frac{2}{5} = 0.4$$

- ∴ The probability of forming a committee in which one chartered accountant is must = $\frac{2}{5}$ (or) 0.4%.

6. A bag contains 8 white and 7 black balls. 4 balls are drawn one by one without replacement. What is the probability that white and black balls appear alternately?

Sol :

There are 8 white and 7 black balls in a bag

Total number of balls in a bag = $8 + 7 = 15$

$$\therefore \text{Total balls} = 15$$

Total balls	Number of white balls	Number of black balls
15	8	7

$$\text{Probability of choosing a white ball in 1st attempt} = \frac{{}^8C_1}{{}^{15}C_1} = \frac{8}{15} \quad [\because {}^nC_1 = n]$$

After 1st Attempt

Total balls	Number of white balls	Number of black balls
14	7	7

$$\text{Probability of choosing a black ball in 2nd attempt} = \frac{{}^7C_1}{{}^{14}C_1} = \frac{7}{14}$$

After 2nd Attempt

Total balls	Number of white balls	Number of black balls
13	7	6

(As the balls are drawn without replacement, after first attempt one ball gets reduced in total balls)

$$\text{Probability of choosing a white ball in 3rd attempt} = \frac{{}^7C_1}{{}^{13}C_1} = \frac{7}{13}$$

After 3rd Attempt

Total balls	Number of white balls	Number of black balls
12	6	6

$$\text{Probability of choosing a black ball in 4th attempt} = \frac{{}^6C_1}{{}^{12}C_1} = \frac{6}{12} = \frac{1}{2}$$

$\therefore$ The probability that white and black balls appear alternatively,

$$= \frac{8}{15} \times \frac{7}{14} \times \frac{7}{13} \times \frac{1}{2} = \frac{14}{195}$$

$$= 0.0717$$

7. There are 5 white and 7 red balls in a box. A box is drawn and then replaced. What is the probability that a white and a red ball are drawn in that order.

- (i) With replacement of balls
(ii) Without replacement of balls.

Sol :

Let 'A' be the event of drawing a white ball and

'B' be the event of drawing a red ball.

- (i) **With Replacement**

When a ball is replaced before the 2nd draw then the outcomes of 2nd draw are found to be independent of the result of 1st draw. Here, A and B are independent events.

∴ Probability of drawing a white and a red ball,

$$P(A \cap B) = P(A) \cdot P(B)$$

[By multiplication theorem]

Total number of balls = 5 + 7 = 12 balls.

Probability of drawing a white ball,

$$P(A) = \frac{{}^5C_1}{{}^{12}C_1} = \frac{5}{12} \quad (\because {}^5C_1 = 5)$$

Probability of drawing a red ball,

$$P(B) = \frac{{}^7C_1}{{}^{12}C_1} = \frac{7}{12}$$

$$P(A \cap B) = \frac{5}{12} \times \frac{7}{12} = \frac{35}{144}$$

Therefore, the probability of drawing a white ball and a red ball with replacement is $\frac{35}{144}$

(ii) Without Replacement

If the ball is drawn and not replaced before the 2nd draw, then the outcomes of 2nd draw are affected, by the results of the 1st experiment. Hence, A and B are said to be dependent events.

Hence, for dependent events,

$$P(A \cap B) = P(A) \cdot P(B/A)$$

[By multiplication theorem for dependent events]

Probability of drawing a white ball,

$P(A) = \frac{{}^5C_1}{{}^{12}C_1} = \frac{5}{12}$ (As the white ball which is drawn already is not replaced, reduce one ball from the total balls).

Probability of drawing a red ball when the white ball drawn is not replaced,

$P(B/A) = \frac{{}^7C_1}{{}^{11}C_1} = \frac{7}{11}$ (Total number of remaining balls = 4 white + 7 red)

$$\therefore P(A \cap B) = P(A) \cdot P(B/A)$$

$$= \frac{5}{12} \times \frac{7}{11} = \frac{35}{132}$$

Therefore, the probability of drawing a white ball and a red ball without replacement is $\frac{35}{132}$.

4.3 STATISTICAL INDEPENDENCE

4.3.1 Conditional, Joint and Marginal Probability

Q9. Write about various probabilities under Statistical Independence.

Ans :

Probability under Conditions of Statistical Independence

Statistical independent events are those events, whose occurrence has no effect on the occurrence of any other event.

The happening or not happening of any event does not affect the happening of another event.

Example

If a coin is tossed twice, the result of second tossing would not be affected by the result of the first tossing,

There are three types of probabilities under statistical independence,

1. Marginal probability
2. Joint probability
3. Conditional probability
 - (a) For independent events
 - (b) For dependent events.

(i) Marginal Probabilities under Statistical Independence

Marginal probability is the simple probability of the occurrence of an event.

It is also known as 'unconditional' probability or 'single probability'.

For example, when a single unbiased coin is tossed the probability that it is a head is $P(H) = \frac{1}{2}$

or probability that it is a tail, $P(T) = \frac{1}{2}$

These two events are independent and do not overlap one another i.e., the events are statistically independent of the outcomes of next coin been tossed.

The individual probabilities obtained in this case i.e., $P(H)$ or $P(T)$ are called as marginal probabilities.

(ii) Joint Probabilities Under Statistical Independence

A joint probability is the probability of occurrence of two or more simple independent events together.

In other words, the product of two marginal probabilities occurring together or in succession is called as 'joint probability'.

For example, let $P(W)$ is the probability of a girl with white complexion and $P(B)$ is the probability of a girl with black hair.

Here, $P(W)$ and $P(B)$ are marginal probability.

Probability of a girl with white complexion and with black hair together is called as 'joint probability'.

Joint probability of these two events can be represented as, $P(W \cap B)$.

$$\therefore P(W \cap B) = P(W) \cdot P(B)$$

(By multiplication theorem)

Where, $P(W \cap B) \rightarrow$ Probability of a girl with white complexion and with black hair

$P(W)$ = Probability of a girl with white complexion

$P(B)$ = Probability of a girl with black hair.

For example, there are 100 girls of whom 60 girls are white and 70 girls possess black hair, the probability of girls with white complexion and black hair is called joint probability.

Mathematically, it can be calculated as follows,

$$P(W) = \frac{60}{100} = 0.6$$

$$P(B) = \frac{70}{100} = 0.7$$

Probability of a girl having white complexion as well as black hair,

$$P(W \cap B) = 0.6 \times 0.7 = 0.42 \text{ or } 42\%$$

Here, joint probability = 0.42.

Similarly, joint probability for three independent events A, B and C can be represented as $P(A \cap B \cap C)$.

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

(By multiplication theorem of probability)

(iii) Conditional Probabilities under Statistical Independence

Conditional probability is the probability of occurrence of second event (B) given that the first event (A) has already occurred.

For statistically independent events, the conditional probability of event B given that event A has already occurred is simply the probability of event B. Symbolically, it can be represented as,

$$P(B/A) = P(B)$$

and is read as the probability of event B, given that the event A has already occurred.

For example, the probability of a second toss of a fair coin given that the tossing of first coin resulted as head.

In the given example, occurrence of first toss is the event A and occurrence of second toss is the event B.

The outcome of first toss is head which has absolutely no effect on the result of the second toss.

In this case, the conditional probability of event B given that A has already occurred is equal to the probability of event B.

Note

Conditional Probability for Statistically Dependent Events.

If the occurrence of one event affects the occurrence of another event then the two events are said to be “statistically dependent events”.

When the probability of one event is dependent or affected by the occurrence of some other event, then it is called as ‘conditional probability for statistically dependent events’.

Let the event A has already occurred and A and B are two dependent events, then the probability of occurrence of B given A has already occurred is represented as $P(B/A)$.

$$\therefore P(B/A) = \frac{P(A \cap B)}{P(A)}$$

$P(A \cap B)$ = Joint probability

= Probability of occurrence of both the events A and B.

$P(A)$ = Marginal probability

= Probability of occurrence of A

Note

Conditional probability for statistically dependent events is joint probability divided by marginal probability in this case,

(or)

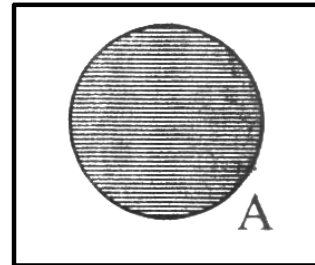
Joint probability is the product of conditional probability and marginal or prior probability.

Example

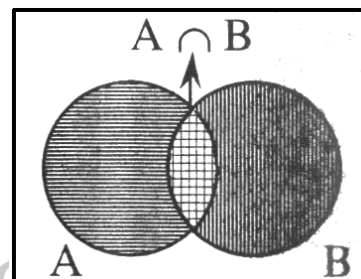
A person has appeared to an interview. The person possesses an MBA degree the probability that the person getting selected in the interview given that the person is a MBA holder can be regarded as the example for conditional probability for dependent events.

1. Marginal Probability

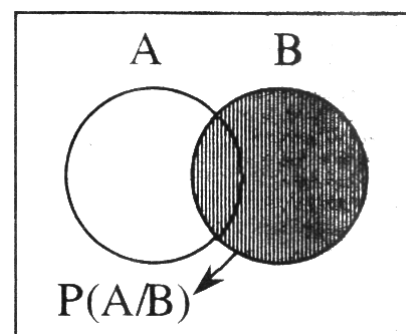
$P(A)$

**2. Joint Probability**

$P(A \cap B)$

**3. Conditional Probability**

(a) For dependent events $P(A/B)$



(b) For independent events $P(A/B) = P(A)$

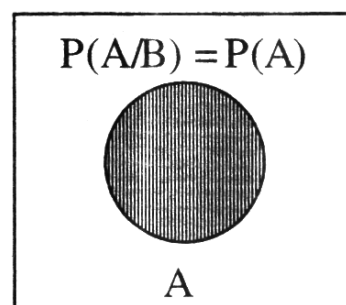


Fig. Diagrammatic Representation of Various Probability

8. A restaurant features two types of lunches special lunch and ordinary lunch. Sixty percent of its men customers order the special lunch and the rest of the men order the ordinary lunch. Ninety percent of its women customers order the ordinary lunch and the rest order the special lunch. Seventy percent of its customers are men. In its preparations, what ratios of special to ordinary lunches should the restaurant plan for? Determine the probability that the man will order a special lunch.

Sol :

Let 'M' denote the event of the customer being a man.

'W' indicate the event of the customer being a women.

'S' indicate the event of ordering a special lunch.

'O' indicate the event of ordering ordinary lunch. Given that,

The probability of men who ordered special lunch, $P(M \cap S) = 60\% = 0.6$

Probability of men ordering ordinary lunch, $P(M \cap O) = 40\% = 0.4$.

Probability of women ordering ordinary lunch, $P(W \cap O) = 90\% = 0.9$

Probability of women ordering special lunch, $P(W \cap S) = 10\% = 0.1$

Probability of customers who are men, $P\{M\} = 70\% = 0.7$

Probability of customers who are women, $P(W) = 1 - P(M)$
 $= 1 - 0.7 = 0.3$.

Let us consider total 100 customers of whom 70 are men and 30 are women. The ratio of men to women = 70 : 30.

With the help of the given information the percentages can be converted to numbers.

Total customers = 100

Total men = 70

Total women = 30

Men ordering special lunch

$$= 60\% \text{ of } 70 = \frac{60}{100} \times 70 = 42$$

Men ordering ordinary lunch

$$= 40\% \text{ of } 70 = \frac{40}{100} \times 70 = 28$$

Women ordering special lunch

$$= 10\% \text{ of } 30 = \frac{10}{100} \times 30 = 3$$

Women ordering ordinary lunch

$$= 90\% \text{ of } 30 = \frac{90}{100} \times 30 = 27.$$

Tabulating the above results,

	S	O	Total
M	42	28	70
W	3	27	30
Total	45	55	100

- (i) Therefore, the ratio of special lunch to ordinary lunch from the above table,

$$S : O = 45 : 55$$

$$= 9 : 11$$

- (ii) Probability that the man will order a special lunch = $P(S/M)$

$$P(S/M) = \frac{P(S \cap M)}{P(M)} \text{ (Conditional probability)}$$

$\therefore P(S \cap M)$ = Probability of a man and a special lunch

$$= \frac{42}{100} = 0.42 \text{ (From table)}$$

$P(M)$ = Probability of a customer being a man

$$= \frac{70}{100} = 0.7 \text{ (From table)}$$

$$\therefore P\{S/M\} = \frac{0.42}{0.70} = 0.6$$

Therefore, the probability that a man will order a special lunch = 0.6 or 60%.

4.4 SAMPLING

Q10. Define the term Sampling.

Ans :

Sampling may be defined as the selection of some part of an aggregate or totality on the basis of which a judgement or inference about the aggregate or totality is made. In other words, it is the process of obtaining information about an entire population by examining only a part of it.

In most of the research work and surveys, the usual approach happens to be to 'make generalizations' or to 'draw inferences' based on samples about the parameters of population from which the samples are taken. The researcher quite often selects only a few items from the universe for his study purpose. All this is done on the assumption that the sample data will enable him to estimate the population parameters.

The items so selected constitute what is technically called a sample, their selection process is called sample design and the survey conducted on the basis of sample is described a sample survey. Sample should be truly representative of population characteristics without any bias so that it may result in valid and reliable conclusions.

Q11. What are the essentials of Sampling.

Ans :

i) Representativeness

A sample should be so selected that it truly represents the universe otherwise the results obtained may be misleading. To ensure representativeness the random method of selection should be used.

(ii) Adequacy

The size of sample should be adequate, otherwise it may not represent the characteristics of the universe.

(iii) Independence

All items of the sample should be selected independently of one another and all items

of the universe should have the same chance of being selected in the sample. By independence of selection we mean that the selection of a particular item in one draw has influence on the probabilities of selection in any other draw.

(iv) Homogeneity

When we talk of homogeneity we mean that there is no basic difference in the nature of units of the universe and that of the sample. If two samples from the same universe are taken, they should give more or less the same unit.

Q12. Explain the need for Sampling.

Ans :

Sampling is used in practice for a variety of reasons like,

- (i) Sampling can save time and money. A sample study is usually less expensive than a census study and produces results at a relatively faster speed.
- (ii) Sampling is the only way when population contains infinitely many members.
- (iii) Sampling may enable more accurate measurements for a sample study is generally conducted by trained and experienced investigators.
- (iv) Sampling remains the only choice when a test involves the destruction of the items under study.
- (v) Sampling usually enables to estimate the sampling error, and thus, assists in obtaining information concerning some characteristics of the population.

4.4.1 Reasons of Sample Survey

Q13. Define Sample Survey. What are the reasons of Sample Survey?

Ans :

A sample survey is defined as a survey which is conducted using a sampling method. This implies that only a portion and not entire population is

surveyed. In addition there are various essentials factors to be considered while drawing a sample. Both questionnaire and schedule are popular used method of collecting data in research surveys.

1. Movement of Population Element

The population of fish, birds, snakes, mosquitoes, etc. are large and are constantly moving, being born and dying. So instead of attempting to count all elements of such populations, it is desirable to make estimates using techniques such as counting birds at a place picked at random, setting nets at predetermined places, etc.

2. Cost and/or Time Required to Contact the Whole Population

A census involves a complete count of every individual member of the population of interest, such as persons in a state, households in a town, shops in a city, students in a college, and so on. Apart from the cost and the large amount of resources (such as enumerators, clerical assistance, etc.) that are required, the main problem is the time required to process the data. Hence the results are known after a big gap of time.

3. Destructive Nature of Certain Tests

The census becomes extremely difficult, if not impossible, when the population of interest is either infinite in terms of size (number); constantly changing; in a state of movement; or observation results required destruction. For example, sometimes it is required to test the strength of some manufactured item by applying a stress until the unit breaks. The amount of stress that results in breakage is the value of the observation that is recorded. If this procedure is applied to an entire population, there would be nothing left. This type of testing is called destructive testing and requires that a sample be used in such cases.

4.4.2 Bias in Survey

Q14. What are the types of bias during sample survey.

Ans :

Not all surveys produce trust worthy results. Results based on a survey are biased when the method used to obtain those results would consistently produce values that are either too high or too low. Following are the common types of bias that might occur in surveys:

1. Undercoverage Bias

This bias occur when a random sample chosen does not represent the population of interest. For instance, passengers at a railway station are surveyed to determine attitude towards buying station ticket, the results are not likely to represent all passengers at railway stations.

2. Non-response Bias

This bias occurs when only a small number of respondents re-pond or return their questionnaire. The sample results would be biased because only those respondent who were particularly concerned about the subject chose to respond.

3. Wording Bias

This bias occurs when respondents respond differently from how they truly feel. It may be due to the reason that the questionnaire contains questions that tend to confuse the respondents. For instance, questions for the survey about the use of drug, payment of income tax, abusive behavior, etc. must be worded and conducted carefully to minimise response bias.

4.4.3 Definitions of Population, Sample, Parameter of Statistics

Q15. Discuss the basic terminology used in sampling.

Ans :

1. Population

Population is a collection of all objects, animate t inanimate under study is known as a

population. It can be a finite population with a finite number of objects and finite population with infinite number of objects.

Example

- (i) The numbers of students in a class is 60. Here $N = 60$ which is finite i.e., the population is finite.
- (ii) The number of stars in the sky. Here N is infinite. Therefore the population is infinite population.

2. Sample

A finite subset of the population is called as sample. It is also a part of the population.

Sample Size : is the number of objects in the sample.

Example

Selecting 2 girls in a class of 20 students is a sample

Note

A sample with less than 30 items is a small sample and a sample with more than 30 items is a large sample.

3. Random Sample

A random sample is one in which each member of the population has equal chance or probability of being included in the sample which is taken from the population.

4. Parameters of Statistics

- (i) Population parameters or parameters: The measures of population are mean (μ), variance (σ^2), standard deviation (σ) and proportions (P).
- (ii) Sample statistics or statistics : The measures computed from the sample observations are mean ($\bar{X}$), variance (s^2) standard deviation (S) and proportions (P).

4.4.4 Principles of Sampling

Q16. What are the various Principles of Sampling ?

Ans :

The following important principles which determine the possibility of arriving at a valid statistical inference about the features of a population or process:

1. Principle of statistical regularity
2. Principle of inertia of large numbers
3. Principle of Optimization
4. Principle of Persistence of Small Numbers
5. Principles of Validity

4.4.4.1 Principle of Statistical Regularity

Q17. Explain the principle of statistical regularity.

Ans :

(May-18)

This principle is based on the mathematical theory of probability. According to King, 'The law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to process the characteristic of the large group.' This principle, emphasises on two factors:

(i) Sample Size Should be Large

As the size of sample increases, it becomes more and more representative of parent population and shows its characteristics. However, in actual practice, large samples are more expensive. Thus a balance has to be maintained between the sample size, degree of accuracy desired and financial resources available

(ii) Samples Must be Drawn Randomly

The random sample is the one in which elements of the population are drawn in a such way that each combination of elements has an equal probability of being selected in the sample. When the term random sample is used without any specification, it usually refers to a simple random sample. The selection of samples based on this principle can reduce the amount of efforts required in arriving at a conclusion about the characteristic of a large population. For example, to understand the book buying habit of students in a college, instead of approaching every student, it is easy to talk to a randomly selected group of students to draw the inference about all students in the college.

4.4.4.2 Principle of Inertia of Large Numbers**Q18. Explain the Principle of Inertia of Large Numbers**

Ans : (May-18, May-17)

This principle states that other things being equal, as the sample size increases, the results tend to be more reliable and accurate. This is based on the fact that the behavior of a phenomena 'en masse' i.e., a large scale is generally stable. By this we mean that if individual events are observed, their behavior may be erratic and unpredictable but when a large number of events, are considered, they tend to behave in a stable pattern.

The difference in the aggregate result is likely to be insignificant when the number in the sample is large, for when large numbers are considered, the variations in the component parts tend to balance each other. Therefore, the variation is insignificant.

Eg. A coin is tossed 20 times, normally we expect an equal number of heads and tails, i.e., 10 each. But in reality, we may not get equal numbers of heads and tails. The result may be of any combination of head and tail. If the same is done 10000 times, there is a chance of getting 5000 heads and 5000 tails i.e., 50-50 chance of head and tail.

The basic reason for such likelihood is that the experiment has been carried out a sufficiently large number of times and the possibility of variation in one direction compensating for other in a different direction is greater. If at one time we get continuously 5 heads, then at the other time, we may get continuously 5 tails, and so on.

4.4.4.3 Principle of Optimization**Q19. Explain the principle of Optimization.**

Ans : (May-18)

This principle impresses upon obtaining optimum results in terms of efficiency and cost of the design with the resources at our disposal. The reciprocal of sampling variance of an estimate provides a measure of its efficiency, which the total expenses incurred in terms of money and man-hours provide a measure of the cost of design. The principle of optimization consists of

- i) Achieving a given level of efficiency at minimum cost and
- ii) Obtaining maximum possible efficiency with a given level of cost.

4.4.4.4 Principle of Persistence of Small Numbers**Q20. Explain the Principle of Persistence of Small Numbers.**

Ans : (May-18)

If some of the items in a population possess markedly distinct characteristics from other items, then this tendency would be revealed in the sampled unit also.

4.4.4.5 Principles of Validity**Q21. Explain the Principles of Validity.**

Ans : (May-18)

By the validity of a sample design we mean that it should enable us to obtain tests and estimates about the parameters of the population. The samples obtained technique of probability sampling satisfy this principle.

4.5 METHODS OF SAMPLING

4.5.1 Probability and Non probability Sampling Methods

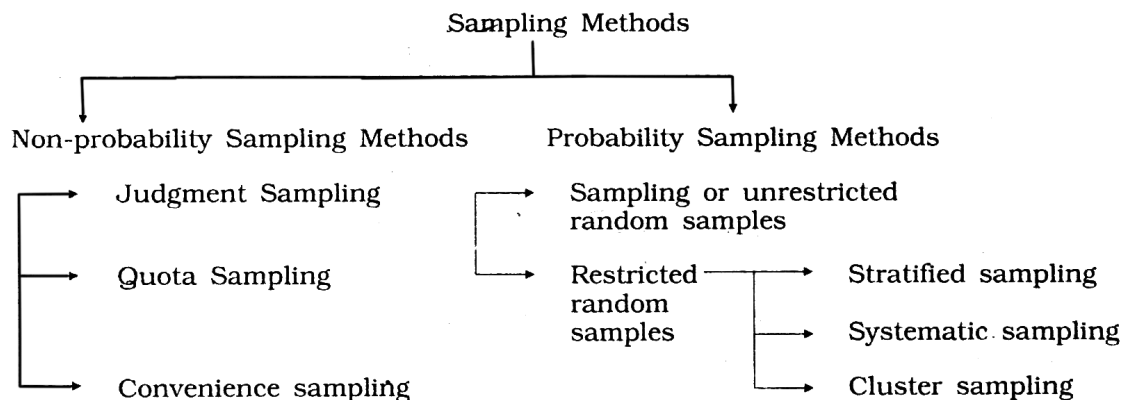
Q22. Explain the various Methods of Sampling.

Ans :

(Jan.-21, May-19)

The various methods of sampling are grouped into two categories :

- (i) Non - Probability Sampling Methods
- (ii) Probability Sampling Methods



Simple random sampling refers to that sampling technique in which each and every unit of the population has an equal opportunity of being selected in the sample. In simple random sampling which items get selected in the sample is just a matter of chance—personal bias of the investigator does not influence the selection.

I. Non - Probability Sampling Methods

(a) Judgment Sampling

In this method of sampling the choice of sample items depends exclusively on the judgment of the investigator. In other words, the investigator exercises his judgment in the choice and includes those items in the sample which he thinks are most typical of the universe with regard to the characteristics under investigation.

For example, if sample of ten students is to be selected from a class of sixty for analysing the spending habits of students, the investigator would select 10 students who, in his opinion, are representative of the class.

Merits

Though the principles of sampling theory are not applicable to judgment sampling, the method is sometimes used in solving many types of economic and business problems. The use of judgment sampling is justified under a variety of circumstances:

- (i) When only a small number of sampling units are in the universe, simple random selection may miss the more important elements, whereas judgment selection would certainly include them in the sample.
- (ii) When we want to study some unknown traits of a population, some of whose characteristics are known, we may then stratify the population according to these known properties and select sampling units from each stratum on the basis of judgment. This method is used to obtain a more representative sample.

- (iii) In solving everyday business problems and making public policy decisions, executives and public officials are often pressed for time and cannot wait for probability sample designs. Judgment sampling is then the only practical method to arrive at solutions to their urgent problems.

Limitations

- (i) This method is not scientific because the population units to be sampled may be affected by the personal prejudice or bias of the investigator. Thus, judgment sampling involves the risk that the investigator may establish foregone conclusions by including those items in the sample which conform to his preconceived notions.

For example, if an investigator holds the view that the wages of workers in a certain establishment are very low, and if he adopts the judgment sampling method, he may include only those workers in the sample whose wages are low and thereby establish his point of view which may be far from the truth. Since an element of subjectiveness is possible, this method cannot be recommended for general use.

- (ii) There is no objective way of evaluating the reliability of sample results. The success of this method depends upon the excellence in judgment. If the individual making decisions is knowledgeable about the population and has good judgment, then the resulting sample may be representative, otherwise the inferences based on the sample may be erroneous. It may be noted that even if a judgment sample is reasonably representative, there is no objective method for determining the size or likelihood of sampling error. This is a big defect of the method.

(b) Quota Sampling

Quota sampling is a type of judgment sampling and is perhaps the most commonly used sampling technique in non-probability category. In a quota sample, quotas are set up according to some specified characteristics such as so many in each of several income groups, so many in each age, so many with certain political or religious affiliations,

and so on. Each interviewer is then told to interview a certain number of persons which constitute his quota. Within the quota, the selection of sample items depends on personal judgments. For example, in a radio listening survey, the interviewers may be told to interview 500 people living in a certain area and that out of every 100 persons interviewed 60 are to be housewives, 25 farmers and 15 children under the age of 15. Within these quotas the interviewer is free to select the people to be interviewed. The cost per person interviewed may be relatively small for a quota sample but there are numerous opportunities for bias which may invalidate the results.

For example, interviewers may miss farmers working in the fields or talk with those housewives who are at home. If a person refuses to respond, the interviewer simply selects someone else. Because of the risk of personal prejudice and bias entering the process of selection, the quota sampling is not widely used in practical work.

Quota sampling and stratified random sampling are similar in as much as in both methods the universe is divided into parts and the total sample is allocated among the parts. However, the two procedures diverge radically. In stratified random sampling the sample with each stratum is chosen at random. In quota sampling, the sampling within each cell is not done at random, the field representatives are given wide latitude in the selection of respondents to meet their quotas.

Quota sampling is often used in public opinion studies. It occasionally provides satisfactory results if the interviewers are carefully trained and if they follow their instructions closely. It is often found that since the choice of respondents within a cell is left to the field representatives, the more accessible and articulate people within a cell will usually be the ones who are interviewed. Slight negligence on the part of interviewers may lead to interviewing ineligible respondents. Even with alert and conscientious field representatives it is often difficult to determine such control category as age, income, educational qualifications, etc.

(c) Convenience Sampling

A Convenience sample is obtained by selecting 'convenient' population units. The method of convenience sampling is also called the chunk. A

chunk refers to that fraction of the population being investigated which is selected neither by probability nor by judgment but by convenience. A sample obtained from readily available lists such as automobile registrations; telephone directories, etc., is a convenience sample and not a random sample even if the sample is drawn at random from the lists. If a person is to submit a project report on labour-management relations in textile industry and he takes a textile mill close to his office and interviews some people over there, he is following the convenience sampling method. Convenience samples are prone to bias by their very nature—selecting population elements which are convenient to choose almost always make them special or different from the best of the elements in the population in some way.

II. Probability sampling methods

(a) Simple or Unrestricted Random Sampling

Simple random sampling refers to that sampling technique in which each and every unit of the population has an equal opportunity of being selected in the sample. In simple random sampling which items get selected in the sample is just a matter of chance—personal bias of the investigator does not influence the selection.

- All n items of the sample are selected independently of one another all N items in the population have the same chance of being included in the sample. By independence of selection we mean that the selection of a particular item in one draw has no influence on the probabilities of selection in any other draw.
- At each selection, all remaining items in the population have the same chance of being drawn. If sampling is made with replacement, i.e., when each unit drawn from the population is returned prior to drawing the next unit, each item has a probability of $1/N$ of being drawn at each selection. If sampling is without replacement, i.e., when each unit drawn from the population is not returned prior to drawing the next unit, the probability of selection of each item remaining in the population at the first draw is $1/N$, at the

second draw is $1/(N-1)$, at the third draw is $1/(N-2)$, and so on. It should be noted that sampling with replacement has very limited and special use for statistics—we are mostly concerned with sampling without replacement.

- All the possible samples of a given n are equally likely to be selected.

(b) Lottery Method

Lottery Method This is a very popular method of taking a random sample. Under this method, all items of the universe are numbered or named on separate slips of paper of identical size and shape. These slips are then folded and mixed up in a container or drum. A blindfold selection is then made of the number of slips required to constitute the desired sample size. The selection of items thus depends entirely on chance. The method would be quite clear with the help of an example. If we want to take a sample of 10 persons out of a population of 100, the procedure is to write the names of the 100 persons on separate slips of paper, fold these slips, mix them thoroughly and then make a blindfold selection of 10 slips.

The above method is very popular in lottery draws where a decision about prizes is to be made. However, while adopting lottery method it is absolutely essential to see that the slips are of identical size, shape and colour, otherwise there is a lot of possibility of personal prejudice and bias affecting the results.

III. Restricted Random Sampling

(a) Stratified Sampling

Stratified random sampling or simply stratified sampling is one of the random methods which, by using the available information concerning the population, attempts to design a more efficient sample than obtained by the simple random procedure.

While applying stratified random sampling technique, the procedure followed is given below:

- (a) The universe to be sampled is sub-divided (or stratified) into groups which are mutually exclusive and include all items in the universe.
- (b) A simple random sample is then chosen independently from each group.

This sampling procedure differs from simple random sampling is that in the latter the sample items are chosen at random from the entire universe. In stratified random sampling the sampling is designed so that a designated number of items is chosen from each stratum. In simple random sampling the distribution of the sample among strata is left entirely to chance.

(b) Systematic Sampling

A systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals until the sample has been formed. This method is popularly used in those cases where a complete list of the population from which sample is to be drawn is available. The list may be prepared in alphabetical, geographical, numerical or some other order. The items are serially numbered. The first item is selected at random generally by following the Lottery method. Subsequent items are selected by taking every k th item from the list where ' k ' refers to the sampling interval or sampling ratio, i.e., the ratio of population size to the size of the sample. Symbolically,

$$k = \frac{N}{n}$$

where k = Sampling interval, N = Universe size, and n = Sample size.

While calculating k , it is possible that we get a fractional value. In such a case we should use approximation procedure, i.e., if the fraction is less than 0.5 it should be omitted and if it is more than 0.5 it should be taken as 1. If it is exactly 0.5 it should be omitted, if the number is even and should be taken as 1, if the number is odd. This is based on the principle that the number after approximation should preferably be even.

(c) Multi-stage Sampling (or) Cluster Sampling

Under this method, the random selection is made of primary, intermediate and final (or the ultimate) units from a given population or stratum. There are several stages in which the sampling process is carried out. At first, the first stage units are sampled by some suitable method, such as simple random sampling. Then, a sample of second stage units is selected from each of the selected first stage units, again by some suitable method which may be the same as or different from the method employed for the first stage units. Further stages may be added as required.

Q23. What are the differences between random sampling and non random sampling.

Ans :

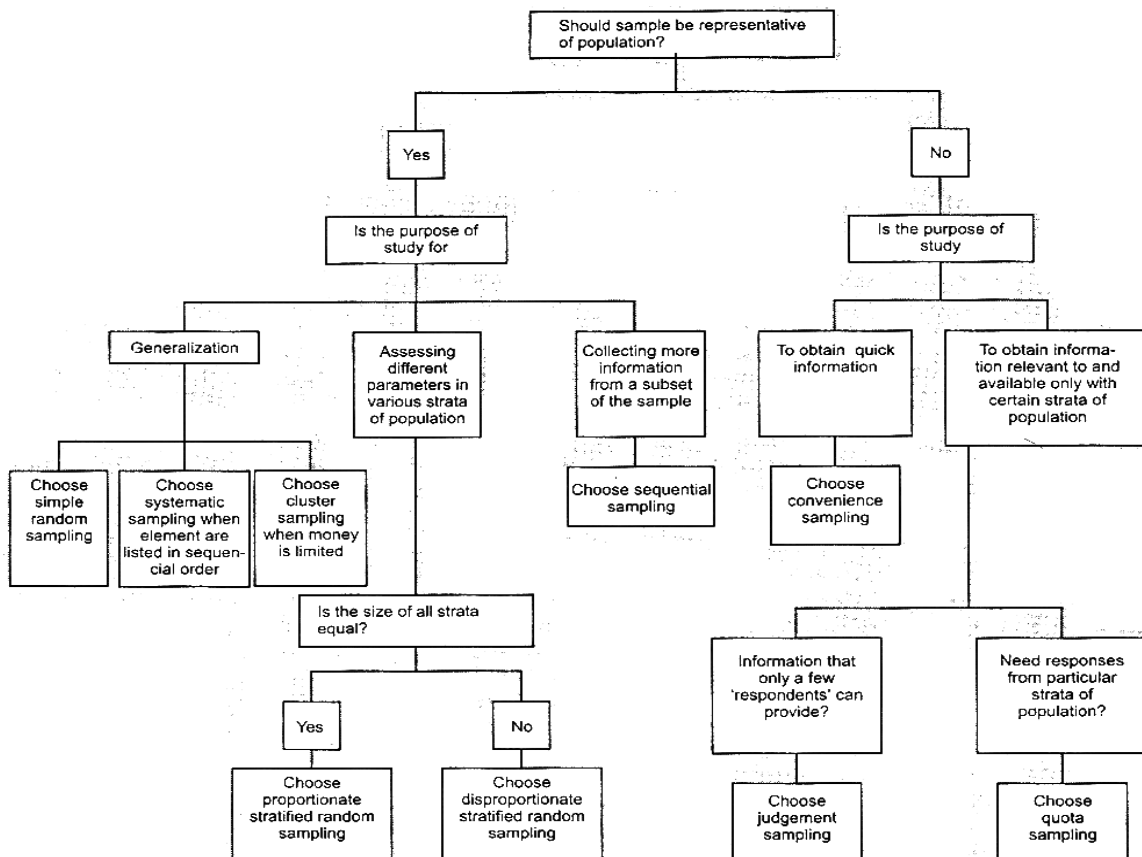
S.No.	Random Sampling	S.No.	Non-random Sampling
1.	Random sampling is also called as probability sampling. As the sampling process is random, the laws of probability are applied.	1.	A non-random sample is selected on the basis other than probability considerations, such as expert judgement, convenience etc.
2.	Random sampling can be of different types such as, simple random sampling, stratified sampling, systematic sampling etc.	2.	Non-random sampling can be of different types such as judgement sampling, quota convenience sampling etc.
3.	The population in random sampling is represented in a better way due to its large size.	3.	Non-random sample is judgemental and depends exclusively on the discretion of the investigator.

4.6 CHOICE OF SAMPLING METHOD

Q24. Explain the choice of sample methods.

Ans :

The choice of particular sampling method (procedure) must be decided according to various factors such as : nature of study, size of the population, size of the sample, availability of resources, degree of precision desired, etc. A choice plan is shown in figure.



Judging the Reliability of a Sample

The reliability of a sample can be determined in the following ways to ensure dependable results:

- (i) A number of samples may be taken from the same population and the results of various samples compared. If there is not much variation in the results of the different samples, it is a measure of its reliability.
- (ii) Sub-sample may be taken from the main sample and studied. If the results of the sub-samples are similar to those given by the main sample it gives a measure of its reliability.
- (iii) If some mathematical properties are found in the distribution under study, the sample result can be compared with expected values obtained on the basis of mathematical relationship and if the difference between them is not significant, the sample has given dependable results.

In probability distributions where binomial, normal, Poisson or any other theoretical probability distribution is applicable, sample results can be compared with the expected values to get an idea about the reliability of the sample.

4.7 SAMPLING DISTRIBUTION

Q25. What is Sampling Distribution. Explain the characteristics of Sampling Distribution.

Ans :

Samples of a given size may be drawn randomly from a population and the statistical constants like mean, standard deviation may be computed for each such sample. The distribution of each such statistic is called sampling distribution.

Characteristic of a Sampling Distribution

Sampling distribution is a frequency distribution representing the means taken from a great many samples, of the same size. The main characteristic of this is that it approaches normal distribution even when the population distribution is not normal provided the sample size is sufficiently large (greater than 30). The significance of the sampling distribution follows from the fact that the mean of sampling distribution is the same as the mean of the population.

Q26. What is sample size determination ?

Ans :

When using a sampling method, size of the sample has to be determined, various experts have different views. The sample size depends on various factors. It could be the time limits, the financial aspect, accuracy point to be achieved. Sampling theory is not of much help to determine size of sample.

There are two considerations to determine sample size. They are,

1. Sample size should be directly proportional to the variation in the individual item. As sample size increases as variation in individual item increases and vice versa.
2. Larger the sample size more is the accuracy i.e. sample size $\propto$ accuracy level.

Q27. What are the factors to be considered to determine the sample size?

Ans :

(i) The Size of the Universe

The larger the size of the universe, the bigger should be the sample size.

(ii) The Resources Available

If the resources available are vast a larger sample size could be taken. However, in most cases resources constitute a big constraint on sample size.

(iii) The Degree of Accuracy (or) Precision Desired

The greater the degree of accuracy desired, the larger should be the sample size. However, it does not necessarily mean that bigger samples always ensure greater accuracy. If a sample is selected by experts by following scientific method, it may ensure better results even when it is small compared to a situation in which a large sample size is selected by inexperienced people.

(iv) Homogeneity (or) Heterogeneity of the Universe

If the universe consists of homogeneous units a small sample may serve the purpose but if the universe consists of heterogeneous units a large sample may be inevitable.

(v) Nature of Study

For an intensive and continuous study a small sample may be suitable. But for studies which are not likely to be repeated and are quite extensive in nature, it may be necessary to take a larger sample size.

(vi) Method of Sampling Adopted

The size of sample is also influenced by the type of sampling plan adopted. For example, if the sample is a simple random sample it may necessitate a bigger sample size. However, in a properly drawn stratified sampling plan, even a small sample may give better results.

(vii) Nature of Respondents

Where it is expected a large number of respondents will not co-operate and send back the questionnaire, a large sample should be selected.

9. Given standard deviation (σ) = 8, sample mean = 24, population mean = 20 and value at a specified level of confidence or desired degree of precision is 95% Find out sample size from the given information.

Sol :

Formula used to determine sample size is given below.

$$n = \left(\frac{z\sigma}{d} \right)^2$$

Where,

- n - Sample size
- z - Desired degree of precision
- d - Standard error or difference between population mean and sample mean
- σ - Standard deviation of the population.

Desired degree of precision given is 95% i.e. 5% (100% - 95%) level of significance. From the normal distribution table, z value is 1.96 at 5% level of significance. Thus, $z = 1.96$, $d = 4$ (24 - 20) and $\sigma = 8$.

Substituting the Values in the Formula

$$n = \left(\frac{1.96 \times 8}{4} \right)^2$$

$$n = (3.92)^2 = 15.37 = 15$$

Thus, sample size is equal to 15.

10. Given $p=0.6$, $q=0.4$ and $\sigma = 0.006$. Determine sample size from the given data

Sol:

Sample size can be determined from standard error formula. Formula for calculating standard error is given below.

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

$$\sigma_p = \sqrt{\frac{pq}{n}} \Rightarrow \sigma_p^2 = \frac{pq}{n}$$

$$\Rightarrow n = \frac{pq}{\sigma_p^2}$$

Substituting the given data i.e., $p=0.6$, $q=0.4$ and $\sigma_p = 0.006$ in the formula.

$$n = \frac{pq}{\sigma_p^2}$$

$$n = \frac{0.6 \times 0.4}{(0.006)^2} = \frac{0.24}{0.000036} = 6667$$

Thus, sample size is 6667.

11. Determine the sample size if $\sigma = 6$, population mean = 25, sample mean = 23 and the desired degree of precision is 90 per cent.

Sol :

$$n = \left(\frac{z\sigma}{d} \right)^2$$

$$\sigma = 6, d = 25 - 23 = 2$$

$$z = 2.576 \text{ (at 1\% level the Z value is 2.576)}$$

Substituting the values :

$$n = \left(\frac{2.576 \times 6}{2} \right)^2 = 7.728^2 = 59.72 \text{ or } 60.$$

4.8 STANDARD ERROR (SE)

Q28. Define Standard Error (SE). What are the uses of Standard Error?

Ans :

(Jan.-21)

The standard deviation of a sampling distribution of a statistic is known as its standard error. It is so called because it measures the sampling variability due to chance or random forces. If the universe distribution is not normal, then the sampling

distribution of sample means approaches normality as the sample size increases. The word 'error' is used in place of 'deviation' to emphasize that variation among sample means is due to sampling errors. The standard error gives us an idea about reliability and precision of an error.

Uses of Standard Error

Standard deviation of sampling distribution of a statistic is known as its standard error. The uses are.

- (i) The standard error helps in testing whether the difference between observed and expected frequencies could arise due to change.
- (ii) Standard error gives an idea about the reliability and precision of a sample.
- (iii) Standard error enables us to specify the limits within which the parameters of the population are expected to lie with a specified degree of confidence.

Q29. Explain the merits and limitations of sampling.

Ans :

Merits

The sampling technique has the following merits over the complete enumeration survey:

(i) Less Time-consuming

Since the sample is a study of a part of the population, considerable time and labour are saved when a sample survey is carried out. Time is saved not only in collecting data but also in processing it. For these reasons a sample provides more timely data in practice than a census.

(ii) Less Cost

Although the amount of effort and expense involved in collecting information is always greater per unit of the sample than a complete census, the total financial burden of a sample survey is generally less than that

of a complete census. This is because of the fact that in sampling, we study only a part of population and the total expense of collecting data is less than that required when the census method is adopted. This is a great advantage particularly in an underdeveloped economy where much of the information would be difficult to collect by the census method for lack of adequate resources.

(iii) More Reliable Results

Although the sampling technique involves certain inaccuracies owing to sampling errors, the result obtained is generally more reliable than that obtained from a complete count. There are several reasons for it. First, it is always possible to determine the extent of sampling errors. Secondly, other types of errors to which a survey is subject, such as inaccuracy of information, incompleteness of returns, etc., are likely to be more serious in a complete census than in a sample survey. This is because more effective precautions can be taken - in a sample survey to ensure that information is accurate and complete. For these reasons not only the total error be expected to be smaller in a sample survey but sample result can also be used with a greater degree of confidence because of our knowledge of the probable size of error. Thirdly, it is possible to avail of the services of experts and to impart thorough training to the investigators in a sample survey which further reduces the possibility of errors. Follow-up work can also be undertaken much more effectively in the sampling method. Indeed, even a complete census can only be tested for accuracy by some type of sampling check.

(iv) More Detailed Information

Since the sampling technique saves time and money, it is possible to collect more detailed information in a sample survey. For example, if the population consists of 1,000 persons in a survey of the consumption pattern of the people, the two alternative techniques available are as follows:

- (a) We may collect the necessary data from each one of the 1,000 people through a questionnaire containing, say, 100 questions (census method): or
- (b) We may take a sample of 100 persons (i.e., 10% of population) and prepare questionnaire containing as many as 10 questions. The expenses involved in the latter case would almost be the same as in the former but it will enable nine times more information to be obtained.

(v) Sampling Method

If is the only Method that can be used in Certain Cases There are some cases in which the census method is inapplicable and the only practicable means is provided by the sample method. For example, if one is interested in testing the breaking strength of chalks manufactured in a factory under the census method all the chalks would be broken in the process of testing. Hence, census method is impracticable and resort must be had to the sample method. Similarly, if the producer wants to find out whether the tensile strength of a lot of steel wires meets the specified standard, he must resort to sample method because census would mean complete destruction of all the wires. Also if the population under investigation is infinite, sampling is the only possible solution.

The Sample Method is often used to Judge the Accuracy of the Information Obtained on a Census Basis.

For example, in the population census which is conducted very often (10 years in our country) the field officers employ the sample method to determine the accuracy of information obtained by the enumerators on the census basis.

Limitations

- A sample survey must be carefully planned and executed otherwise the results obtained may be inaccurate and misleading. Of course, even for a complete count care must be taken but serious errors may arise in sampling, if the sampling procedure is not perfect.
- Sampling generally requires the services of experts, if only for consultation purposes. In the absence of qualified and experienced persons, the information obtained from sample surveys cannot be relied upon. In India, shortage of experts in the sampling field is a serious hurdle in the way of reliable statistics.
- At times the sampling plan may be so complicated that it requires more time, labour and money than a complete count. This is so if the size of the sample is a large proportion of the total population and if complicated weighted procedures are used. With each additional complication in the survey, the chances of error multiply and greater care has to be taken, which in turn, means more time and labour.
- If the information is required for each and every unit in the domain of study, a complete enumeration survey is necessary.

Exercise Problems

1. In a certain class there are 21 students in subject X, 17 in subject Y and 10 in subject Z. Of these 12 attend subjects X and Y, 5 attend subjects Y and Z, 6 attend subject X and Z. These include 2 students who attend all the three subjects. Find the probability that a student studies one subject alone.

[Ans : 8/27]

2. If $P(A) = 0.3$, $P(B) = 0.2$ and $P(C) = 0.1$, where A, B, C are independent events. Then find the probability of occurrence of at least one of the three events A, B and C.

[Ans : 0.496]

3. A box contains 4 red pens and 5 black pens. Find the probability of drawing 3 black pens one by one,

- (i) With replacement
- (ii) Without replacement.

[Ans: (i) 125/729, (ii) 5/42]

4. Two aeroplanes hit a target in succession. The probabilities of each correctly hitting the target are 0.3 and 0.2 respectively. The second will hit the target if the first misses the target. Find the probability that,

- (a) Target is hit
- (b) Both fail to hit the target.

[Ans: (a) 0.44, (b) 0.56]

5. Consider a standard deck of cards. One card is drawn at random,

- (i) Find the probability of drawing either a king or a red card.
- (ii) Find the probability of drawing a king which is also a Knave.

[Ans: 7/13]

Short Question & Answers

1. Probability.

Ans :

The term 'probability' was coined by an Italian mathematician, 'Galileo'. He was the first man who measured probability quantitatively while dealing with the problems associated with rolling of dice in gambling.

The word 'probability' or 'chance' is the most common word used in our day-to-day life.

For example, in our daily life we use certain statements like 'Probably he may win the elections', 'It is likely that India may win the match', 'She may score above 90% in the coming examination' etc.

A probability is a quantitative measure of uncertain events. It helps in determining the chances of the occurrence of an event.

The systematic and extensive study of probability theory was made by 'B.Pascal'. Pierre de Fermat', Jacques Bernoulli' in mid-seventeenth century.

Probability is a part of our daily lives, as life is uncertain. In personal and managerial decisions, the use of probability theory helps the individual to make educated guess in matters where the outcome is uncertain. Probability helps in making decisions under uncertain conditions and predicts the outcome of an event.

2. Random Experiment.

Ans :

Random experiment refers to a situation where erratic, nonspecific and arbitrary results can be Obtained.

If an experiment is conducted repeatedly for 'n' number of times under identical conditions and if the outcome of an experiment varies in each case, then such an experiment is called as 'random experiment'.

3. Mutually Exclusive Events.

Ans :

When two events cannot occur simultaneously in a single trial then such events are called as mutually exclusive or incompatible events.

Example

- (a) When a single coin is tossed, either a head or a tail can turn up, both cannot come at the same time
- (b) At a single point of time, a person can be alive or dead.

4. Non-mutually Exclusive Events.

Ans :

When two events can occur simultaneously in a single trial then such events are said to be non-mutually exclusive events.

Example

From a pack of cards, drawing a red card and drawing a queen are the two events. These two events can occur simultaneously while drawing a red queen.

Hence, these two events are said to be non-mutually exclusive events which can occur at the same time.

5. Importance of probability.

Ans :

The various practical applications of the theory of probability are as follows,

- (i) The fundamental laws of statistics, viz the law of statistical regularity and law of inertia of large numbers depends mainly on theory of probability
- (ii) Probability helps in taking effective decisions under uncertain conditions

- (iii) The main objective of the theory of probability is to reduce the complication in betting and games. It helps in deciding whether it is worthwhile betting at a game, by calculating the expected value of an action with likely return on each of the outcome, given the chances of success attached to each outcome.
- (iv) The use of probability theory is increasing in economic decision making. It is very helpful in situations of risk and uncertainty.

6. Multiplication Theorem of Probability.

Ans :

Multiplication Theorem

If 'A' and 'B' are two independent events then the probability of occurrence of both the events is equal to the product of their individual probabilities.

For independent events,

$$P(A \cap B) = P(A) \cdot P(B)$$

Similarly,

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C) \text{ and so on}$$

If 'A' and 'B' are two dependent events, in such a case multiplication theorem is altered and is given as follows. For dependent events,

$$\begin{aligned} P(A \cap B) &= P(A/B) \cdot P(B) \\ &= P(B/A) \cdot P(A) \end{aligned}$$

Where, $P(A/B)$ is a conditional probability of A given that B has occurred (The probability of occurrence of event A when event B has already occurred is the conditional probability of A given B).

7. Marginal probability.

Ans :

Marginal probability is the simple probability of the occurrence of an event.

It is also known as 'unconditional' probability or 'single probability'.

For example, when a single unbiased coin is tossed the probability that it is a head is $P(H) = \frac{1}{2}$

or probability that it is a tail, $P(T) = \frac{1}{2}$

These two events are independent and do not overlap one another i.e., the events are statistically independent of the outcomes of next coin been tossed.

The individual probabilities obtained in this case i.e., $P(H)$ or $P(T)$ are called as marginal probabilities.

8. Joint probability.

Ans :

A joint probability is the probability of occurrence of two or more simple independent events together.

In other words, the product of two marginal probabilities occurring together or in succession is called as 'joint probability'.

For example, let $P(W)$ is the probability of a girl with white complexion and $P(B)$ is the probability of a girl with black hair.

Here, $P(W)$ and $P(B)$ are marginal probability.

Probability of a girl with white complexion and with black hair together is called as 'joint probability'.

Joint probability of these two events can be represented as, $P(W \cap B)$.

$$\therefore P(W \cap B) = P(W) \cdot P(B)$$

(By multiplication theorem)

Where, $P(W \cap B) \rightarrow$ Probability of a girl with white complexion and with black hair

$P(W)$ = Probability of a girl with white complexion

$P(B)$ = Probability of a girl with black hair.

9. Define the term Sampling.

Ans :

Sampling may be defined as the selection of some part of an aggregate or totality on the basis of which a judgement or inference about the aggregate or totality is made. In other words, it is the process of obtaining information about an entire population by examining only a part of it.

In most of the research work and surveys, the usual approach happens to be to 'make generalizations' or to 'draw inferences' based on samples about the parameters of population from which the samples are taken. The researcher quite often selects only a few items from the universe for his study purpose. All this is done on the assumption that the sample data will enable him to estimate the population parameters.

The items so selected constitute what is technically called a sample, their selection process is called sample design and the survey conducted on the basis of sample is described a sample survey. Sample should be truly representative of population characteristics without any bias so that it may result in valid and reliable conclusions.

10. Need for Sampling.

Ans :

Sampling is used in practice for a variety of reasons like,

- (i) Sampling can save time and money. A sample study is usually less expensive than a census study and produces results at a relatively faster speed.
- (ii) Sampling is the only way when population contains infinitely many members.
- (iii) Sampling may enable more accurate measurements for a sample study is generally conducted by trained and experienced investigators.

- (iv) Sampling remains the only choice when a test involves the destruction of the items under study.
- (v) Sampling usually enables to estimate the sampling error, and thus, assists in obtaining information concerning some characteristics of the population.

11. What is Sampling Distribution.

Ans :

Samples of a given size may be drawn randomly from a population and the statistical constants like mean, standard deviation may be computed for each such sample. The distribution of each such statistic is called sampling distribution.

Characteristic of a Sampling Distribution

Sampling distribution is a frequency distribution representing the means taken from a great many samples, of the same size. The main characteristic of this is that it approaches normal distribution even when the population distribution is not normal provided the sample size is sufficiently large (greater than 30). The significance of the sampling distribution follows from the fact that the mean of sampling distribution is the same as the mean of the population.

Choose the Correct Answer

1. Most of the women in this sample felt that their actual weight was [c]
 - (a) about the same as their ideal weight
 - (b) less than their ideal weight.
 - (c) greater than their ideal weight
 - (d) no more than 2 pounds different from their ideal weight.
2. A chi-square test of the relationship between personal perception of emotional health and marital status led to rejection of the null hypothesis, indicating that there is a relationship between these two variables. One conclusion that can be drawn is : [d]
 - (a) Marriage leads to better emotional health.
 - (b) Better emotional health leads to marriage.
 - (c) The more emotionally healthy someone is, the more likely they are to be married
 - (d) There are likely to be confounding variables related to both emotional health and marital status.
3. A chi-square test involves a set of counts called "expected counts." What are the expected counts?
 - (a) Hypothetical counts that would occur if the alternative hypothesis were true. [b]
 - (b) Hypothetical counts that would occur if the null hypothesis were true.
 - (c) The actual counts that did occur in the observed data
 - (d) The long-run counts that would be expected if the observed counts are representative.
4. Pick the choice that best completes the following sentence. If a relationship between two variables is called statistically significant, it means the investigators think the variables are [a]
 - (a) related in the population represented by the sample.
 - (b) not related in the population represented by the sample.
 - (c) related in the sample due to chance alone.
 - (d) very important.
5. Simpson's Paradox occurs when [c]
 - (a) No baseline risk is given, so it is not known whether or not a high relative risk has practical importance
 - (b) A confounding variable rather than the explanatory variable is responsible for a change in the response variable.
 - (c) The direction of the relationship between two variables changes when the categories of a confounding variable are taken into account.
 - (d) The results of a test are statistically significant but are really due to chance.

6. What is the approximate shape of the distribution ? [c]
(a) Nearly symmetric (b) Skewed to the left.
(c) Skewed to the right (d) Bimodal (has more than one peak).
7. Of the following, which is the most important additional information that would be useful before making a decision about participation in school sports ? [c]
(a) Where was the study conducted?
(b) How many students in the study participated in after-school sports?
(c) What is the baseline risk for getting asthma?
(d) Who funded the study?
8. One use of a regression line is [d]
(a) to determine if any x-values are outliers.
(b) to determine if any y-values are outliers.
(c) to determine if a change in x causes a change in y.
(d) to estimate the change in y for a one-unit change in x.
9. Which of the following would indicate that a dataset is not bell-shaped? [c]
(a) The range is equal to 5 standard deviations.
(b) The range is larger than the interquartile range.
(c) The mean is much smaller than the median.
(d) There are no outliers.
10. The value of a correlation is reported by a researcher to be $r = 0.5$. Which of the following statements is correct ? [a]
(a) The x-variable explains 25% of the variability in the y-variable.
(b) The x-variable explains 25% of the variability in the y-variable.
(c) The x-variable explains 50% of the variability in the y-variable.
(d) The x-variable explains 50% of the variability in the y-variable.

Fill in the blanks

1. Which of the following is the explanatory variable in this study _____ .
2. Which of the following is a confounding variable in this study _____ .
3. A magazine printed a survey in its monthly issue and asked readers to fill it out and send it in. Over 1000 readers did so. This type of sample is called _____ .
4. A polling agency conducted a survey of 100 doctors on the question "Are you willing to treat women patients with the recently approved pill RU-486"? The conservative margin of error associated with the 95% confidence interval for the percent who say 'yes' is _____ .
5. Which one of these statistics is unaffected by outliers _____ .
6. A list of 5 pulse rates is: 70, 64, 80, 74, 92. What is the median for this list _____ .
7. Among people with age over 30, what's the "risk" of always exceeding the speed limit _____ .
8. Among people with age under 30 what are the odds that they always exceed the speed limit _____ .
9. What is the relative risk of always exceeding the speed limit for people under 30 compared to people over 30 _____ .
10. The median of the distribution is approximately _____ .

ANSWERS

1. Occupation
2. Exercise
3. a self-selected sample.
4. 10%
5. Interquartile range
6. 74
7. 0.20
8. 1 to 1
9. 2.5
10. 10 pounds.

UNIT V

Correlation Analysis

Scatter diagram, Positive and negative correlation, limits for coefficient of correlation, Karl Pearson's coefficient of correlation, Spearman's Rank correlation.

Regression Analysis: Concept, least square fit of a linear regression, two lines of regression, properties of regression coefficients(Simple problems only)

Time Series Analysis: Components, Models of Time Series – Additive, Multiplicative and Mixed models; Trend analysis- Free hand curve, Semi averages, moving averages, Least Square methods(Simple problems only).

5.1 CORRELATION ANALYSIS

Q1. Define Correlation. Explain the significance of Correlation.

Ans :

Meaning

Correlation is the study of the linear relationship between two variables. When there is a relationship of 'quantitative measure between two set of variables, the appropriate statistical tool for measuring the relationship and expressing each in a precise way is known as correlation.

For example, there is a relationship between the heights and weights of persons, demand and prices of commodities etc.

Correlation analysis is the statistical tool we can use to describe the degree to which one variable is linearly related to another.

Definitions

- i) **According to Croxton and Cowden,** "The appropriate statistical tool for discovering and measuring the relationship of quantitative nature and expressing it in brief formula is known as correlation".
- ii) **According to Tippet,** "The effects of correlation are to reduce the range of uncertainty of our prediction".

The coefficient of correlation measures the degree of relationship between two set of figures. As the reliability of estimates depend upon the closeness of the relationship it is imperative that utmost care be taken while interpreting the value

of coefficient of correlation, otherwise wrong conclusion can be drawn.

Significance of Measuring Correlation

1. Correction is very useful to economists to study the relationship between variables, like price and quantity demanded. To businessmen, it helps to estimate costs, sales, price and other related variables.
2. In economic theory we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply and quantity demanded; convenience, amenities, and service standards are related to customer retention; yield a crop related to quantity of fertilizer applied, type of soil, quality of seeds, rainfall and so on. Correlation analysis helps in measuring the degree of association and direction of such relationship.
3. The relation between variables can be verified and tested for significance, with the help of the correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.
4. The coefficient of correlation is a relative measure and we can compare the relationship between variables, which are expressed in different units.
5. Correlations are useful in the areas of health care such as determining the validity and reliability of clinical measures or in expressing how health problems are related to certain biological or environmental factors. For example, correlation coefficient can be used

to determine the degree of inter-observer reliability for two doctors who are assessing a patient's disease.

6. Sampling error can also be calculated.
7. Correlation is the basis for the concept of regression and ratio of variation.
8. The decision making is heavily facilitated by reducing the range of uncertainty and hence empowering the predictions.

Q2. Explain the scope of correlation analysis.

Ans :

The existence of correlation between two (or more) variables only implies that these variables

- (a) either tend to increase or decrease together or
- (b) an increase (or decrease) in one is accompanied by the corresponding decrease (or increase) in the other. The questions of the type, whether changes in a variable are due to changes in the other, i.e., whether a cause and effect type relationship exists between them, are not answered by the study of correlation analysis. If there is a correlation between two variables, it may be due to any of the following situations:

(i) One of the variable may be affecting the other

A correlation coefficient calculated from the data on quantity demanded and corresponding price of tea would only reveal that the degree of association between them is very high. It will not give us any idea about whether price is affecting demand of tea or vice-versa. In order to know this, we need to have some additional information apart from the study of correlation.

For example if, on the basis of some additional information, we say that the price of tea affects its demand, then price will be the cause and quantity will be the effect. The causal variable is also termed as independent variable while the other variable is termed as dependent variable.

(ii) The two variables may act upon each other

Cause and effect relation exists in this case also but it may be very difficult to find out which of the two variables is independent.

For example, if we have data on price of wheat and its cost of production, the correlation between them may be very high because higher price of wheat may attract farmers to produce more wheat and more production of wheat may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of wheat. For the purpose of determining a relationship between the two variables in such situations, we can take any one of them as independent variable.

(iii) The two variables may be acted upon by the outside influences:

In this case we might get a high value of correlation between the two variables, however, apparently no cause and effect type relation seems to exist between them. For example, the demands of the two commodities, say X and Y, may be positively correlated because the incomes of the consumers are rising. Coefficient of correlation obtained in such a situation is called a spurious or nonsense correlation.

(iv) A high value of the correlation coefficient may be obtained due to sheer coincidence (or pure chance)

This is another situation of spurious correlation. Given the data on any two variables, one may obtain a high value of correlation coefficient when in fact they do not have any relationship.

For example, a high value of correlation coefficient may be obtained between the size of shoe and the income of persons of a locality.

Q3. Explain the Properties of Correlation.**Ans :**

- The value of correlation 'r' varies between $[-1, +1]$. This indicates that the r values does not exceed unity.
- Sign of the correlation sign of the Covariance.
- If $r = -1$ variables are perfectly negatively correlated.
- If $r = +1$ variables are perfectly positively correlated.

If $r = 0$ variables are not correlated in a linear fashion. There may be non-linear relationship between variables.

Correlation coefficient is independent of change of scale and shifting of origin. In other words, Shifting the origin and change the scale do not have any effect on the value of correlation.

5.2 TYPES OF CORRELATION**5.2.1 Positive and Negative Correlation****Q4. Explain various types of Correlation.****(or)**

What are the different types of correlations.

Ans :

Broadly speaking, there are four types of correlation, namely,

- Positive correlation,
- Negative correlation,
- Linear correlation and
- Non-Linear Correlation.

A) Positive correlation

If the values of two variables deviate in the same direction i.e., if increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the

other variable, the corresponding correlation is said to be positive or direct.

Examples

- Sales revenue of a product and expenditure on Advertising.
- Amount of rain fall and yield of a crop (up to a point)
- Price of a commodity and quantity of supply of a commodity
- Height of the Parent and the height of the Child.
- Number of patients admitted into a Hospital and Revenue of the Hospital.
- Number of workers and output of a factory.

i) Perfect Positive Correlation

If the variables X and Y are perfectly positively related to each other then, we get a graph as shown in fig. below.

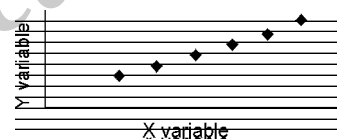


Fig.: Perfect Positive Correlation ($r = +1$)

ii) Very High Positive Correlation

If the variables X and Y are related to each other with a very high degree of positive relationship then we can notice a graph as in figure below.

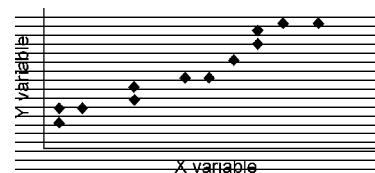
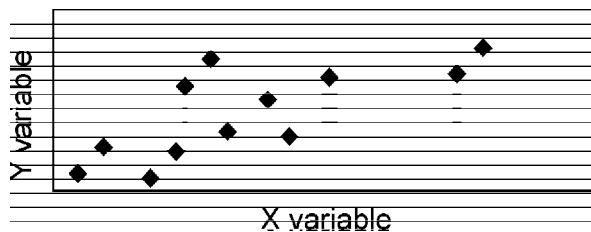


Fig.: Very High Positive Correlation ($r = \text{nearly } +1$)

iii) Very Low Positive Correlation

If the variables X and Y are related to each other with a very low degree of positive relationship then we can notice a graph as in fig. below.

Fig.: Very Low Positive Correlation ($r = \text{near to } +0$)**B) Negative Correlation**

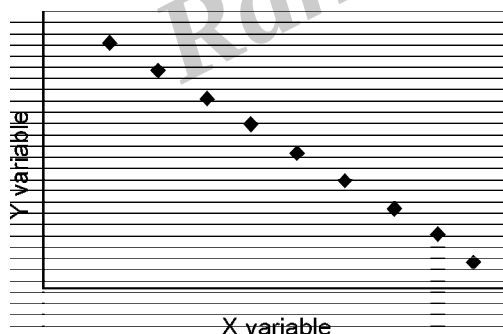
Correlation is said to be negative or inverse if the variables deviate in the opposite direction i.e., if the increase (decrease) in the values of one variable results, on the average, in a corresponding decrease (increase) in the values of the other variable.

Examples

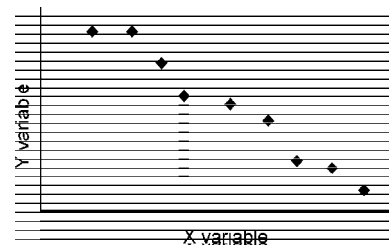
1. Price and demand of a commodity.
2. Sales of Woolen garments and the day temperature.

i) Perfect Negative Correlation

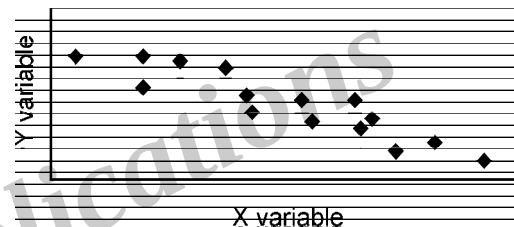
If the variables X and Y are perfectly negatively related to each other then, we get a graph as shown in fig. below.

Fig.: Perfect Negative Correlation ($r = -1$)**ii) Very High Negative Correlation**

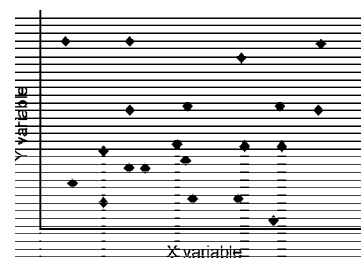
If the variables X and Y are related to each other with a very high degree of negative relationship then we can notice a graph as in fig. below.

Fig.: Very High Negative Correlation ($r = \text{near to } -1$)**iii) Very low Negative Correlation**

If the variables X and Y are related to each other with a very low degree of negative relationship then we can notice a graph as in fig. below.

Fig.: Very Low Negative Correlation
($r = \text{near to } 0 \text{ but negative}$)

- iv) No Correlation :** If the scatter diagram show the points which are highly spread over and show no trend or patterns we can say that there is no correlation between the variables.

Fig.: No Correlation ($r = 0$)**C) Linear Correlation**

Two variables are said to be linearly related if corresponding to a unit change in one variable there is a constant change in the other variable over the entire range of the values.

If two variables are related linearly, then we can express the relationship as

$$Y = a + bX$$

where 'a' is called as the "intercept" (If $X = 0$, then $Y = a$) and 'b' is called as the "rate of change" or slope.

If we plot the values of X and the corresponding values of Y on a graph, then the graph would be a straight line as shown in fig. below.

Example

X	1	2	3	4	5
Y	8	11	14	17	20

For a unit change in the value of x , a constant 3 units changes in the value of y can be noticed. The above can be expressed as : $Y = 5 + 3x$.

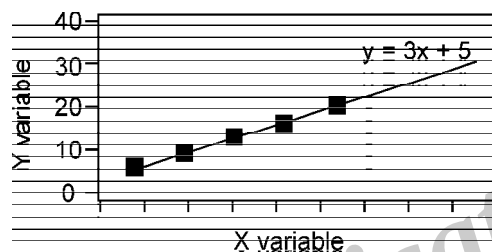


Fig.: Linear Correlation

D) Non Linear (Curvilinear) Correlation

If corresponding to a unit change in one variable, the other variable does not change in a constant rate, but change at varying rates, then the relationship between two variables is said to be nonlinear or curvilinear as shown in fig. below. In this case, if the data are plotted on the graph, we do not get a straight line curve.

Mathematically, the correlation is non-linear if the slope of the plotted curve is not constant. Data relating to Economics, Social Science and Business Management do exhibit often non-linear relationship. We confine ourselves to linear correlation only.

Example

X	-3	-2	-1	0	1	2	3
Y	9	4	1	0	1	4	9

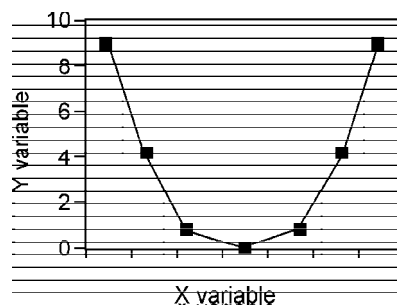


Fig.: Non Linear Correlation

5.2.2 Limits for Coefficient of Correlation**Q5. State the Limits for Coefficient of Correlation.***Ans :***Limits for Co-efficient of Correlation for (x, y)**

The value of the coefficient of correlation should lie between +1 and -1. If $r = +1$, the correlation is perfect and positive and if $r = -1$ the correlation is perfect and negative. When $r = 0$, it means that there is no relationship between the two variables.

$$\text{Hence, } -1 < r(x, y) \leq 1$$

Note

The correlation coefficient describes not only the magnitude of correlation but also its direction. Thus, +1 would mean that correlation is positive and the magnitude of correlation is 1.

Similarly -1 means correlation is negative and the magnitude of correlation is again 1.

Degree of Correlation	The Value of $r(x,y)$ (Positive and Negative)
1. Perfect correlation	Exactly 1
2. Very high degree of correlation	0.90 and above but less than 1
3. Fairly high degree of correlation	0.75 and above but less than 0.90
4. Moderate degree of correlation	0.50 and above but less than 0.75
5. Low degree of correlation	0.25 and above but less than 0.50
6. Very high degree of correlation	Below 0.25
7. Absence of correlation	Equal to 0

Q6. Explain various methods of Correlation.*Ans :*

The various method of studying linear correlation can be shown diagrammatically as follows:

1. Scatter Diagram
2. Two way frequency table
3. Spearman's Rank Correlation Method
4. Method of Concurrent Deviations
5. Method of Least Squares.

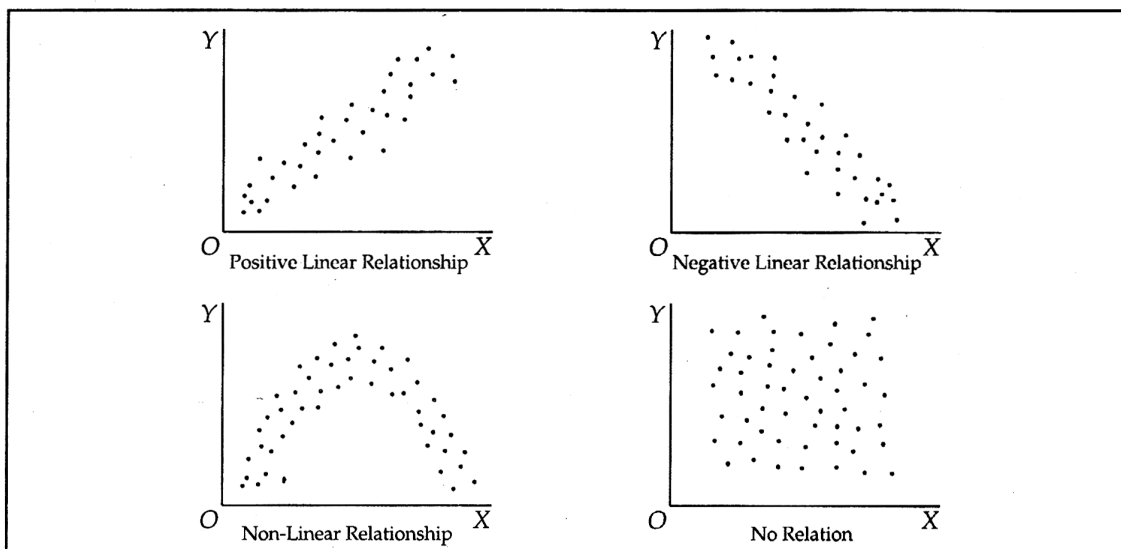
5.3 SCATTER DIAGRAM

Q7. Define Scatter Diagram.

Ans :

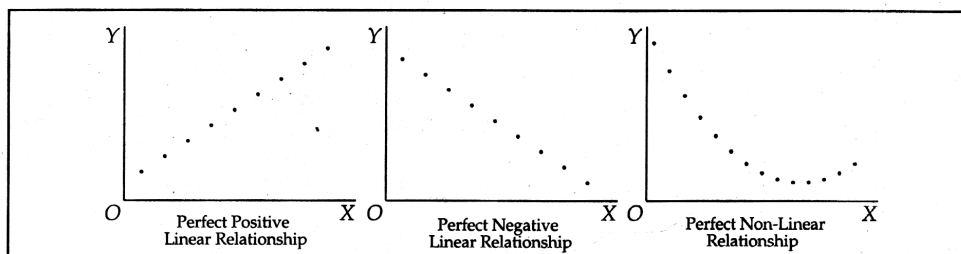
Let the bivariate data be denoted by (X_i, Y_i) , where $i = 1, 2, \dots, n$. In order to have some idea about the extent of association between variables X and Y , each pair (X_i, Y_i) , $i = 1, 2, \dots, n$, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

Each pair of values (X_i, Y_i) is denoted by a point on the graph. The set of such points (also known as dots of the diagram) may cluster around a straight line or a curve or may not show any tendency of association. Various possible situations are shown with the help of following diagrams:



Figure

If all the points or dots lie exactly on a straight line or a curve, the association between the variables is said to be perfect. This is shown below:



A scatter diagram of the data helps in having a visual idea about the nature of association between two variables. If the points cluster along a straight line, the association between variables is linear. Further, if the points cluster along a curve, the corresponding association is non-linear or curvilinear. Finally, if the points neither cluster along a straight line nor along a curve, there is absence of any association between the variables.

It is also obvious from the above figure that when low (high) values of X are associated with low (high) value of Y, the association between them is said to be positive. Contrary to this, when low (high) values of X are associated with high (low) values of Y, the association between them is said to be negative.

5.4 KARL PEARSON'S COEFFICIENT OF CORRELATION

Q8. What is Karl Pearson's Coefficient of Correlation? Explain properties of Coefficient of Correlation.

Ans : (Imp.)

Karl Pearson's Coefficient of Correlation is arrived at with the help of a statistical formula that takes into account the mean and standard deviation of the two variables, the number of such observations and the covariance between them. Since Karl Pearson's coefficient of correlation is a number, it can describe the strength of the correlation in greater detail and more objectively. A value of -1 signifies "absolute" negative correlation, a value between -1 and -0.5 signifies strong negative correlation, a value between -0.5 and -0.25 signifies moderate negative correlation and a value between -0.25 and 0 signifies weak negative correlation. Similarly, a value of $+1$ signifies "absolute" positive correlation, a value between $+1$ and $+0.5$ signifies strong positive correlation, a value between $+0.5$ and $+0.25$ signifies moderate positive correlation and a value between $+0.25$ and 0 signifies weak positive correlation.

Properties of Karl Pearson's Coefficient of Correlation

1. It is based on Arithmetic Mean and Standard Deviation.
2. It lies between -1
3. It measures both direction as well as degree of change. If r is less than 0 , there is negative correlation, which means the direction of change of the two variables will be opposite. If r is more than 0 , there is positive correlation, which means the direction of change of the two variables will be same.

Higher the value of r , greater is the degree of correlation. Hence, Karl Pearson's coefficient of correlation is said to be the ideal measure of correlation.

4. It is independent of change in scale. In other words, if a constant amount is added/subtracted from all values of a variable, the value of r does not change.
5. It is independent of change in origin. Thus, if a constant amount is multiplied with or divides all values of a variable, r does not change.
6. It is independent of direction. In other words, Correlation of X and Y is same as Correlation of Y and X.
7. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.
8. It takes into account all items of the variable(s).
9. It does not prove causation but is simply a measure of co-variation.
10. Correlation coefficient of two variables X and Y is the Geometric Mean of two regression coefficients, regression coefficient of X on Y and regression coefficient of Y on X. Symbolically,

$$r = \text{Square root of } (b_{xy} * b_{yx})$$
11. Correlation coefficient can be calculated between two unrelated variables and such a number can be misleading. Such correlation is called accidental correlation, spurious correlation or non sense correlation.

Q9. Explain merits and demerits of Coefficient of Correlation.

Ans :

Merits of Karl Pearson's Coefficient of Correlation

1. It takes into account all items of the variable(s).
2. It is a numerical measure and hence more objective.
3. It measures both direction as well as degree of change.

4. It facilitates comparisons between two series.
5. It is capable of further Algebraic treatment
6. It is more practical and hence popular and is more commonly used.

Demerits of Karl Pearson's Coefficient of Correlation

1. It is not easy to calculate as complex formulae are involved.
2. It is more time consuming compared to methods such as rank correlation
3. It assumes a linear relationship between the two variables which may not be correct
4. It is impacted by extreme values as it is based of mean and standard deviation.
5. It is not easy to interpret.

Q10. Explain the methods of Coefficient of Correlation.

Ans :

i) Direct Method when deviations are taken from actual mean

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

However, this formula is transformed in the following form

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Where

$$x = X - \bar{X} \text{ and } y = Y - \bar{Y}$$

Steps :

1. Find the means of the two series ($\bar{X}$, $\bar{Y}$)
2. Take the deviations of X series from the mean of X and denote these deviations as x.
3. Square these deviations and obtain the total. Denote it as $\sum x^2$.
4. Take the deviations of Y series from the Mean of Y and denote these deviations as y.
5. Square these deviations, obtain the total and denote it as $\sum y^2$.
6. Multiply the deviations of X and Y series, obtain the total and denote it $\sum xy$.
7. Substitute the above values in the formula.

ii) Short-Cut Method

When deviations are taken from assumed mean.

When actual mean is in fraction, then the above formula becomes tedious. In such cases, assumed mean is used for calculating correlation. The formula is.

$$r = \frac{\sum dxdy - \frac{\sum dx \cdot \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}} \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

Where

$\Sigma dx dy$ = Sum of the product of the deviations of X and Y series from their assumed means.

Σdx^2 = Sum of the squares of the deviations of X series from an assumed mean.

Σdy^2 = Sum of the squares of the deviations of Y series from an assumed mean.

Σdx = Sum of the deviations of X series from an assumed mean.

Σdy = Sum of the deviations of Y Series from an assumed mean.

N = No. of Pairs of observations.

The values of coefficient of correlation as obtained by above formulae will always lie between ± 1 . When there is perfect positive correlation its value is +1 and when there is perfect negative correlation, its value is -1. When $r = 0$ means that there is no relationship between the two variables. We normally get values which lie between +1 and -1.

Probable Error of the Coefficient of correlation and its interpretation.

The probable error of the coefficient of correlation helps in interpretation. The probable error of the coefficient of correlation is obtained as follows:

$$\text{P.E. of } r = 0.6745 \frac{1-r^2}{\sqrt{N}}$$

Where

r = Coefficient of correlation;

N = Number of pairs of observations.

If the probable error is added to and subtracted from the coefficient of correlation, it would give two such limits within which we can reasonably expect the value of coefficient of correlation to vary.

Symbolically $P(\rho) = r \pm \text{P.E.}$

Where 'P' denotes the correlation in the population. Suppose, the Coefficient of correlation for a pair of 10 observations is 0.8 and its P.E. is 0.05. the limits of the correlation in the population would be $r \pm \text{P.E.}$ i.e. 0.8 ± 0.05 or 0.75 – 0.85.

If the value of r is less than the probable error then r is not at all significant, i.e. there is no evidence of correlation. If the value of r is more than six times the probable error, it is significant. Hence it can be said that r is significant, when

$$r > 6 \text{ P.E. or } \frac{r}{\text{P.E.}} > 6$$

PROBLEMS

1. Calculate coefficient of correlation and find the probable error.

X	14	19	24	21	26	22	15	20	19
Y	31	36	48	37	50	45	33	41	39

Sol :

$$\text{Average of X} = \bar{X} = \frac{\Sigma X}{N} = \frac{180}{9} = 20$$

$$\text{Average of } Y = \bar{Y} = \frac{\Sigma Y}{N} = \frac{360}{9} = 40$$

Karl Pearson's correlation coefficient,

$$r = \frac{\Sigma dx.dy}{\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}}$$

Where $x = X - \bar{X}$, $y = Y - \bar{Y}$

(i.e., deviations taken from actual mean)

X	Y	dx (X - 20)	dy (Y - 40)	dx ²	dy ²	dx dy
14	31	-6	-9	36	81	54
19	36	-1	-4	1	16	4
24	48	4	8	16	64	32
21	37	1	-3	1	9	-3
26	50	6	10	36	100	60
22	45	2	5	4	25	10
15	33	-5	-7	25	49	35
20	41	0	1	0	1	0
19	39	-1	-1	1	1	1
$\Sigma X = 180$	$\Sigma Y = 360$	$\Sigma dx = 0$	$\Sigma dy = 0$	$\Sigma x^2 = 120$	$\Sigma y^2 = 346$	$\Sigma dx dy = 193$

$$r = \frac{\Sigma dx.dy}{\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}} = \frac{193}{\sqrt{120} \cdot \sqrt{346}}$$

$$= \frac{193}{10.95 \times 18.60} = \frac{193}{203.67} = 0.947$$

Karl Pearson's coefficient of correlation = 0.947

Probable Error

$$P.E = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$$

$$= 0.6745 \times \frac{1-(0.947)^2}{\sqrt{9}}$$

$$= 0.6745 \times 0.0344$$

$$= \pm 0.0232$$

2. Calculate Pearson's coefficient of correlation for the following data using 44 and 26 assumed means of X and Y respectively

X	43	44	46	40	44	42	45	42	38	40	42	57
Y	29	31	19	18	19	27	27	29	41	30	26	10

Sol :

X	(X - 44) dx	dx ²	Y	(Y - 26) dy	dy ²	dx dy
43	-1	1	29	3	9	-3
44	0	0	31	5	25	0
46	2	4	19	-7	49	-14
40	-4	16	18	-8	64	32
44	0	0	19	-7	49	0
42	-2	2	27	1	1	-2
45	1	1	27	1	1	1
42	-2	4	29	2	4	-4
38	-6	36	41	15	225	-90
40	-4	16	30	4	16	-16
42	-2	4	26	0	0	0
57	13	169	10	-16	256	-208
$\Sigma X = 523$	$\Sigma dx = -5$	$\Sigma dx^2 = 255$	$\Sigma Y = 306$	$\Sigma dy = -7$	$\Sigma dy^2 = 699$	$\Sigma dx dy = -304$

$$r = \frac{N \Sigma dx dy - (\Sigma dx)(\Sigma dy)}{\sqrt{N \Sigma dx^2 - (\Sigma dx)^2} \sqrt{N \Sigma dy^2 - (\Sigma dy)^2}}$$

$$\Sigma dx dy = -304, \Sigma dx = -5, \Sigma dy = -7$$

$$\Sigma dx^2 = 255, \Sigma dy^2 = 699, N = 12$$

$$r = \frac{(12)(-304) - (-5)(-7)}{\sqrt{(12)(255) - (-5)^2} \sqrt{(12)(699) - (-7)^2}}$$

$$= \frac{-3648 - 35}{\sqrt{3060 - (-5)^2} \sqrt{(12)(699) - (-7)^2}} = \frac{-3683}{\sqrt{3035} \sqrt{8412}} = -0.728$$

Probable Error

$$\begin{aligned}
 PE_r &= 0.6745 \times \frac{1-r^2}{\sqrt{n}} \\
 &= 0.6745 \times \frac{1-(-0.728)^2}{\sqrt{12}} \\
 &= \frac{0.6745 \times 0.464}{3.464} = \frac{0.313}{3.464} = 0.090
 \end{aligned}$$

3. The following results are obtained between two series from their respective means. Compute the coefficient of correlation.

Particulars	X Series	Y Series
Number of items	7	7
Arithmetic Mean	4	8
Sum of the squares of deviations from Mean	28	76
Summation of product of deviations of X and Y series from their respective means	46	

Sol :

Given $N = 7$, $\bar{X} = 4$, $\bar{Y} = 8$, $\sum x^2 = 28$, $\sum y^2 = 76$, $\sum xy = 46$

Karl Pearson's Coefficient of Correlation

$$\begin{aligned}
 r &= \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}} \\
 &= \frac{46}{\sqrt{28 \times 76}} = \frac{46}{\sqrt{2128}} = \frac{26}{46.13} = 0.997
 \end{aligned}$$

4. Test the significance of correlation for the following values based on the number of observation (i) 10 and (ii) 100. The coefficient of correlation + 0.4 and + 0.9

Sol :

The coefficient of correlation significant if it is more than 6 times of its Probable Error i.e.

$$r > 6 \text{ P.E. or } \frac{6}{\text{P.E.}} > 6$$

Now let us study whether is significant in each of the case given above.

1. $N = 10$, $r = + 0.4$

$$\text{P.E.} = 0.6745 \cdot \frac{1-r^2}{\sqrt{N}} = 0.6745 \frac{1-(4)^2}{\sqrt{10}} = 0.18$$

$$\text{So, } \frac{4}{\text{P.E.}} = \frac{0.4}{0.18} = 2.22$$

Since $r/\text{P.E.}$ is less than 6, Coefficient of Correlation is not significant

2. $N = 100$ $r = + 0.4$

$$\text{P.E.} = 0.6745 \frac{1-(4)^2}{\sqrt{100}} = 0.06$$

$$\frac{r}{\text{P.E.}} = \frac{0.4}{0.06} = 6.67$$

Since $r/\text{P.E.}$ is more than 6, Coefficient of Correlation is significant.

3. $N=10$, $r = +0.9$

$$\text{P.E.} = 0.6745 \frac{1-(9)^2}{\sqrt{10}} = 0.04$$

$$\frac{r}{\text{P.E.}} = \frac{0.9}{0.04} = 22.5$$

r is 22.5 times its P.E. Hence it is highly significant

4. $N = 100$, $r = + 0.9$

$$\text{P.E.} = 0.6745 \frac{1-(9)^2}{\sqrt{100}} = 0.0128$$

$$\frac{r}{\text{P.E.}} = \frac{0.9}{0.0128} = 70.3$$

r is 70.3 times its P.E. hence it is highly significant.

5. Given

Total of Product of deviations of X and Y series = +85

Number of pairs of observations = 10

Total of the deviations of X series = +6

Total of the deviations of Y series = +13

Total of the square of deviations of X series 162

Total of the square of deviations of Y series 211

Find out the coefficient of correlation when arbitrary means of X series and Y series are 35 and 30 respectively.

Sol :

Given

$\Sigma dx dy = 85$, $\Sigma dx = 6$, $\Sigma dy = 13$, $\Sigma dx^2 = 162$, $\Sigma dy^2 = 211$, $N = 10$

Karl Pearson's Coefficient of Correlation

$$r = \frac{\Sigma dx \cdot dy - \frac{\Sigma dx \cdot \Sigma dy}{N}}{\sqrt{\left(\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}\right) \left(\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}\right)}} = \frac{85 - \frac{6 \times 13}{10}}{\sqrt{162 - \frac{(6)^2}{10}} \sqrt{211 - \frac{(13)^2}{10}}}$$

$$= \frac{85 - 7.8}{\sqrt{162 - 3.6} \sqrt{211 - 16.9}} = \frac{77.2}{\sqrt{158.4} \times \sqrt{194.1}}$$

$$= \frac{77.2}{12.586 \times 13.932} = \frac{772}{175.348}$$

$$= +0.44$$

5.5 SPEARMAN'S RANK CORRELATION

Q11. Explain about rank correlation and its features.

Ans :

(Imp.)

The Karl Pearson's method is based on the assumption that the population being studied is normally distributed. When it is known that the population is not normal or when the shape of the distribution is not known, there is need for a measure of correlation that involves no assumption about the parameter of the population.

It is possible to avoid making any assumptions about the populations being studied by ranking the observations according to size and basing the calculations on the ranks rather than upon the original observations. It does not matter which way the items are ranked, item number one may be the largest or it may be the smallest. Using ranks rather than actual observations gives the coefficient of rank correlation.

This method of finding out variability or the lack of it between two variables was developed by the British Psychologist Charles Edward Spearman in 1904. This measure is especially useful when quantitative measures for certain factors (such as in the evaluation of leadership ability or the judgment of female beauty) cannot be fixed, but the individual in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank in the group. Spearman's rank correlation coefficient is defined as :

$$R = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)} \text{ or } 1 - \frac{6 \Sigma D^2}{N^3 - N}$$

where R denotes rank coefficient of correlation and D refers to the difference of rank between paired items in two series.

Features of Spearman's Correlation Coefficient

- The sum of the differences of ranks between two variables shall be zero. Symbolically, $\sum D = 0$
- Spearman's correlation coefficient is distribution-free or non-parametric because no strict assumptions are made about the form of population from which sample observations are drawn.
- The Spearman's correlation coefficient is nothing but Karl Pearson's correlation coefficient between the ranks. Hence, it can be interpreted in the same manner as Pearsonian correlation coefficient.

In rank correlation we may have two types of problems :

- Where ranks are given.
- Where ranks are not given.

i) Where Ranks are Given**Steps**

- (i) Take the differences of the two ranks, i.e. $(R_1 - R_2)$ and denote these differences by D .
- (ii) Square these differences and obtain the total $\sum D^2$.
- (iii) Apply the formula $R = 1 - \frac{6\sum D^2}{N^3 - N}$

Example

The ranking of 10 students in two subjects A and B are as follows :

A	B
6	3
5	8
3	4
10	9
2	1
4	6
9	10
7	7
8	5
1	2

Calculate rank correlation coefficient.

Sol. :

Calculation of Rank Correlation Coefficient

R_1	R_2	D	$(R_1 - R_2)^2$ D^2
6	3	3	9
5	8	-3	9
3	4	-1	1
10	9	1	1
2	1	1	1
4	6	2	4
9	10	-1	1
7	7	0	0
8	5	3	9
1	2	-1	1
			$\Sigma D^2 = 36$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 36}{10^3 - 10}$$

$$= 1 - \frac{216}{990} = 0.782$$

ii) Where Ranks are Not Given

Steps :

- 1) to assign the ranks.

Note : Ranks can be as-signed by taking either highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value we must follow the same method in case of both the variables.

- 2) Take the differences of the two ranks, i.e. $(R_1 - R_2)$ and denote these differences by D.
- 3) Square these differences and obtain the total ΣD^2 .

- 4) Apply the formula $R = 1 - \frac{6\Sigma D^2}{N^3 - N}$

Example

Quotations of Index Numbers of security prices of a certain joint stock company are given below :

Year	Debenture price	Share price
1	97.8	73.2
2	99.2	85.8
3	98.8	78.9
4	98.3	75.8
5	98.4	77.2
6	96.7	87.2
7	97.1	83.8

Using rank correlation method, determine the relationship between debenture prices and share prices.

Sol :

Calculation of Rank Correlation Coefficient

X	R _x	Y	R _y	(R _x - R _y) D	(R _x - R _y) ² D ²
97.8	3	73.2	2	2	4
99.2	7	85.8	6	1	1
98.8	6	78.9	4	2	4
98.3	4	75.8	2	2	4
98.4	5	77.2	3	2	4
96.7	1	87.2	7	-6	36
97.1	2	83.8	5	-3	9
					ΣD ² = 62

$$R = 1 - \frac{6 \sum D^2}{N^3 - N} = 1 - \frac{6 \times 62}{7^3 - 7} = 1 - 1.107 = -0.107.$$

Equal Ranks

In some cases it may be found necessary to rank two or more individuals or entries as equal. In such a case it is customary to give each individual an average rank. Thus, if two individuals are ranked equal at fifth place, they are each given the rank $\frac{5+6}{2}$, that is 5.5 while, if three are ranked equal at fifth place, they are given the rank $\frac{5+6+7}{3} = 6$. In other words, where two or more items are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of the ranks which these individuals would have got had they differed slightly from each other.

Where equal ranks are assigned to some entries an adjustment in the above formula for calculating the rank coefficient of correlation is made.

The adjustment consists of adding $\frac{1}{12}(m^3 - m)$ to the value of $\sum D^2$, where M stands for the number of items whose ranks are common. If there are more than one such group of items with common rank, this value is added as many times the number of such groups. The formula can thus be written

$$R = 1 - \frac{6 \left\{ \sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \dots \right\}}{N^3 - N}$$

Example

Obtain the rank correlation coefficient between the variables X and Y from the following pairs of observed values.

X:	50	55	65	50	55	60	50	65	70	75
Y:	110	110	115	125	140	115	130	120	115	160

Sol :

For finding ranks correlation coefficient first rank two various values. Taking lowest as 1 and next higher as 2, etc..

X	Rank X R_1	Y	Rank Y R_2	D	$(R_1 - R_2)^2$ D^2
50	2	110	1.5	-0.5	0.25
55	4.5	110	1.5	3.0	9.00
65	7.5	115	4	3.5	12.25
50	2	125	7	-5	25.00
55	4.5	140	9	-4.5	20.25
60	6	115	4	2	4.00
50	2	130	8	6	36.00
65	7.5	120	6	1.5	2.25
70	9	115	4	5.00	25.00
75	10	160	10	0.00	00.00
					$\sum D^2 = 134$

It may be noted that in series X, 50 has repeated thrice ($m = 3$), 55 has been repeated twice ($m = 2$), 65 has been repeated twice ($m = 2$). In series Y, 110 has been repeated twice ($m = 2$) and 115 thrice ($m = 3$).

$$R = 1 - \frac{6 \left\{ \sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) \right\}}{N^3 - N}$$

$$\begin{aligned}
 R &= 1 - \frac{6 \left\{ 134 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3) \right\}}{10^3 - 10} \\
 &= 1 - \frac{6[134 + 2 + .5 + .5 + 2]}{990} \\
 &= 1 - \frac{6(139.5)}{990} = 1 - \frac{837}{990} = 1 - .845 = 0.155
 \end{aligned}$$

Q12. What are the merits and demerits of rank correlation ?

(or)

State the merits and demerits of rank correlation ?

Ans :

Merits

1. It is easy to calculate and understand as compared to Pearson's r .
2. When the ranks of different values of the variables are given, it is then the only method left to calculate the degree of correlation.
3. When actual values are given and we are interested in using this formula then, we have to give ranks to calculate correlation.
4. This method is employed usefully when the data is given in a qualitative nature like beauty, honesty, intelligence etc.

Demerits

1. This method cannot be employed in a grouped frequency distribution.
2. If the items exceed 30, it is then difficult to find out ranks and their differences.
3. This method lacks precision as compared to Pearson's co-efficient of correlation, as all the information concerning the variables is not used. It is just possible that the difference between r_k and r may be very insignificant.

PROBLEMS ON RANK CORRELATION

6. Twelve entries in painting competition were ranked by two judges as shown below :

Judge	1	5	2	3	4	1	6	8	7	10	9	12	11
Judge	2	4	5	2	1	6	7	10	9	11	12	3	8

Sol :

Judge 1	Judge 2	$R_1 - R_2 = d$	d^2
5	4	1	1
2	5	-3	9
3	2	1	1
4	1	3	9
1	6	-5	25
6	7	-1	1
8	10	-2	4
7	9	-2	4
10	11	-1	1
9	12	-3	9
12	3	9	81
11	8	3	9
			154

$$r_s = 1 - \frac{6 \sum d^2}{(N)^3 - N}$$

$$r_s = 1 - \frac{6 \times 154}{(12)^3 - 12}$$

$$r_s = 1 - \frac{924}{1728 - 12}$$

$$r_s = 1 - \frac{924}{1716}$$

$$r_s = 1 - 0.5385$$

$$r_s = 1 - 0.4615 \text{ or } 0.46$$

7. Ten competitors in a beauty contest are ranked by three judges in the following order.

1 st Judge	1	5	4	8	9	6	10	7	3	2
2 nd Judge	4	8	7	6	5	9	10	3	2	1
3 rd Judge	6	7	8	1	5	10	9	2	3	4

Use the rank correlation coefficient to discuss which pair of judges has the nearest approach to beauty.

Sol :

Judge 1	Judge 2	Judge 3	$R_1 - R_2$	$R_2 - R_3$		$R_1 - R_3$		
(R_1)	(R_2)	(R_3)	D_1	D_1^2	D_2	D_2^2	D_3	D_3^2
1	4	6	-3	9	-2	4	-5	25
5	8	7	-3	9	1	1	-2	4
4	7	8	-3	9	1	1	-4	16
8	6	1	2	4	5	25	7	49
9	5	5	4	16	0	0	4	16
6	9	10	-3	9	-1	1	-4	16
10	10	9	0	0	1	1	1	1
7	3	2	4	16	1	1	5	25
3	2	3	1	1	-1	1	0	0
2	1	4	1	1	-3	9	-2	4
				74		44		156

i) Between 1st and 2nd judge

$$r_s = 1 - \frac{6\sum(d)^2}{N^3 - N} = 1 - \frac{6(74)}{(10)^3 - 10}$$

$$r_s = 1 - \frac{444}{1000 - 10}$$

$$r_s = 1 - \frac{444}{990}$$

$$r_s = 1 - 0.4485$$

$$r_s = 0.5515$$

ii) Between 2nd and 3rd judge

$$r_s = 1 - \frac{6\sum(d)^2}{N^3 - N}$$

$$r_s = 1 - \frac{6 \times (44)}{(10)^3 - 10}$$

$$r_s = 1 - \frac{264}{1000 - 10}$$

$$r_s = 1 - \frac{264}{990}$$

$$r_s = 1 - 0.267$$

$$r_s = 0.733$$

iii) Between 1st judge and 3rd judge

$$r_s = 1 - \frac{6\sum(d)^2}{N^3 - N}$$

$$r_s = 1 - \frac{6 \times 156}{(10^3) - 10}$$

$$r_s = 1 - \frac{936}{1000 - 10}$$

$$r_s = 1 - \frac{936}{990}$$

$$r_s = 1 - 0.945$$

$$r_s = 0.0545$$

Since coefficient of correlation is maximum in the judgement of 2nd and 3rd judges. It is to be concluded that they have the nearest approach to common tastes in beauty.

8. Calculate the coefficient of correlation from the following data by the spearman's Rank Differences method :

Prices of Tea (₹)	75	88	95	70	60	80	81	50
Prices of Coffee(₹)	120	134	150	55	110	140	142	100

Sol :

Price of Tea (X)	R ₁	Price of Coffee (Y)	R ₂	R ₁ -R ₂ =d	d ²
75	4	120	3	1	1
88	7	134	4	3	9
95	8	150	7	1	1
70	3	55	1	-2	4
60	2	110	2	0	0
80	5	140	5	0	0
81	6	142	6	0	0
50	1	100	8	-7	49
					64

$$r_s = 1 - \frac{6\sum(d)^2}{(N)^3 - N} = 1 - \frac{6 \times 64}{(8)^3 - 8}$$

$$r_s = 1 - \frac{384}{512 - 8} = 1 - \frac{384}{504}$$

$$r_s = 1 - 0.762$$

$$r_s = 0.238$$

9. Compute Rank coefficient of correlation for the following data :

X	15	20	28	12	40	60	20	80
Y	40	30	50	30	20	10	30	60

Sol :

X	R ₁	Y	R ₂	R ₁ - R ₂ = d	d ²
15	2	40	6	-4	16
20	3.5	30	4	-0.5	0.25
28	5	50	7	-2	4
12	1	30	4	-3	9
40	6	20	2	4	16
60	7	10	1	6	36
20	3.5	30	4	-0.5	0.25
80	8	60	8	0	0
					81.5

$$r_s = 1 - \frac{6 \left[\sum d^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) \right]}{(N)^3 - N}$$

$$r_s = 1 - \frac{6 \left[81.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3) \right]}{(8)^3 - 8}$$

$$r_s = 1 - \frac{6[81.5 + 0.5 + 2]}{512 - 8}$$

$$r_s = 1 - \frac{6(84)}{504}$$

$$r_s = 1 - \frac{504}{504}$$

$$r_s = 1 - 1$$

$$= 0$$

10. Find the coefficient of rank correlation between the marks obtained in Mathematics (X) and those in Statistics (Y) by 10 students of certain class out of a total of 50 marks in each subject.

Students No.	1	2	3	4	5	6	7	8	9	10
X	12	18	32	18	25	24	25	40	38	22
Y	16	15	28	16	24	22	28	36	34	19

Sol :

X	R ₁	Y	R ₂	R ₁ - R ₂ = d	d ²
12	1	16	1.5	-0.5	0.25
18	1.5	15	1	0.5	0.25
32	8	28	7.5	0.5	0.25
18	1.5	16	1.5	0	0
25	6.5	24	6	0.5	0.25
24	5	22	5	0	0
25	6.5	28	7.5	1	1
40	10	36	10	0	0
38	9	34	9	0	0
22	4	19	4	0	0

$$r_s = 1 - \frac{6 \left[\sum d^2 \frac{1}{12} (m^3 - m) + \frac{1}{12} (m^3 - m) \right]}{N^3 - N}$$

$$r_s = 1 - \frac{6 \left[2 \times \frac{1}{12} (4^3 - 4) + \frac{1}{12} (4^3 - 4) \right]}{(10)^3 - 10}$$

$$r_s = 1 - \frac{6 \left[2 \times \frac{1}{12} (64 - 4) + \frac{1}{12} (64 - 4) \right]}{1000 - 10}$$

$$r_s = 1 - \frac{6 [2 \times 5 \times 5]}{990}$$

$$r_s = 1 - \frac{300}{990}$$

$$r_s = 1 - 0.303$$

$$r_s = 0.697$$

5.6 REGRESSION ANALYSIS

5.6.1 Concept of Regression

Q13. Define the term Regression. Explain the utility of Regression Analysis.

Ans :

Meaning

The dictionary meaning of the term 'regression' is the act of the returning or going back. The term 'regression' was first used by Sir Francis Galton in 1877 while studying the relationship between the heights of father and sons. Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables.

1. **Dependant Variable** is the single variable being explained/ predicted by the regression model (response variable).
2. **Independent Variable** is the explanatory variable(s) used to predict the dependant variable (Predictor variable).

Definitions

- (i) "Regression is the measure of the average relationship between two or more variables in terms of the original units of data."
- (ii) **According to Blair**, "Regression is the measure of the average relationship between two or more variable in terms of the original units of the data."
- (iii) **According to Taro Yamane**, "One of the most frequently used techniques in economics and business research, to find a relation between two or more variable that are related causally, is regression analysis."

Utility of Regression Analysis

1. Regression analysis helps in establishing a functional relationship between two or more variables. Once this is established it can be used for various advanced analytical purposes.
2. Since most of the problems of economic analysis are based on cause and effect

relationship, the regression analysis is a highly valuable tool in economics and business research.

3. This can be used for prediction or estimation of future production, prices, sales, investments, income, profits and population which are indispensable for efficient planning of an economy and are of paramount importance to a businessman or an economist.
4. Regression analysis is widely used in statistical estimation of demand curves, supply curves, production functions, cost functions, consumption functions, etc. Economists have discovered many types of production functions by fitting regression lines to input and output data.

Q14. What are the objectives of Regression Analysis.

Ans :

1. The first objective of regression analysis is to provide estimates of values of the dependent variable from values of independent variable. This is done with the help of the regression line. The regression line describes the average relationship existing between X and Y variables, more precisely, it is a line which displays mean values of Y for given values of X.
2. The second objective of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimation. For this purpose standard error of estimate is obtained. This helps in understanding the correlation existing between X and Y.
3. In general, we can model the expected value of y as an n^{th} order polynomial, yielding the general polynomial regression model

$$Y = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots + a_nx^n + \varepsilon$$

Conveniently, these models are all linear from the point of view of estimation, since the regression function is linear in terms of the unknown parameters $a_0, a_1, \dots$. Therefore, for least squares

analysis, the computational and inferential problems of polynomial regression can be completely addressed using the techniques of multiple regressions. This is done by treating $x, x^2, \dots$ as being distinct independent variables in a multiple regression model.

Q15. What are the assumptions of Regression Analysis.

Ans :

The following assumptions are made while making use of the regression technique:

1. There exists an actual relationship between the dependent and independent variables.
2. The regression analysis is used to estimate the values within the range for which it is valid and not for the values outside its range.
3. The relationship that existed between the dependent and independent variables remains the same till the regression equation is calculated.
4. The dependent variable takes any random value but the values of the independent variables are fixed quantities without error and are chosen by the analyst or the user.
5. In regression, we have only one dependent variable in our estimating equation. However, we can use more than one independent variable.

Q16. What are the limitations of Regression Analysis.

Ans :

1. It assumes a linear relationship between two variables which need not be the case always.
2. It assumes a static relationship between the two variables over a period of time. However, relationships between variables can change with a change in other factors. For example, the change in demand for a given change in price can be estimated using regression. However, the impact of price on demand will be different when a family or a nation is poor and when such a family or nation has abundance of wealth or resources.

3. Regression analysis provides meaningful insights only up to a certain limit. For example, increasing production results in a decrease in marginal cost. However, beyond a certain point, increase in production can result in the costs going up.

Q17. Explain different types of Regression.

Ans :

The various types of Regression are as follows:

1. Simple Regression

In statistics, simple regression is the least squares estimator of a linear regression model with a single predictor variable. In other words, simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.

2. Multiple Regression

Multiple regression analysis represents a logical extension of two-variable regression analysis. Instead of a single independent variable, two or more independent variables are used to estimate the values of a dependent variable. However, the fundamental concepts in the analysis remain the same.

For example, a college admissions officer wishing to predict the future grades of college applicants might use three variables (High School GPA, SAT and Quality of letters of recommendation) to predict college GPA. The applicants with the highest predicted college GPA would be admitted. The prediction method would be developed based on students already attending college and then used on subsequent classes. Predicted scores from multiple regression are linear combinations of the predictor variables. Therefore, the general form of a prediction equation from multiple regression is:

$$Y' = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + A$$

where Y' is the predicted score, X_1 is the score on the first predictor variable, X_2 is the score on the second, etc. The Y intercept is A . The regression coefficients (b_1, b_2 , etc.) are analogous to the slope in simple regression.

3. Curvilinear Regression

The analysis of the linear regression model can be extended in a straightforward way to cover situations in which the dependent variable is affected by several controlled variables or in which it is affected non-linearly by one controlled variable.

For example, suppose that there are three controlled variables, x_1, x_2 and x_3 . A linear regression equation is of the form,

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

4. Polynomial Regression

Suppose that the dependent variable is a polynomial function of a single controlled variable. For example, in cubic regression, the regression equation is given by,

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

This type of regression can be approached in the same way as multiple regressions. In the case of cubic regression we can substitute $x_1 = x, x_2 = x^2$ and $x_3 = x^3$. The least squares estimates of a_0, a_1, a_2 and a_3 can then be obtained.

If the observations are taken in such a way that there are an equal number of observations on y at a series of equally spaced values of x , then it is computationally more efficient to use the method of orthogonal polynomials.

5.6.2 Least Square Fit of a Linear Regression

Q18. Discuss about least square fit of a linear regression.

Ans :

The principle consists of minimizing the sum of the squares of the residuals or the errors of estimates, i.e., the deviations between the given observed values of the variable and their

corresponding estimated values as given by the line of best fit. Lines of regression uses the principle of least squares to give the best fit line for estimating the value of one variable given the value of another variable.

Standard Error of Estimate

The estimates obtained using the regression equations may not be perfect. A measure of precision of these estimate is given by the standard error of the estimate. Standard deviation gives us measure of dispersion of the observations about the mean of the distribution whereas standard error of estimate gives us a measure of the observations about the line of regression.

5.6.3 Two Lines of Regression

Q19. Explain briefly about Regression Equations.

Ans :

i) Regression equation of X on Y

$$X - \bar{X} = r \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y})$$

$\bar{X}$ = Mean of X series, $\bar{Y}$ = Mean of Y series

$r \frac{\sigma_X}{\sigma_Y}$ is called regression co-efficient of X on Y and it is denoted by the symbol b_{xy} .

If the deviations are taken from actual meant, we need not calculate coefficient of correlation (r), σ_X and σ_Y . In such cases.

$$r \frac{\sigma_X}{\sigma_Y} \text{ (or } b_{xy}) = \Sigma xy / \Sigma y^2 \text{ where } x = X - \bar{X},$$

$$y = Y - \bar{Y}$$

If the deviations are taken from assumed mean.

$$\frac{\sigma_X}{\sigma_Y} = \frac{\Sigma dx dy - \frac{\Sigma dx \cdot \Sigma dy}{N}}{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}$$

where $dx = X - A$ and $dy = Y - A$

ii) Regression Equation of Y on X

$$Y - \bar{Y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$$

$r \frac{\sigma_Y}{\sigma_X}$ is called regression coefficient of Y on X and it is denoted by the symbol b_{yx}

If deviations are taken from actual mean.

$$r \frac{\sigma_Y}{\sigma_X} \text{ (or } b_{yx} \text{)} = \frac{\sum xy}{\sum x^2}$$

If the deviations are taken from assumed mean.

$$r = \frac{\sigma_Y}{\sigma_X} = \frac{\sum dx dy - \frac{\sum dx \cdot \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}}$$

The steps for calculation of regression coefficients are same as in case of correlation.

With the help of two regression coefficients, it is possible to calculate correlation coefficient (r) symbolically.

$$r = \sqrt{b_{xy} \times b_{yx}}$$

The following Points should be noted, about regression coefficients.

1. Both the regression coefficients must be less than one.
2. Both the regression coefficients will have the same sign, i.e. either they will be positive or negative
3. The coefficient of correlation will have the same sign as that of regression coefficients.

Note : Regression Equations can be obtained directly i.e. without taking deviations from actual mean or assumed mean.

In such a case two normal equations are to be solved as follows :

Regression Equation of X on Y i.e.

$$X = a + bY$$

$$\sum x = Na + b \sum y$$

$$\sum xy = a \sum y + b \sum y^2$$

Regression Equation of Y on X i.e. $Y = a + bX$

$$\sum y = Na + b \sum x, \sum xy = a \sum x + b \sum x^2$$

5.6.4 Properties of Regression Coefficients**Q20. Explain the Properties of Regression Coefficients.**

Ans :

Properties of Regression Coefficient(s)

1. $r^2 = b_{xy} * b_{yx}$. In other words, r is the Geometric mean between the two regression coefficients b_{xy} and b_{yx} .
2. Both the regression coefficients will have the same sign, i.e. either they will be positive or negative. Also, the coefficient of correlation will have the same sign as that of regression coefficients.
3. The arithmetic mean of the two regression coefficients is greater than the correlation coefficient. In other words, $(b_{xy} + b_{yx}) / 2 > r$.
4. If one regression coefficient is greater than 1, the other has to be less than one. This is an extension of the first property as the product of the two coefficients is equal to square of the correlation coefficient r . Since r lies between -1 and $+1$, r^2 cannot be greater than 1. Thus $b_{xy} * b_{yx}$ cannot be greater than 1. Thus, if one regression coefficient is greater than 1, the other has to be less than one.

Q21. What are the differences between correlation and regression analysis ?

Ans :

(Imp.)

S.No.	Correlation Analysis	Regression Analysis
1.	Correlation is a measure of the 'degree and direction' of relationship between the variables.	Regression studies 'nature' of relationship between the variables.
2.	Correlation means the relationship between two or more variables which vary in sympathy so that the movements in one variable tend to be accompanied by the corresponding movements in others.	Regression means stepping back or returning to average value and is a mathematical measure expressing the average relationship between the variables.
3.	Correlation does not indicate the cause and effect relationship between the variables.	Regression clearly indicates the cause and effect relationship between the variables.
4.	Correlation cannot say which variable is the dependent variable and which is the independent variable.	In regression, the variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.
5.	Correlation coefficient is a relative measure of the linear relationship.	Regression coefficients are absolute measures indicating the change in the value of one variable for a unit change in the value of the other variable.
6.	Correlation analysis cannot be used for predicting or estimating value.	Regression analysis is very helpful in predicting and estimating value of one variable given the value of another variable.
7.	Correlation coefficients are symmetric i.e., $r_{yx} = r_{xy}$.	Regression coefficients are asymmetric i.e., $b_{xy} \neq b_{yx}$.
8.	The range of r is $+1$ to -1 .	The range of b_{xy} and b_{yx} is not restricted.
9.	Correlation coefficient can be calculated from regression coefficients.	Regression coefficients cannot be directly compared from correlation coefficient.
10.	There may be non-sense correlation between variables due to chance.	There is no such thing in regression.

PROBLEMS

11. From the following data obtain the two regression equations :

X :	6	2	10	4	8
Y :	9	11	5	8	7

Sol :

X	Y	XY	X ²	Y ²
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma XY = 214$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$

Table : Obtaining Regression Equations

Regression equation of Y on X, $Y = a + bX$

To determine the values of a and b the following two normal equations are to be solved.

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values

$$40 = 5a + 30b \quad \dots (1)$$

$$214 = 30a + 220b \quad \dots (2)$$

$$\text{Multiplying equation (1) by 6, } 240 = 30a + 180b \quad \dots (3)$$

$$214 = 30a + 220b \quad \dots (4)$$

Deducing equation (4) from (3) $-40b = 26$ or $b = -0.65$

Substituting the value of b in equation (1)

$$40 = 5a + 30(-0.65)$$

$$\text{or } 5a = 40 + 19.5 = 59.5 \text{ or } a = 11.9$$

Putting the values of a and b in the equation

Regression equation of Y on X

$$Y = 11.9 - 0.65X$$

Regression equation of X on Y $X = a + bY$ and the two normal equations are:

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values

$$30 = 5a + 40b \quad \dots (5)$$

$$214 = 40a + 340b \quad \dots (6)$$

Multiplying equation (5) by 8:

$$240 = 40a + 320b \quad \dots (7)$$

$$214 = 40a + 340b \quad \dots (8)$$

From equation (7) and (8)

$$-20b = 26 \text{ or } b = -1.3$$

Substituting the value of b in equation (5);

$$30 = 5a + 40(-1.3)$$

$$5a = 30 + 52 = 82$$

$$\therefore a = 16.4$$

Putting the value of a and b in the equation.

The regression equation of X of Y :

$$X = 16.4 - 1.3 Y.$$

- 12. From the following data, calculate the regression equations taking deviation of items from the mean of X and Y series.**

X	6	2	10	4	8
Y	9	11	5	8	7

Sol :

Regression Equation of X on Y

$$X - \bar{X} = b_{xy} (Y - \bar{Y})$$

$$\bar{X} = \frac{\sum X}{N} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{40}{5} = 8$$

$$\therefore \bar{X} = 6, \bar{Y} = 8$$

X	Y	$X - \bar{X} (x)$	$Y - \bar{Y} (y)$	$(X - \bar{X})^2 = x^2$	$(Y - \bar{Y})^2 = y^2$	xy
6	9	0	1	0	1	0
2	11	-4	3	16	9	-12
10	5	4	-3	16	9	-12
4	8	-2	0	4	0	0
8	7	2	-1	4	1	-2
$\sum X = 30$	$\sum Y = 40$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 40$	$\sum y^2 = 20$	$\sum xy = 26$

Number of Pairs $N = 5$

Regression Coefficient

$$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{-26}{20} = -1.3$$

$$\therefore X - \bar{X} = b_{xy} (Y - \bar{Y})$$

$$X - 6 = 1.3 (Y - 8)$$

$$X - 6 = 1.3 Y + 10.4$$

$$X = -1.3Y + 10.4 + 6$$

$$X = 1.3Y + 16.4$$

(or)

$$X = 16.4 - 1.3 Y$$

Regression Equation of Y on X

$$(Y - \bar{Y}) = b_{yx} (X - \bar{X})$$

Regression Coefficient

$$b_{yx} = \frac{\sum x^2}{\sum y^2} = \frac{-26}{40} = -0.65$$

$$\therefore Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$Y - 8 = -0.65 (X - 6)$$

$$Y - 8 = -0.65x + 3.9$$

$$Y = -0.65x + 3.9 + 8$$

$$Y = -0.65x + 11.9$$

or

$$Y = 11.9 - 0.65x$$

5.7 TIME SERIES ANALYSIS

Q22. Define Time Series.

Ans :

Introduction

An arrangement of statistical data in accordance with time of occurrence or in a chronological order is called a time series. The numerical data which we get at different points of time-the set of observations is known as time series.

Examples:

- The Annual Production of Steel in India over the last 10 years;
- The Monthly Sales of a Chemical Industry for the last 6 months;
- The daily closing price of a share in the Calculate Stock Exchange;
- Hourly temperature recorded by the Meteorological office in a city;
- Yearly Price or Quantity Index Numbers.

Mathematically, a Time Series is defined by the values $Y_1, Y_2, \dots, Y_n$ of a variable Y at times $t_1, t_2, \dots, t_n$

Here Y is a function of time t and Y_t denotes the value of the variable Y at time t . The basic assumption is that changes witnessed over time in a sample group will be extrapolated to population.

Definitions

- According to Morris Hamburg**, "A time series is a set of observations arranged in chronological order".
- According to Kenny and Keeping**, "A set of data depending on the time is called time series".
- According to Croxton and Cowden**, "A time series consists of data arranged chronologically".
- According to Patterson**, "Time series consists of statistical data which are collected, recorded or observed over successive increments".

Q23. Explain the Utility of Time Series Analysis

Ans : (Imp.)

- It Helps in the Analysis of past Behaviour of a Variable**

Analysis of past data discloses the effect of various factors on the variable under study. These studies isolate and analyze the effects of various sets of homogenous factors on the problem under study.

ii) It Helps in Forecasting

The analysis of past condition is the basis of forecasting the future behaviour of the variable under study. For example, the analysis of a time series relating to income or wages or production would be the basis for forecasting future income or wages or production. This helps in making future plan of action.

iii) It helps in Evaluation of Current Achievement

The review and evaluation of progress made on the basis of a plan are done on the basis of time series data. For example the progress of our Five-year plans is judged by the annual growth rates in the gross national product.

iv) It Helps in Making Comparative studies

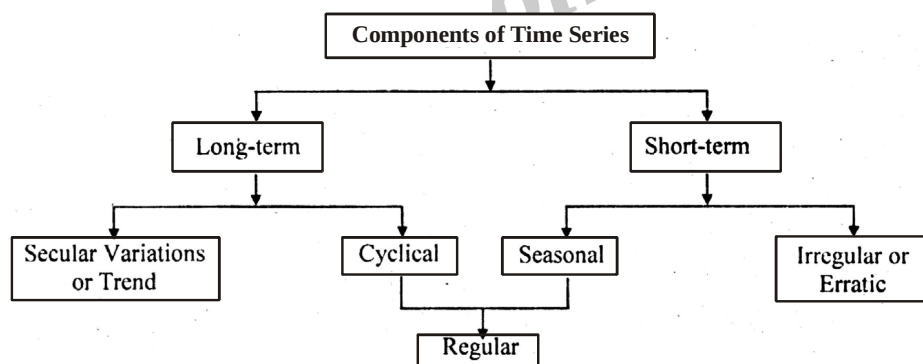
Once the data are arranged chronologically comparison between one time period and another is facilitated. It provides a scientific basis for making comparisons by studying and isolating the effects of various components of a time of series. It also helps in making regional comparison amongst data collected on the basis of time.

5.7.1 Components of Time Series**Q24. What are the Components of Time Series ?**

Ans :

(May-18, May-17)

A time series is a historical series of statistical data. Since these statistical data are historical, they are subject to the influences of all the changes that may occur over any period of time. The usual classification of the influences affecting statistical data recording over time is one based upon the nature of the influence. Classified in this manner, these influences are as follows :

**1) Secular Variations (or) Trend (T)**

The general tendency of the time series data to increase or decrease or stagnate during a long-period of time is called the secular trend or simply trend.

According to Simpson and Kafka

“Trend, also called secular or long-term trend, is the basic tendency of a series to grow or decline over a period of time. The concept of trend does not include short-range oscillations, but rather the steady movements over a long time”.

The trend thus refers to the general direction and the movement of time series considering a fairly long period of time. Most business and economic series are influenced by some secular forces of change in which the underlying tendency is one of growth or decline, for example, in series concerning population, national income, agricultural production, currency in circulation, bank deposits, etc.,

an upward tendency is usually observed while a downward tendency is noticed in time series relating 10 birth and death rates specially in deaths by epidemics, tuberculosis, etc.

2) Seasonal Variations (S)

Seasonal variations refers to rhythmic forces of change inherent in most time series showing a regular or a periodic pattern of movement over a span of less than a year and has the same or almost the same pattern year after year.

According to Hirsch “a recurrent pattern of charge within the period that results from the operation of forces connected with climate or custom at different times of the period.”

Reasons for Seasonal Variations

i) Climate and Natural Forces

The result of natural forces like climate is causing seasonal variation. Umbrellas are sold more in rainy season. In winter season, sale of the woolen clothes will increase. In hot season, the sales of ice, ice-cream, fruit salad, etc., will increase. Thus climate and weather play an important role in seasonal movement. Agricultural production depends upon the monsoon.

ii) Customs and Habits

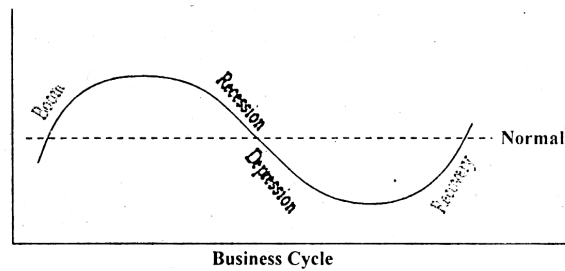
Man-made conventions are the customs, habits, fashion, etc. There is the custom of wearing new clothes, preparing sweets and buying crackers for Deepavali, Onam, Christmas, etc. At that time, there is more demand for cloth, sweets and crackers. It will happen every year. In marriage season, the price of gold will increase.

Purposes of Studying Seasonal Variations

- i) To analyze seasonal pattern in a short-period time series.
- ii) Once the seasonal factor is known it can be used for separating cyclical and irregular forces from the residual forces of changes.
- iii) The seasonal factor can be used for adjustment in the value projected on the basis of trend. Short-term forecasts are always affected by seasonal factors.
- iv) The observed value can be appraised in terms of seasonal factor and adjusted so as to get correct idea of the trend, if any, in an economic phenomenon. The study of seasonal variations is, therefore, necessary for proper appraisal of business facts influenced by a seasonal variation.
- v) The seasonal factor can also be used for pricing of articles and services so as to level up the seasonal variations in demand.
- vi) A study of the seasonal patterns is extremely useful in planning future operations and in formulation of policy decisions regarding purchase, production, inventory control, personnel requirement, selling and advertising programme, etc.

3) Cyclical Variations (C)

As the economy expands during a period of boom, we would expect to find that such data as sales, output or consumer expenditure also show a rising trend; and during a period of slump, we would expect them to show a downward trend. Thus a wavelike motion may be observed in the pattern of our data. They are prosperity (boom) recession, depression and recovery; and are shown below:



The study of cyclical variation is helpful to businessmen and economist for framing suitable policies and stabilizing the level of business activities.

Purposes of Studying Cyclical Variation

- i) One can easily study the character of business fluctuations. Good policies can be formulated at stabilizing the level of business activity.
- ii) Businessman can take timely steps in maintaining business during booms and depressions.
- iii) A careful study of cyclical variations facilitates a businessman to face the recession period and make them ready to reap the benefits during booms.

4) Irregular Variations (I)

Irregular variations are the effect of random factors. These are generally mixed up with seasonal and cyclical variations and are caused by irregular and accidental factors like floods, famines, wars, strikes, lockouts, etc. There is no regular period or time of their occurrence and that is why they are called random or chance fluctuations. Sometimes these factors are very effective and may even give rise to cyclical fluctuations. However since these are not regular factors, no advance preparation can be done to meet their consequences. Their effects are also unpredictable and irregular.

5.8 MODELS OF TIME SERIES

5.8.1 Additive, Multiplicative and Mixed Models

Q25. Explain the various models are used in Time Series Analysis.

Ans :

There are two mathematical models which are commonly used for the decomposition of a time series into its components viz.

- (a) Decomposition by Additive model
- (b) Decomposition by Multiplicative model.
- (c) Decomposition by Mixed Model

(a) Decomposition by Additive Model

According to the additive mode, a time series can be expressed as

$$Y_t = T_t + S_t + C_t + R_t$$

where Y_t = is the time series value at time t,

T_t = represent the trend value.

S_t = represents the seasonal variations.

C_t = represents the cyclic movements and

R_t = represents the random fluctuations at time t .

Obviously, the term S_t will not appear in a series of annual data. The additive model implicitly implies that seasonal forces (in different years), cyclical forces (in different cycles) and irregular forces (in different long time periods) operate with equal absolute effect irrespective of the long trend value. As such C_t (and S_t) will have positive or negative values according as whether we are in an above normal or below normal phase of the cycle (and year) and the total of positive and negative values for any cycle (and any year) will be zero. R_t will also have positive or negative values and in the long run R_t will be zero. Occasionally, there may be a few isolated occurrences of extreme R_t of episodic nature.

(b) Decomposition by Multiplicative Model

On the other hand if we have reasons to assume that the various components in a time series operate proportionately to the general level of the series, the traditional or classical multiplicative model is appropriate. According to the multiplicative model

$$Y_t = T_t \times S_t \times C_t \times R_t$$

where S_t , C_t and R_t instead of assuming positive and negative values are indices fluctuating above or below unity and the geometric means of S_t in a year, C_t in a cycle and R_t in a long-term period are unity. In a time series with both positive and negative values the multiplicative model can not be applied unless the time series is translated by adding a suitable positive value. It may be pointed out that the multiplicative decomposition of a time series is same as the additive decomposition of the logarithmic values of the original time series.

In practice, most of the series relating to economic data confirm to the multiplicative model.

(c) Decomposition by Mixed Models

In addition to the additive and multiplicative models discussed above, the components in a time series may be combined in large number of ways. The different models, defined under different assumptions, will yield different results. Some of the defined models can be:

$$Y_t = (T_t \times S_t \times C_t) + R_t$$

$$Y_t = (T_t \times C_t) + (S_t \times R_t)$$

$$Y_t = T_t + (S_t \times C_t \times R_t)$$

$$Y_t = T_t + S_t + (C_t \times R_t)$$

$$Y_t = T_t \times (S_t + C_t) \times R_t$$

5.9 TREND ANALYSIS

Q26. Define Trend Analysis. Explain the purpose of measuring trend.

Ans :

The term trend analysis refers to the concept of collecting information and attempting to spot a pattern, or trend, in the information. In some fields of study, the term "trend analysis" has more formally-defined meanings.

Although trend analysis is often used to predict future events, it could be used to estimate uncertain events in the past, such as how many ancient kings probably ruled between two dates, based on data such as the average years which other known kings reigned.

Trend analysis uses a technique called least squares to fit a trend line to a set of time series data and then project the line into the future for a forecast.

Purposes of Measuring Trend

There are three basic purposes of measuring secular trend :

- 1) The first purpose is to study the past growth or decline of a series. The secular trend describes the basic growth tendency ignoring short-term fluctuations.

- 2) The second and most important purpose of measuring secular trend is to project the curve into the future as a long-term forecast. If the past growth has been steady and if the conditions that determine this growth may reasonably be expected to persist in the future, a trend curve may be projected over five to ten years into the future as a preliminary forecast.
- 3) The third purpose of measuring secular trend is to eliminate it, in order to clarify the cycles and other short-term movements in the data. A steep trend may observe minor cycles. Dividing the data, by the trend values yield ratios which make the curve fluctuate around a horizontal line, thus bringing the cycles into clear relief.

Q27. Explain the uses of trend analysis.

Ans :

Uses of Trend

- 1) The trend describes the basic growth tendency ignoring short term fluctuations.
- 2) It describes the pattern of behavior which has characterized the series in the past.
- 3) Future behavior can be forecasted in the assumption that past behavior will continue in the future also.
- 4) Trend analysis facilitate us to compare two or more time series over different period of time and this helps to draw conclusions about them,.

Q28. Explain the various methods for measuring trend.

Ans :

Methods / Measurement of Trend

There are Your methods for determining trend in time series :

1. Freehand (or Graphical) Method,
2. Semi-Average Method,
3. Moving Average Method.
4. Least Squares Method;

5.9.1 Free Hand Curve**Q29. What is Free Hand Curve? Explain with illustration.**

Ans :

A trend is determined by just inspecting the plotted points on a graph sheet. Observe the up and down movements of the-points. Smooth out the irregularities by drawing a freehand curve or line through the scatter points. The curve so drawn would give a general notion of the direction of the change. Such a freehand smoothed curve eliminates the short- time swings and shows the long period general tendency of the changes in the data.

Drawing a smooth freehand curve requires a personal skill and judgement. The drawn curve should pass through the plotted points in such a manner that the variations in one direction are approximately equal to the variation in other direction. Different persons, however, drawn different curves at different directions, with different slopes and in different styles. This may lead to different conclusions. To overcome these limitations, we can use the semi-average method of measuring the trend.

Example

Fit a trend line to the following data by the freehand method,

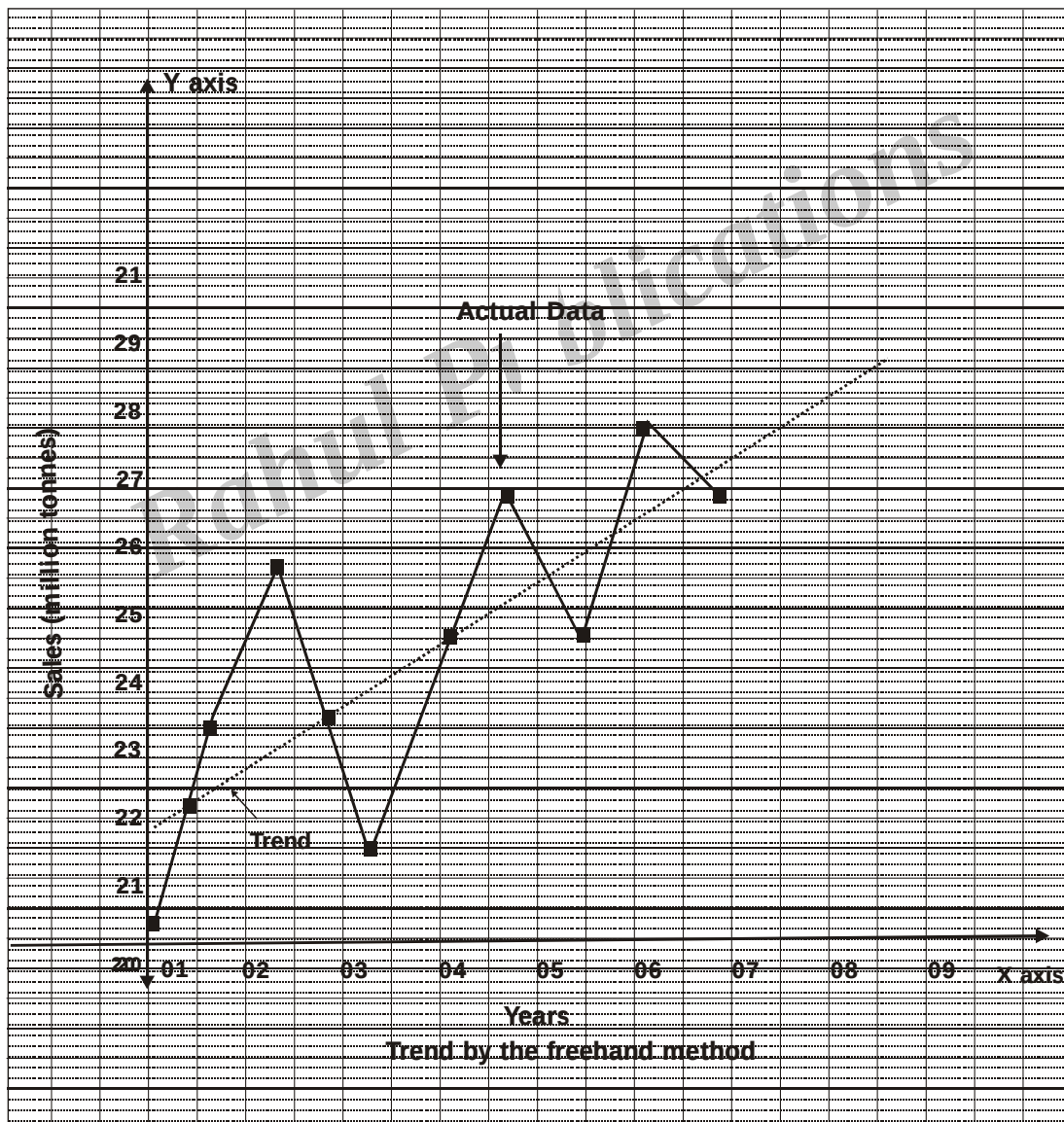
Year	2001	2002	2003	2004	2005	2006	2007	2008	2009
Sales (million tonnes)	19	22	24	20	23	25	23	26	25

Sol :

Steps

1. Time series data is plotted on the graph
2. The direction of the trend is examined on the basis of the plotted data (dots)
3. A straight line is drawn which shows the direction of the trend.

The actual data and the trend line are shown in the following graph.



5.9.2 Semi Averages Method

Q30. Explain the procedure of “semi-average method”. What are its merits and demerits? Explain with an illustration.

Ans :

Procedure of Semi-Average Method

The following procedure is followed for semi-average method,

- (i) The entire time series is classified into two equal parts with respect to time. For even period, equal split. For odd period, equal parts obtained by omitting middle period.
- (ii) Compute the arithmetic mean of time series values for each half separately. These means are called semi-averages.
- (iii) Semi averages are plotted as points against the middle point of the respective time period covered by each part.
- (iv) The line joining these points gives the straight line trend fitting the given data.

Merits of Semi-Average Method

The following are some of the merits of semi-average method,

1. Objectivity
2. Ease of apply and understandability
3. Extend both ways the line i.e., we can get past and future estimates.

Demerits of Semi-Average Method

Some of the demerits of semi-average method are,

1. Linear trend assumption may not exist.
2. A men may be questioned.
3. Thus, values of trend are not precise and reliable.

Example

Using the following data, fit a trend line by using the method of semi-averages,

Year	1996	1997	1998	1999	2000	2001	2002
Output	700	900	1100	900	1300	1000	1600

Sol :

Step 1

The data provided in the problem is of seven years i.e., (an odd number). Thus, the middle year [1999] shall be ignored and the remaining years are divided into two equal time periods and their arithmetic averages is computed as follows,

$$\text{Average of the first three years} = \frac{700 + 900 + 1100}{3} = \frac{2700}{3} = 900$$

$$\text{Average of the last three years} = \frac{1300 + 1000 + 1600}{3} = \frac{3900}{3} = 1300$$

Therefore, the semi-averages are 900 and 1300

Step 2

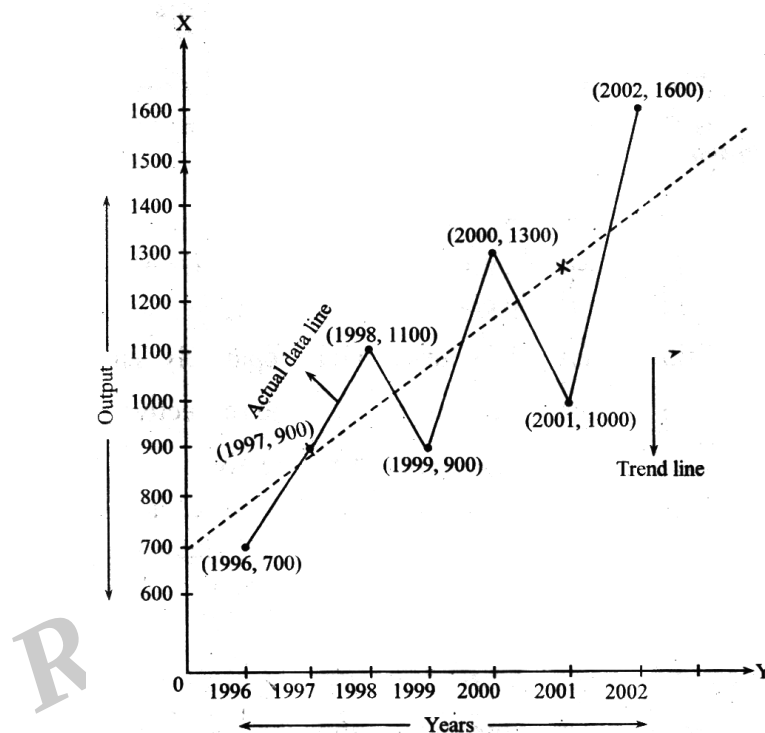
The next step is to plot the semi-averages against the mid-point (middle year) of each time period. Thus, it would be year 1997 and 2001 respectively.

Step 3

The plotted points are joined in order to derive the trend line using the semi average method.

Step 4

The original data and the trend line is plotted on a graph as follows,



5.9.3 Moving Averages Method

Q31. Discuss the method of moving averages in measuring trend. What are its merits and limitations of moving average method?

Ans :

(Nov.-20)

Moving Average Method

In moving average method, the average value for a number of years (month or weeks) is secured and this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the calculation of the average.

The effect of averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend.

The period of moving average is decided in the light of the length of the cycle. More applicable to data with cyclical movements.

Formula for 3 yearly moving average will be,

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3} \dots$$

Formula for 5 yearly moving average will be,

$$\frac{a+b+c+d+e}{5}, \frac{b+c+d+e+f}{5}, \frac{c+d+e+f+g}{5} \dots$$

Methods of Moving Average

The following two methods are followed in moving averages,

a) Odd Yearly Method

- i) Calculate 3/5...yearly totals
- ii) Now compute 3/5 yearly average by dividing the totals calculated in step (i) by the respective number of years, i.e. 3/5/...
- iii) Short term oscillations are calculated using the formula, $Y - Y_c$

Where, Y - Actual value and Y_c - Estimated value.

b) Even Yearly Method

Example : 4 years

- i) Calculate 4 yearly moving totals and place at the centre of middle two years of the four years considered.
- ii) Divide 4 yearly moving totals by 4 to get 4 yearly average.
- iii) Take a 2 period moving average of the moving average which gives the 4 yearly moving average centered.

Merits of Moving Average

The merits of moving average are as follows,

1. Of all the mathematical methods of fitting a trend, this method is the simplest.
2. The method is flexible so that even if a few more observations are to be added, the entire calculations are not changed.
3. If the period of the moving average happens to coincide with the period of the cycle, the cyclical fluctuations are automatically eliminated.
4. The shape of the curve in case of moving average method is determined by the data rather than the statisticians choice of mathematical function.

Limitations of Moving Average

The following are the limitations of moving averages,

1. Trend values cannot be computed for all the years. For example, in a 5 yearly moving we cannot compute trend values for the first two and the last two years.
2. It is difficult to decide the period of moving average since there is no hard and fast rule for the purpose.

3. Moving average cannot be used in forecasting as it is not represented by any mathematical function.
4. When the trend is not linear, the moving average lies either above or below the true sweep of the data.

PROBLEMS ON MOVING AVERAGE METHOD

13. Calculate the 3-yearly moving averages of the production figures given below and draw the trend:

Year	Production (in met. tonnes)	Year	Production (in met. tonnes)
1985	15	1993	63
1986	21	1994	70
1987	30	1995	74
1988	36	1996	82
1989	42	1997	90
1990	46	1998	95
1991	50	1999	102
1992	56	—	—

Sol.

Calculation Of 3-yearly Moving Averages

Year	Production Y	3-Yearly totals	3-Yearly Moving average
1985	15	—	—
1986	21	66	22.00
1987	30	87	29.00
1988	36	108	36.00
1989	42	124	41.33
1990	46	138	46.00
1991	50	152	50.67
1992	56	169	56.33
1993	63	189	63.00
1994	70	207	69.00
1995	74	226	75.33
1996	82	246	82.00
1997	90	267	89.00
1998	95	287	95.67
1999	102	—	—

14. Construct 5-yearly moving averages of the number of students studying in a college shown below :

Year	No.of students	Year	No. of students
1990	332	1995	405
1991	317	1996	410
1992	357	1997	427
1993	392	1998	405
1994	402	1999	438

Sol :

Calculation Of 5-yearly Moving Averages

Year	No.of Students	5-Yearly Total	5-Yearly Moving Average
1990	332	—	—
1991	317	—	—
1992	357	1800	360.0
1993	392	1873	374.6
1994	402	1966	393.2
1995	405	2036	407.2
1996	410	2049	409.8
1997	427	2085	417.0
1998	405	—	—
1999	438	—	—

15. Assume a four-yearly cycle and calculate the trend by the method of moving averages from the following data relating to the production of tea in India :

Year	Production (mn.lbs.)	Year	Production (mn.lbs.)
1987	464	1992	540
1988	515	1993	557
1989	518	1994	571
1990	467	1995	586
1991	502	1996	612

Sol :

Calculation Of Trend By The Moving Average Method

Year	Production (mn.lbs)	4-yearly moving totals	4-yearly moving average	4-yearly moving average centered
1987	464	—	—	—
1988	515	1964	491.00	—
		←		←
1989	518	2002	500.50	495.75
		←		←
1990	467	2027	506.75	503.62
		←		←
1991	502	2066	516.50	511.62
		←		←
1992	540	2170	542.50	529.50
		←		←
1993	557	2254	563.50	553.00
		←		←
1994	571	2326	581.50	572.50
		←		←
1995	586	← —	—	—
1996	612	— —	—	—

16. Use of four-yearly moving average method to calculate trend for the following data and comment.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
production	133	248	267	299	321	500	350	450	399	550

Sol :

In four yearly moving averages, the moving average values are placed in the middle of 2 time periods. An alternative is made to synchronize them with the original data by taking 2 period average of the between corresponding time periods. This technique is called centering. The corresponding moving average values are called centered moving averages.

Year (1)	(4) Production (2)	(4) Yearly Moving Totals (3)	Yearly Moving Averages (4) = (3) ÷ 4	2 Point Moving Total (5)	(4) 4.m Average Centered (6) = (5) ÷ 2
2005	133	–	–		
2006	248	947	236.75	521	261
2007	267	1135	283.75	631	316
2008	299	1387	346.75	715	358
2009	321	1470	367.75	773	387
2010	500	1621	405.25	830	415
2011	350	1699	424.75	862	431
2012	450	1749	437.25	–	–
2013	399	–			
2014	550	–			

17. From the following data, calculate trend values using four yearly moving average.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007
Production (in tons)	506	620	1036	673	588	696	1116	738	663

Calculation of trend by 4-yearly moving averages

Sol :

Calculation of trend by 4-yearly moving averages

Year (1)	Production (2)	4 yearly moving Totals (3)	4 yearly moving average (4) = (3) ÷ 4	2 points moving total (5)	4 yearly moving average centred (6) = (5) ÷ 4
1999	506	–			
2000	620	–			
2001	1036		2835	708.75	
		2917	729.25	1438	719
2002	673		–	1477.50	738.75
2003	588	2993	748.25		
		3073	768.25	1516.50	758.25
2004	696			1552.75	776.375
		3138	784.5		
2005	3213			1587.75	793.875
		3213	803.25		
2006	738	–			
2007	663	–			

5.9.4 Least Square Methods

Q32. Define least square method. Explain merits and demerits of least square method.

Ans :

(Imp.)

Least square method is the most widely used method and provides us with a mathematical device to obtain an objective fit to the trend of a given time series. This method is so called because a trend line computed by this method is such that the sum of the squares of the deviation between the original data and the corresponding computed trend values is minimum. This method can be used to fit either a straight line trend or a parabolic trend.

The straight line trend equation is in the form of $Y = a + bX$

Where, Y denotes the trend value of the dependent variable

X denotes the independent variable.

a and b are constants.

The values of a and b are obtained by solving the following normal equations.

$$\sum Y = Na + b \sum X$$

$$\sum XY = a \sum X + b \sum X^2$$

Where, N represents the number of years in the series.

When $\sum X = 0$ the above normal equations are simplified to

$$a = \frac{\sum Y}{N}$$

$$b = \frac{\sum XY}{\sum X^2}$$

By substituting a and b values in straight line trend equation $Y = a + bX$, we get the straight line equation which can be used for estimation of future values.

Merits of Least Squares Method

The following are the merits of least squares method,

1. The method of least squares is a mathematical method of measuring trend and is free from subjectiveness.
2. This method provides the line of best fit since it is this line from where the sum of positive and negative deviations is zero and the sum of square of deviations is the least.
3. This method enables us to compute the trend values for all the given time periods in the series.
4. The trend equation can be used to estimate the values of the variable for any given time period 't' in future and the forecasted values are quite reliable.
5. This method is the only technique which enables us to obtain the rate of growth per annum for yearly data in case of linear trend.

Demerits of Least Squares Method

Some of the demerits of least squares are as follows,

1. Fresh calculations become necessary even if a single new observation is added.
2. Calculations required in this method are quite tedious and time consuming as compared with other methods.
3. Future predictions based on this method completely ignore the cyclical, seasonal and erratic fluctuations.
4. This method cannot be used to fit growth curves, gomper t_2 curve, logistic curve etc. to which most of the business and economic time series conform.

PROBLEMS

- 18. Annual trend of Milk Consumption (Y) is $18.6 + 1.8 X$. Convert the equation into monthly basis.**

Sol :

Given that

$$Y = 18.6 + 1.8 X$$

The annual trend equation $Y_e = a + bX$

The formula for converting annual trend into monthly trend is as follows,

$$Y_e = \frac{a}{12} + \frac{b}{12 \times 12} X$$

Substituting the values in the above equation, we get,

$$Y_e = \frac{18.6}{12} + \frac{1.8}{144} X$$

$$Y_e = 1.55 + 0.0125 X$$

- 19. Given $Y = 150 + 12X$ (origin 2009) X - unit = 1 year, Y - unit = annual profits ('000 `). Convert into monthly trend equation.**

Sol :

The formula for converting annual trend into monthly trend is as follows,

Given that,

X unit = 1 year i.e., 12 months

$$Y_e = \frac{a}{12} + \frac{b}{12 \times 12} X$$

Substituting the values in the above equation, we get,

$$Y_e = \frac{150}{12} + \frac{12}{144} X$$

$$Y_e = 12.5 + 0.08333X$$

Where, X unit = 1 month and Y unit = monthly value of the variable.

20. Given below is the data of production of a certain company in lakhs of units.

Year	1995	1996	1997	1998	1999	2000	2001
Production	15	14	18	20	17	24	27

Compute the Linear trend by the method of least squares. predict the production for the year "2004".

Sol :

The Normal Equation are,

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

By solving these equations, we get a and b. This method is called the method of least squares.

Year	Production Y	Deviation from 1998 X	X ²	XY
1995	15	-3	9	-45
1996	14	-2	4	-28
1997	18	-1	1	-18
1998	20	0	0	0
1999	17	1	1	17
2000	24	2	4	48
2001	27	3	9	81
Total	135	0	28	55

Since $\Sigma X = 0$, the above equations simplify to,

$$a = \frac{\Sigma Y}{N} = \frac{135}{7} = 19.29$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{55}{28} = 1.96$$

∴ The straight line trend, $Y_e = 19.29 + 1.96X$.

Prediction of Production for the Year 2004

For year 2004,

$X = 6$ (deviation from mid year 1998).

$$\therefore Y_c = 19.29 + 1.96(6) = 19.29 + 11.76 = 31.05$$

The production of the company for the year 2004 is predicated to be 31.05 lakhs of units.

21. Calculate 3 yearly moving averages from the following data.

Year	2007	2008	2009	2010	2011	2012	2013
Sales (in million `)	41	44	45	46	47	48	49

Sol :

Calculation of 3-Yearly Moving Averages

Year	Sales (in Million `)	3-Yearly Total (in Million `)	3-Yearly Moving Average
2007	41	—	—
2008	44	41 + 44 + 45 = 130	$\frac{130}{3} = 43.33$
2009	45	44 + 45 + 46 = 135	$\frac{135}{3} = 45$
2010	46	45 + 46 + 47 = 138	$\frac{138}{3} = 46$
2011	47	46 + 47 + 48 = 141	$\frac{141}{3} = 47$
2012	48	47 + 48 + 49 = 144	$\frac{144}{3} = 48$
2013	49	—	—

22. The sales of a company in lakhs of rupees for the years 1994-2001 are given below:

Year	1994	1995	1996	1997	1998	1999	2000	2001
Production	550	560	555	585	540	525	545	585

Compute the Linear trend by the method of least squares. predict the production for the year "2002".

Sol :

The model of least squares helps in solving this problem.

The straight line trend is $y_c = a + bX$

By solving these normal equations, we get a and b.

Normal equation are,

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b \Sigma X^2$$

Since there are even number of observations, deviations are taken from both of the middle periods i.e., 1997 and 1998.

Year	Sales Y	Deviations From mid years X	X ²	XY
1994	550	-7	49	-3850
1995	560	-5	25	-2800
1996	555	-3	9	-1665
1997	585	-1	1	-585
1998	540	1	1	540
1999	525	3	9	1575
2000	545	5	25	2725
2001	585	7	49	4095
N = 8	ΣY = 4445	ΣX = 0	ΣX² = 168	ΣXY = 35

Since $\Sigma Y = 0$, the normal equation simplify to,

$$a = \frac{\Sigma Y}{N} = \frac{4445}{8} = 555.63$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{35}{168} = 0.21$$

∴ The straight line trend is $Y_e = 555.63 + 0.21X$

Estimation of Sales for the Year 2002

For year 2002, deviation $X = 9$,

$$\begin{aligned} \therefore Y &= 555.63 + 0.21(9) \\ &= 555.63 + 1.89 \\ &= 557.52 \end{aligned}$$

For year 2002, the sales of the company is estimated to be ` 557.52 lakhs.

23. Below table shows the figures of production (in thousand tons) of a sugar factory.

Year	1979	1980	1981	1982	1983	1984	1985
Production	77	88	94	85	91	98	90

Fit a straight line trend by the method of least squares and show trend values.

Sol :

Computation of Straight Line Trend

Year (Y)	Production	X	XY	X ²	Trend $Y_c = a + bx$
1979	77	-3	-231	9	83
1980	88	-2	-176	4	85
1981	94	-1	-94	1	87
1982	85	0	0	0	89
1983	91	+1	+91	1	91
1984	98	+ 2	+196	4	93
1985	90	+ 3	+ 270	9	95
	623	0	56	28	

Since

$$\Sigma X = 0; a = \frac{\Sigma Y}{N} = \frac{623}{7} = 89$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{56}{28} = 2$$

Σ The equation of Trend line is $Y_c = 89 + 2x$

On the basis of the above equation, we can calculate the value of Y_c for all the seven years. For example, Y_c for 1979 = $89 + 2(-3) = 89 - 6 = 83$.

$$1980 = 89 + 2(-2) = 89 - 4 = 85$$

$$1981 = 89 + 2(-1) = 89 - 2 = 87$$

$$1982 = 89 + 2(0) = 89 + 0 = 89$$

$$1983 = 89 + 2(1) = 89 + 2 = 91$$

$$1984 = 89 + 2(2) = 89 + 4 = 93$$

$$1985 = 89 + 2(3) = 89 + 6 = 95$$

24. Fit a straight line trend for the following series. Estimate the value for 1997.

Year	1990	1991	1992	1993	1994	1995	1995
Production of steel (in tonnes)	60	72	75	60	80	85	95

Sol. :

Fitting Straight Line Trend

Year	Production of steel (m. tonnes) Y	Deviations from 1993 X	XY	X ²
1990	60	-3	-180	9
1991	72	-2	-144	4
1992	75	-1	-75	1
1993	65	0	0	0
1994	80	+1	+80	1
1995	85	+2	+170	4
1996	95	+3	+285	9
N = 7	ΣY = 532	ΣX = 0	ΣXY = 136	ΣX² = 28

The equation of the straight line trend is

$$Y_c = a + bX$$

$$\Sigma X = 0, a = \frac{\Sigma Y}{\Sigma N} = \frac{532}{7} = 76$$

and

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{136}{28} = 4.857$$

Hence, $Y_c = 76 + 4.857X$

For 1997, X will be $Y_{1997} = 76 + 4.857(4)$
 $= 95.428$

Thus, the likely sales for the year 1997 is 95.428 m tonnes

25. Fit a straight line trend by the method of least squares to the following data. Assuming that the same rate of change continues, what would be the predicted earning for the year 1998 ?

Year	1989	1990	1991	1992	1993	1994	1995	1996
Earnings (Rs. lakhs)	38	40	65	72	69	60	87	95

Sol. :

**Fitting Of Straight Line Trend By
The Method Of Least Squares**

Year	Earnings (Rs.lakhs) Y	Deviations from 1992.5	Deviations multiplied by 2 X	XY	X ²
1989	38	-3.5	-7	-266	49
1990	40	-2.5	-5	-200	25
1991	65	-1.5	-3	-195	9
1992	72	-0.5	1	-72	1
1993	69	+0.5	+1	+69	1
1994	60	+1.5	+3	+180	9
1995	87	+2.5	+5	+435	25
1996	95	+3.5	+7	+665	49
N = 8	ΣY = 526		ΣX = 0	ΣXY = 616	ΣX² = 168

The equation of the straight line trend is

$$a = \frac{\sum Y}{N} = \frac{526}{8} = 65.75$$

$$b = \frac{\sum XY}{\sum X^2} = \frac{616}{168} = 3.667 \quad Y_c = 65.75 + 3.667(X)$$

For 1998, X will be +11.

When X is 11, Y will be

$$Y_c = 65.75 + 3.667(11) = 65.75 + 40.337 = 106.087.$$

Thus, the estimated earnings for the year 1998 are Rs.106.087 lakhs.

- 26. The following data relate to the number of passenger cars (in millions) sold from 2004 to 2011.**

Years	2004	2005	2006	2007	2008	2009	2010	2011
Number	6.7	5.3	4.3	6.1	5.6	7.9	5.8	6.1

- Fit a straight line trend to the data through 2009.
- Use your result in, (a) To estimate production in 2011 and compare with the actual production.

Sol :

Let, the equation of the straight line trend be $Y_c = a + b_x$, then,

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

a) Calculations to fit a straight line

Year	No. of Passenger cars (in millions) (Y)	Deviations from 2009 (X)	X ² (X)	XY
2004	6.7	-5	25	-33.5
2005	5.3	-4	16	-21.2
2006	4.3	-3	9	-12.9
2007	6.1	-2	4	-12.2
2008	5.6	-1	1	-5.6
2009	7.9	0	0	0
2010	5.8	1	1	5.8
2011	6.1	2	4	12.2
N = 8	$\Sigma Y = 47.8$	$\Sigma X = -12$	$\Sigma X^2 = 60$	$\Sigma XY = -67.4$

$$\Sigma Y = Na + b\Sigma X \quad \text{or} \quad 47.8 = 8a - 12b \quad \dots (1)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \text{or} \quad -67.4 = -12a + 60b \quad \dots (2)$$

Simplifying both the equations by multiplying equation (1) by 3 and equation (2) by 2, we get.

$$143.4 = 24a - 36b$$

$$-134.8 = -24a + 120b$$

$$8.6 = 84b$$

$$\text{i.e.,} \quad b = \frac{8.6}{84} = 0.102$$

Substituting the value of b in equation (1), we get,

$$47.8 = 8a - 12b$$

$$47.8 = 8a - 12(0.102)$$

$$47.8 = 8a - 1.224$$

$$8a = 49.024$$

$$a = \frac{49.024}{8} = 6.128$$

Hence, the required straight line equation is $Y_C = 6.128 + 0.102 X$

(b) Estimating the sale of passenger cars in 2011 by substituting the value of $X = 2$.

$$Y_C = 6.128 + 0.102 (2)$$

$$= 6.128 + 0.204 = 6.332$$

Thus, the estimated sale for 2011 is 6.332 million passenger cars whereas actual sales are 6.1 million passenger cars. There is some difference in actual figure and estimated figure because estimates depend on few assumptions.

Note : As number of passenger cars sold is given in question. So, estimated sales and actual sales are calculated instead of estimated production are calculated instead of estimated production and actual production.

Rahul Publications

Exercise Problems

1. Calculate Correlation for the data given below:

X	42	36	48	43	55	52	38
Y	132	120	140	143	142	148	122

(Ans: $r = 0.856$)

2. Calculate Co-efficient of Correlation and interpret the value:

Marks in QT	57	42	40	38	42	45	42	44	40	46	44	43
Marks in IOM	10	26	30	41	29	27	27	19	18	19	31	29

(Ans: $r = -0.7285$).

3. Calculate the coefficient of correlation and probable error from the following - data:

X	1	2	3	4	5	6	7	8	9	10
Y	20	16	14	10	10	9	8	7	6	5

(Ans: $r = -0.95$; P.E. = 0.0208)

4. Fit a straight line regression equation of Y on X from the following data.

X	10	12	13	16	17	20	25	29
Y	10	12	24	27	29	33	37	42

(Ans: $Y = 1.6 X - 1.65$)

5. Find the two regression equations from the following data:

X	1	2	3	4	5
Y	2	3	5	4	6

(Ans: $X = 0.9 Y - 0.6$; $Y = 0.9 X + 1.3$)

6. From the following data, obtain the two regression lines.

X	2	6	8	11	13	13	13	14
Y	8	6	10	12	12	14	14	20

(Ans: $X = 0.8125 Y + 0.25$; $Y = 0.8125 X + 3.875$)

7. Calculate the trend values by the method of least squares from the data given below and estimate the sales for the year 1985.

Year :	1976	1977	1978	1979	1980
Sales of T.V. ('000) :	9	12	14	16	20

(Ans: $Y_c = 20 + 5x$, 44.5)

8. Fit a straight line trend by the method of least of least squares from the following data and estimate 1986 sales.

Year :	1979	1980	1981	1982	1983	1984	1985
Sales :	100	105	96	102	108	109	112

(Ans: $Y = 104.57 + 2x$) Sales = 112.57

9. Fit a trend equation $Y = a + bx$:

Years :	1997	1998	1999	2000	2001	2002	2003	2004
Production (M.T) :	20	37	46	55	62	57	60	82

(Ans : $69.83 - 7.01x$)

Rahul Publications

Short Question and Answers

1. Define Correlation.

Ans :

Meaning

Correlation is the study of the linear relationship between two variables. When there is a relationship of 'quantitative measure between two set of variables, the appropriate statistical tool for measuring the relationship and expressing each in a precise way is known as correlation.

For example, there is a relationship between the heights and weights of persons, demand and prices of commodities etc.

Correlation analysis is the statistical tool we can use to describe the degree to which one variable is linearly related to another.

Definitions

- i) **According to Croxton and Cowden**, "The appropriate statistical tool for discovering and measuring the relationship of quantitative nature and expressing it in brief formula is known as correlation".
- ii) **According to Tippet**, "The effects of correlation are to reduce the range of uncertainty of our prediction".

The coefficient of correlation measures the degree of relationship between two set of figures. As the reliability of estimates depend upon the closeness of the relationship it is imperative that utmost care be taken while interpreting the value of coefficient of correlation, otherwise wrong conclusion can be drawn.

2. Positive correlation.

Ans :

If the values of two variables deviate in the same direction i.e., if increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, the corresponding correlation is said to be positive or direct.

Examples

- i) Sales revenue of a product and expenditure on Advertising.
- ii) Amount of rain fall and yield of a crop (up to a point)
- iii) Price of a commodity and quantity of supply of a commodity
- iv) Height of the Parent and the height of the Child.
- v) Number of patients admitted into a Hospital and Revenue of the Hospital.
- vi) Number of workers and output of a factory.

i) Perfect Positive Correlation

If the variables X and Y are perfectly positively related to each other then, we get a graph as shown in fig. below.

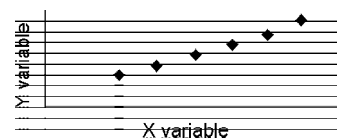


Fig.: Perfect Positive Correlation ($r = +1$)

ii) Very High Positive Correlation

If the variables X and Y are related to each other with a very high degree of positive relationship then we can notice a graph as in figure below.

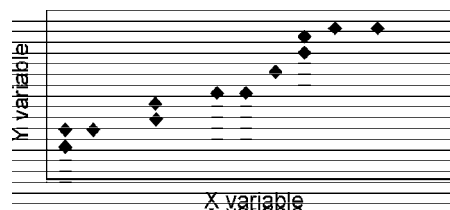


Fig.: Very High Positive Correlation ($r = \text{nearly } +1$)

iii) Very Low Positive Correlation

If the variables X and Y are related to each other with a very low degree of positive relationship then we can notice a graph as in fig.below.

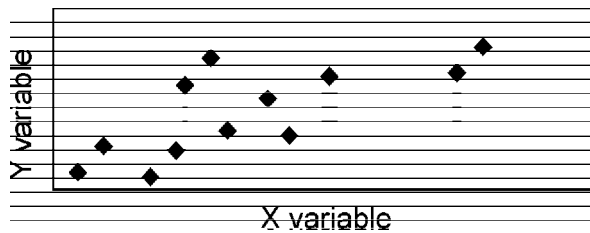


Fig.: Very Low Positive Correlation ($r = \text{near to } +0$)

3. Negative Correlation.

Ans :

Correlation is said to be negative or inverse if the variables deviate in the opposite direction i.e., if the increase (decrease) in the values of one variable results, on the average, in a corresponding decrease (increase) in the values of the other variable.

Examples

1. Price and demand of a commodity.
2. Sales of Woolen garments and the day temperature.

i) Perfect Negative Correlation

If the variables X and Y are perfectly negatively related to each other then, we get a graph as shown in fig.below.

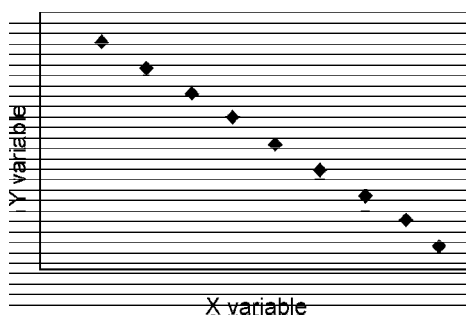


Fig.: Perfect Negative Correlation ($r = -1$)

ii) Very High Negative Correlation

If the variables X and Y are related to each other with a very high degree of negative relationship then we can notice a graph as in fig.below.

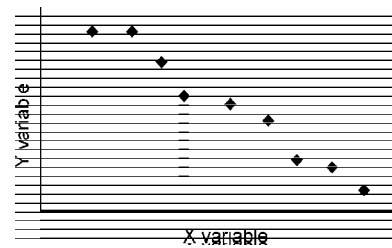


Fig.: Very High Negative Correlation ($r = \text{near to } -1$)

iii) Very low Negative Correlation

If the variables X and Y are related to each other with a very low degree of negative relationship then we can notice a graph as in fig.below.

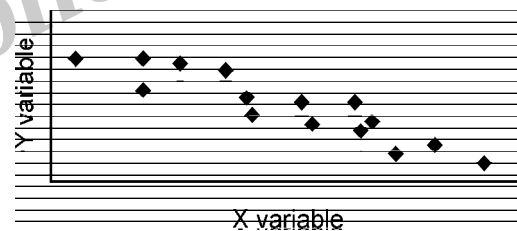


Fig.: Very Low Negative Correlation

4. State the Limits for Coefficient of Correlation.

Ans :

Limits for Co-efficient of Correlation for (x, y)

The value of the coefficient of correlation should lie between +1 and -1. If $r = +1$, the correlation is perfect and positive and if $r = -1$ the correlation is perfect and negative. When $r = 0$, it means that there is no relationship between the two variables.

$$\text{Hence, } -1 < r(x, y) \leq 1$$

Note

The correlation coefficient describes not only the magnitude of correlation but also its direction. Thus, +1 would mean that correlation is positive and the magnitude of correlation is 1.

Similarly -1 means correlation is negative and the magnitude of correlation is again 1.

S.No.	Degree of Correlation	The Value of $r(x,y)$ (Positive and Negative)
1.	Perfect correlation	Exactly 1
2.	Very high degree of correlation	0.90 and above but less than 1
3.	Fairly high degree of correlation	0.75 and above but less than 0.90
4.	Moderate degree of correlation	0.50 and above but less than 0.75
5.	Low degree of correlation	0.25 and above but less than 0.50
6.	Very high degree of correlation	Below 0.25
7.	Absence of correlation	Equal to 0

5. Properties of Karl Pearson's Coefficient of Correlation

Ans :

1. It is based on Arithmetic Mean and Standard Deviation.
2. It lies between -1
3. It measures both direction as well as degree of change. If r is less than 0, there is negative correlation, which means the direction of change of the two variables will be opposite. If r is more than 0, there is positive correlation, which means the direction of change of the two, variables will be same. Higher the value of r , greater is the degree of correlation. Hence, Karl Pearson's coefficient of correlation is said to be the ideal measure of correlation.
4. It is independent of change in scale. In other words, if a constant amount is added/ subtracted from all values of a variable, the value of r does not change.
5. It is independent of change in origin. Thus, if a constant amount is multiplied with or divides all values of a variable, r does not change.
6. It is independent of direction. In other words, Correlation of X and Y is same as Correlation of Y and X .
7. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.
8. It takes into account all items of the variable(s).

6. Spearman's Rank Correlation.

Ans :

The Karl Pearson's method is based on the assumption that the population being studied is normally distributed. When it is known that the population is not normal or when the shape of the distribution is not known, there is need for a measure of correlation that involves no assumption about the parameter of the population.

It is possible to avoid making any assumptions about the populations being studied by ranking the observations according to size and basing the calculations on the ranks rather than upon the original observations. It does not matter which way the items are ranked, item number one may be the largest or it may be the smallest. Using ranks rather than actual observations gives the coefficient of rank correlation.

This method of finding out variability or the lack of it between two variables was developed by the British Psychologist Charles Edward Spearman in 1904. This measure is especially useful when quantitative measures for certain factors (such as in the evaluation of leadership ability or the judgment of female beauty) cannot be fixed, but the individual in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank in the group. Spearman's rank correlation coefficient is defined as :

$$R = 1 - \frac{6\sum D^2}{N(N^2 - 1)} \text{ or } 1 - \frac{6\sum D^2}{N^3 - N}$$

where R denotes rank coefficient of correlation and D refers to the difference of rank between paired items in two series.

Features of Spearman's Correlation Coefficient

- The sum of the differences of ranks between two variables shall be zero. Symbolically, $\sum D = 0$
- Spearman's correlation coefficient is distribution-free or non-parametric because no strict assumptions are made about the form of population from which sample observations are drawn.
- The Spearman's correlation coefficient is nothing but Karl Pearson's correlation coefficient between the ranks. Hence, it can be interpreted in the same manner as Pearsonian correlation coefficient.

7. Define the term Regression.

Ans :

Meaning

The dictionary meaning of the term 'regression' is the act of the returning or going back. The term 'regression' was first used by Sir Francis Galton in 1877 while studying the relationship between the heights of father and sons. Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables.

1. **Dependant Variable** is the single variable being explained/ predicted by the regression model (response variable).
2. **Independent Variable** is the explanatory variable(s) used to predict the dependant variable (Predictor variable).

Definitions

- (i) "Regression is the measure of the average relationship between two or more variables in terms of the original units of data."
- (ii) **According to Blair**, "Regression is the measure of the average relationship between two or more variable in terms of the original units of the data."
- (iii) **According to Taro Yamane**, "One of the most frequently used techniques in economics and business research, to find a relation between two or more variable that are related causally, is regression analysis."

8. Utility of Regression Analysis.

Ans :

1. Regression analysis helps in establishing a functional relationship between two or more variables. Once this is established it can be used for various advanced analytical purposes.
2. Since most of the problems of economic analysis are based on cause and effect relationship, the regression analysis is a highly valuable tool in economics and business research.

3. This can be used for prediction or estimation of future production, prices, sales, investments, income, profits and population which are indispensable for efficient planning of an economy and are of paramount importance to a businessman or an economist.
4. Regression analysis is widely used in statistical estimation of demand curves, supply curves, production functions, cost functions, consumption functions, etc. Economists have discovered many types of production functions by fitting regression lines to input and output data.

9. Define Correlation.

Ans :

The following assumptions are made while making use of the regression technique:

1. There exists an actual relationship between the dependent and independent variables.
2. The regression analysis is used to estimate the values within the range for which it is valid and not for the values outside its range.
3. The relationship that existed between the dependent and independent variables remains the same till the regression equation is calculated.
4. The dependent variable takes any random value but the values of the independent variables are fixed quantities without error and are chosen by the analyst or the user.
5. In regression, we have only one dependent variable in our estimating equation. However, we can use more than one independent variable.

10. What are the limitations of Regression Analysis.

Ans :

1. It assumes a linear relationship between two variables which need not be the case always.
2. It assumes a static relationship between the two variables over a period of time. However, relationships between variables can change

with a change in other factors. For example, the change in demand for a given change in price can be estimated using regression. However, the impact of price on demand will be different when a family or a nation is poor and when such a family or nation has abundance of wealth or resources.

3. Regression analysis provides meaningful insights only up to a certain limit. For example, increasing production results in a decrease in marginal cost. However, beyond a certain point, increase in production can result in the costs going up.

11. Define Time Series.

Ans :

Introduction

An arrangement of statistical data in accordance with time of occurrence or in a chronological order is called a time series. The numerical data which we get at different points of time-the set of observations is known as time series.

Examples:

- i) The Annual Production of Steel in India over the last 10 years;
- ii) The Monthly Sales of a Chemical Industry for the last 6 months;
- iii) The daily closing price of a share in the Calculate Stock Exchange;
- iv) Hourly temperature recorded by the Meteorological office in a city;
- v) Yearly Price or Quantity Index Numbers.

Mathematically, a Time Series is defined by the values $Y_1, Y_2, \dots, Y_n$ of a variable Y at times $t_1, t_2, \dots, t_n$

Here Y is a function of time t and Y_t denotes the value of the variable Y at time t . The basic assumption is that changes witnessed over time in a sample group will be extrapolated to population.

Definitions

- i) **According to Morris Hamburg**, "A time series is a set of observations arranged in chronological order".

- ii) **According to Kenny and Keeping**, "A set of data depending on the time is called time series".
 - iii) **According to Croxton and Cowden**, "A time series consists of data arranged chronologically".
 - iii) **According to Patterson**, "Time series consists of statistical data which are collected, recorded or observed over successive increments".
-

12. Explain the uses of trend analysis.

Ans :

Uses of Trend

- 1) The trend describes the basic growth tendency ignoring short term fluctuations.
- 2) It describes the pattern of behavior which has characterized the series in the past.
- 3) Future behavior can be forecasted in the assumption that past behavior will continue in the future also.
- 4) Trend analysis facilitate us to compare two or more time series over different period of time and this helps to draw conclusions about them.

Rahul Publications

Choose the Correct Answers

1. The strength (degree) of the correlation between a set of independent variables X and a dependent variable Y is measured by (d)
(a) Coefficient of Correlation (b) Coefficient of Determination
(c) Standard error of estimate (d) All of the above
2. The percent of total variation of the dependent variable Y explained by the set of independent variables X is measured by (c)
(a) Coefficient of Correlation (b) Coefficient of Skewness
(c) Coefficient of Determination (d) Standard Error or Estimate
3. A coefficient of correlation is computed to be -0.95 means that (c)
(a) The relationship between two variables is weak
(b) The relationship between two variables is strong and positive
(c) The relationship between two variables is strong and but negative
(d) Correlation coefficient cannot have this value
4. Let the coefficient of determination computed to be 0.39 in a problem involving one independent variable and one dependent variable. This result means that (c)
(a) The relationship between two variables is negative
(b) The correlation coefficient is 0.39 also
(c) 39% of the total variation is explained by the independent variable
(d) 39% of the total variation is explained by the dependent variable
5. Relationship between correlation coefficient and coefficient of determination is that (b)
(a) both are unrelated
(b) The coefficient of determination is the coefficient of correlation squared
(c) The coefficient of determination is the square root of the coefficient of correlation
(d) both are equal
6. Multicollinearity exists when (a)
(a) Independent variables are correlated less than -0.70 or more than 0.70
(b) An independent variables is strongly correlated with a dependent variable
(c) There is only one independent variable
(d) The relationship between dependent and independent variable is non-linear
7. If "time" is used as the independent variable in a simple linear regression analysis, then which of the following assumption could be violated (d)
(a) There is a linear relationship between the independent and dependent variables
(b) The residual variation is the same for all fitted values of Y
(c) The residuals are normally distributed
(d) Successive observations of the dependent variable are uncorrelated

8. In multiple regression, when the global test of significance is rejected, we can conclude that (c)
- (a) All of the net sample regression coefficient are equal to zero
 - (b) All of the sample regression coefficient are not equal to zero
 - (c) At least one sample regression coefficient is not equal to zero
 - (d) The regression equation intersects the Y-axis at zero.
9. A residual is defined as (a)
- (a) $Y - \hat{Y}$
 - (b) Error sum of square
 - (c) Regression sum of squares
 - (d) Type I Error
10. What test statistic is used for a global test of significance ? (d)
- (a) Z test
 - (b) t test
 - (c) Chi-square test
 - (d) F test

Rahul Publications

Fill in the blanks

1. An orderly set of data arranged in accordance with their time of occurrence is called _____
2. A time series consists of _____
3. The graph of time series is called _____
4. Secular trend can be measured by _____
5. The secular trend is measured by the method of semi-averages when _____
6. Increase in the number of patients in the hospital due to heat stroke is _____
7. The systematic components of time series which follow regular pattern of variations are called _____
8. The unsystematic sequence which follows irregular pattern of variations is called _____
9. In time series seasonal variations can occur within a period of _____
10. Wheat crops badly damaged on account of rains is _____

ANSWERS

1. Time series
2. All of the above
3. Historigram
4. Four methods
5. Trend is linear
6. Seasonal variation
7. Signal
8. Noise
9. One year
10. Random movement

FACULTY OF MANAGEMENT
B.B.A. II - Semester (CBCS) Examination
January - 2021
BUSINESS STATISTICS

Time : 2 Hours]

[Max. Marks : 80

PART - A (4 × 5 = 20 Marks)

Note : Answer any four questions.

1. Statistical tools

Ans :

Statistical methods involved in carrying out a study include planning, designing, collecting data, analyzing, drawing meaningful interpretation and reporting of the research findings. The statistical analysis gives meaning to the meaningless numbers, thereby breathing life into a lifeless data. The results and inferences are precise only if proper statistical tests are used.

2. Primary data (Unit-I, Q.No. 11, (1st point))
3. Graphical presentation (Unit-I, Q.No. 32)
4. Probability (Unit-V, SQA-1)
5. Class Interval (CI) (Unit-I, Q.No. 23 (iii-point))
6. Mean = 35.4, Median = 34.3 find the mode ?

Sol :

$$\begin{aligned}\text{Mode} &= 3 \text{ Median} - 2 \text{ Mean} \\ &= 3 \times 34.3 - 2 \times 35.4 \\ &= 102.9 - 70.8 \\ &= 32.1\end{aligned}$$

7. Standard Error (Unit-IV, Q.No. 28)
8. Sampling method. (Unit-IV, Q.No. 22)

PART - B (4 × 15 = 60 Marks)

Note : Answer any four questions.

9. What are the functions of graphs? Advantages and limitations of graphs. (Unit-I, Q.No. 34, 35)
10. In a sample study about coffee habit in two towns, the following information was received.

Town A: Females 40%, total coffee drinkers were 45% and males noncoffee drinkers were 20%.

Town B: Males were 55%, male non-coffee drinkers 30% and females coffee drinkers were 15%

Present the above data in a tabular form.

Ans :

Coffee Drinking Habits in two towns

Coffee Drinking Habits	Town A			Town B		
	Male	Female	Total	Male	Female	Total
Coffee Drinkers	40%	5%	45%	25%	15%	40%
Non Coffee Drinkers	20%	35%	55%	30%	30%	60%
Total	60%	40%	100%	55%	45%	100%

11. Let 10, 5, 2, 3, 5, 20, 40, 10, 5 find the variance

Ans :

$$\text{Mean } \bar{x} = \frac{10 + 5 + 2 + 3 + 5 + 20 + 40 + 10 + 5}{9}$$

$$= \frac{100}{9}$$

$$= 11.11$$

$$\text{Variance} = \frac{(X - \bar{X})^2}{n}$$

X	(X - $\bar{X}$)	(X - $\bar{X}$) ²
10	- 1.11	1.2321
5	- 6.11	37.3321
2	- 9.11	82.9921
3	- 8.11	65.7721
5	- 6.11	37.3321
20	8.89	79.0321
40	28.89	834.6321
10	- 1.11	1.2321
5	- 6.11	37.3321
		<u>1176.889</u>

$$= \frac{1176.889}{9} = 130.765$$

12. Find the Mean, Median for the following data :

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Frequency	6	5	8	15	7	6	3

Sol :

Calculation of Mean and Median

Marks	Frequency (F)	Mid-point (M)	FM	Cumulative Frequency
0 – 10	6	5	30	6
10 – 20	5	15	75	11
20 – 30	8	25	200	19 f
30 – 40	15 F	35	525	34
40 – 50	7	45	315	41
50 – 60	6	55	330	47
60 – 70	3	65	195	50 → N
	<u>50</u>		<u>1670</u>	

$$\text{Mean} = \frac{\Sigma FM}{N} = \frac{1670}{50} = 33.4$$

$$\text{Median} = L + \frac{\frac{N}{2} - F}{f} \times C$$

$$= \frac{50}{2} = 25$$

$$= 30 + \frac{25 - 15}{19} \times 10$$

$$= 30 + \frac{100}{19}$$

$$= 30 + 5.26 = 35.26$$

13. From the following data calculate index number by simple aggregate method

Commodity	Price in 1997 Rs.	Price in 1998 Rs.
A	162	171
B	256	164
C	257	180
D	132	145

Ans :

Commodity	Price in 1997 (p ₀)	Price in 1998 (p ₁)
A	162	171
B	256	164
C	257	180
D	132	145
	<u>807</u>	<u>660</u>

$$P_{0,1} = \frac{\sum p_1}{\sum p_0} \times 100$$

$$= \frac{660}{807} \times 100$$

$$= 81.78$$

 14. Calculate the consumer price index for the following information.

Commodity	A	B	C	D	E
Base year quantity	2	4	1	4	1
Base year price	10	14	15	6	12
Current year price	12	16	20	8	16

Ans :

Commodity	Weights (w)	p ₀	p ₁	$R = \frac{p_1}{p_0} \times 100$	RW
A	2	10	12	120	240
B	4	14	16	114.28	457.12
C	1	15	20	133.33	133.33
D	4	6	8	133.33	533.32
E	1	12	16	133.33	133.33
	<u>12</u>				<u>1497.1</u>

$$\frac{1497.1}{12} = 124.75$$

15. Three unbalanced coins are tossed. What is the probability of obtaining (i) All heads (ii) Two heads (iii) One head (iv) at least one head (v) at least two heads (vi) All tails.

Ans :

(i) Getting all heads

Let E_1 = Event of getting all heads, Then

$$E_1 = \{HHH\} \text{ and therefore } n(E_1) = 1$$

$$\text{Therefore, } p \{\text{getting 011 heads}\} = p(E_1) = \frac{n(E_1)}{n(S)} = \frac{1}{8}$$

(ii) Getting two heads

Let E_2 = Event of getting two heads, Then

$$E_2 = \{HHT, HTH, THH\} \text{ and therefore } n(E_2) = 3$$

$$\text{Therefore, } p \{\text{getting 2 heads}\} = p(E_2) = \frac{n(E_2)}{n(S)} = \frac{3}{8}$$

(iii) One head

Let E_3 = Event of getting one heads, Then

$$E_3 = \{HTT, THT, TTH\} \quad n(E_3) = 3$$

$$P(E_3) = \frac{n(E_3)}{n(S)} = \frac{3}{8}$$

(iv) Atleast one head

Let E_4 = Getting atleast 1 head. Then

$$E_4 = \{HTT, THT, TTH, HHT, HTH, THH, HHH\}$$

$$n(E_4) = 7 \quad p(E_4) = \frac{n(E_4)}{n(S)} = \frac{7}{8}$$

(v) Atleast two heads

Let E_5 = Getting atleast 2 heads. Then

$$E_5 = \{HHT, HTH, THH, HHH\}$$

$$\text{Therefore } n(E_5) = 4$$

$$p(E_5) = \frac{n(E_5)}{n(S)} = \frac{4}{8} = \frac{1}{2}$$

(vi) All tails

E_6 = Getting all tails

$$E_6 = \{TTT\}$$

$$n(E_6) = 1$$

$$p(E_6) = \frac{n(E_6)}{n(S)} = \frac{1}{8}$$

16. What are the merits and limitations of stratified Random sampling. (Unit-IV, Q.No. 22)
17. Calculate Spearman's Rank correlation and comment.

X	480	330	400	90	160	160	650	240	460	570
Y	130	130	240	60	150	40	200	90	60	190

Ans :

X	R ₁	Y	R ₂	R ₁ - R ₂	D ²
480	3	130	5.5	-2.5	6.25
330	6	130	5.5	0.5	0.25
400	5	240	1	4	16
90	10	60	8.5	1.5	2.25
160	8.5	150	4	4.5	20.25
160	8.5	40	10	-1.5	2.25
650	1	200	2	1	1
240	7	90	7	0	0
460	4	60	8.5	-4.5	20.25
570	2	190	3	-1	1
					<u>69.5</u>

$$1 - \frac{6 \left(\sum d^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) \right)}{N^3 - N}$$

$$1 - \frac{6 \left[69.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) \right]}{10^3 - 10}$$

$$1 - \frac{6 \times [69.5 + 0.5 + 0.5 + 0.5]}{1000 - 10}$$

$$1 - \frac{6 \times 71}{990}$$

$$1 - \frac{426}{990}$$

$$1 - 0.430 = 0.57$$

18. Find a trend line to the following data using least squares method and estimate the sales volume for the year 2005,

Year	1994	1995	1996	1997	1998	1999	2000	2001
Sales Volume	425	375	298	390	468	456	575	660

Ans :

Year	Sales	X	X ²	-XY
1994	425	-7	49	-2975
1995	375	-5	25	-1875
1996	298	-3	9	-894
1997	390	-1	1	-390
1998	468	1	1	468
1999	456	3	9	1368
2000	575	5	25	2875
2001	660	7	49	4620
	<u>3647</u>		<u>168</u>	<u>3197</u>

$$Y_c = a + bX$$

$$a = \frac{\Sigma y}{n} = \frac{3647}{8} = 455.87$$

$$b = \frac{\Sigma xy}{\Sigma X^2} = \frac{3197}{168} = 19.03$$

Straight line trend is $Y_c = 455.87 + 19.03X$

Estimation of sales volume for the year

$$\begin{aligned}
 &2005 \ X = 15 \\
 &= 455.87 + 19.03 (15) \\
 &= 455.87 + 285.45 \\
 &= 741.32
 \end{aligned}$$

FACULTY OF MANAGEMENT
B.B.A. II - Semester (CBCS) Examination
November - 2020
BUSINESS STATISTICS

Time : 2 Hours]

[Max. Marks : 80

PART - A (4 × 5 = 20 Marks)

Note : Answer any four questions.**Answers**

1. Drawing Diagrams

Ans :

A diagram is a visual form for presenting statistical data for highlighting the basic facts and relationship which are inherent in the data. The diagrammatic presentation is more understandable and it is appreciated by everyone. It attracts the attention and it is a quicker way of grasping the results saving the time. It is very much required, particularly, in presenting qualitative data.

2. Statistical Tools	(Refer Jan.-21, Q.No. 1)
3. Standard Deviation	(Unit-II, Q.No. 48)
4. Definitions of Probability	(Unit-IV, SQA-1)
5. Arithmetic Mean (AM)	(Unit-II, Q.No. 6)
6. Primary Data	(Unit-I, Q.No. 11, (1 st Point))
7. Spearman's Rank Correlation	(Unit-V, SQA-6)
8. Moving Averages.	(Unit-V, Q.No. 31)

PART - B (4 × 15 = 60 Marks)

Note : Answer any four questions.

9. What is Classification ? Explain various bases of classification. (Unit-I, Q.No. 19, 20)
10. There are 17 balls numbered from 1 to 17 in a bag. If a person selects one at random, what is the probability that the number printed on the ball will be an even number greater than 9?

Ans :

There are 17 balls numbered from 1 to 17 in a bag.

The number printed on the ball will be an even number greater than 9

The probability getting chance of even number greater than 9

$$= 10, 12, 14, 16$$

$$= \frac{4}{17}$$

11. Find the sum of "n" terms of the following series.

4, 7, 11, 14 (where $n = 10$)

Ans :

Here, $a_1 = 4$, $a_2 = 7$

Now, $n = 10$. (given)

$$d = a_2 - a_1 = 7 - 4 = 3$$

$$a = a_1 = 4$$

$$\begin{aligned} \text{Also, } S_n &= n/2 [2a + (n - 1)d] \\ &= 10/2 [(2 \times 4) + (10 - 1)3] \\ &= 5[8 + (9) 3] \\ &= 5[8 + 27] \\ &= 5 \times 35 = 175 \end{aligned}$$

So, $S_n = 175$

12. Find the mean, mode, median of the following data.

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Frequency	6	5	8	15	7	6	3

Ans :

Calculation of Mean, Median and Mode

Marks	Frequency	Cumulative frequency	Mid point (M)	FM
0 – 10	6	6	5	30
10 – 20	5	11	15	75
20 – 30	8 f_1	19 f	25	200
30 – 40	15 $F f$	34	35	525
40 – 50	7 f_2	41	55	330
50 – 60	6	47	66	330
60 – 70	3	50	65	195
	<u>50</u>			<u>1670</u>

$$\text{Mean} = \frac{\Sigma FM}{N} = \frac{1670}{50} = 33.4$$

$$\text{Median} = L + \frac{\frac{N}{2} - F}{f} \times C$$

$$\frac{N}{2} = \frac{50}{2} = 25$$

$$30 + \frac{25-15}{19} \times 10$$

$$30 + \frac{100}{19} \\ = 30 + 5.26 = 35.26$$

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

$$\Delta_1 = f - f_1 \\ = 15 - 8 = 7$$

$$\Delta_2 = f - f_2 \\ = 15 - 7 = 8$$

$$= 30 + \frac{7}{7+8} \times 10$$

$$= 30 + \frac{70}{15}$$

$$= 30 + 4.6 = 34.6$$

13. Discuss the uses and limitation of Index Numbers.

(Unit-III, Q.No. 3, 4)

14. Compute consumer price index for the given data.

Group	Group Index No.	Weight
Food	277	33
Clothing	283	7
Shelter	322	10
Education	107	12
Fuel	355	4

Sol :

Group	Group Index No.	Weight	WI
Food	277	33	9141
Clothing	283	7	1981
Shelter	322	10	3220
Education	107	12	1284
Fuel	355	4	1420
		<u>66</u>	<u>17046</u>

$$= \frac{\sum IW}{\sum W} = \frac{17046}{66} = 258.27$$

15. A dice was thrown 9000 times and either 3 or 4 found to have been occurred 3220 times. Is this consistent with the hypothesis that the dice is unbiased ? Use $\alpha = 1\%$.

Ans :

$$n = 9000$$

$$p(3 \text{ or } 4) = 1/6 + 1/6 = 2/6 = 1/3$$

$$(1 - p) = 2/3$$

$$\text{Mean} = n * p = 9000 * 1/3 = 3000.$$

Standard error equal $\sqrt{n * p * (1 - p)} = \sqrt{9000 * 1/3 * 2/3} = \sqrt{2000} = \text{Somewhere between } 44.72 \text{ and } 44.73, \text{ the exact value not being important for this problem.}$

A raw score of 3240 will generate a z-score of $(x - m) / se$ which would be equal to:

$$z = (3,240 - 3,000) / 44.72 \text{ or } 44.73.$$

Therefore $z = 5.367 \text{ or } 5.366.$

Both are more than 3 standard deviations from the mean.

The probability of getting a z-score greater than 3 standard deviations from the mean would be equal to .00135.

This is less than .05 and also less than .01 which are the probabilities normally used as criteria.

The .05 is normally used at the .95 confidence level for a one tailed test.

The .01 is normally used as the .99 confidence level for a one tailed test.

Bottom line is the probability of getting such a score out of 9000 tosses just based on random differences between different samples is very remote and therefore not considered feasible.

This indicates that the dice are probably biased in some way.

16. In a random sample of 100 persons taken from village A, 45 are found to be consuming tea. In another sample of 200 persons taken from village B, 120 persons are found to be consuming tea. Do you conclude that people in village A are less consumers of tea than people in village B? Use $\alpha = 10\%$.

Sol :

Given that $n_1 = 100$ $x_1 = 45$ and $n_2 = 200$ $x_2 = 120$

$$p_1 = \text{Sample proportion of consuming tea} = \frac{x_1}{n_1} = \frac{45}{100} = 0.45 \text{ of } q_1 = 0.55$$

$$p_2 = \text{Sample proportion of consuming tea} = \frac{x_2}{n_2} = \frac{120}{200} = 0.6 \text{ of } q_2 = 0.4$$

Null Hypothesis

$H_0 = p_1 = p_2$ i.e., there is no significant differences between A and B

Alternate Hypothesis

$H_1 = p_1 \neq p_2$ i.e., there is a significant differences between A and B

Here population proportion p is not know

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + p_1}$$

$$P = \frac{100 \times 0.45 + 200 \times 0.6}{100 + 200}$$

$$= \frac{45 + 120}{300} = \frac{165}{300} = 0.55$$

$$P = 0.55 \text{ and } Q = 0.45$$

The test statistic is given by

$$Z = \frac{p_1 - p_2}{PQ \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$Z = \frac{0.45 - 0.60}{\sqrt{0.55 \times 0.45 \left(\frac{1}{100} + \frac{1}{200} \right)}} = \frac{-0.15}{\sqrt{0.2475 \left(\frac{2+1}{200} \right)}}$$

$$= \frac{-0.15}{\sqrt{0.2475 \times \frac{3}{200}}}$$

$$= \frac{-0.15}{\sqrt{0.2475 \times 0.015}}$$

$$= \frac{-0.15}{0.0609} = -2.46$$

The tabulated value $Z_{\alpha} = 1.96$

$2.46 > 1.96$ Reject H_0

We accept Alternate Hypothesis

17. Fit a linear trend line to the following data using "Least Squares Principle" and estimate the sales volume for the year 2007.

Year	1994	1995	1996	1997	1998
Sales Volume	425	378	298	390	468

Sol :

Year	Sales Volume (y)	X	X ²	XY
1994	425	-2	4	-850
1995	378	-1	1	-378
1996	298	0	0	0
1997	390	1	1	390
1998	468	2	4	936
	<u>1959</u>		<u>10</u>	<u>98</u>

$$Y_c = a + bX$$

$$a = \frac{\Sigma y}{n} \quad b = \frac{\Sigma xy}{\Sigma X^2}$$

$$a = \frac{1959}{5} \quad b = \frac{98}{10}$$

$$a = 391.8 \quad b = 9.8$$

$$Y_c = 391.8 + 9.8 X$$

The estimate the sales volume for the $Y_c = 2007$ $X = 11$

$$Y_c = 391.8 + 9.8 (11)$$

$$= 381.8 + 107.8 = 499.6$$

18. Determine the Karl Pearson's coefficient of correlation and comment.

X	5	15	25	35	45	55	65	75
Y	55	67	100	111	150	200	300	500

Sol :

X	Y	X - $\bar{X}$ (x)	Y - $\bar{Y}$ (y)	xy	x ²	y ²
5	55	-35	-142	4970	1225	20,164
15	67	-25	-130	3250	625	16,900
25	100	-15	-97	1455	225	9409
35	111	-5	86	430	25	7396
45	150	+5	-47	-235	25	2209
55	290	+15	+93	1395	225	8649
65	300	+25	+103	2575	625	10,609
75	500	+35	+303	10,605	1,225	91,809
<u>320</u>	<u>1573</u>			24,445	4,200	1,67,145

$$\bar{X} = \frac{320}{8} = 40$$

$$\bar{Y} = \frac{1573}{8} = 197$$

$$= \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$= \frac{24445}{\sqrt{4200 \times 1,67,145}}$$

$$= \frac{24,445}{26,495} = 0.92$$

FACULTY OF MANAGEMENT**Bachelor of Business Administration (BBA) II - Semester (CBCS) Examination****May - 2019****BUSINESS STATISTICS**

Time : 3 Hours]

[Max. Marks : 80

Note: Answer all the questions.**PART – A (5 × 4 = 20 Marks)****[Short Answer Type]****ANSWERS**

1. Answer any five of the following questions.

(a) Time Series Analysis	(Unit-V, SQA - 11)
(b) Scatter diagram	(Unit-V, SQA - 7)
(c) Sampling method	(Unit-IV, Q.No. 22)
(d) Skewness	(Unit-II, Q.No. 54)
(e) Geometric Mean (GM)	(Unit-II, SQA - 4)
(f) Variance Meaning	(Unit-II, Q.No. 52)
(g) Secondary Data	(Unit-I, Q.No. 15)
(h) Price Index	

Ans :**Weighted Price Indexes**

At the time of constructing the weighted price indexes or indices, the rational weights are allocated in an explicit manner. These rational weights show the relative significance of items or commodities which are related with the computation of an index. Quantity weights and value weights are used in this weighted indexes or indices. Weighted price indexes, or indices are further divided into two types as follows,

- (a) Weighted aggregate price index
- (b) Weighted average of price relative index.

(a) Weighted Aggregate Price Index

In a weighted aggregate price index, certain weight is assigned to each and every commodity or item of group in accordance with its significance. This helps in gathering more information and improving accuracy of the estimates. The following methods are used in weighted aggregate price index:

- (i) Laspeyre's method
- (ii) Pasche's method
- (iii) Fisher's ideal method
- (iv) Kelly's method
- (v) Walsch's method
- (vi) Marshal Edgeworth's method
- (vii) Dorbish and Bowley's method

(b) Weighted Average of Price Relative Index

In weighted average of price relative index, value of each commodity or item related with the calculation of composite index is ascertained by multiplying the price of each item with its quantity consumed. Quantity consumed is considered for computing the weighted average of price relative. The formula for weighted average of price relative index is as follows,

$$P_{01} = \frac{\sum((p_1 \div p_0) \times 100)(p_0 q_0)}{\sum p_0 q_0} = \frac{\sum PV}{\sum V} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Where,

$V (= p_0 q_0)$ = Base prices and quantities determining values

$$P \left(= \frac{p_1}{p_0} \times 100 \right) = \text{Price relative}$$

This formula is equivalent to the formula of Laspeyre's index formula,

If 'V' is taken as p_0, q_1 , then the formula would,

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Then it would be equal to Paasche's index method.

PART – B**[Essay Answer Type]**

Note: Answer all the questions using the internal choice.

2. (a) What are the functions of statistics ? **(Unit-I, Q.No. 2)**

OR

- (b) An urn contains 10 white, 4 red and 5 black balls. If 3 balls are drawn at random find the probability that :
- (i) Two of the balls drawn are white
 - (ii) One is of each colour
 - (iii) None is red
 - (iv) Atleast one is white.

Ans :

Total No. of balls = 19

$n(s) = \text{No. of ways choosing 3 from 19} = {}^{19}C_3$

- (i) Two of the balls drawn are white

Suppose '2' balls are selected

This can be done in ${}^{10}C_2$ ways

Here $n(\epsilon) = {}^{10}C_2$

$$p(\epsilon) = \frac{n(\epsilon)}{n(s)} = \frac{{}^{10}C_2}{{}^{19}C_3}$$

(ii) One is of each colour

$$n(\epsilon) = {}^{10}C_1 \times {}^4C_1 \times {}^5C_1$$

$$n(s) = {}^{19}C_3$$

$$p(\epsilon) = \frac{{}^{10}C_1 \times {}^4C_1 \times {}^5C_1}{{}^{19}C_3}$$

(iii) None is red

$$n(\epsilon) = {}^4C_0$$

$$n(s) = {}^{19}C_3$$

$$p(\epsilon) = \frac{{}^4C_0}{{}^{19}C_3}$$

(iv) $n(\epsilon) = {}^{10}C_1$

$$n(s) = {}^{19}C_3$$

$$= \frac{{}^{10}C_1}{{}^{19}C_3}$$

3. (a) Find the sum of “n” terms of the following progression

7, 14, 28, 56, ... (where $n = 10$)

(Out of syllabus)

OR

(b) Find the sum of “n” terms of the following series

4, 7, 11, 14, ... (where $n = 10$)

(Out of syllabus)

4. (a) Compute consumer price index for the given data :

Group Index No.	Group	Weight
Food	277	53
Clothing	283	7
Shelter	322	10
Education	107	12
Fuel	355	4

Ans :

Computation of Consumer Price Index

Group	Group Index No (I)	Weight (W)	WI
Food	277	53	14681
Clothing	283	7	1981
Shelter	322	10	3220
Education	108	12	1284
Fuel	355	4	1420
		<u>86</u>	<u>22,586</u>

$$\text{Consumer Price Index} = \frac{\sum IW}{\sum W}$$

$$\begin{aligned}\text{Using Arithmetic Mean} &= \frac{22,586}{86} \\ &= 260.62\end{aligned}$$

OR

- (b) Explain splicing of index numbers and its importance.

(Unit-III, Q.No. 17)

5. (a) In a random sample of 100 persons taken from village A, 45 are found to be consuming tea. In another sample of 200 persons taken from village B, 120 persons are found to be consuming tea. Do you conclude that people in village A are less consumers of tea than people in village B ? Use $\alpha = 0.10\%$

(Incomplete Question)

OR

- (b) How does small sample tests differ from large ?

Ans :

Sl. No.	Large Sample	Small Sample
1.	The sample size is greater than 30.	The sample size is 30 or less than 30.
2.	The value of a statistic obtain from the sample can be taken as an estimate of the population parameter.	The value of a statistic obtain from the sample can not be taken as an estimate of the population parameter.
3.	Normal distribution is used for testing.	Sampling distribution like t, F etc. are used for testing.

6. (a) Find the coefficient of correlation for the following data :

X	1	2	3	4	5	6	7	8	9	10
Y	20	10	14	10	10	9	8	7	6	5

Ans :

X	Y	(X - $\bar{X}$) (x)	Y - $\bar{Y}$ (y)	x ²	y ²	xy
1	20	- 4.5	10.1	20.25	102.01	- 45.45
2	10	- 3.5	0.1	12.25	0.01	- 0.35
3	14	- 2.5	4.1	6.25	16.81	- 10.25
4	10	- 1.5	0.1	2.25	0.01	- 0.15
5	10	- 0.5	0.1	0.25	0.01	- 0.05
6	9	+0.5	-0.9	0.25	0.81	- 0.45
7	8	+1.5	- 1.9	2.25	3.61	- 2.85
8	7	+2.5	- 2.9	6.25	8.41	- 7.25
9	6	+3.5	- 3.9	12.25	15.21	- 13.65
10	5	+4.5	- 4.9	20.25	24.01	- 22.05
				<u>82.5</u>	<u>170.9</u>	<u>- 102.5</u>

$$\bar{X} = \frac{1+2+3+4+5+6+7+8+9+10}{10} = \frac{55}{10} = 5.5$$

$$\bar{Y} = \frac{20+10+14+10+10+9+8+7+6+5}{10} = \frac{99}{10} = 9.9$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-102.5}{\sqrt{82.5 \times 170.9}}$$

$$= \frac{-102.5}{\sqrt{14099.25}} = \frac{-102.5}{118.74} = -0.86$$

OR

- (b) Find the Regression Analysis for the following data. Find :

- (i) X on Y
(ii) Y on X

X	2	4	6	8
Y	10	20	25	30

Ans :

X	Y	X - $\bar{X}$ (x)	Y - $\bar{Y}$ (y)	x ²	y ²	xy
2	10	-3	-11.25	9	26.56	33.75
4	20	-1	-1.25	1	1.56	1.25
6	25	1	3.75	1	14.06	3.75
8	30	3	8.75	9	76.56	26.25
20	85			20	218.74	65

(i) X on Y

$$X - \bar{X} = b_{xy} (y - \bar{y})$$

(ii) Y on X

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

$$\bar{X} = \frac{20}{4} = 5$$

$$\bar{Y} = \frac{85}{4} = 21.25$$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{65}{218.74} = 0.29$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{65}{20} = 3.25$$

(i) X on Y

$$X - \bar{X} = b_{xy} (y - \bar{y})$$

$$X - 5 = 0.29 (y - 21.25)$$

$$X - 5 = 0.29y - 6.16$$

$$X = 0.29y - 6.16 + 5$$

$$X = 0.29y - 1.16$$

(ii) Y on X

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$Y - 21.25 = 3.25 (X - 5)$$

$$Y - 21.25 = 3.25 X - 16.25$$

$$Y = 3.25 X - 16.25 + 21.25$$

$$Y = 3.25 X + 5$$

FACULTIES OF MANAGEMENT
B.Com. II – Semester (CBCS) Examination
May - 2018
BUSINESS STATISTICS

Time : 3 Hours]

[Max. Marks : 80

PART - A (5 × 4 = 20 Marks)**[Short Answer Type]**

1. Answer any five of the following questions.

a) Define Statistics	(Unit - I, SQA. 1)
b) What do you mean by secondary data ?	(Unit - I, Q.No. 15)
c) Merits of Arithmetic Mean	(Unit - II, Q.No. 6)
d) Significance of measures of dispersion	(Unit - II, Q.No. 35)
e) Types of Index Numbers	(Unit - III, Q.No. 6)
f) Independent and Dependent Events	

Ans :

i) Independent Events

Two or more events are considered as independent events, when the outcome of one event does not influence and is not influenced by the other event.

Example

When a student has appeared in physics and chemistry examinations, his marks obtained in physics is independent of the marks obtained in chemistry.

ii) Dependent Events

Dependent events are the events in which the occurrence or non-occurrence of one event in any one trial influences the probability of occurrence of other events in other trials.

Example

When a card is drawn from a pack of playing cards and is not replaced then this changes the probability of occurrence of the second card.

Here probability of occurrence of second event is dependent on the occurrence of first event.

g) Types of Correlation	(Unit - V, Q.No. 4)
h) Components of Time Series	(Unit - V, Q.No. 24)

PART - B (5 × 12 = 60 Marks)**[Essay Answer Type]**

2. (a) What do you mean by statistics ? Explain the importance and scope of the statistics.

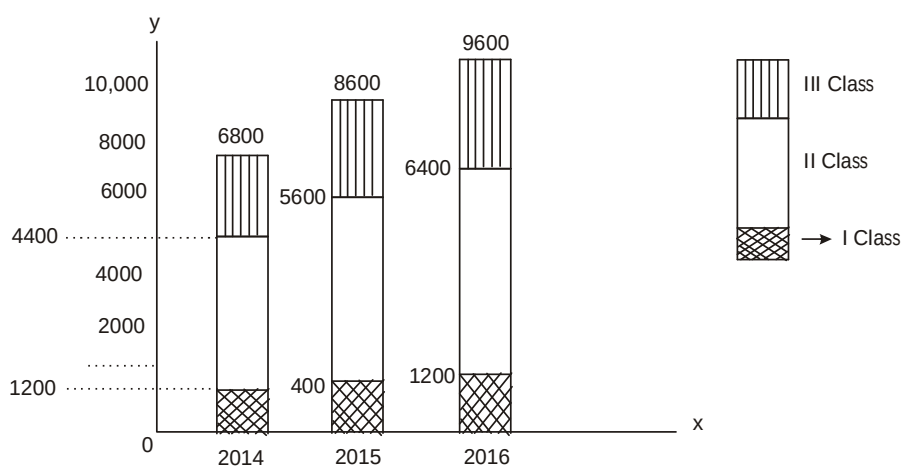
(Unit - I, Q.No. 1, 5, 6)

(OR)

- (b) The following table shows the result of B.B.A students for the last 3 years. Represent the data by sub-divided bar diagram.

Year	I Class	II Class	III Class
2014	1200	3,200	2,400
2015	1400	4,200	3,000
2016	1,200	5,200	3,200

Ans :



3. (a) Calculate Arithmetic Mean from the following data.

Marks	Below	20	30	40	50	60	70	80	90	100
No. of student		10	18	25	32	43	61	67	85	100

Ans :

C.I	f	Midpoint (m)	fm
10 - 20	10	15	150
20 - 30	8	25	200
30 - 40	7	35	245
40 - 50	7	45	315
50 - 60	11	55	605
60 - 70	18	65	1170
70 - 80	6	75	450
80 - 90	18	85	1530
90 - 100	15	95	1425
	N 100		6090

$$\bar{X} = \frac{\sum fm}{N} = \frac{6090}{100} = 60.9$$

(OR)

(b) Calculate the Quartile Deviation and its co-efficient.

Value	15-25	25-35	35-45	45-55	55-65	65-75	75-85	85-95
Frequency	32	38	45	98	122	80	50	25

	C.I	F	Cumulative frequency
	15 - 25	32	32
	25 - 35	38	70
	35 - 45	45	115
L	45 - 55	98	213
		F	(Q ₁)
	55 - 65	122	335
L	65 - 75	80	415
		F	(Q ₃)
	75 - 85	50	465
	85 - 95	25	490
			→ N

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

$$Q_3 = L + \frac{\frac{3N}{4} - F}{f} \times C$$

$$= \frac{3 \times 490}{4} = 367.5$$

$$65 + \frac{367.50 - 335}{80} \times 10$$

$$65 + 4.60 = 69.06$$

$$Q_1 = L + \frac{\frac{N}{4} - F}{f} \times C$$

$$\frac{N}{4} = \frac{490}{4} = 122.50$$

$$45 + \frac{122.50 - 115}{98} \times 10$$

$$45 + 0.76 = 45.76$$

$$\begin{aligned} Q.D &= \frac{69.06 - 45.76}{2} \\ &= 11.65 \end{aligned}$$

4. (a) Calculate Fisher's Ideal Index numbers from the following data.

Items	Price		Quantity	
	2016	2017	2016	2017
A	16	40	100	120
B	4	12	30	20
C	2	4	40	50
D	4	10	20	16
E	2	10	80	60

Ans :

Item	P ₀	P ₁	Q ₀	Q ₁	P ₁ q ₀	P ₀ q ₁	P ₁ q ₁	P ₀ q ₀
A	16	40	100	120	4000	1600	4800	1920
B	4	12	30	20	360	120	240	80
C	2	4	40	50	160	80	200	100
D	4	10	20	16	200	80	160	64
E	2	10	80	60	800	160	600	120
					5520	2040	6000	2284

$$= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100$$

$$= \sqrt{\frac{5520}{2040} \times \frac{6000}{2284}} \times 100$$

$$= \sqrt{2.7058 \times 2.6269} \times 100$$

$$= \sqrt{7.1078} \times 100$$

$$= 2.666 \times 100 = 266.60$$

(OR)

(b) Define Index Numbers and explain its uses and limitations.

(Unit - III, Q.No. 1,3,4)

5. (a) From 30 tickets marked with first 30 numerals, 1 ticket is drawn at random. It is then replaced and a second drawn is made. Find the probability that in the first drawn it is a multiple of 5 or 7 and in second drawn it is a multiple of 3 (or) 7.

Ans :

- (i) Let A be an event that first drawn is a multiple of 5

B is an event multiple of 7

Then $A = \{5, 10, 15, 20, 25, 30\}$ $B = \{7, 14, 21, 28\}$

$$P(A) = \frac{6}{30}, \quad P(B) = \frac{4}{30}$$

By using Additional Theorem (Non - Mutually exclusive events)

$$P(A \cup B) = P(A) + P(B)$$

$$P(A \cup B) = \frac{10}{30} = 0.30$$

- ii) Let A be an event is a multiple of '3' and

Let B be an event is a multiple of 7.

Then $A = \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$

$B = \{7, 14, 21, 28\}$

$$P(A) = \frac{10}{30}, \quad P(B) = \frac{4}{30}$$

$$P(A \cap B) = \frac{1}{30}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\frac{10}{30} + \frac{4}{30} - \frac{1}{30}$$

$$= \frac{13}{30}$$

(OR)

(b) Explain the principles of sampling.

(Unit - IV, Q.No. 17, 18, 19, 20, 21)

6. (a) Compute Karl Pearson's Co-efficient of correlation from the following data and calculate its probable error and interpret result.

Marks in Accountancy	77	54	27	52	14	35	90	25	56	60
Marks in Economics	35	58	60	40	50	40	35	56	34	42

Ans :

X	Y	X - $\bar{X}$ (x)	Y - $\bar{Y}$ (y)	x ²	y ²	xy
77	35	28	-10	784	100	- 280
54	58	5	13	25	169	65
27	60	-22	15	484	225	- 330
52	40	3	- 5	9	25	- 15
14	50	- 35	+ 5	1225	25	- 175
35	40	- 14	- 5	196	25	+ 70
90	35	45	- 10	2025	100	- 450
25	56	- 34	11	1156	121	- 374
56	34	7	- 11	49	121	- 77
60	42	11	- 3	121	9	- 33
490	450			6074	920	- 1599

$$\bar{x} = \frac{490}{10} = 49 \quad \bar{y} = \frac{450}{10} = 45$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{-1599}{\sqrt{6074 \times 920}}$$

$$= \frac{-1599}{\sqrt{5588080}} = \frac{-1599}{2363.9} = -0.67$$

(OR)

- (b) Karl Pearsons Co-efficient of correlation between the ages of brothers and sisters in a community was found to be 0.8. Average age of the brothers was 25 and that of sisters 22 years. Their standard deviations were 4 and 5 respectively. Find

- The expected age of brother when sister's age is 12 years.
- The expected age of sister when brother's age is 33 years.

Ans :

Given $r_{xy} = 0.8$

$$\bar{x} = 25, \bar{y} = 22$$

$$\sigma_x = 4 \quad \sigma_y = 5$$

- i) The expected age of brother when sister's age is 12 years

$$(x - \bar{x}) = r_{xy} \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$(x - 25) = 0.8 \left(\frac{4}{5} \right) (y - 22)$$

$$(x - 25) = 0.8 (0.8) (12 - 22)$$

$$(x - 25) = 0.64(-10)$$

$$(x - 25) = -6.4$$

$$x = -6.4 + 25$$

$$x = 18.6$$

ii) **The expected age of sister when brother's age is 33 years.**

$$(y - \bar{y}) = r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$(y - 22) = 0.8 \left(\frac{5}{4} \right) (x - 25)$$

$$(y - 22) = 0.8 (1.25) (33 - 25)$$

$$(y - 22) = 1(8)$$

$$y - 22 = 8$$

$$y = 22 + 8$$

$$y = 30$$

FACULTIES OF MANAGEMENT**B.B.A (CBCS) I-Year II-Semester Examination****May - 2017****BUSINESS STATISTICS****Time : 3 Hours]****[Max. Marks : 80****SECTION - A (5 × 4 = 20 M)***Answer any 5 Questions***ANSWERS**

1. Scope of statistics.	(Unit-I, SQA-2)
2. Requisites of measures of central tendency.	(Unit-II, Q.No. 5)
3. Significance of measure of dispersion	(Unit-II, Q.No. 35)
4. Characteristics of index numbers	(Unit- III, Q.No. 2)
5. Axiomatic approach to probability	(Unit- IV, Q.No. 5)
6. Inertia of large numbers	(Unit- IV, Q.No. 18)
7. Components of time series	(Unit-V, Q.No. 24)
8. Primary Data Vs Secondary Data	(Unit-I, Q.No. 18)

SECTION - B (12 × 5 = 60 M)*Answer all the Questions*

9. (a) The following table shows the net revenue earned by railways during nine years period as given below.

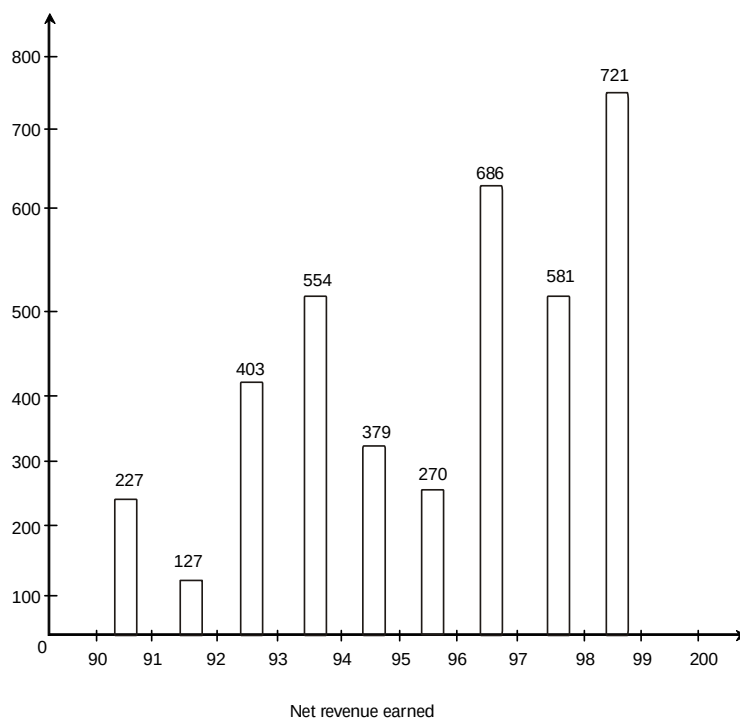
Year	Amount
1990 - 91	227
1991 - 92	127
1992 - 93	403
1993 - 94	554
1994 - 95	379
1995 - 96	270
1996 - 97	686
1997 - 98	581
1998 - 99	721

Draw a line diagram from the above data.

Ans :

Line Diagram

It is the simplest form of all the diagrams. It consists in drawing vertical lines, each vertical line being equal to frequency. The variate (x) values are presented on a suitable scale along the x-axis and the corresponding frequencies are presented on a suitable scale along y-axis. Line diagrams facilitate comparisons though they are not attractive or appealing to eye.



OR

- (b) Explain the functions, advantages and limitations of graphs.
10. (a) Calculate the Harmonic mean from the given data:

(Unit-I, Q.No. 34, 35)

X :	10	20	40	60	120
F :	1	3	6	5	4

Ans :

Computation of Harmonic Mean

X	F	Reciprocal $\times \left(\frac{1}{X}\right)$	F $\times$ Rec. X
10	1	0.10000	0.10000
20	3	0.05000	0.15000
40	6	0.02500	0.15000
60	5	0.01666	0.08330
120	4	0.00833	0.03332
	19		0.51662

$$\text{H.M} = \text{Rec.} \left(\frac{0.51662}{19} \right) = 36.77$$

OR

- (b) From the following data find out the Karl Pearson's Co-efficient of Skewness

X :	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5
F :	35	40	48	100	125	87	43	22

Ans :

Mid value (m)	Frequency (f)	d'	fd'	fd ¹²
4.5	35	- 3	- 105	- 11025
5.5	40	- 2	- 80	- 6400
6.5	48	- 1	- 48	- 2304
7.5	100	0	0	0
8.5	125	+ 1	125	15625
9.5	87	2	174	30276
10.5	43	3	129	16641
11.5	22	4	88	7744
	Total 500		283	50557

$$\text{Mean or } \bar{X} = A + \frac{\sum fd'}{N} \times h$$

$$= 7.5 + \frac{283}{500} \times 1 = 7.5 + 0.566 = 8.066$$

Mode = 7.5 By inspection

$$\text{S.Dev. } \sigma = h \times \sqrt{\left\{ \frac{\sum fd^{12}}{N} - \left(\frac{\sum fd'}{N} \right)^2 \right\}} = 1 \times \sqrt{\left\{ \frac{50557}{500} - \left(\frac{283}{500} \right)^2 \right\}}$$

$$= 1 \times \sqrt{101.144 - 0.3203} = \sqrt{100.7937} = 10.8$$

$$\text{S.K or } j = \frac{\bar{X} - H_0}{\sigma}$$

$$= \frac{8.066 - 7.5}{10.8} = \frac{0.566}{10.8} = 0.524$$

11. (a) Construct index number by chain base method for the following data:

Years	1992	1993	1994	1995	1996	1997
Price of wheat per 100 kg	530	600	770	800	840	900

Ans :

Chain Base Method

Year	Price	Link relatives	Chain base Index No.	1992 fixed base Index No.
1992	530	100	100	100
1993	600	$\frac{600}{530} \times 100 = 113.21$	$\frac{113.21}{100} \times 100 = 113.21$	$\frac{600}{530} \times 100 = 113.21$
1994	700	$\frac{770}{600} \times 100 = 128.33$	$\frac{128.33 \times 113.21}{100} = 145.28$	$\frac{770}{530} \times 100 = 145.28$
1995	800	$\frac{800}{770} \times 100 = 103.89$	$\frac{103.89 \times 145.28}{100} = 150.94$	$\frac{800}{530} \times 100 = 150.94$
1996	840	$\frac{840}{800} \times 100 = 105.00$	$\frac{105 \times 150.93}{100} = 158.47$	$\frac{840}{530} \times 100 = 158.49$
1997	900	$\frac{900}{840} \times 100 = 107.14$	$\frac{107.14 \times 158.47}{100} = 169.78$	$\frac{900}{530} \times 100 = 169.81$

Note : It may be noted that chain index numbers are the same as fixed base index no.

OR

(b) Explain the types and characteristics of index numbers.

(Unit-III, Q.No. 2, 6)

12. (a) Calculate the probability of getting 52 Sundays in a ordinary year and a leap year.

Ans :

The probability of getting 53 Sundays in an ordinary year = $\frac{1}{7}$

In a leap year (which consists of 364 days) there are 52 complete weeks and 2 more days. The following are the possible combinations for two more days.

1. Sunday and Monday
2. Monday and Tuesday
3. Tuesday and Wednesday
4. Wednesday and Thursday
5. Thursday and Friday
6. Friday and Saturday
7. Saturday and Sunday

In order that a leap year selected at random should contain 53 Sundays, one of the two 'more' day must be a Sunday. Since out of the above 7 possibilities, 2 namely (1) and (7) are favourable to this event, that required.

$$\text{Probability} = \frac{2}{7}$$

OR

(b) Write briefly about the various sampling methods.

(Unit-IV, Q.No. 22)

13. (a) The following figures relate to the profits of a commercial concern for 8 years. Calculate three yearly moving averages.

Years	1976	1977	1978	1979	1980	1981	1982	1983
Profits(Rs.)	15,420	14,470	15,520	21,020	26,120	31,950	35,370	34,670

Ans :

1 Year	2 Profits	3 3 yearly moving total	4 3 yearly moving Avg (4) $4 = (3) \div (3)$	5 Short item fluctuations $y - y_c$ 5 = (2) – (4)
1996	15,420	-	-	-
1977	14,470	$15,420 + 14,470 + 15,520 = 45,410$	15136.67	- 666.67
1978	15,520	$14,470 + 15,520 + 21,020 = 51,010$	17003.34	-1483.34
1979	21,020	$15,520 + 21,020 + 26,120 = 62,660$	20886.67	133.33
1980	26,120	$21,020 + 26,120 + 31,950 = 79,090$	26363.33	- 243.33
1981	31,950	$26,120 + 31,950 + 35,370 = 93,440$	31146.67	803.33
1982	35,370	$31,950 + 35,370 + 34,670 = 1,01,990$	33996.67	1373.33
1983	34,670	-	-	-

(b) Calculate rank correlation co-efficient from the following data

X	48	58	72	62	56	45	39	52	25
F	63	78	65	70	38	54	60	32	28

Ans :

Calculation for Rank correlation co-efficient

S.No.	X	F	R_1	R_2	$d = R_1 - R_2$	d^2
1	48	63	6	4	2	4
2	58	78	3	1	2	4
3	72	65	1	3	- 2	4
4	62	70	2	2	0	0
5	56	38	4	7	- 3	9
6	45	54	7	6	1	1
7	39	60	8	5	3	9
8	52	32	5	8	- 3	9
9	25	28	9	9	0	0
					0	40

The spearman's rank correlation coefficient is given by

$$r = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 40}{9(9^2-1)}$$

$$= 1 - \frac{240}{9(81-1)}$$

$$= 1 - \frac{240}{720}$$

$$= 1 - 0.333$$

$$= 0.667$$

FACULTY OF MANAGEMENT
B.B.A I Year II Semester (CBCS) Examination
MODEL PAPER - I
BUSINESS STATISTICS

Time : 3 Hours]

[Max. Marks : 80

SECTION - A (5 × 4 = 20 M)

Answer any 5 Questions

ANSWERS

1. (a) Scope of statistics (Unit-I, SQA. 2)
- (b) Merits of Arithmetic Average. (Unit-II, SQA. 2)
- (c) Define Geometric Mean. (Unit-II, SQA. 4)
- (d) Characteristics of Index Numbers (Unit-III, SQA. 2)
- (e) Explain Paasche's Method of index number. (Unit-III, SQA. 5)
- (f) Probability. (Unit-IV, SQA. 1)
- (g) Define Correlation. (Unit-V, SQA. 1)
- (h) State the Limits for Coefficient of Correlation. (Unit-V, SQA. 4)

SECTION - B (5 × 12 = 60 M)

Answer all the Questions

2. (a) Explain the importance of statistics in economic analysis and planning. (Unit - I, Q.No. 5)
- OR
- (b) What do you mean by Graphical Presentation ? Explain the techniques of construction of graphs. (Unit - I, Q.No. 32)
3. (a) From the following calculate skewness based on quartiles.

Income (₹)	Below 200	200 - 400	400 - 600	600 - 800	800 - 1000	1000 - Above
No. of Persons	25	40	80	75	20	16

(Unit - II, Prob. 49)

OR

- (b) Calculate mode for the following data.

Class Intervals	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Frequency	4	13	21	44	33	22	7

(Unit - II, Prob. 33)

4. (a) Explain the Characteristics of Index Numbers. (Unit - III, Q.No. 2)

OR

- (b) The following are the index numbers of wholesale prices of a commodity 'based on 1990 :

Year	Index Numbers
1993	100
1994	108
1995	120
1996	150
1997	210
1998	225
1999	240

Prepare new Index Numbers taking 1995 as base.

(Unit - III, Prob. 14)

5. (a) What are the various approaches of probability?

(Unit - IV, Q.No. 5)

OR

- (b) There are 5 white and 7 red balls in a box. A box is drawn and then replaced. What is the probability that a white and a red ball are drawn in that order.

(i) With replacement of balls

(ii) Without replacement of balls.

(Unit - IV, Prob. 7)

6. (a) Fit a straight line trend for the following series. Estimate the value for 1997.

Year	1990	1991	1992	1993	1994	1995	1995
Production of steel (in tonnes)	60	72	75	60	80	85	95

(Unit - V, Prob. 24)

OR

- (b) From the following data, calculate the regression equations taking deviation of items from the mean of X and Y series.

X	6	2	10	4	8
Y	9	11	5	8	7

(Unit - V, Prob. 12)

FACULTY OF MANAGEMENT
B.B.A I Year II Semester (CBCS) Examination
MODEL PAPER - II
BUSINESS STATISTICS

Time : 3 Hours]

[Max. Marks : 80

SECTION - A (5 × 4 = 20 M)

Answer any 5 Questions

ANSWERS

1.	(a) What is Principals of Measurement.	(Unit-I, SQA. 3)
	(b) State the merits and demerits of geometric mean.	(Unit-II, SQA. 5)
	(c) Discuss briefly about deciles.	(Unit-II, SQA. 8)
	(d) Define Base Shifting.	(Unit-III, SQA. 7)
	(e) Define deflating.	(Unit-III, SQA. 8)
	(f) Multiplication Theorem of Probability.	(Unit-IV, SQA. 6)
	(g) Spearman's Rank Correlation.	(Unit-V, SQA. 6)
	(h) Define Time Series.	(Unit-V, SQA. 11)

SECTION - B (5 × 12 = 60 M)

Answer all the Questions

2.	(a) What are the difference between Primary and Secondary Data ?	(Unit - I, Q.No. 18)

OR

	(b) What are the functions of statistics ?	(Unit - I, Q.No. 2)

3.	(a) Following is the distribution of marks obtained by 50 students in Mercantile law	

Marks more than	0	10	20	30	40	50
No. of Students	50	46	40	20	10	3

Calculate the Median marks.

(Unit - II, Prob. 27)

OR

	(b) Find the mode of the following data. 0, 1, 6, 7, 2, 3, 7, 6, 6, 2, 6, 0, 5, 6, 0	(Unit - II, Prob. 31)

4.	(a) Explain the various methods of constructing index numbers.	(Unit - III, Q.No. 6)

OR

	(b) In the construction of a certain cost of living number, the following group index numbers are found. Calculate the Cost of Living Index Number by using (i) Weighted Arithmetic Mean and (ii) Weighted Geometric Mean.	

Group	Index Numbers	Weights
Food	350	5
Fuel and Lighting	200	1
Clothing	240	1
House Rent	160	1
Miscellaneous	250	2

(Unit - III, Prob. 18)

5. (a) Explain the various Methods of Sampling.

(Unit - IV, Q.No. 22)

OR

- (b) A committee of 4 people is to be appointed from 3 officers of the production department, 4 officers of the purchase department, two officers of the sales department and one chartered accountant. Find the probability of forming the committee in the following manner.

(i) There must be one from each category

(ii) It should have at least one from the purchase department

(iii) The chartered accountant must be in the committee.

(Unit - IV, Prob. 5)

6. (a) What are the differences between correlation and regression analysis ?

(Unit - V, Q.No. 21)

OR

- (b) Given below is the data of production of a certain company in lakhs of units.

Year	1995	1996	1997	1998	1999	2000	2001
Production	15	14	18	20	17	24	27

Compute the Linear trend by the method of least squares. predict the production for the year "2004".

(Unit - V, Prob. 20)

FACULTY OF MANAGEMENT
B.B.A I Year II Semester (CBCS) Examination
MODEL PAPER - III
BUSINESS STATISTICS

Time : 3 Hours]

[Max. Marks : 80

SECTION - A ($5 \times 4 = 20$ M)

Answer any 5 Questions

ANSWERS

1. (a) What is tabulation ? (Unit-I, SQA. 9)
- (b) What are the properties of dispersion ? (Unit-II, SQA. 10)
- (c) What is quartile deviations ? (Unit-II, SQA. 13)
- (d) Consumer Price Index (Unit-III, SQA. 9)
- (e) Define Chain Index Numbers. (Unit-III, SQA. 6)
- (f) Define the term Sampling. (Unit-IV, SQA. 9)
- (g) Properties of Karl Pearson's Coefficient of Correlation (Unit-V, SQA. 5)
- (h) Define the term Regression. (Unit-V, SQA. 7)

SECTION - B ($5 \times 12 = 60$ M)

Answer all the Questions

2. (a) Explain the various methods of collecting secondary data. (Unit - I, Q.No. 1)
- OR
- (b) Explain different types of classification? (Unit - I, Q.No. 20)
 3. (a) The following table gives the marks obtained by a group of 80 students in an examination. Calculate the variance.

Marks obtained	No. of students
10 – 14	2
14 – 18	4
18 – 22	4
22 – 26	8
26 – 30	12
30 – 34	16
34 – 38	10
38 – 42	8
42 – 46	4
46 – 50	6
50 – 54	2
54 – 58	4

(Unit - II, Prob. 47)

OR

- (b) Define Dispersion. Explain the objectives of Measures of Dispersion. (Unit - II, Q.No. 34)
4. (a) What are the various Problems in the Construction of Index Numbers? (Unit - III, Q.No. 21)

OR

- (b) Compute Fisher's Index Number and apply both factor reversal test and time reversal test from the following data :

Commodity	Base year		Current year	
	Price	Expenditure	Price	Expenditure
	Rs.	Rs.	Rs.	Rs.
A	5	25	10	60
B	1	10	2	24
C	4	16	8	40
D	2	40	5	75

(Unit - III, Prob. 7)

5. (a) What are the various Principles of Sampling ? (Unit - IV, Q.No. 16)

OR

- (b) Define Standard Error (SE). What are the uses of Standard Error? (Unit - IV, Q.No. 28)
6. (a) What is Karl Pearson's Coefficient of Correlation? Explain properties of Coefficient of Correlation. (Unit - V, Q.No. 8)

OR

- (b) Calculate coefficient of correlation and find the probable error.

X	14	19	24	21	26	22	15	20	19
Y	31	36	48	37	50	45	33	41	39

(Unit - V, Prob. 1)