Topics relating to Statistics and Sampling for PC-2 of Subordinate Audit /Accounts Service (SAS) Examination

Collection of Data

Organization of Data

Presentation of Data

Measures of Central Tendencies-Average

Measures of Dispersion etc.

Correlation and Regression

Sampling techniques and Estimation

Highly useful guide to Statistics and Sampling in Audit with lot of Audit examples

Edited and complied by
Pawan Kumar Dhamija
Statistical Advisor

Preface to the book

The book has actually been compiled for the examinees of PC-2 of Subordinate Audit /Accounts Service (SAS) so most of the topics have been covered as per the syllabus of this Exam. Some topics not related to SAS Exam have also been added to make the book useful as a handbook on Statistics and Sampling for different types of Audits.

This book may however be very useful for all the officers of IA&AD to understand the basics of Statistics and apply Statistics and Sampling in the different kinds of Audits. Wherever possible, examples from Audit have been given alongwith the explanation of the various terms.

Though the solved examples/calculations may not be much relevant for the examination as such, sufficient solved examples and unsolved questions have been added at the end of each chapter to make the concept clear and to help apply Statistics and Sampling in Audit Work. Every chapter contains a large number of multiple choice questions to make it easy to revise the chapter and test the learning of the reader.

Pawan Dhamija
Statistical Advisor
O/o the CAG of India
October 2019

	Table of contents	
Para No.	Title	Page No
	Chapter I **Introduction to Statistics**	
1.1	Meaning of Statistics	1
1.2	Stages of Statistical Study and Statistical Tools	2
1.3	Descriptive and Inferential Statistics	3
1.4	Functions of Statistics	3
1.5	Importance of Statistics	4
1.6	Limitations of Statistics	5
	Exercise	7
	Chapter 2 **Collection of Data**	
2.1	Data and its Types	9
2.2	Sources of data	10
2.3	Variable and its Types	12
2.4	Methods of Collecting Primary Data	13
2.5	Main Sources of Errors in collection of data	14
2.6	Important points to be kept in mind while drafting the questionnaire	14
2.7	Two important sources of Secondary data	15
	Exercise	16
	Chapter 3 **Organization of Data**	
3.1	Classification of data	20
3.2	Purpose of Classification	20
3.3	Methods of Classification	20
3.4	Characteristics of good classification	21
3.5	Types of Classification	21
3.6	Frequency Distribution	21
3.7	Some Definitions	26
3.8	Processing Data (Normalisation and standardization of data)	27
3.9	Analysis of data	27
	Exercise	29

Chapter 1
Introduction to Statistics

Statistics is not a new discipline but is as old as the human activity itself. Its sphere of utility, however, has been increasing over the years. In the olden days, it was considered as a by-product of the administrative activity of the State thereby limiting its scope. The governments in those days used to keep records of population, birth, deaths, etc., for administrative purposes. Statistical methods are now widely used in various diversified fields such as agriculture, economics, sociology, business management, health, audit, etc.

1.1. Meaning of Statistics

The word 'statistics' has been used in the plural sense to refer to numerical statements of facts or data. On the other hand it is also used in the singular sense to refer to a subject of study like any other subject such as (mathematics, economics, etc.)

(a) ***Statistics Defined in Plural Sense:*** Statistics has been defined differently by different writers. According to Yule and Kendall statistics means "Quantitative data affected to a marked extent by multiplicity of causes."

A more comprehensive definition of statistics was given by Horace Secrist. According to him statistics means "Aggregate of facts affected to marked extent by 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 is quite comprehensive and points out the characteristics that numerical facts (data) must possess so that they may be called statistics. Let us discuss about these characteristics one by one.

Characteristics of Statistics defined in plural sense:

(i) ***Aggregate of facts***: Individual figures cannot be called statistics. They should form a part of aggregate of facts relating to any particular field of enquiry. For example, Ram's monthly income is Rs. 2,000. This is not a statistical statement. However, when we are given monthly income of 10 persons working in an organisation, they will be called statistics.

(ii) ***Affected by multiplicity of factors****:* There are several factors that affect a phenomenon. For instance, the consumption of a household on any item is affected by income, taste, education, etc. The data relating to such phenomenon can be called statistics. But if we write the numbers one to ten along with their squares, then these figures though more than one, cannot be called statistics. These figures are not affected by the various causes.

(iii) ***Numerically expressed***: Qualitative characteristics such as beauty, colour of eyes, etc., cannot be measured directly and hence, in general, they do not fall under the purview of statistics. We have

to quantify these characteristics before they become statistics. For example, in college we may count the number of girls having black eyes or blue eyes or brown eyes, then this is statistics.

(iv) ***Estimated according to a reasonable standard of accuracy:*** Statistics are either enumerated or estimated, but reasonable standards of accuracy must be maintained. Suppose, one is interested in understanding the average level of performance of the students, who take admission to B.Com class. For this purpose one must collect the marks obtained by the students at the senior secondary level. It may be done by complete enumeration of the marks of all the students or by taking a sample of students. On the basis of the result of the sample, one may estimate the average level of performance of all students. Thus, statistics may be obtained by enumeration or estimation

(v**)** ***Collected in a systematic manner for a predetermined purpose****:* The data should be collected in a systematic manner. The purpose for which data is collected, must be decided in advance. The purpose should be specific and well-defined.

(vi**)** ***Placed in relation to each other***: The numerical facts should be comparable if they are to be called statistics. For instance, statistics on production and export of an item during a year are related so when they are put together they constitute statistics. But if we have three figures: 1) production of rice in India in 1986 2) number of children born in USA in 1987, and 3) number of cars registered in UK in 1988. These figures may be facts alright, but taken together they cannot be called statistics as they have no relation among themselves.

It is thus clear that all statistics are numerical statements of facts but all numerical statements of facts are not statistics. They will be called statistics only if the above characteristics are present in them.

(b) *Statistics Defined in Singular Sense*: Statistics, when used in the singular sense, has been defined as a body of methods which provides tools for data collection, analysis and interpretation. Some of these definitions of statistics in singular sense are:

According to Croxton and Cowden, "Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data."

The definition given by Seligman is - "Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry."

Thus, we can conclude that the word 'statistics' may be used either in plural sense to refer to data or in singular sense to refer to a body of methods for making wise decisions in the face of uncertainty.

1.2. Stages of Statistical Study and Statistical Tools

Studying Statistics as a singular noun implies the knowledge of various stages of statistical study. At the first stage statistical data are collected. Second stage involves organization of data in systematic order. In the third stage the data are presented in the form of graphs, diagrams or tables. Fourth stage involves analysis of data in terms of averages, percentages, etc. Final stage contains interpretation of data to arrive at certain conclusions.

Each stage of statistical study involves the use of certain standard techniques or methods called statistical tools. Stages of Statistical Study and the Related Statistical Tools are listed as under:

Stages	**Statistical Study**	**Statistical Tools**
Stage I	Collection of Data	Census or Sample Techniques
Stage II	Organisation of Data	Classification of data, making data array or frequency tables
Stage III	Presentation of Data	Tables, Graphs and Diagrams
Stage IV	Analysis of Data	Calculation of Percentages, Averages, Dispersion, Correlation, Regression, etc.
Stage V	Interpretation of Data	Based on analysis of data finding degree of Relationship between different economic variables in order to arrive at certain conclusions about the variables under study.

1.3. Descriptive and Inferential Statistics

Descriptive Statistics refer to various measures that are used to describe the characteristic features of the data. Such measures include measures of central tendency, measures of dispersion, etc. Graphs, tables and charts that display data are also examples of descriptive statistics. We use descriptive statistics, while computing the average marks of a sample of 10 out of 40 students from the same class without attempting any generalisation about the entire class.

Inferential Statistics on the other hand refer to statistical process of drawing valid inferences about the characteristics of population data on the basis of sample data. If the average marks of the entire class are estimated on the basis of the sample average, we would say that we are using inferential statistics.

Thus, Descriptive statistics is the field of Study concerned with Collection, Organisation, Summarisation and Analysis of Data while Inferential Statistics in addition involves drawing of inference about a set of data (Population) when only a part of data (Sample) is observed/audited.

1.4. Functions of Statistics

Some of the important functions of statistics are:

a) ***To present facts in a proper form:*** Statistical methods present general statements in a precise and definite form. For example, saying that in India 30,000 persons died in road accidents last year is more convincing than saying that many persons died in road accidents last year.

b) ***To simplify bulky and complex data*****:** Statistical methods simplify bulky and complex data so that they could easily be understood. The raw data is often meaningless.

c) ***To facilitate comparison***: The primary purpose of statistics is to facilitate comparison of different phenomena over time or space. For instance, comparison of the national income over time gives an idea of the development of the country over the years. This comparison can be made among the various nations also to know where a country stands viz. a viz. the other countries so far as the growth of the economy is concerned.

d) ***Helps in formulation of policies***: Statistical methods are also useful in formulating various policies in social, economic, and business fields. The government for instance, uses enrolment data to know the progress of education in the various states. This helps in concentrating on the states where the enrolment is low. It also helps in changing the policy to achieve the targets.

e) ***To study relationship between different phenomena***: Statistical measures such as correlation and regression are used to study relationships between variables. Such relationships are important for making decisions. For instance, one may find a relationship between the increase in demand of the number of diesel vehicles and pollution level.

f) **To forecast future *values***: Some of the statistical techniques are used for forecasting future values of a variable. On the basis of weather data for a number of years, it is possible to predict the rainfall for a year.

g) ***To draw valid inferences*:** Statistical methods are also useful in drawing inferences regarding the characteristics of the universe (population) on the basis of sample data.

1.5. Importance of Statistics

In the ancient times statistics was used as the science of state affairs only. Data on a wide range of activities such as population, births and deaths were collected by the State for administrative purposes. However, in recent years, the scope of statistics has widened considerably to bring to its fold social and economic phenomena. Now-a-days statistics is applied in almost all spheres of human activity.

1.5.1 *Statistics and State*: In earlier times, the role of the State was confined to the maintenance of law and order. For that purpose, it used to collect data relating to manpower, crimes, income, etc., with the objective of formulating suitable military and fiscal policies. Now a days, developing countries such as India are following the policy of planned economic development. For that purpose the government must base its decisions/policies on correct analysis of statistical data. For instance, in formulating its five year plans, the government must have an idea about the availability of raw materials, capital goods, financial resources, the distribution of population etc., to evolve various policies.

1.5.2. *Statistics in Economics*: Statistical analysis helps in study of a number of economic problems such as production, consumption, distribution, etc. Data on prices, wages, consumption, savings and investment, etc., are vital in formulating various economic policies. Likewise data on national income are useful in formulating policies for reducing disparities of income. Index numbers, time series analysis, regression analysis, etc., are vital in economic planning. For example, the consumer price index is used for grant of dearness allowance (DA) or bonus to workers. Demand forecasting could be made by business firms using time series analysis. For testing various economic hypotheses also statistical data are used.

1.5.3. *Statistics in Business and Management*: With the growing size and increasing competition, the activities of modern business enterprises are becoming more complex. The success of the managerial decision-making depends upon the timely availability of relevant information much of which comes from statistical data. Statistical data have been increasingly used in business and industry in its various operations like sales, purchases, production, marketing, finance, etc. Statistical methods are now widely applied in market and production research, investment policies, quality control of manufactured products, economic forecasting, **auditing** and many other fields. One element common to all problems faced by managers is the need to take decisions under uncertainty; statistical methods provide techniques to deal with such situations.

1.5.4. *Statistics in Audit and Accounts*: In an age of Big Data, statistical analysis is becoming an increasingly powerful tool for accountants and auditors. Taking a course in introductory statistics will help accountants improve their efficiency and help their clients make better decisions.

(a) Auditors: Accountants who perform audits benefit greatly from understanding and using statistical analysis. For example, when conducting a **reliability assessment**, one of the accountant's first tasks is to gather evidence. Auditors know that the easiest way to do this is by looking at a portion of the whole, rather than gathering every bit of data available. Statistically representative samples are preferred in this area as they help auditors work more efficiently and objectively. Moreover the statistical skills enable an auditor to intelligently collect, organise and analyze data relevant to decision making and to make appropriate audit observations by helping them:

- To develop an appreciation about averages, variability, correlation, regression, etc.
- To make data into information. Data becomes information when it is used in arriving at a wise decision about what to audit, when to audit and how to audit.
- To develop understanding of ideas of statistical reliability/precision, probability, Risk/errors, etc.
- To use these ideas to develop a proper sampling design including taking decision about sample size and for drawing valid inferences based on sample.

***(b) Accountants*:** Accounting standards are front and center when managers determine retirement and other benefits. Accountants set premium adjustments to account for future risk and account for artificial fluctuations in short-term interest rates using statistical models and methods. Accountants and others use historical statistical data to develop policy recommendations to help control defined benefit plans and promote retirement security.

1.6. Limitations of Statistics

In modern times, Statistics has emerged to be one of the most useful subject in all walks of life. However, it has certain limitations. Following are some important limitations of Statistics:

(a) *Study of Numerical Facts only*: Statistics studies only such facts as can be expressed in numerical terms. It does not study qualitative phenomena like honesty, friendship, wisdom, health, patriotism, justice, etc.

(b) *Study of Aggregates only*: Statistics studies only the aggregates of quantitative facts. It does not study statistical facts relating to any particular unit. For e.g. it may be a statistical fact that the price of bread rose by 20% last year. But, as this fact relates to an individual item, it is not a subject matter of Statistics. However, it becomes a subject matter of Statistics if we study general increase in price level of major commodities in the form of index numbers.

(c) *Results are true only on an Average*: Most statistical findings are true only as averages. Unlike the laws of natural sciences, statistical laws are not exact. They are not always valid under all conditions. For instance, if per capita income in India is Rs. 1,00,000 per annum; it does not mean that the income of each and every Indian is Rs. 1,00,000 per annum. Some may have much more income and others may have very less.

(d) ***Without Reference, Results may prove to be wrong*****:** In order to understand the conclusions precisely, it is necessary that the circumstances and conditions under which these conclusions have been drawn are also studied. Otherwise, one may draw wrong conclusions. For e.g. in the business of steel, profits earned during three years are Rs. 5 lakh, Rs. 6 lakh and Rs. 7 lakh respectively. On the other hand, in the cement business profits earned during the same three years is Rs. 7 lakh, Rs. 6 lakh and Rs. 5 lakh respectively. Thus, the average profit in both the businesses comes out to Rs. 6 lakh per annum. It may lead to the conclusion that both the businesses have similar economic status, but it is not true. We may see that steel-business is making progress while cement-business is on the decline.

(e) ***Can be used only by Experts:*** Statistics can be used only by those persons who have special knowledge of statistical methods. Those who are ignorant about these methods cannot make sensible use of statistics. It can, therefore, be said that data in the hands of an unqualified person is like a medicine in the hands of a quack who may abuse it, leading to disastrous consequences. In the words of Yule and Kendall, "Statistical methods are most dangerous tools in the hand of an inexpert."

(f) ***Prone to Misuse*****:** Misuse of Statistics is very common. Statistics may be used to support a pre-drawn conclusion even when it is absolutely false. It is usually said, "Statistics are like clay by which you can make a god or a devil, as you please." Misuse of statistics is indeed its greatest limitation. The results obtained through Statistics can be manipulated according to one's own interest and such manipulated results can mislead the community. Suppose you are told that during a year the number of car accidents in a city by women drivers is 10 while those committed by men drivers is 40. On the basis of this information, you may conclude that women are safe drivers. However, you must also know the total number of drivers of both types before you could arrive at a correct conclusion.

Multiple choice Questions:

Q1. Which of the following is correct regarding Statistics?

a) Aggregate of facts
b) Numerically expressed
c) Affected by multiplicity of causes
d) All of these [Ans. (d)]

Q2. In singular sense Statistics means:

a) Statistics science
b) Statistical law
c) Both (a) and (b)
d) None of these [Ans. (c)]

Q3. The aggregate of data is called:

a) Statistics
b) Editing of data
c) Analysis of data
d) Collection of data [Ans. (a)]

Q4. Which of the following indicates a stage of statistical study?

a) Collection of data
b) Presentation of data
c) Analysis of data
d) All of these [Ans. (d)]

Q5. In plural sense, which of the following is not a characteristic of Statistics?

a) Aggregate of data
b) Only expressed in words
c) Affected by multiplicity of causes
d) Collected in a systematic manner [Ans. (b)]

Q6. In inferential statistics, we study

a) the methods to make decisions about population based on sample results
b) how to make decisions about mean, median, or mode
c) how a sample is obtained from a population
d) None of the above [Ans. (a)]

Q7. In descriptive statistics, we study

a) The description of decision making process
b) The methods for organizing, displaying, and describing data
c) How to describe the probability distribution
d) None of the above [Ans. (b)]

Answer the following Questions

Q1. Define Statistics as a singular noun and as a plural noun. [60 words]

Q2. Write three sentences highlighting the importance of Statistics in Economics. [60 words]

Q3. State the main limitations of Statistics. [4 points]

Q4. "Statistical methods are dangerous weapons in the hands of an unqualified person." Explain.

Q5. Define Statistics. What are its basic characteristics? [100 words]

Q6. Tabulate the main stages of Statistics and the related tools.

Q7. There are three kinds of lies-lies, damned lies and Statistics. Explain the statement. [100 words]

Q8. The government and policy makers use statistical data to formulate suitable policies of economic development'. Illustrate with two examples. [60 words]

Q9. Statistical methods are no substitute for common sense. Comment. [60 words]

Q10. Explain with examples the use of Statistics in Audit.

Chapter 2
Collection of Data

'Data' is the raw material in any statistical procedure. Once we have data, we may use a descriptive number, a table or a graph to summarize the same and tell the audience in clear terms what the data say.

2.1. Data and its Types

Data means (i) Raw Material of Statistics (ii) Set of related numbers. Examples of data are amounts mentioned on different vouchers of a department for a month, no. of vouchers examined/audited per day for 30 days by an Audit Officer or no. of Gram Panchayats (GPs) per District. Data can be categorized on different basis:

(a) ***Qualitative or Categorical data****:* The data which have no notion of magnitude or size of the characteristics as the same can't be measured are called Qualitative data. It has three types as under:

(i) ***In binary data*** the variable can take only two values like Sex (M/F), Error in voucher (Y/ N), value mentioned in voucher greater than Rupees 3 lakh (Y/N), etc.

(ii) ***In Nominal data*** a variable can take more than two values i.e. the information fits into more than two categories but the categories can't be ranked. It means nominal data can't be put into any meaningful order; for e.g. Name of the state, Name of Department, etc.

(iii) ***In Ordinal data*** a variable can take a number of values that can be ranked through some gradient/Scale; for e.g. Division in exam (First, Second, Third), Audit Risk of different Departments (low, moderate, high) of an organisation.

(b) ***Quantitative [Numerical] data***: are measures of values or counts and are expressed as numbers. It is of two kinds:

(i) ***Discrete data*** where the possible values of the variable are quite distinct and separated and normally values/magnitude have no decimals. For e.g. Number of vouchers received in a treasury in the various months, no. of schools audited in different Districts of a State.

(ii) ***Continuous data*** where the variable **c**an assume continuous/uninterrupted range of values and the values may have decimals. These data normally arise as a result of some type of measurement. For example values (in Rupees) mentioned on the vouchers audited, service tax paid by the various assesses of a Commissionerate, weight or height of a person.

2.1.2. ***Primary and Secondary Data*****:** Data is categorized as Primary and Secondary Data depending upon its source:

a) ***Primary Data***: Data collected originally by the investigator/researcher himself or by some agency on his/her behalf and used by him/her for statistical analysis. For example data collected in the Department through Beneficiary Survey.

b) ***Secondary Data***: Data already collected and possessed by some agency and taken over from there for analysis by the researcher. Most of the data used for Audit are the secondary data as it is not collected by the Department directly and is obtained from the Auditee or from other sources.

c) ***Considerations for choice between Primary and Secondary Data***: While choosing between Primary and Secondary Data, we consider: (i) Nature, objective and Scope of Enquiry/ Audit. (ii) Time and Finances available (iii) Degree of precision aimed (iv) Reliability of the agency which collected data.

d) ***Differences between Primary and Secondary Data:***

Primary Data	Secondary Data
(i) Data collected originally by the investigator/ auditor himself or by some agency on his/her behalf and used by him for statistical analysis. (ii) Originality: The Data collected by this method is original as it has been collected by the investigator/ researcher himself or on his behalf. (iii)Time/money: It involves more money and time in collecting original data. (iv) Editing of Data: It does not need much Editing as it has been collected keeping the purpose of enquiry in mind.	(i) Data already collected and possessed by some agency/person and taken over from there for analysis by the researcher/auditor. (ii) Originality: The Data collected by this method is not original as it has been collected by some other agency and taken over for analysis by the researcher. (iii) Time/money: It involves much less money and time since the data are already available. (iv) Editing of Data: It needs a lot of Editing as it has been taken over for analysis by the researcher from some source.

If we are planning to study an economic or social problem, we would require data on certain variables. For e.g. if we want to study, how the demand for a certain commodity reacts to change in its price, we would require data on quantity demanded of that commodity and its price, in several markets and at different points of time.

We may collect our own data by conducting market survey or an enquiry into household budgets; this would be the Primary Data. We may also use already available published or unpublished data; it would be called Secondary Data.

2.2. Sources of Data

2.2.1. ***Primary Source***: Primary data can be collected through questionnaires, interview, case studies, experimentation and observation.

2.2.2. ***Secondary Sources*****:** Secondary Sources consist of Published sources and Un-published sources; Secondary Sources can also be categorized as Internal and External Sources.

(a) ***Internal Sources of Data****:* These sources are available within the organization. If available, internal secondary data may be obtained with less time, effort and money than the external secondary data. In addition, they may also be more relevant to the situation at hand since they are from within the organization. The internal sources include sources like; Accounting resources; Sales Force Report; Views of Internal Experts and Miscellaneous Reports brought out by the organisation.

(b) ***External Sources of Data***: These sources are available outside the organization in a larger environment. Collection of external data is more difficult because the data have much greater

variety and the sources are much more numerous. External data can be divided into following classes.

Government Publications- Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data. These are:

(i) ***Registrar General of India-*** It is an office which generate demographic data. It includes details of gender, age, occupation, etc.

(ii) ***Central Statistical Organization-*** This organization publishes the national accounts statistics. It contains estimates of national income for several years, growth rate and rate of major economic activities. Annual survey of Industries is also published by the CSO. It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.

(iii) ***Director General of Commercial Intelligence-*** This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.

(iv) ***Ministry of Commerce and Industries-*** This ministry through the office of economic advisor provides information on wholesale price indices. These indices may be related to a number of sectors like food, fuel, power, food grains etc.

(v) ***Planning Commission-*** It provides the basic statistics of Indian Economy.

(vi) ***Reserve Bank of India-*** This provides information on Banking, Savings and investment. RBI also prepares currency and finance reports.

(vii) ***Labour Bureau-*** It provides information on skilled, unskilled, white collared jobs etc.

(viii) ***National Sample Survey Office -*** This is a part of Ministry of Statistics and PI and it provides social, economic, demographic, industrial and agricultural statistics based on regular Sample Surveys conducted by it.

(ix) ***Department of Economic Affairs-*** It conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.

(x) ***State Statistical Abstract-*** This gives information on various types of activities related to the state like - commercial activities, education, occupation etc.

Non-Government Publications- These includes publications of various industrial and trade associations, such as; Indian Cotton Mill Association, chambers of commerce, Bombay Stock Exchange, Confederation of Indian Industries (CII) and Small Industries Development Board of India

Unpublished Sources: The statistical data needn't always be published. There are various sources of unpublished statistical material such as the records maintained by private firms, business enterprises, scholars, research workers, etc. They may not like to release their data to any outside agency.

Other source**:** web-sites of the Ministries, Departments and other organisations as also those of various International Organisations like World Bank, WHO, UNICEF also provide huge volume of data for the various countries.

2.2.3. ***Precautions required while using secondary data***: The secondary data should not be used as such. It is risky to use statistics collected by others unless they have been properly scrutinized and found suitable, reliable and adequate. Before using the secondary data, an investigator should consider the points like; (i) Are the data reliable? (ii) Are the data suitable for the purpose of investigation? (iii) Are the data adequate? (iv) Are the data collected by the proper method? (v) Whether the data are unbiased? (vi) The period to which the data pertains (vii) Reliability of the agency which has collected the data.

2.3. Variable and its Types

A characteristic taking different values in different persons, places or things is called a variable. The variables can be of two types:

2.3.1. ***Quantitative variable***: A quantitative variable is the one that can be measured; for e.g. heights of individuals, values (in Rupees) mentioned on the different vouchers, service tax paid by the different assesses. The quantitative variable can take the forms:

(a) ***Discrete variable***: When the variable takes on a countable number of values. Most often these variables represent some kind of count such as the number of vouchers received in different treasuries in a month.

(b) ***Continuous variable***: When the variable can take any value in some range of values. Common examples would be height (inches), weight (pounds), or time to recovery (days).

2.3.2. ***Qualitative variable***: A qualitative variable is the one that can't be measured and convey information about category/attribute only. For e.g. no. of vouchers of different categories, no. of Gram Panchayats having more than one Secondary schools. Qualitative variables are often further classified as either:

(a) ***Nominal Variable*****:** A nominal variable is a categorical variable which has two or more categories without having any kind of natural order. They are variables with no numeric value. A pie chart displays data in categories with nominal variables. Examples of Nominal Variables are gender (Male, Female, and Transgender); Nationality (Indian, Pakistani, and Bangla Deshi); Blood Group (A, B, AB and O**).**

(b) ***Ordinal Variable*****:** On ordinal variable is a special type of categorical variable for which the levels can be naturally ordered. Some examples of ordinal variable are High school class rankings: 1st, 2nd, 3rd; Social economic classes: working, middle, upper; Risk of Audit (High, moderate or low).

Often categorical variables are disguised as quantitative variables. For example, one might record gender information coded as 0 = Male, 1 = Female. Still the variable is categorical; it is not naturally measured as a number.

In some cases it's tougher to make the distinction between quantitative and qualitative variable. An Auditor may collect survey data of the following nature:

How do you rank the department with an Audit Risk Perspective?

1	2	3	4	5
Very High	High	Moderate	Low	No risk

Technically the numbers are artificial. But, the Auditor will work with these numbers as though they had meaning. For instance, two people might respond "Very High" (1) and "Moderate" (3) for a department risk, the Auditor would, perhaps, compute an average of 1 and 3 which is 2; he may thus categorize the Department into high risk category.

2.4. Methods of Collecting Primary Data

Data help in reaching a sound conclusion by providing information. For statistical investigation, collection of data is the first and main step. Following are the Methods of Collecting Primary Data:

2.4.1. ***Direct Personal Investigation***: In this method data are collected personally by the investigator (organising agency) from the source concerned i.e. from the person about whom information is being collected. If the information is required for a student, it is obtained from him only.

Merits: (i) Reliable and accurate first-hand information; (ii) likely to have encouraging response

(iii) Extraction of exact information possible

Demerits: (i) Suitable for intensive (local) studies; (ii) personal biases may have adverse effect; (iii) Sensitive questions can't be asked (iv) investigator has to be intelligent to extract the exact information.

2.4.2. ***Indirect Oral Interviews***: In this method data are collected by the investigator from the persons who have the knowledge about the person/unit about whom/which the information is being collected. For e.g., if the information for a student is required, it may be collected from his/her teacher.

Merits: (i) Sensitive questions can be asked (ii) Compared to first method it is less expensive & time saving

Demerits: (i) Less reliable than the first method (ii) Investigator's biases are present

2.4.3. ***Information Received through Local Correspondents/agencies***: In this method local agents/ correspondents are appointed to collect information; normally used by newspapers/government to collect information about the various events. Used for reporting case(s) of a disease or death/birth in a village; collection of prices for construction of indices, etc.

Merits: (i) Suitable for extensive investigation. (ii) It is quite economical in terms of time and money.

Demerits: (i) Personal prejudice and bias of correspondents are present. (ii) The data collected by this method are not very accurate.

2.4.4. ***Mailed Questionnaire method*****:** A list of questions relating to the field of enquiry having space for answers to be filled by the respondents constitutes a questionnaire. List of Questions (Questionnaire) is prepared and sent to the informants with covering note by post with a request to supply the relevant information.

Merits: (i) Useful when the field of investigation is very large (ii) free from the biases of investigator (iii) an economic and expeditious method.

Demerits: (i) Can't be used if the informants are illiterate (ii) high degree of non-response (iii) difficult to verify the information so collected. (iii) There is a possibility of personal bias of the respondent.

2.4.5. ***Schedule Sent through Enumerators***: In this method enumerators get replies from the informants and fill in the schedule [Questionnaire filled by enumerator] in their own handwriting.

Merits: (i) Can be used even when the informants are illiterate (ii) very little non-response (iii) information can be checked by supplementary questions.

Demerits: (i) It is a costly and time consuming method (ii) It requires proper training of the enumerators. (iii) Investigator bias may be present.

2.5. Main Sources of Errors in Collection of Data

Following are the sources of errors while collecting the data:

(i) Errors related to the measurement of objects which may occur when: (a) the scales of measurement are different for different enumerators, and (b) different enumerators allow different degree of approximation in their measurements, even while using identical scales.

(ii) Errors due to wrong responses simply because the respondents are not able to handle/understand the questions precisely.

(iii) Some respondents may not respond; it introduces an element of error known as error due to non-response.

(iv) Errors due to miscalculations i.e. arithmetical errors.

(v) Errors due to 'communication gap' or due to lack of proper recording of the information.

2.6. Important points to be kept in mind while drafting the questionnaire

Following are the desired/requisite characteristics of a good questionnaire:

(i) Covering Letter giving objective, assuring confidentiality, inducements like gifts should be enclosed.

(ii) Questions should be (a) Small in no. (b) Logically arranged (c) short & simple (d) not require calculations.

(iii) Questions of Sensitive nature should be avoided.

(iv) Clarifications in the form of foot-notes may be given wherever necessary

(v) Instructions for informants may be a part of the questionnaire, if required

(vi) Objective questions with options are better than the descriptive questions.

(vii) For large studies, the questionnaire should be pre-tested by conducting pilot survey.

(viii) Cross-checks can be introduced to check the correctness of responses.

2.6.*1.* *Pilot Survey*: A pre-test or a guiding survey known as pilot survey is conducted on a small scale before the main survey. This is done to test the questionnaire and the field conditions.

2.7. Two important Sources of Secondary Data

Two important sources of data on Social and Economic parameters are 'Census of India' and 'Reports & Publications of 'National Sample Survey Office'

2.7.1. ***Census of India***: Census of India is a decennial (conducted every 10^{th} year) activity of the Government of India. It is conducted by the Registrar General & Census Commissioner, India. It is a very comprehensive source of secondary data. It collects information about population size and the various aspects of demographic changes in India. Broadly, it includes statistical information on the following parameters:

(i) Size, growth rate and distribution of population;

(ii) Population projections;

(iii) Density of population;

(iv) Sex composition of population;

(v) State of literacy;

(vi) Nature of Employment.

Information on these parameters relates to country as a whole as well as different states and union territories of the country. As the name suggests, Census of India coves each and every household of the country.

2.7.2 ***Reports and Publications of National Sample Survey Office (NSSO):*** Reports and publications of NSSO is another important source of secondary data in India. NSSO is a government organization under the Ministry of Statistics and Programme Implementation. This organization conducts regular sample surveys to collect basic statistical information relating to a variety of economic activities in rural as well as urban parts of the country. Broadly, reports and publications of NSSO contain statistical information of the following parameters of economic change:

(i) Land and Livestock Holdings;

(ii) Housing Conditions and Migration with special emphasis on slum dwellers;

(iii) Employment and Unemployment status in India;

(iv) Level and pattern of Consumer Expenditure of diverse categories of the people.

Unlike Census of India, Reports and Publications of National Sample Survey Office are based on Sample Surveys.

Multiple choice Questions:

Q1. Data collected for the first time from the source of origin is called:

(a) Primary data

(b) Secondary data

(c) Internal data

(d) None of these [Ans. (a)]

Q2. Which of the following methods is used when an investigator collects the required information from the persons who have the knowledge about the person about whom the information is being collected?

(a) Direct Personal Investigation

(b) Indirect Oral investigation

(c) Mailing method

(d) Enumerator's Method [Ans. (b)]

Q3. Which of the following is a source of secondary data?

(a) Government publication

(b) Private publication

(c) Report published by the State Bank of India

(d) All of these. [Ans. (d)]

Q4. Which of the following is a method of primary data collection?

(a) Direct personal investigation

(b) Indirect oral investigation

(c) Collection of information through questionnaire

(d) All of these [Ans. (d)]

Q5. Reports on quality control, production and financial accounts issued by companies are considered as

(a) external secondary data sources

(b) internal secondary data sources

(c) external primary data sources

(d) internal primary data sources [Ans. (b)]

Q6. Government and non-government publications are considered as

(a) external secondary data sources

(b) internal secondary data sources

(c) external primary data sources

(d) internal primary data sources [Ans. (a)]

Q7. Data which is generated within company such as routine business activities is classified as

(a) external primary data sources
(b) internal primary data sources
(c) external secondary data sources
(d) internal secondary data sources [Ans. (d)]

Q8. Type of questions included in questionnaire to record responses in which respondent can answer in any way are classified as

(a) multiple choices questions
(b) itemized question
(c) open ended questions
(d) close ended questions [Ans. (c)]

Q9. What are secondary data?

(a) Unimportant data
(b) Ordinary data
(c) Existing data
(d) Ordinal data [Ans. (c)]

Q10. Secondary data are LEAST helpful to

(a) formulate hypotheses
(b) develop questionnaires
(c) evaluate new products
(d) interpret tables [Ans. (c)]

Q11. Which ONE is a disadvantage of secondary data?

(a) Inexpensive.
(b) Fast to obtain.
(c) Addresses a fresh topic
(d) Already exist [Ans. (c)]

Q12. Which ONE is an advantage of secondary data?

(a) May be outdated
(b) Expensive
(c) May not be accurate
(d) Inexpensive [Ans. (d)]

Q13. Which of these is not a method of data collection.

(a) Questionnaires
(b) Interviews
(c) Experiments
(d) Observations [Ans. (c)]

Q14. Secondary/existing data may include which of the following?

(a) Official documents
(b) Personal documents
(c) Archived research data
(d) All of the above [Ans. (d)]

Q15. Which of the following terms best describes data that were originally collected at an earlier time by a different person for a different purpose?

(a) Primary data
(b) Secondary data
(c) Experimental data
(d) Field Data [Ans. (b)]

Q16. Which ONE of these methods is the fastest way to collect data?

(a) Online
(b) Personal
(c) Phone
(d) Postal [Ans. (a)]

Q17. Which ONE of these methods is the most expensive way to collect data per respondent?

(a) Online
(b) Personal
(c) Phone
(d) Postal [Ans. (b)]

Q18. Which ONE of these methods has the highest response rate?

(a) Online
(b) Personal
(c) Phone
(d) Postal [Ans. (b)]

Answer in one line

Q1. Give two examples of Govt. Publication of secondary data.

Q2. What do you mean by Internal Sources of data? Are they really useful in Audit?

Q3. What is a pilot survey?

Q4. What is Primary data?

Q5. What do you mean by data?

Q6. Define a questionnaire.

Answer in 80-100 words

Q1. Write in brief about the sources of secondary data.

Q2. Differentiate between primary and secondary data on the basis of originality, time & money needed and editing.

Q3. What are the desired/requisite characteristics of a good questionnaire?

Q4. Mention the various methods of collection of primary data. Explain any one of them.

Q5. Write a short note on NSSO.

Q6. What are the precautions required while using secondary data?

Q7. State 3 merits and 3 demerits of collecting data by 'Personal Interview". (3)

Q8. Distinguish between 'quantitative' and 'qualitative' data, and give some examples of both.

Chapter 3
Organisation of Data

The Collected Data in unorganised form is called raw data; it is organised or classified for making it useful for comparison, further analysis, presentation, etc.

3.1. Classification of Data: Classification means arranging the data into different classes or groups on the basis of their similarities. All similar items of data are put in one class and all dissimilar items of data are put in different classes. Statistical data is classified according to its characteristics. For example, if we have collected data regarding the number of students admitted in a university in a year; we can classify them on the basis of gender, Division in the last exam, Marital Status, etc. The set of characteristics we choose for the classification of the data depends upon the objective of the study. For example, if we want to study the educational background of the students, we classify them on the basis of their Division in the last exam.

3.2 Purpose of Classification:

Classification helps in achieving the following objectives:

a) It helps in presenting the data in a concise and simple form.
b) It divides the voluminous data on the basis of similarities and dissimilarities so as to enable a comparison.
c) It is a process of presenting raw data in a systematic manner enabling us to draw meaningful conclusions.
d) It provides a basis for tabulation and analysis of data.
e) It provides us a meaningful pattern in the data and enables us to identify the possible characteristics in the data.
f) It also helps in bringing out relationship between the variables and to present a mental picture of the information conveyed by the data.

3.3 Methods of Classification:

Classification can be done according to attributes or according to variables.

3.3.1. ***Classification According to Attributes***: An attribute is a qualitative characteristic which cannot be expressed numerically. Only the presence or absence of an attribute can be known. For example intelligence, religion, caste, sex, etc., are attributes. These can't be quantified. When classification is to be done on the basis of attributes, groups are differentiated either by the presence or absence of the attribute (e.g. school with or without girl's toilet) or by its differing qualities. Based on difference in quality, we can determine the group into which a particular item is placed. For instance, if we select audit risk as the basis of classification, there will be departments with high, moderate or low risk. There are two types of classification based on attributes.

a) ***Simple Classification***: In simple classification the data are classified on the basis of only one attribute. The data classified on the basis of sex or blood group are examples of simple classification.

b) ***Manifold Classification***: In this classification the data are classified on the basis of more than one attributes. For example, the data relating to the number of students in a university can be classified on the basis of their sex and marital status simultaneously.

3.3.2. Classification According to Variables: Variables refer to quantifiable characteristics that can be expressed numerically. Examples of variables are wages, age, height, weight, marks, distance, etc. All these variables can be expressed in quantitative terms. In this form of classification, the data are shown in the form of a **frequency distribution**. A frequency distribution is a tabular presentation that generally organises data into quantifiable classes and shows the number of observations (frequencies) falling into each of these classes.

3.4. Characteristics of good classification

Following are the general guiding principles for good classifications

(i) ***Exhaustive:*** Classification should be exhaustive. It means each and every item in data must belong to one of the classes. Introduction of residual class (i.e. other, miscellaneous, etc.) should be avoided.

(ii) ***Mutually exclusive***: The classification should be such that each item can be placed in one and only one class.

(iii) ***Suitability:*** The classification should confirm to object of inquiry i.e. the purpose for which it is done.

(iv) ***Stability***: Only one principle must be maintained throughout the classification and analysis.

(v) ***Homogeneity***: The items included in each class must be homogeneous i.e. similar to each other.

(vi) ***Flexibility***: A good classification should be flexible enough to accommodate new situation or changed situations.

(vii) ***Arithmetic accuracy***: A good classification should have arithmetical accuracy wherever numerical/quantitative data are used.

3.5 Types of Classification

a) According to time i.e. **chronological**; for e.g. number of assesses over different years during 2012-13 to 2018-19.

b) According to area i.e. **geographical** for e.g. number of Gram Panchayats audited in the different states or different districts.

c) According to attribute i.e. **qualitative** for e.g. number of schools having separate girl's toilets.

d) According to magnitude of variables i.e. **quantitative** i.e. classification of service tax assesses as per amount of tax paid in a year.

3.6. Frequency Distribution

Series of individual observations where the data items are listed one after the other without any order is called raw data. For better understanding, these observations could be arranged in ascending or descending

order. Then it is called an **array**. For example marks obtained by 10 students out of 25 marks are given under: 12, 17, 06, 15, 11, 00, 18, 11, 07, 10.

The above data list gives raw data. The presentation of data in above form doesn't reveal much information. Array for the above data becomes: 00, 06, 07, 10, 11, 11, 12, 15, 17, 18.

When large number of observations are available on a single characteristic, often it is useful to condense data without losing any information of interest. Frequency Distribution is one such method of condensing data. A Frequency Distribution is a tabular statement with two columns; first column describes the variable category/values and the second column represents the frequencies (i.e. the number of times the variable is taking a particular value). There can be un-grouped or grouped frequency distributions (or frequency distribution tables). Some of the graphs that can be used with frequency distributions are histograms, bar charts and pie charts. Frequency distributions can be used for **both qualitative and quantitative data**.

A frequency distribution can be classified as **a)** Series of individual observations **b)** Discrete frequency distribution **c)** Continuous frequency distribution.

3.6.1. *Construction of Frequency Tables:*

(a) ***Frequency Distribution of a Discrete Variable****:* Since, a discrete variable can take some or discrete values, it will be natural to take a separate class for each distinct value of the discrete variable.

To construct a frequency table or frequency distribution of a discrete variable, we use tally marking method. According to this method, for each observation in the data set we mark a vertical bar (|) called tally mark in the frequency table. A running tally is kept till the last observation; every fifth tally is a crossed tally. Thus **Tally Marking Method** is used for constructing a frequency distribution table. After counting the tallies against an observation, we can find the frequencies, relative frequency, cumulative frequency etc. as required.

Let us consider the following example relating to the weekly number of car accidents during 30 weeks.
3, 4, 4, 5, 5, 3, 4, 3, 5, 7, 6, 4, 4, 3, 4, 5, 5, 5, 5, 5, 3, 5, 6, 4, 5, 4, 4, 6, 5, 6.

Table: Frequency distribution for weekly number of car accidents.

Number of car accidents	**Tally Marks**	**No. of weeks/Frequency**
3	𝍸	5
4	𝍸 ||||	9
5	𝍸 𝍸 1	11
6	||||	4
7	1	1
Total		30

(b) ***Frequency Distribution of a Continuous Variable*****:** Continuous data series is one where the measurements are only approximations and are expressed in class intervals within certain limits. All the

classes taken together must cover at least the lowest value to the highest value in the data set. Though equal class intervals are preferred in frequency distribution, unequal class intervals may be necessary in certain situations to avoid a large number of empty or almost empty classes. Following are the steps for construction of a frequency table:

1. Normally 7 - 10 class intervals are used for freq. table, we may however use the formula: k = 1 + 3.322*log N to decide the number of classes, where N = number of items/observations in the data set and k number of classes.
2. Decide the individual class limits and select a suitable starting point of the first class which is arbitrary, it must be less than or equal to the minimum value.
3. Take an observation and mark a vertical bar (|) called **tally mark** against the class to which it belongs. A running tally is kept till the last observation; every fifth tally is a crossed tally. Thus **Tally Marking Method** is used for constructing a frequency distribution table.
4. By counting the tallies, find frequencies, relative frequency or cumulative frequency as required.

Some guidelines that should be followed while dividing continuous data into classes are as follows:

1. The classes should be mutually exclusive, i.e., non-overlapping. No two classes should have a common value.
2. The classes should be exhaustive, i.e., they must cover the entire range of the data.
3. The number of classes and the width of each class should neither be too small nor too large.

Let us construct a frequency table of daily maximum temperatures in ^{0}C in a city for 50 days:

28, 28, 31, 29, 35, 33, 28, 31, 34, 29, 25, 27, 29, 33, 30, 31, 32, 26, 26, 21, 21, 20, 22, 24, 28, 30, 34, 33, 35, 29, 23, 21, 20, 19, 19, 18, 19, 17, 20, 19, 18, 18, 19, 27, 17, 18, 20, 21, 18, 19

Minimum Value= 17; Maximum Value=35; Range= Max. Value – Min. Value = 35-17 = 18

Number of classes = 6 (say)

So width of each class = 18/6 = 3

Table: Frequency distribution of temperature in a city for 50 days.

Class Intervals (Temperatures in $^\circ C$)	**Tally Marks**	No. of days (Frequency)
17-20	𝍸 𝍸 lll	13
20-23	𝍸 llll	9
23-26	lll	3
26-29	𝍸 lll	8
29-32	𝍸 llll	9
32-35	𝍸 l	6
35-38	ll	2
Total		**50**

(c) *Inclusive and Exclusive series*: Continuous (grouped) frequency distribution can take two forms as:

Inclusive Series

Class Intervals	Frequency(f)
1-3	15
4-6	11
7-9	14
10-12	10
Total	**50**

Exclusive Series

Class Intervals	Frequency(f)
1-4	15
4-7	11
7-10	14
10-13	10
Total	**50**

Differences between inclusive and exclusive series: (i) In exclusive series the upper limit of the class interval is not included while in inclusive series it is included. (ii) In Exclusive series upper limit of a class and the lower limit of the next class are the same but is it not in the case of inclusive series.

The Freq. Distribution with classes of the type: 1-4, 4-7, 7-10, etc. is Exclusive Frequency Distribution (in class 1-4, 1 is included but 4 is not) and the Freq. Distribution with classes of the type: 1-3 (both 1 and 3 included), 4-6, 7- 9 is Inclusive Frequency Distribution.

(d) *Univariate, Bivariate and Multivariate frequency tables*: Based on the number of variables used, there are three categories of frequency distributions:

(i) *Uni-variate Frequency Distribution*: The frequency distribution with one variable is called a uni-variate frequency distribution. For example, the students in a class may be classified on the basis of marks obtained by them.

(ii) *Bi-variate Frequency Distribution*: The frequency table which describes the values of two variables simultaneously is known as bi-variate frequency distribution/table. For example classifying students on the basis of age and height.

(iii) *Multi-variate Frequency Distribution*: The frequency distribution with more than two variables is called multivariate frequency distribution. For example, the students in a class may be classified on the basis of marks, age and sex.

Univariate frequency tables

Rank	Degree of Risk	Number of Units
1	Very High	20
2	High	30
3	Moderate	20
4	Low	15
5	Very low	15
	Total	**100**

Joint or Bivariate Frequency Distribution: Bivariate joint frequency distributions are often presented as two-way table:

Two-way table

	Dance	Sports	TV	Total
Men	2	10	8	20
Women	16	6	8	30
Total	18	16	16	50

(e) ***Applications of Frequency Tables:*** Managing and operating data present in frequency distribution tables is much simpler than operation on raw data. Moreover, there are simple algorithms to calculate median, mean, standard deviation etc. from these tables.

(f) ***Cumulative Frequency Table*****:** Cumulative frequency is defined as a running total of frequencies. Cumulative frequency can also be defined as the sum of all previous frequencies up to the current point. The cumulative frequencies are important when analyzing data and calculation of partitioning values like median, Quartiles, Percentiles, etc. The cumulative frequency is also useful when representing data using diagrams like ogives.

The cumulative frequency is usually expressed by constructing a cumulative frequency table. The cumulative frequency table takes two forms as explained below:

Frequency Table		Less than Cumulative Frequency Table		More than Cumulative Frequency Table	
Age (years)	**No. of persons**	**Age (years)**	**Cumulative Frequency/No. of persons**	**Age (years)**	**Cumulative Frequency/No. of persons**
00-05	3	Less than 5	3	More than 00	67+3 = 80
05-10	18	Less than 10	3+18 = 21	More than 05	59+18 = 77
10-15	13	Less than 15	21+13 = 34	More than 10	46+13 = 59
15-20	12	Less than 20	34+12 = 46	More than 15	34+12 = 46
20-25	7	Less than 25	46+7 = 53	More than 20	27+7 = 34
25 -30	27	Less than 30	53+27 = 80	More than 25	27

3.7. Some Definitions:

(i) *Relative Frequency:* Relative frequency of a particular value of the variable or a class of values of the variable is obtained by dividing the frequency corresponding to that particular value or that particular class by the total number of observations in the data set, i.e., the total frequency.

Relative Frequency = $\frac{frequency\ or\ class\ frequency}{total\ frequency}$

Relative frequency of any value or any class lies between 0 and 1. We calculate relative frequency if we want an idea about the relative importance of the particular value or class in relation to the total frequency.

(ii) *Class limits and real class limits (Class boundaries):* Class limits are the two end-points of a class interval which are used for the construction of a frequency distribution. These are called lower class limit and upper class limit of that class interval. These are not the real limits or end-points of a class interval.

For example, the real class intervals for the inclusive frequency distribution 17-20, 21-24, 25-28, 29-32, 33-36 will be 16.5 - 20.5; 20.5 – 24.5, 24.5 – 28.5, 28.5 – 32.5, 32.5 – 36.5. Thus, real class intervals or class boundaries are obtained by subtracting half from the lower limits and adding half to the upper limits for inclusive series. For calculation of Median, Mode or partition values like quartiles, percentiles or decile; it is necessary to convert class limits into real class limits for inclusive frequency distribution. Likewise for construction of histogram real class intervals are needed.

(iii) *Open-end classes*: It may be so that some values in the data set are extremely small compared to the other values of the data set or some values are extremely large in comparison. Then we do not use the lower limit of the first class and the upper limit of the last class. Such classes where one of the class limits (lower or upper) is not fixed are called open end classes.

(iv) *Class width*: The length of the class is called the class width. It is also known as class size.

Class width = Upper Class Boundary (real limit) - Lower Class Boundary (real limit)

(v) *Class mark*: The midpoint of a class interval is called class mark. It is the representative value of the entire class.

Class mark = $\frac{Lower\ class\ Limit + Upper\ Class\ Limit}{2}$

(vi) *Frequency Density*: It is the frequency per unit width of the class. It is given by:

Frequency Density of a class = $\frac{frequency\ of\ a\ class}{width\ of\ a\ class}$

Frequency densities are essential to compare two classes of unequal width.

3.8. Preprocessing Data (Normalisation and standardization of Data)

Before Data processing, Data preprocessing plays a crucial role. It usually involves normalization and standardization of data. This step is very important when dealing with parameters of different units and scales. For e.g. to calculate Human Development Index the parameters used are income, health and education for which the units as also the scales are different. So these parameters are combined using the normalization of data. Two methods are usually well known for rescaling/preprocessing data.

Normalization scales all numeric variables in the range [0, 1]. One of the formulae commonly used for normalization of data is:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization transforms the data set to have zero mean and unit variance, for example using the equation below:

$$x_{new} = \frac{x - \mu}{\sigma}$$

Both of these techniques have their drawbacks. If you have outliers in your data set, normalizing your data will certainly scale the "normal" data to a very small interval and may not give any indication of outliers. When using standardization, your new data aren't bounded (unlike normalization) and go upto infinite.

3.9. Analysis of data

Analysis of Data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making. In other words analysis of data means moving from **data to Information; Information to Facts and Facts to Knowledge.** The various steps of Data Analysis are explained as under:

***Data requirements*:** The data necessary as inputs to the Audit are specified based upon the objectives of the audit.

***Data collection*:** Data are collected from a variety of sources. The requirements may be communicated by auditor to custodians of the data in the Auditee Department; data from other available sources like National sample Survey Office (NSSO) and Census may also be collected and used by the auditor.

***Data processing*:** Data initially obtained must be processed or organized for analysis. For instance, this may involve placing data into rows and columns in a table format for further analysis.

***Data cleaning*:** Once processed and organized, the data may be incomplete, contain duplicates or contain errors. Data cleaning is the process of preventing and correcting these errors. Common tasks in data cleaning include record matching and de-duplication. Quantitative data methods for outlier detection can be used to get rid of likely incorrectly entered data. For textual data spellcheckers can be used to lessen the amount of mistyped words.

3.9.*1 Exploratory data analysis*: Analysts may apply a variety of techniques to understand the messages contained in the data. Descriptive statistics such as measures of averages or variation may be calculated to help understand the data. Data visualization may also be used to examine the data in graphical format so as to obtain additional insight regarding the messages within the data.

Author Stephen Few described eight types of quantitative messages that users may communicate from a set of data and the associated graphs. These are:

1. ***Time-series***: A single variable is captured over a period of time, such as the unemployment rate over a 10-year period. A line chart may be used to demonstrate the trend.
2. ***Ranking:*** Categorical data are ranked in ascending or descending order, such as a ranking of sales performance of different salesmen (during some period). A bar chart may be used to show ranking.
3. ***Part-to-whole***: Categorical subdivisions are measured as a ratio to the whole (i.e., a percentage out of 100%). A pie chart or sub divided bar chart can show the comparison of ratios, such as the market share represented by competitors in a market.
4. ***Deviation***: Categorical subdivisions are compared against a reference, such as a comparison of actual vs. budget expenses for several departments of a business for a given time period. A multiple bar chart can show comparison of the actual versus the reference amount.
5. ***Frequency distribution***: Shows the number of observations of a particular variable for given interval, such as the number of years in which the stock market return is between intervals such as 0-10%, 11-20%, etc. A histogram, may be used for exhibiting frequency distribution.
6. ***Correlation:*** Comparison between observations represented by two variables (X, Y) to determine if they tend to move in the same or opposite directions. For example, plotting unemployment (X) and inflation (Y) for a sample of months. A scatter plot is typically used for correlation.
7. ***Nominal comparison***: Comparing categorical subdivisions in no particular order, such as the sales volume by product code. A bar chart may be used for this comparison.
8. ***Geographic or geospatial comparison***: Comparison of a variable across geographical areas such as the unemployment rate by states, percentage of women delivered in institutions in different Districts of a state, etc.; a cartogram is a typical graphic used here.

Multiple choice questions: choose the correct answer

1. Which of the following is the objective of classification?
(a) Simplification
(b) Briefness
(c) Comparability
(d) All of these [Ans. (d)]

2. Classification of data on the basis of time period is called:
(a) Geographical classification
(b) Chronological classification
(c) Qualitative classification
(d) Quantitative classification [Ans. (b)]

3. A series in which every class internal excludes items corresponding to its upper limit is called
(a) Exclusive series
(b) Inclusive series
(c) Continuous Series
(d) None of these [Ans. (a)]

4. In a series, the number of times an item occurs is known as:
(a) Number
(b) Class frequency
(c) Frequency
(d) Cumulative frequency [Ans. (c)]

5. The class mid-point is equal to:
(a) The average of the upper class limit and the lower class limit
(b) The product of upper class limit and the lower class limit
(c) The ratio of the upper class limit to the lower class limit
(d) None of the above. [Ans. (a)]

6. Under inclusive method:
(a) The upper class limit of a class is excluded in the class interval
(b) Both the upper and lower class limits of a class are excluded in the class interval
(c) The lower class limit of a class is excluded in the class interval
(d) None of these [Ans. (d)]

7. Type of cumulative frequency distribution in which frequencies are added from top to bottom order is classified as

(a) variation distribution
(b) less than type distribution
(c) more than type distribution
(d) marginal distribution [Ans. (b)]

8. Type of cumulative frequency distribution in which frequencies are added from bottom to top order is classified as

(a) more than type distribution
(b) marginal distribution
(c) variation distribution
(d) less than type distribution [Ans. (a)]

9. 'Less than type distribution' and 'more than type distribution' are types of

(a) class distribution
(b) cumulative class distribution
(c) cumulative frequency distribution
(d) upper limit distribution [Ans. (c)]

10. Class frequency is divided by total number of observations in frequency distribution to convert it into

(a) relative margin distribution
(b) relative variable distribution
(c) relative frequency distribution
(d) cumulative frequency distribution [Ans. (c)]

11. Total of frequency up to an upper class limit or boundary is known as

(a) average frequency
(b) cumulative frequency
(c) frequency distribution
(d) frequency polygon [Ans. (b)]

12. Table which shows frequency of each score is called a

(a) cumulative frequency table
(b) Pi table
(c) Histogram
(d) frequency distribution table [Ans. (d)]

13. Number of times each value appears in a frequency table is called _____ of the value

(a) range
(b) mode
(c) frequency
(d) standard Deviation [Ans. (c)]

14. A frequency distribution is

(a) a tabular summary of a set of data showing the relative frequency
(b) a graphical form of representing data
(c) a tabular summary of a set of data showing the frequency of items in each of several non-overlapping classes
(d) a graphical device for presenting qualitative data [Ans. (c)]

15. The relative frequency of a class is computed by

(a) dividing the midpoint of the class by the sample size
(b) dividing the frequency of the class by the midpoint
(c) dividing the sample size by the frequency of the class
(d) dividing the frequency of the class by the sample size [Ans. (d)]

16. The sum of frequencies for all classes will always equal

(a) 1
(b) the number of elements in a data set
(c) the number of classes
(d) a value between 0 and 1 [Ans. (b)]

17. If several frequency distributions are constructed from the same data set, the distribution with the widest class width will have the

(a) fewest classes
(b) most classes
(c) same number of classes as the other distributions since all are constructed from the same data
(d) Question is incorrect [Ans. (a)]

18. The sum of the percent frequencies for all classes will always equal

(a) one
(b) the number of classes
(c) the number of items in the sample
(d) 100 [Ans. (d)]

19. In constructing a frequency distribution, as the number of classes are decreased, the class width

(a) decreases
(b) remains unchanged
(c) increases
(d) can increase or decrease depending on the data values [Ans. (c)]

20. The difference between the lower class limits of adjacent classes provides the

(a) number of classes
(b) class limits
(c) class midpoint
(d) class width [Ans. (d)]

21. In a cumulative frequency distribution, the last class will always have a cumulative frequency equal to

(a) One
(b) 100%
(c) the total number of elements in the data set
(d) Question is incorrect [Ans. (c)]

TRY

Q1. What is meant by classification of data? State its objectives.

Q2. What are the four main merits of classification?

Q3. The following data gives age of 25 students of a Class. Arrange these data in the form of a frequency table.

15	16	16	17	18	18	17	15	15	16	16	17	15
16	15	16	16	18	15	17	17	18	10	16	15	

Q4. Weight of 20 students is given in kilograms. Using class interval of 4, make a frequency distribution table.

30	45	26	25	42	33	15	35	45	45
45	39	42	40	18	35	41	20	36	48

Q5. Convert the following data in a simple frequency distribution:

12 students obtained less than 10 marks
18 students obtained less than 20 marks
25 students obtained less than 30 marks
40 students obtained less than 40 marks

Q6. Prepare a frequency table with class intervals such that their mid-values are 17, 22, 27 and 32:

32	41	30	47	42	48	14	17	51	44	25	41
36	27	42	36	28	28	37	54	42	31	36	40
30	22	30	31	21	48	16	41	30	21	22	40
37	40	19	16	17	36	33	42	46	54	52	53

Q7. Explain the 'exclusive' and 'inclusive' methods used in classification of data.

Q8. Prepare **(a)** an Ungrouped **(b)** Grouped frequency distribution table using class intervals (i) 1-4, 5-8, 9-12 etc. (ii) 0-4, 4-8, 8-12 etc. for the following data: 12, 4, 11, 5, 7, 4, 3, 6, 3, 8, 12, 9, 2, 5, 1, 4, 11, 2, 9, 8, 5, 2, 7, 4, 3, 12, 7, 8, 2, 10, 9, 1, 8, 6, 10, 7, 3, 2, 1, 11, 5, 8, 1, 2, 6, 8, 3, 9, 12, 10

Q9. Define classification. Convert the following information into ordinary frequency table/series: (4)

(i) 5 students get less than 3 marks (ii) 12 students get less than 6 marks

(ii) 25 students get less than 9 marks (iv) 30 students get less than 12 marks

Q10. The following data are the weights in kilograms of a group of 55 students. Prepare a frequency table taking the magnitude of each class interval at 4 kg and the first class as 40-44.

72, 74, 40, 60, 82, 96, 41, 61, 75, 88, 63, 53, 62, 77, 63, 65, 95, 69, 64, 88, 79, 54, 73, 59, 61, 60, 66, 42, 64, 69, 70, 80, 72, 50, 79, 52, 73, 86, 96, 76, 94, 60, 76, 59, 61, 84, 77, 65, 50, 90, 69, 67, 84, 68, 76.

Q11. Represent the following data in the form of a discrete (ungrouped) Frequency Distribution table.

12, 4, 11, 5, 7, 4, 3, 6, 3, 8, 12, 9, 2, 5, 1, 4, 11, 2, 9, 8, 5, 2, 7, 4, 3, 12, 7, 8, 2, 10, 9, 1, 8, 6, 10, 7, 3, 2, 1, 11, 5, 8, 1, 2, 6, 8, 3, 9, 12, 10.

Q12. The following data show the number of hours worked by 40 statistics students. Refer this table and answer the following questions:

Number of Hours	Frequency
10-19	05
20-29	08
30-39	17
40-49	10

(i) The class width for this distribution is

(a) 9

(b) 10

(c) 11

(d) varies from class to class [Ans. (b)]

(ii) The number of students working 29 hours or less is

(a) 05

(b) 04

(c) 13

(d) cannot be determined without the original data [Ans. (c)]

(iii) The relative frequency of students working 39 hours or less is

(a) 3/4

(b) 2/4

(c) ¼

(d) cannot be determined without the original data [Ans. (a)]

(iv) The cumulative relative frequency for the class of 30 – 39 is

(a) 13

(b) 30

(c) 30/40

(d) cannot be determined without the original data [Ans. (b)]

(v) The class mark of the class 30-39 is

(a) 35

(b) 34

(c) 34.5

(d) 39 [Ans. (c)]

Chapter 4
Presentation of Data using Tables

The tabular presentation of data is one of the techniques of presentation of data, the two other techniques being diagrammatic presentation and graphic presentation.

4.1. Tabulation

The tabular presentation means arranging the collected data in an orderly manner in rows and columns. The horizontal arrangement of the data is known as **rows,** whereas the vertical arrangement **columns.** The tabulation simplifies presentation of data for the purpose of analysis and statistical inferences.

4.2. Major Objectives of Tabulation

(i) ***To simplify the complex data***: Huge data can be presented in a concise and simple manner by means of statistical tables; tabulation helps in presenting the data in an orderly manner.

(ii) ***To Facilitate Comparison of Data***: Data in raw form is difficult to compare. Comparison becomes possible when the related data are presented in tabular form. Tabulation facilitates the comparison of the various aspects of the data.

(iii) ***Useful in further analysis***: Presentation of data in tabular form provides a basis for analysis of data. A systematic presentation of data in tabular form is a precondition for the analysis of data.

(iv) ***Exhibits Trend of Data***: By looking at statistical tables, one can identify the overall pattern of the data i.e. how the data are moving with respect to time.

(v) ***To economise the space***: When the huge data are presented through tables, it not only makes the data comprehensible, it saves space too.

4.3. Differences between Classification and Tabulation

Classification and tabulation appear to convey the same meaning and also serve the same objectives. However, there is a difference between the two. In classification, the data is divided on the basis of similarity and resemblance, whereas tabulation is the process of recording the information in rows and columns. Tabulation begins where classification ends. In fact, classification provides a basis for tabular presentation.

4.4. Classification of tables

Depending upon the use and objectives of the data to be presented, there are different types of statistical tables. They can be classified under the following broad heads:

4.4.1. *On the basis of Coverage (Simple and complex table)*

(a) ***Simple Tables***: Simple tables are also known as **one way tables**. These tables are prepared on the basis of only one characteristic of the collected data. The table showing the data relating to the number of students in a college in different years is an example of simple or one way table.

Table 4.1: Details of interest earned

Sl. No.	Name of the State	Amount (Rs. Crore)
1	Bihar	54.46
2	Haryana	5.80
3	Himachal Pradesh	0.21
4	Jharkhand	1.60
5	Madhya Pradesh	26.55
6	Meghalaya	2.72
Source: Report No. 36 of 2015 of CAG of India- Performance Audit of Mid-Day Meal Scheme		

(b) *Complex Tables*: When a table shows more than one characteristic of the data, it is called a complex table. We may have a two-fold table showing two characteristics or a many-fold table showing several characteristics of the data. The table showing the number of students in a college on the basis of gender and marital status during different years is an example of a complex table.

Table 4.2: Age and Sex wise no. of patients in a hospital (in 00) during 2017

Age/Sex	Male	Female	All
0-15	7	2	9
15-30	10	5	15
30-45	8	4	12
45 and above	6	3	9
Total	31	14	45

Table: 4.3 Rates of cooking Cost (Amount in Rs.)

Period	Primary Level				Upper Primary Level			
	Non NER States		NER states		Non NER States		NER states	
	Centre	State	Centre	State	Centre	State	Centre	State
From Apr. 2011	2.17	0.72	2.60	0.29	3.25	1.09	3.91	0.4
From July 2012	2.33	0.78	2.80	0.31	3.49	1.16	4.19	0.46
From July 2013	2.51	0.83	3.01	0.33	3.75	1.25	4.50	0.50
Source: Report No. 36 of 2015 of CAG of India - Performance Audit of Mid-Day Meal Scheme								

4.4.2. General or Specific Purpose Tables: On the basis of purpose the tables are classified as:

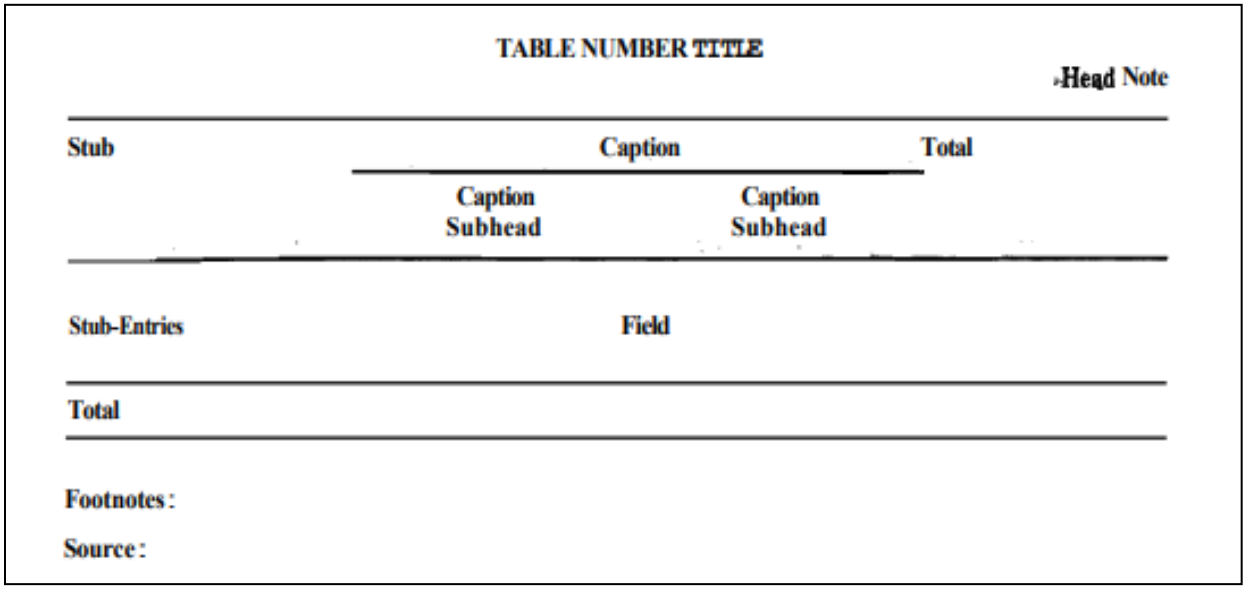
(a) ***General Purpose or Reference Tables:*** This type of table contains wide range of information relating to a specified subject. Such tables are complex tables and are generally found as appendices to various reports. They should be prepared in a systematic manner so as to render references easier. The tables appended to the various Performance Audit Reports may be called general purpose or reference tables. Primarily the sole purpose of a reference table is to present data in such a manner that individual items may be found readily by a reader.

(b) ***Special Purpose or Summary Tables*****:** These tables show a specific point relating to data and are helpful in statistical analysis. They provide a basis for comparison by indicating specific answers to given questions. These tables indicate rates, percentages, averages, etc. For instance table indicating the no. of deaths due to diseases like cholera, typhoid or malaria over some specified period or table indicating the number of vouchers received in a treasury during different months of a year.

4.5. Parts of a Statistical Table

A statistical table, in general, should have the following parts:

a) ***Title:*** There should be a title at the top of every statistical table. The title should be clear, concise and adequate. It should clearly indicate the description of the data being presented in the table. The title should



indicate what, where and when.

b) ***Table Number***: Every table is identified by a number. It facilitates easy reference. It can be placed at the beginning of the title of the table or can be centred above the title of the table.

c) ***Head Note***: Head note is written just below the title, preferably on the right hand corner. It normally indicates the units of data.

d) ***Stub or row title***: Stubs clearly describe the data presented in the rows of the table.

e) ***Caption or column title***. The caption labels the data presented in a column of the table. There may be sub-heads or sub-captions in each caption.

f) ***Body or Field***: The body of the table is the most important part. The information given in the rows and columns forms the body of the table. It contains the quantitative information to be presented.

g) ***Footnote***: Any explanatory notes about the table placed beneath the table, is called 'footnote'. The main purpose of footnote is to clarify some of the specific items given in the table or to explain the ambiguities, omissions, if any, about the data shown in the table. It may also disclose the sources of Secondary data or the abbreviations used.

h) ***Totals***: The totals and sub-totals of all rows and columns should also be given in a table.

4.6. Essential rules for creating tables or Requisites of a Good Statistical Table:

1. Keep them simple. Two or three small tables are better than a single large table.
2. All tables should be self-explanatory.
3. Titles should be clear and concise telling what, when and where.
4. Rows and columns should be clearly labelled/named.
5. Unit of measurement should be clearly stated.
6. Codes, abbreviations and symbols; wherever used, should be footnoted.
7. Totals should be shown for quantitative/numerical data.
8. If data are not original, source(s) should be footnoted.

Multiple choice questions: choose the correct answer

1. Which of these are the component (s) of a table:

(a) Table number

(b) Title

(c) Head note

(d) All of these Ans. (d)]

2. The process of systematic arrangement of data into rows and columns is called:

(a) Bar chart

(b) Classification

(c) Tabulation

(d) None of these [Ans. (c)]

3. The process of presentation of data in the form of a table is called:

(a) Organisation of Data

(b) Classification of Data

(c) Tabulation of Data

(d) None of these [Ans. (c)]

4. Which of the following are the titles of the rows of a table:

(a) Title

(b) Stub

(c) Caption

(d) None of these [Ans. (b)]

5. Which of the following are the titles of the columns of a table:

(a) Title

(b) Stub

(c) Caption

(d) None of these [Ans. (c)]

TRY

Q1. Describe the major parts of statistical table. Draw a format of a table showing all these parts.

Q2. In 2016-17, the contribution of primary, secondary and tertiary sectors to India's GDP was 17.6%, 28.2% and 54.2% respectively. In 2015-16 these shares were 17.7%, 27.0% and 55.3% respectively. This information is based on NAS 2017-18. Present this information in the form of a table.

Q3. Prepare a blank table showing **the marks, age and sex** of the students of a college in 2018. The mark groups should be taken as 0-10, 10-20, 20-30, 30-40 and 40-50 whereas the age should be taken in years as 17, 18**,** 19 and 20.

Q4. The data on manufacture of cars by different manufacturers in India showed that in 2015-16; the share of Maruti Udyog was 38.4 %, Hyundai 18.5 % and other cars 43.1%. In 206-17, the share of Tata Motors dropped to 18% whereas it went up to 48.2% for Maruti Udyog. Present this data in a tabular form.

Q5. Prepare a suitable table from the following information: In Bombay 80% of the total population is tea drinkers. Out of which 62% are males and 18 % are females. Rest are non-tea-drinkers out of which 12% are females. Give a suitable title also.

Q6. What is a Statistical table? Describe the essential parts of a table.

Q7. Write 4 features of a good table.

Q8. Explain the precautions to be observed while constructing a good table.

Chapter 5
Diagrammatic and Graphical Presentation

Presentation of data in tabular form facilitates comparison as it presents mass of data in simple and orderly manner. It is easier to establish trend and patterns when the data is in tabular form. Besides presenting in tabular form, the data can also be presented in the form of diagrams and graphs. The presentation of **data** in the form of diagrams and graphs is also called **visual presentation** of data. Compared to tabular presentation, data presented in diagrams and figures is more impressive and it is easier to draw conclusions based on graphical presentation.

5.1. Rules for constructing graphs

Different from a table, graphs are much more impressive to the reader or audience. The essential rules of creating graphs are almost similar to those for creating tables. These are:

1. Keep them simple and do not try to put in too much information.
2. Every graph should be self- explanatory.
3. There should be an appropriate proportion (normally 2:3) between width and height.
4. Equal quantities should be represented by equal intervals for e.g. value 10 should be twice as long as value 5.
5. Footnotes and source notes should be given wherever required.
6. Legends for axes/variables/attributes must be given wherever needed.

5.2. TYPES OF DIAGRAMS

Diagrams are generally classified on the basis of the number of dimensions viz. length, breadth and height they contain. Broadly, diagrams are classified as: one dimensional, two dimensional and three dimensional diagrams. Besides these diagrams, the data can also be presented in the form of maps and pictographs. However, we shall discuss one and two dimensional diagrams only.

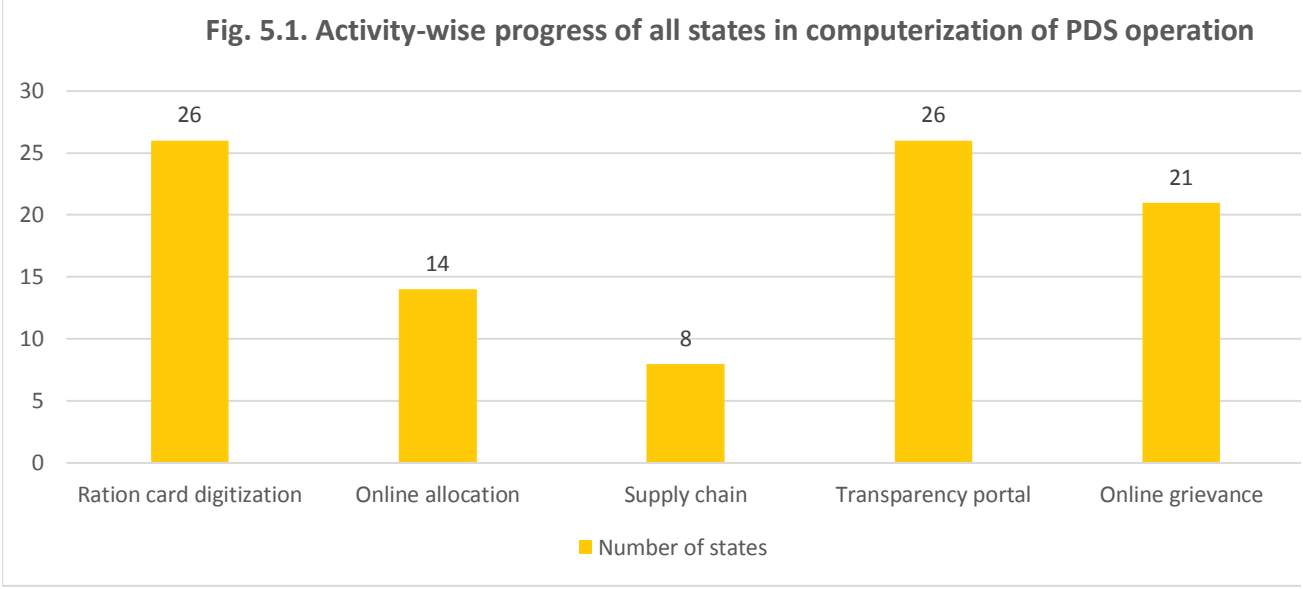
One Dimensional Diagram: **A** diagram prepared on the basis of just one dimension length or breadth is known as one dimensional diagram. One dimensional diagrams can be in the form of simple bar diagrams, multiple bar diagrams, sub-divided bar diagrams and percentage sub-divided bar diagrams.

Two Dimensional Diagram: **A** diagram prepared on the basis of two of the three dimensions viz. length, width and height is known as two dimensional diagram. Two dimensional diagrams can be of the form rectangles, sub-divided rectangles, squares and circles/pie diagrams.

Three Dimensional Diagram: **A** diagram prepared on the basis of three dimensions i.e. length, width and height is known as three dimensional diagram. They are also known as volume diagrams; they consist of cubes, cylinders, spheres, etc. It is difficult to interpret such diagrams so they are not recommended for statistical presentation.

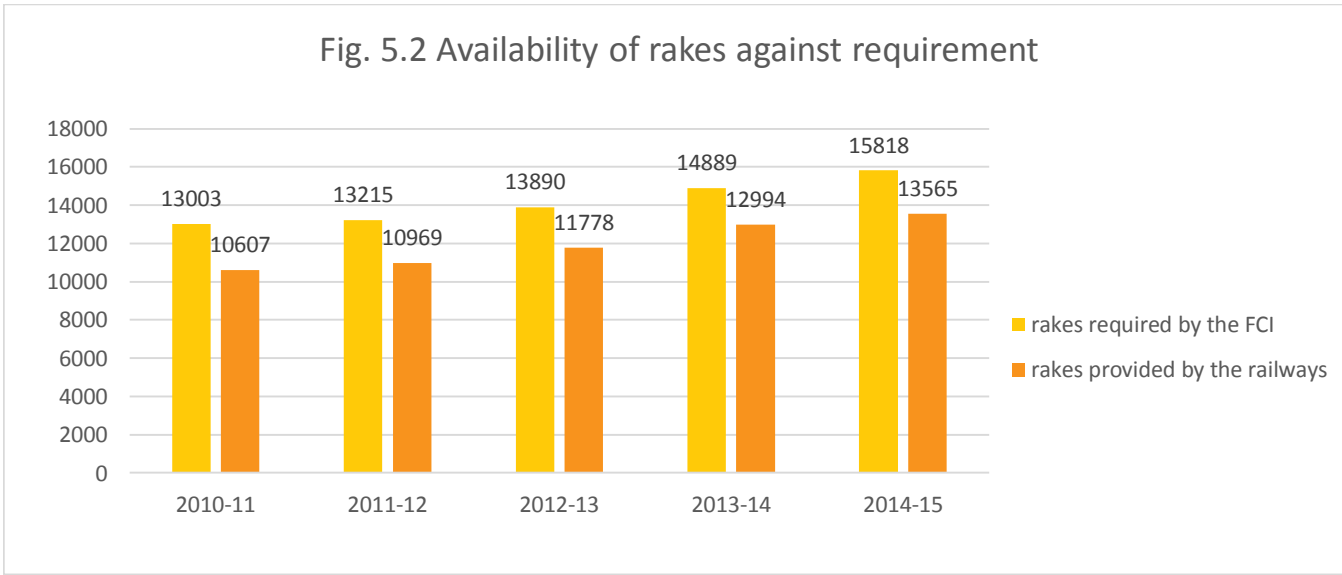
5.2.1. ***Types of Bar Diagrams:*** A bar graph/diagram is a one dimensional diagram; it is used to show absolute or relative frequency of two or more categories of a qualitative variable. Bar graph can be drawn in the form of horizontal or vertical bars; they can also be in the form of single (Simple), grouped (multiple), or stacked (subdivided) bars.

(i) ***A Simple Bar diagram***



Source: Report no. 54 of CAG of India: 2015- Preparedness for Implementation of National Food Security Act, 2013 for the year ended March, 2015

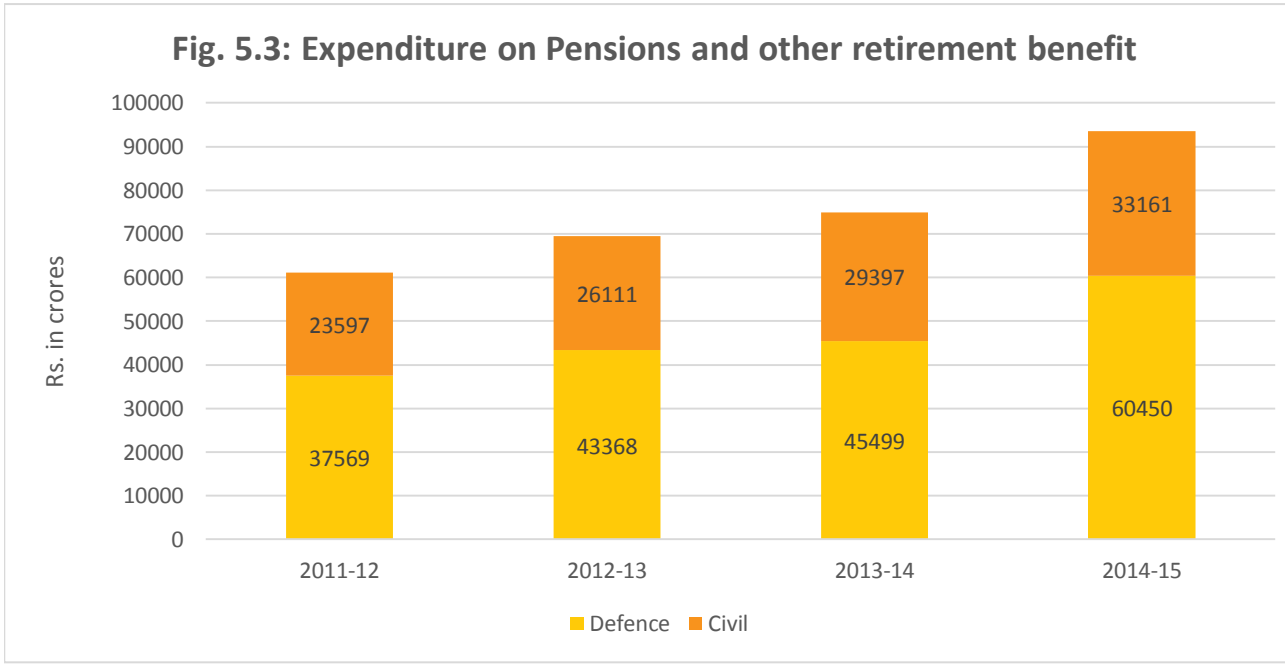
(ii) ***Multiple bar Diagram*****:** In this type of diagram two or more bars are constructed adjoining each other. These bars normally represent two or more related variables or value of the same variable over different



Source: Report no. 54 of CAG of India: 2015 - Preparedness for Implementation of National Food Security Act, 2013 for the year ended March, 2015

times or places. The bars are constructed adjoining each other and identical gap is left in between the bars of different variables. Just like simple bar diagrams, the length of the various bars varies in the ratio of the magnitude of the given-values. The width of the different bars is kept same. This diagram, facilitates comparison of the values of different variables in a set or comparison of the values of the same variable over a period of time.

iii) *Sub-divided or Stacked Bar (100%) Diagram*: This type of bar diagram is prepared to represent the different components of the same variable. It is also called **component diagram**. In this diagram one bar is constructed for the total value of the variable, and then the bar is sub-divided in proportion to the values of the various components of that variable. Actually, the values of the different components are cumulated for constructing this bar diagram and the bar is to be sub-divided at these cumulated points. It can be drawn for percentages of the component values of the variable instead of the actual values.



Source: Report 50 of CAG of India: 2015 - Accounts of Union Government (Financial Audit)

5.2.2. *A pie graph*: A pie graph is a Two Dimensional Diagram. It is another way to present data graphically in the form of a circle. The circle or pie (360^0) represents 100% population which is divided into sections/segments in accordance with the magnitude of each category of qualitative or quantitative data. Thus a pie graph shows comparison of relative frequencies of each category of data. Sector angles for a pie diagram are calculated by the formula:

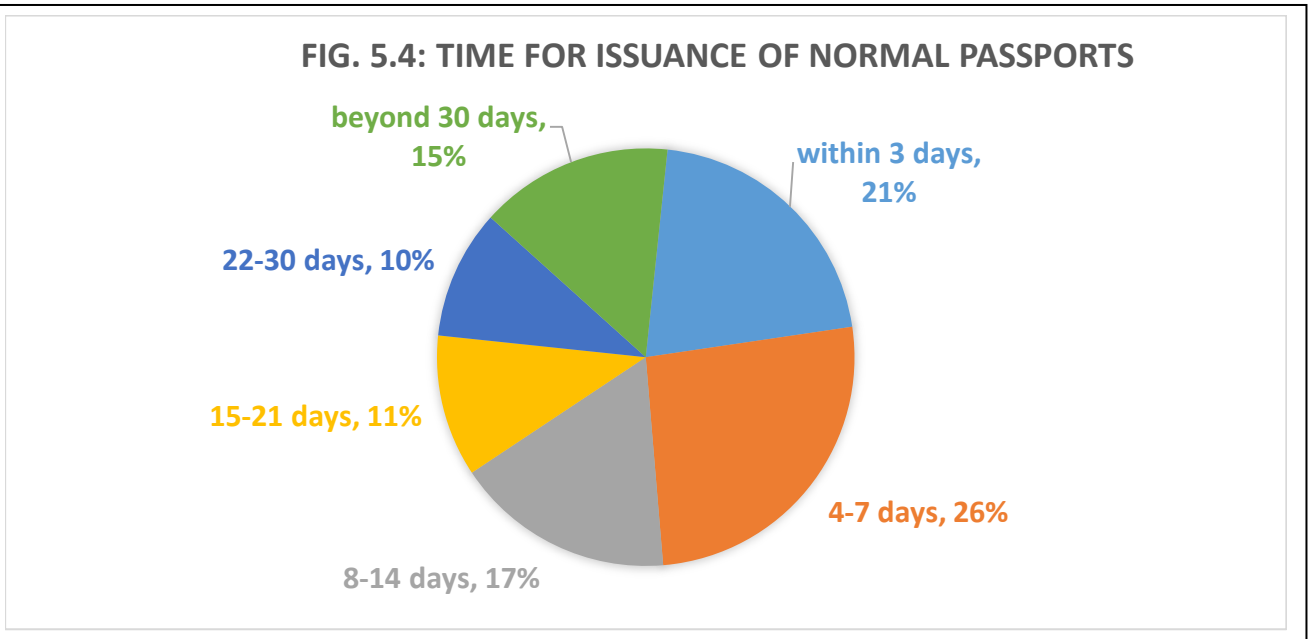
$$\text{Sector angle} = \frac{\text{Class frequency}}{\text{Total frequency}} * 360$$

When percentages are given instead of the absolute values, the sector angle is calculated as:

Sector angle = given percentage *3.6

Limitations of Pie Chart: If there are many segments of a similar size, the chart is hard to interpret and is confusing. For example, a chart with four slices is easy to read; one with more than 10 becomes confusing, especially if it contains many similar sized slices. The confusion increases as actual data are not shown. It can overemphasis large values. It may lead readers to draw inaccurate conclusions. Moreover, it is difficult and time consuming to draw.

Other charts and graphs may be a better option, especially if you are handling many pieces of data or want to make comparisons between data sets. Doughnut charts share the circular shape and overall functionality of pie charts but add the ability to display multiple data sets. Bar graphs represent data by length, allowing for quick comparison and measurement. They may be easier to read if you need to present many pieces of data at a time or want to compare different sets of data in a single chart.



Source: Report 7 of CAG of India of 2016 PA on Implementation of Passport Seva Project

5.3. Difference between Diagrammatic and Graphic Presentation

The frequencies of various characteristics can be presented by Graphs and Diagrams. Following are the differences between graphs and diagrams.

(i) There is no clear cut distinction between the two;

(ii) For Constructing Graph we generally use Graph Paper;

(iii) A Graph represents a mathematical relationship between the two variables;

(iv) Diagrams are more attractive to the eye and as such are better suited for publicity and propaganda;

(v) For representing Frequency Distributions and Time Series Data graphs are more useful.

5.4. False Base line

One of the fundamental rules while constructing graphs is that the scale on the Y-axis should begin from zero. Where the lowest value to be plotted on the Y scale is relatively high and a detailed scale is required to bring out the variations in all the data, starting the Y scale with zero introduces difficulties. For example, if we have a series of production figures over a number of years ranging from 15000 units to 25000 units, then starting with a zero origin would have one of two undesirable consequences: either (i) the necessarily large intervals (say 5000 units) on the Y scale would make us lose sight of the extent of fluctuations in the curve : (ii) a necessarily large graph to permit small intervals (say 1000 units) would entail a waste of a large part of the graph, in addition to poor visual communication.

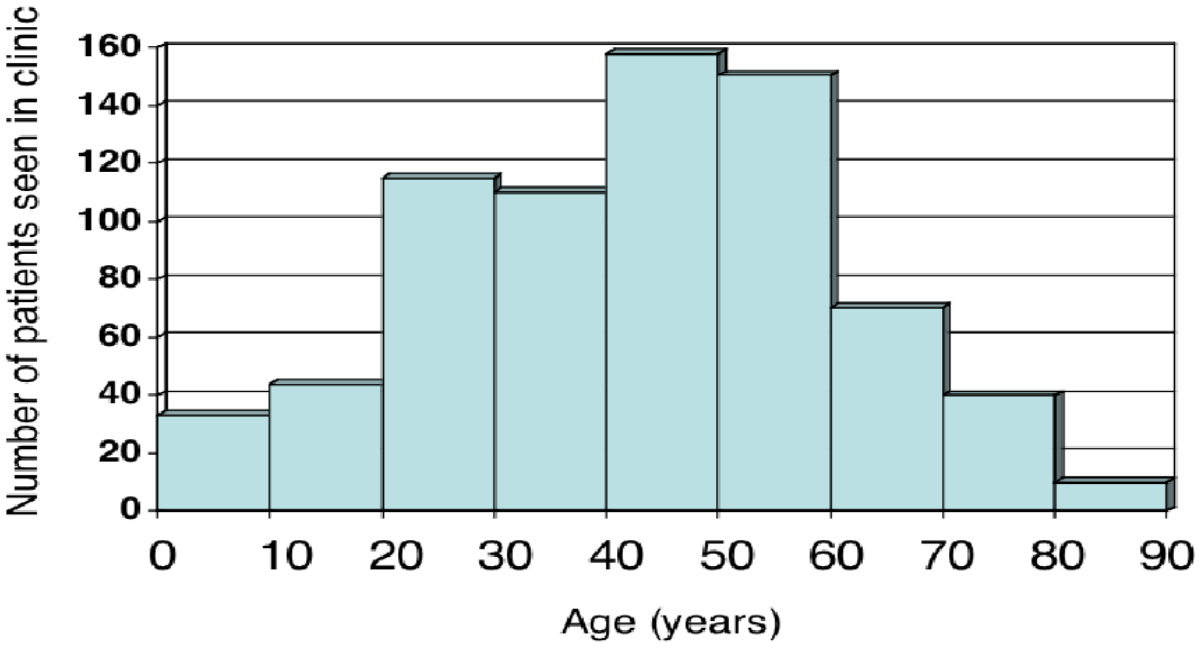
The solution is to break the Y scale. If the zero origin is shown then the scale is broken by drawing a horizontal wavy line (also called kinked or zig-zag line) or a vertical wavy line between zero and the first unit on the Y scale which in our illustration would be 15000 units. These lines are drawn to make the reader aware of the fact that false base has been used. Three important objects of false base line are:

1. Variations in the data are clearly shown
2. A large part of the graph is not wasted or space is saved by using false base.
3. The graph provides a better visual communication.

5.5. Different types of Graphs

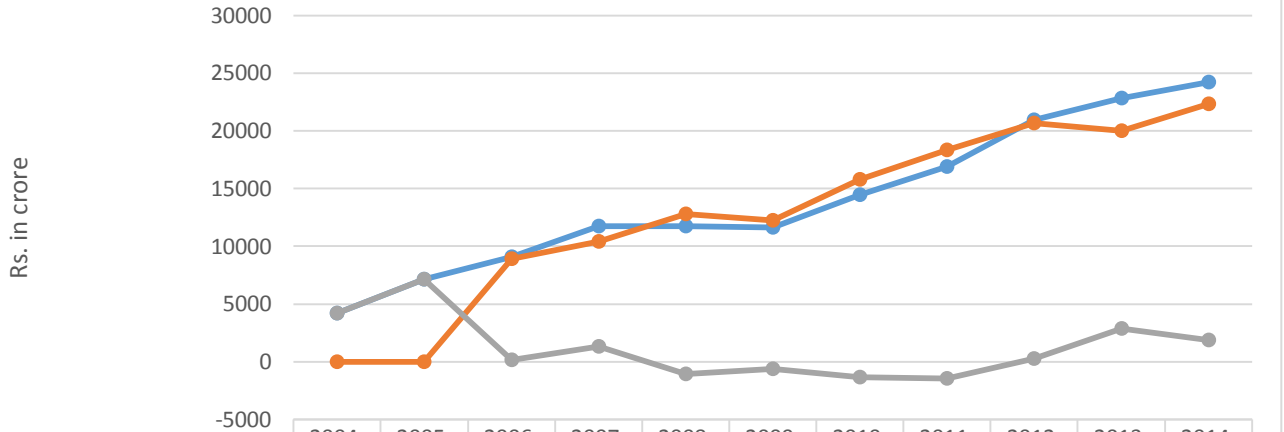
(a) ***A histogram*** is graphical presentation of a frequency distribution. The variable presented in a histogram must be a **continuous quantitative variable**. So data like number of vouchers of different value categories, time intervals for different activities, etc. which are presented in continuous frequency table, can also be represented by drawing a histogram.

Fig. 5.5: Age Group-wise no. of patients seen in a clink



(b) ***A line graph*** is particularly useful if we want to show the change of more than one **quantity over time**. Thus, it is normally used to indicate **time trend** in the given data.

Fig. 5.6: Collection of primary education cess and its transfer to PSK

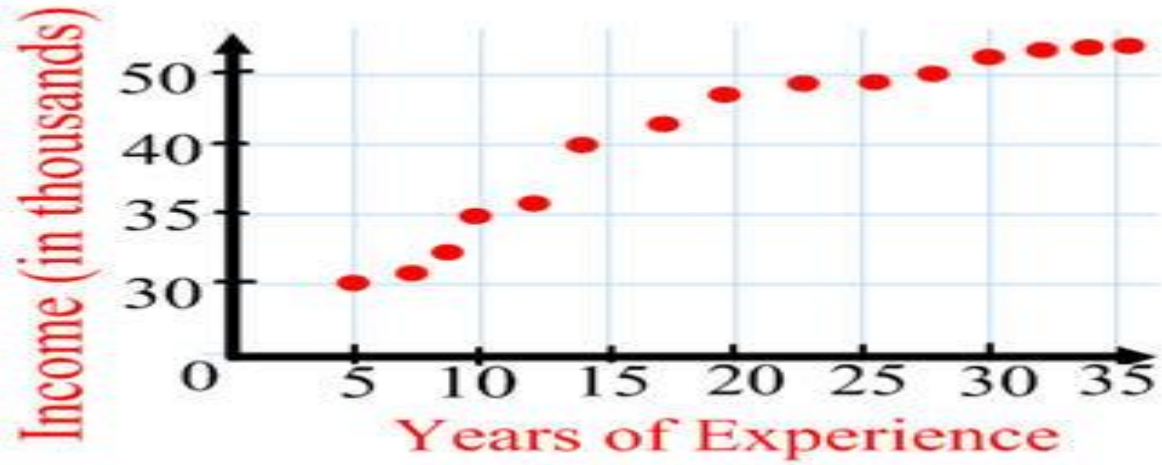


	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15
Total cess collection	4177	7136	9069	11722	11741	11623	14451	16899	20946	22837	24219
Transfer to PSK	0	0	8935	10393	12817	12258	15805	18334	20667	19988	22323
Short/excess transfer	4177	7136	134	1329	-1076	-635	-1354	-1435	279	2849	1896

Source: Report 50 of 2015 of CAG of India - Accounts of Union Government (Financial Audit)

(c) A scatter plot (scatter *diagram*) is a graphical presentation that is used to show how one variable changes in relation to a change in the other variable. It can be used to detect whether there is a correlation between two variables, for instance: correlation between weight and height. Hence, it is also called Correlation Diagram. The independent variable is plotted along the X - axis whereas the dependent variable is plotted along the Y-axis. The following scatter diagram shows that there is a positive relationship between years of experience and income of a person.

Fig. 5.7: Income v/s Years of Experience



5.6. Advantages of diagrammatic/Graphical Presentation

Visual presentation of data means presentation of data in the form of diagrams, curves and lines. It is useful in:

1) The data when presented in the form of diagrams and graphs, creates interest and leaves an impression on the mind of the reader for a longer period.
2) Comparison of data is much easier if it is presented in the form of diagrams and graphs. In several cases, careful glance at the diagram or graph renders the comparison of the complex data much easier.
3) The location of various statistical measures is possible with the help of graphs. Several measures of central value such as Median, Quartiles, Mode, etc., can be located with the help of graphs. Median, Quartiles and percentiles can be obtained by drawing ogives (cumulative frequency curves) while Mode is obtained from Histogram.
4) The trends of the past performance can be established with the help of graphs. Graphical presentation of trends helps in forecasting also.
5) Diagrams and graphs have become an integral part of the advertisement campaign of business firms. Any advertisement, without visual effect looks incomplete.

5.7. Limitations of Diagrammatic Presentations:

Although diagrams and graphs are powerful and effective media for presenting statistical data, they are not complete substitute for tabular presentation. The main limitations of diagrams and graphs are

1. They can present only approximate values.
2. They can represent only limited amount of information.
3. They are intended mostly to explain quantitative facts to the general public. From the point of view of the statistician, they are not of much help in analysing data.
4. They can be easily misinterpreted and, therefore, can be used for selfish motives during advertisement, propaganda and electioneering. As such the diagrams should never be accepted without a close inspection of the intention of the presenter.
5. The two-dimensional diagrams and the three-dimensional diagrams cannot be accurately appraised visually and therefore, as far as possible they should be used carefully.

Multiple Choice Questions

1. Bar diagram is a

(a) One-dimensional diagram
(b) Two-dimensional diagram
(c) Diagram with no dimension
(d) None of the above [Ans. (a)]

2. A Histogram is a graphical presentation of a frequency distribution of a :

(a) Individual series
(b) Discrete series
(c) Continuous series
(d) None of these [Ans. (c)]

3. In the first quadrant, the values of X and Y are:

(a) +ve
(b) –ve
(c) X is +ve and Y is-ve
(d) None of these [Ans. (a)]

4. If the values in a series are very large and the difference between the smallest value and zero is high, then we use -------- base line

(a) Original
(b) False
(c) True
(d) None of these [Ans. (b)]

5. A graph of cumulative frequency distribution is called _____

(a) Histogram
(b) Frequency Polygon
(c) Ogive
(d) None of above [Ans. (c)]

6. In a Pie we calculate the angles for each sectors by the formula

(a) Sector angle = $\frac{\text{Class frequency}}{\text{Total frequency}} * 100$

(b) Sector angle = $\frac{\text{Class frequency}}{\text{Total frequency}} * 180$

(c) Sector angle = $\frac{\text{Total frequency}}{\text{Class frequency}} * 360$

(d) Sector angle = $\frac{\text{Class frequency}}{\text{Total frequency}} * 360$ [Ans. (d)]

7. A circle in which sectors represents various quantities is called

(a) Histogram
(b) Frequency Polygon
(c) Pie Chart
(d) Component Bar chart [Ans. (c)]

8. A Histogram contains a set of

(a) Adjacent rectangles
(b) Non Adjacent Rectangles
(c) Adjacent squares
(d) Adjacent triangles [Ans. (a)]

9. In a histogram the area of each rectangle is proportional to

(a) the class mark of the corresponding class interval
(b) the class size of the corresponding class interval
(c) frequency of the corresponding class interval
(d) cumulative frequency of the corresponding class interval [Ans. (c)]

10. Which of the following statements about histogram rectangles is correct?

(a) The rectangles are proportional in height to the number of items falling in the classes.
(b) There are generally five rectangles in every histogram.
(c) The area in a rectangle depends on the number of items in the class as compared to the number of items in all other classes.
(d) All of these. [Ans. (a)]

TRY

Q1. Prepare a pie chart to show the percentage distribution of exports:

Country	USA	JAPAN	UK	China	Others
Percentage	25	15	30	20	10

Q2. Draw the following on the different graphs: (a) Histogram (ii) Less than ogive.

Mid-point	10	20	30	40	50
No. of students	8	18	15	22	14

Q3. Represent the following data by a sub-divided bar diagram. (No. of students)

Faculty	Arts	Commerce	Science	Total
(i) 1990-91	300	480	420	1200
(ii) 2001-01	320	300	480	1100

Q4. Present the following information about Expenditures of 2 families as a subdivided bar diagram

Item	food	clothes	Education	others	Total
family 'A'	500	200	150	150	1000
family 'A'	300	60	40	100	500

Q5. Draw a histogram and frequency polygon from the following data.

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

Q6. The table shows the percentage distribution of NDP in different sectors of two years. Represent the data in percentage bar diagram:

	Year	
Sector	2002-03	2007-08
Primary	52	33
Secondary	20	27
Tertiary	28	40

Q7. What kind of diagrams are more effective in representing the following? (i) Monthly rainfall in a year (ii) Composition of the population of Delhi by religion (iii) Components of cost in a factory
Ans. (i) Time series graphs/Line Graph (ii) Pie Diagram (iii) Sub-divided bar diagram or pie chart

Q8. The Indian Sugar Mills Association reported that, 'Sugar production during the first fortnight of December 2011 was about 7,87,000 tonnes, as against 7,78,000 tonnes during the same fortnight last year. The off-take of sugar from factories during the first fortnight of December 2011 was 6,83,000 tonnes for internal consumption and 41,000 tonnes for exports as against 6,54,000 tonnes for internal consumption and 11,000 tonnes for exports during the same fortnight last season.

(i) Present the data in tabular form.

(ii) Present these data diagrammatically and justify the type of diagram used by you.

Chapter 6

Measures of Central Tendencies - Averages

We have already studied how the data are classified and presented in the form of tables, diagrams and graphs. If the characteristics of the data are to be properly understood, it is necessary to summarise and analyse the data further. The first step in that direction is the computation of Averages or Measures of Central Tendency, which gives a bird's-eye view of the entire data.

6.1. Central Tendency

A single value that summarizes the observed values of a variable is called measure of Central Tendency. Alternatively, the single value which can be taken as the representative of the entire data set is called the measure of the Central Tendency or **Measures of Location**. The measures of Central Tendency most often used are Arithmetic Mean (Average), Median, Mode and Geometric Mean.

6.2. Requisites/Desired Characteristics of a measure of Central tendency

As suggested by the eminent statisticians Yule and Kendall, an ideal average should possess the following characteristics:

(a) ***Easy to understand and simple to compute***: It should be easy to understand an average and its computation should also be simple.

(b) ***Rigidly defined:*** An average should be rigidly defined by a mathematical formula so that the same answer is derived by different persons who compute it. It should not depend on the bias of a person computing it.

(c) ***Based on all items in the data***: For calculating an average, each and every item of the data set should be included.

(d) ***Not unduly affected by extreme items***: Extreme value i.e. maximum or minimum values, should not unduly affect the average. If the average changes with the inclusion or exclusion of an extreme item, then it is not a true representative of the data set.

(e) ***Capable of further mathematical treatment***: An average should be amenable to further algebraic treatment. It means combining values of measures for different groups, calculations of missing values, adjustment for wrong entries, etc., is possible without the knowledge of actual values of all items.

(f) ***Sampling stability***: The average should have 'sampling stability'. This means, if we take different samples from the aggregate, the average of any sample should approximately be the same as those of other samples.

6.3. Arithmetic Mean

The arithmetic mean is commonly known as mean or average. It is a measure of central tendency because other figures of the data gather around it. Arithmetic mean is defined as the sum of the given observations divided by the number of observations. It is the most commonly used statistical average in commerce, management, economics, finance, production, etc. The arithmetic mean is also called as simple Arithmetic Mean.

(a*) Calculation of Arithmetic Mean (AM) for raw data:* AM for raw data is calculated by adding up all the given values and dividing the sum by number of values in the set. The formula of the mean is:

Mean $(\bar{x}) = \frac{\sum \mathbf{Xi}}{n}$; where ∑Xi is the sum of all the given items and 'n' the number of items in data set.

Ex. 1: Calculation of Mean for raw data:

No. of Assesses (in lakhs) who paid Income Tax during 2015-16 in a State

Month	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.
No. of Assesses	38	39	40	36	40	37	38	41	42	40

The Arithmetic mean of the number of assesses is:

$$\text{Mean } (\bar{x}) = \frac{\sum \mathbf{Xi})}{n} \qquad = \frac{38 + 39 + 40 + 36 + 40 + 37 + 38 + 41 + 42 + 40}{10} = 39.1 \text{ lakh}$$

(b) *Calculation of Mean for Frequency Distribution*: For data arranged in a frequency distribution table the mean is given by the formula:

Mean $(\bar{x}) = \frac{\sum(\mathbf{fi} * \mathbf{Xi})}{\mathbf{N}}$; Where fi means frequencies of x_i values and N = ∑fi. To calculate mean for frequency table, we first obtain the product of fi and xi as fi*xi and its sum $\sum(\text{fi} * \text{Xi})$; finally the sum $\sum(\text{fi} * \text{Xi})$ is divided by N = ∑fi

Ex. 2: Calculation of Mean for Discrete Frequency Distribution [Direct Method]

Xi	**fi**	**fi*Xi**
1	4	4
2	6	12
3	5	15
4	4	16
5	4	20
6	3	18

Xi	**fi**	**fi*Xi**
7	4	28
8	6	48
9	4	36
10	3	30
11	3	33
12	4	48
Total ∑	50	308

$$\text{Mean } (\bar{x}) = \frac{\sum(\mathbf{fi} * \mathbf{Xi})}{\sum \mathbf{fi}} \quad = \frac{308}{50} = 6.16$$

(c) *Calculation of Mean for Continuous Frequency Distribution*: To find the mean of the frequency distribution with class intervals [continuous frequency distribution], first we obtain the mid value of the

class interval. The mid value is also called the **class mark**; it is obtained as ½*(lower limit + upper limit). For the first class it is ½*(20 + 25) = 22.5. After calculation of class mark, we proceed as above.

Ex. 3: Calculation of Mean for Continuous Frequency Distribution [Direct Method]

Class	20-25	25-30	30-35	35-40	40-45	45-50	50-55
f	10	12	8	20	11	4	5

Solution:

Class	**f**	**x**	**fx**
20 - 25	10	22.5	225.0
25 - 30	12	27.5	330.0
30 - 35	8	32.5	260.0
35 - 40	20	37.5	750.0
40 - 45	11	42.5	467.5
45 - 50	4	47.5	190.0
50 - 55	5	52.5	262.5
Total ∑	**70**		**2485.0**

$$\textbf{Mean } (\bar{x}) = \frac{\sum(\mathbf{fi} * \mathbf{Xi})}{\mathbf{N}} = \frac{\mathbf{2485}}{\mathbf{70}} = \mathbf{35.5}$$

(ii) Mean for frequency table using Short cut method: $\bar{\mathbf{x}} = \boldsymbol{a} + \frac{\sum(\mathbf{fi} * \mathbf{di})}{\mathbf{N}}$

where di $= x_i - a$; `a' is assumed mean; normally the middle value of x is taken as assumed mean.

(iii) Mean for frequency table by Step Deviation Method:

$\bar{\mathbf{x}} = \boldsymbol{a} + \frac{\sum(\mathbf{fi} * \mathbf{di}')}{\mathbf{N}} * h$ where; $di' = \frac{x_i - a}{h}$, 'a' is assumed mean and h is the class height.

Ex. 4: Calculation of Mean for Continuous Frequency Distribution by (i) Short cut method (ii) Step Deviation method

Class	20-25	25-30	30-35	35-40	40-45	45-50	50-55
f	10	12	8	20	11	4	5

Solution: (i) Short cut method [assumed mean a = 37.5]

Class	**f**	**x**	d = x – a = x – 37.5	F*d
20 - 25	10	22.5	-15	-150
25 - 30	12	27.5	-10	-120
30 - 35	8	32.5	-5	-40
35 - 40	20	**37.5**	0	00
40 - 45	11	42.5	5	+55
45 - 50	4	47.5	10	+40
50 - 55	5	52.5	15	+75
Total ∑	**70**			**-140**

$$\text{Mean } (\bar{x}) = a + \frac{\sum(fi*di)}{N}$$

$$= 37.5 + \frac{-140}{70}$$

$$= 37.5 - 2 = 35.5$$

(ii) Step Deviation Method: [assumed mean a = 37.5 and h = class height = 5]

Class	**f**	**x**	$d' = \frac{x-a}{h} = \frac{x-37.5}{5}$	f*d'
20 - 25	10	22.5	-3	-30
25 - 30	12	27.5	-2	-24
30 - 35	8	32.5	-1	-8
35 - 40	20	**37.5**	0	00
40 - 45	11	42.5	1	+11
45 - 50	4	47.5	2	+8
50 - 55	5	52.5	3	+15
Total ∑	**70**			**-28**

$$\textbf{Mean } (\bar{x}) = a + \frac{\sum(fi*di')}{N} * h =$$

$$= 37.5 + \frac{-28}{70} * 5$$

$$= 37.5 - 2 = 35.5$$

6.3.1. ***Properties of Arithmetic Mean*** The main properties of arithmetic mean are:

(a) The sum of the deviations of the individual items from the arithmetic mean is always zero i.e.

$\sum(\mathbf{Xi} - \bar{x}) = 0$. This is explained in the following illustration.

Sl. No.	Xi	$\mathbf{Xi} - \bar{x}$
1	5	-2
2	6	-1
3	7	0
4	8	1
5	9	2
Total $\sum$	35	0

Mean $(\bar{x}) = \frac{\sum \mathbf{Xi}}{\mathbf{n}} = \frac{\mathbf{35}}{\mathbf{5}} = 7$

In this illustration we note that the sum of positive deviations from the mean is equal to the sum of negative deviations. Precisely, therefore, **mean is also known as the centre of gravity of the given data** set. This is true for all kinds of data with class intervals or without class intervals.

(b) The sum of the square of deviations from the arithmetic mean is minimum i.e. the sum of the square of deviations from the arithmetic mean is always less than the sum of squares of deviations of the items taken from any other value. In other words $\sum(\mathbf{Xi} - \bar{x})^2$ is always minimum.

(c) If the number of items and mean are known, the total of the items can be obtained by multiplying the mean by the number of items, i.e., $\sum x = n\bar{x}$, where 'n' is the number of items. This property has a great practical significance. For example, if we know the number of workers in a factory, say 50, and average monthly wage as Rs. 8200, we can obtain the total monthly wage bill as Rs. 50 x 8200 = Rs. 41,00,00.

(d) If in the given set of values, we add or remove an observation which is equal to mean, the arithmetic mean remains unaffected.

(e) If each of the values of a variable 'x' is increased or decreased by some constant C, the arithmetic mean also increases or decreases by C. Similarly, when all the values of a variable '**x**' are multiplied by a constant, say k, the arithmetic mean is also multiplied by the same quantity k. For example, if the mean of a certain number of observations is 24 and 6 is added to each and every observation; then the mean of the resultant observations will be 24 + 6 = 30.

(f) If we have the arithmetic mean and number of items of two or more related groups, we can have a combined mean of these groups as follows:

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

where $\bar{x}_1, \bar{x}_2 \; and \; \bar{x}_3$ are the arithmetic means of groups 1, 2 and 3 respectively, and $n_1, n_2 \; and \; n_3$ are the number of items in groups 1, 2 and 3 respectively.

Ex. 5: Calculation of combined mean: Arithmetic mean of marks of two sections of students are 18 and 16; these sections have 35 and 40 students respectively. Calculate the combined mean of the 75 students of both the sections taken together.

Combined mean $\overline{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{35*18 + 40*16}{35 + 40} = \frac{1270}{75} = \mathbf{16.93}$

6.3.2. Merits and Demerits of Arithmetic Mean (AM): Mean fulfills all the conditions of a good average. However, it is largely affected by the extreme values so it must not be used for skewed data i.e. the data which have extreme values or outliers. When the number of observations in the data set is small, the mean is very sensitive to extreme values. For this reason, the mean is misleading when applied to skewed (non-symmetric) data. Following are the merits and demerits of arithmetic mean:

MERITS:

(i) ***Easy to understand and simple to compute***: AM is easy to understand and calculate.

(ii) ***Rigidly defined***: AM is rigidly defined by a mathematical formula so that the same answer is derived by different persons who compute it. It does not depend on the bias of a person computing it.

(iii) ***Based on all items in the data***: Arithmetic Mean is based on all the observations i.e. it considers all the given items of the data set in its calculation.

(iv) ***Capable of further mathematical treatment***: AM is amenable to further algebraic treatment. It means, if we are given the AM of three data sets of similar type, it is possible to obtain the combined AM of all those three data sets. Moreover, if mean is given, we can find the missing item of the data set.

(v) ***Sampling stability***: The AM has 'sampling stability'. This means, if we take different samples from the aggregate/population, the AM of any sample is approximately the same as those of the AM of other samples.

DEMERITS:

(i) ***It cannot be located graphically***: AM can't be located on graph like median and mode which can be located on graph also.

(ii) ***Not useful for skewed data***: AM is effective only if the frequency is normally distributed. In case skewness is more, the AM become ineffective.

(iii) ***Can't be calculated for open end classes:*** In case of open end class intervals we have to assume the limits of such intervals otherwise we can't calculate AM. For calculation of Median and mode open end classes pose no problem.

(iv) ***Can't be calculated for qualitative variables***: AM can't be calculated for qualitative variables such as Cleverness, Riches etc. as they can't be expressed numerically.

(v) ***Impossible or laughable conclusions***: AM can give strange results; for e.g. if there are 60, 50 and 42 students in three classes then average number of students in a class is $\frac{60+50+42}{3} = 50.67$, which is impossible as students can't be in fractions.

6.4. Weighted Arithmetic Mean:

Sometimes, some items of the data set are of greater importance than the others. For example, while constructing the cost of living index (indices) for a particular class of people, the commodities they consume have varying importance. The simple arithmetic mean of the prices of such commodities will not depict a true picture of their living pattern. Different commodities are to be assigned weights and a weighted arithmetic mean is to be worked out in such situations. Likewise for admission in a course, different subjects have different importance. For e.g. for admission in B.Sc. Physics and Chemistry may be more important than English. In such cases where some items are more important; weighted mean is preferred over Arithmetic mean.

To compute weighted arithmetic mean, different values of the variable x (viz. x_1, x_2 ..... x_n) are assigned different weights (viz. w_1, w_2,..... w_n) respectively. These values are then multiplied by their respective weights. The products so arrived are added and a total $\sum(\boldsymbol{W_i} * \boldsymbol{X_i})$ is obtained. It is then divided by the total of weights $\sum$**wi** and the quotient is the weighted arithmetic mean.

The main difficulty in the computation of weighted arithmetic mean is with regard to assigning of weights. These weights may be either actual or estimated. If actual weights are available, they must be used. If they are not available, some arbitrary weights may be assigned depending upon the situation.

Ex. 6: ***Calculation of Weighted Average/Mean*****:** Prices of three commodities A, B & C increased by 20 %, 30 % and 50 % respectively. Commodity A is 4 times more important than C, and B is three times more important than C. What is the mean/average rise in price of these three commodities? Compare it with Simple Arithmetic Mean.

Commodity	**Percentage rise in prices (x)**	**Weights (w)**	**w*x**
A	20	4	80
B	30	3	90
C	50	1	50
Total	100	8	220

Weighted Mean $(x_w) = \frac{\sum(wi * Xi)}{\sum wi}$

$= \frac{220}{8} = 27.5\%$

Simple Mean $(\bar{x}) = \frac{\sum Xi}{n} = \frac{100}{3} = 33.3\%$

6.4.1 ***Uses of Weighted Arithmetic Mean*****:** Weighted arithmetic mean is mainly useful under the following situations:

(i) When the given items are of unequal importance

(ii) When averaging percentages which have been computed by taking different number of items in the denominator

To be more specific, weighted arithmetic mean is used for:

(i) Construction of Index Numbers.

(ii) Computation of standardised birth and death rates.

(iii) Finding out an average output per machine, where machines are of varying capacities.

(iv) Determining the average wages of skilled, semi-skilled and unskilled workers of a factory.

6.5. Median

The median is also a measure of central tendency. Unlike arithmetic mean, median is based on the position of a given observation in a series arranged in ascending or descending order. Therefore, it is called a **positional average.** Median is the middle value of the variable when the values are arranged in an ordered array [ascending or descending order]. An equal number of items lie on either side of the median. Thus Median divides the given series (data set) into two equal parts. The median is usually denoted by 'Md'

(a) ***Calculation of Median for Raw data:*** For raw data, having arranged the data in ascending order or descending order, the median for a series with 'n' items is calculated as:

Median = $(\frac{n+1}{2})^{th}$ observation, **if n is odd** and

Arithmetic Mean of $\left(\frac{n}{2}\right)^{th}$ and $(\frac{n}{2}+1)^{th}$ observations, **if n is even.** This formula can be used to calculate median for **ungrouped frequency table** also.

Ex. 7: Calculation of Median for Raw data (Individual series) – odd number of items

Imagine that an athlete in a typical 200-metre training session runs in the following times: 26.1, 25.6, 25.7, 25.2 and 25.0 seconds. Calculate his median time?

Ans. Here n = no. of items = 5 **(Odd)**

Items in ascending order: 25.0, 25.2, 25.6, 25.7, 26.1.

Median = $(\frac{n+1}{2})^{th}$ observation = $\frac{5+1}{2}$ the item = 3^{rd} item

The third value in the data set in ascending order i.e. 25.6 seconds is the median time.

Ex. 8: Calculation of Median for Raw data (Individual series) – even number of items

Find the median of 24.7, 28.2, 25.0, 25.2, 25.6, 23.1, 25.7, 26.1

Ans. Here n = no. of items = 6 **(Even)**

Again, data in ascending order are: 23.1, 24.7, 25.0, 25.2, 25.6, 25.7, 26.1, 28.2

Median = Arithmetic Mean of $\left(\frac{n}{2}\right)^{th}$ and $(\frac{n}{2}+1)^{th}$ observations; here $\frac{n}{2} = \frac{8}{2} = 4$

So Median = Arithmetic Mean of 4^{th} and 5^{th} observations for items arranged in an ascending order

Median = $\frac{25.2+25.6}{2} = 25.4$

(b) ***Calculation of Median for Discrete Series: [Ungrouped Frequency Table]*:** In this case also we first arrange the data in ascending or descending order. Then find out the cumulative frequencies. As median being $(\frac{N+1}{2})^{th}$ observation, N is the sum of given frequencies; locate the value corresponding to $(\frac{N+1}{2})^{th}$ observation or next higher value in the column of cumulated frequencies; the corresponding value of the variable (x) gives the Median.

Ex. 9: Calculation of Median for Discrete Series: [Ungrouped Frequency Table]:

Marks out of 10 (x)	No. of students (f)	Cumulative Frequency
4	5	5
5	7	12
6	**10**	**22**
7	8	30
8	4	34
9	1	35

Here N = 35, sum of all the frequencies

Median = $(\frac{N+1}{2})^{th}$ observation

$= (\frac{35+1}{2})^{th}$ observation

$= 18^{th}$ observation

In the c.f. column number after 18 is 22, so the corresponding value 6 is the median.

- Cumulative frequency (c.f.) is obtained by successive addition of the frequencies

(c) ***Calculation of Median for Continuous Series: [Grouped Frequency distribution table*]:** In this case also we first arrange the data in ascending or descending order. Then find out the cumulative frequencies. Locate the class corresponding to $(\frac{N}{2})$ or next higher value in the column of cumulated frequencies; N is the sum of given frequencies. Thus, having determined a median class, the following formula gives us the median.

$$\text{Median} = l + \frac{\frac{N}{2} - cf_{m-1}}{fm} * h$$

Where: l = Lower limit of the median (c.f. = N/2) class

cf_{m-1} = Cumulative frequency of the class **prior to the median class**

fm = Frequency of the median class

h = Width/height of the class interval

Ex. 10: Calculation of Median for Continuous Series: [**Grouped Frequency distribution table]**

Height of Grade 10 Boys in a government school

Height (cm)	No. of boys (f)	Cumulative frequency
150 - 155	4	4
155 - 160	7	11
160 - 165	**18**	**29**
165 - 170	11	40
170 - 175	6	46
175 - 180	4	50

Here: $\frac{N}{2} = \frac{50}{2} = 25$ so the median class is 160-165

Using the formula: $$\text{Median} = l + \frac{\frac{N}{2} - cf_{m-1}}{fm} * h$$

$$\text{Median} = 160 + \frac{25 - 11}{18} * 5$$

$$= 160 + \frac{14}{18} * 5 = 160 + \frac{70}{18} = 163.89$$

6.5.1. ***Properties of Median*****:** The important properties of median are:

(i) The sum of the absolute deviations (i.e. deviations ignoring signs) from the median is minimum i.e. |x – Md| is the minimum.

(ii) Median is not affected by the extreme items i.e. outliers.

(iii) For an open-ended distribution, median is more suitable average than that of mean. For example, the income distribution is an open-ended distribution, median income would be a more representative figure.

(iv) For the qualitative information, median is the most suitable measure of central tendency. For example, a respondent may be asked to rate different departments, in the order of their risk as very high risk, high risk, moderate risk, low risk and no risk. Suppose he ranks them exactly as given here, the third adjective viz. moderate risk is the median of his five ratings.

(iv) The median can be located graphically (using cumulative frequency curves)

(vi) It is easy to compute and easy to understand. In some cases it can be obtained just by inspection.

6.5.*2. Merits and Demerits of Median:* Following are the merits and limitations of median.

MERITS

(i) For an open-ended distribution, such as income distribution, median gives more representative value.

(ii) Since median is not distorted by the extreme items, in some cases it is preferred over mean as the latter is likely to be distorted by extreme values.

(iii) For dealing the qualitative phenomena, median is the most suitable average.

(iv) Since median minimises the total absolute deviations, median is preferred in the situations wherein the total geographical distance is to be minimised. For example, there is a conference of five top executives from five different cities of India lying almost in a straight line. The city located at a median distance would be a more proper place for the conference.

DEMERITS

(i) Median is not capable of algebraic treatment. That means we cannot have a combined median of two or more groups, unless all the items of the groups are known. It is because median does not depend on all the items of the given data set.

(ii) It is described, sometimes, as an insensitive measure as it is not based on all items of the series.

(iii) It is affected more by sampling fluctuations than the value of mean.

(iv) The computational formula of a median is in a way an interpolation under the assumption that the items in the median class are uniformly distributed, which is not very true.

6.6. Mode

Mode is also a measure of central tendency. Mode is the value of a variable which is repeated most often in the data set i.e. the value of the variable which occurs most frequently. It shows the centre of concentration of the frequency in and around a given value. It is not the centre of gravity like mean. It is a positional measure similar to median. It is commonly denoted by Mo. For example, take the case of a shopkeeper who sells shoes/readymade garments. He is interested to know the sizes of shoes/ garments which are commonly demanded. Here, mean would indicate a size that may not fit any person. Median may not provide a representative size because of the unevenness in the distribution. It is the mode which will help in making a choice of approximate size for which an order can be placed. Mode can also be used for Qualitative data like intelligence, beauty, honesty, etc. If 2 values occur most often, there are 2 modes for the distribution and it is called a **bimodal distribution**. When Mode is ill-defined (either more than one mode or mode occurs near the beginning/end of the data set) mode can be calculated using the empirical relationship:

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

(a) ***Calculation of Mode for raw data*****:** For raw data the mode can be obtained by arranging the data in an ascending/descending order and then finding the item appearing the maximum number of times. However, if the number of observations are large the mode can be obtained by arranging the data in a frequency table. The value/observation with highest freq. gives mode.

For e.g. for observations: 1, 3, 5, 7, 9, 13, 5, 7, 8, 2, 10, 5

In ascending order: 1, 2, 3, 5, 5, 5, 7, 7, 8, 9, 10, 13;

Observation 5 occurs the maximum number of times (i.e. no. 5 has the highest frequency) so mode = 5

(b) Calculation of Mode: for ungrouped frequency table:

Marks out of 10 (x)	No. of students (f)
4	5
5	7
6	**10**
7	8
8	4
9	1

Here, mode = 6 which has the highest frequency.

(c) Mode for grouped frequency distribution

Mode = l + $\frac{f_1 - f_o}{2f_1 - f_0 - f_2}$ *** h where**

l = Lower limit of the modal class

f_0 = Frequency of the class prior to the modal class

f_1 = Frequency of the modal class

f_2 = Frequency of the class following the modal class

h = Width of the class interval

Modal class is the class which has the highest frequency

Ex. 11: Calculation of Mode for grouped frequency table:

Height of Grade 10 Boys in a government school

Height (cm)	No. of boys (f)
150 - 155	4
155 - 160	7
160 - 165	**18**
165 - 170	11
170 - 175	6
175 - 180	4

Here modal class is the class 160-165 because the highest frequency occurs here.

Thus, l = 160, f_1 = 18, f_0 = 7, f_2 = 11, h = class height = 5

So Mode = $l + \frac{f_1 - f_o}{2f_1 - f_0 - f_2} * h$

$= 160 + \frac{18-7}{2*18-7-11} * 5$

$= 160 + \frac{11}{18} * 5$

$= 160 + \frac{55}{18}$

Note: Remember that Median always lies between mean and mode.

6.6.1. *Merits and Demerits of Mode*

MERITS

(i) In certain situations mode is the only suitable average, e.g., modal size of garments, modal size of shoes, modal wages, etc.

(ii) Mode can be used to describe qualitative phenomena. For instance, if a printing press turns out five impressions which we rate very sharp, sharp, sharp, blurred and sharp, then the modal value is sharp.

(iii) In the case of skewed distribution, mode is the indicator of the point of heaviest concentration.

(iv) Even if one or more classes are open-ended, mode can be calculated.

DEMERITS

(i) Too often, there is no modal value. It is a useless measure, when there are more than one mode.

(ii) It is not capable of further algebraic treatment.

(iii) It is an ill-defined measure since different methods may yield somewhat different answers.

(iv) It is not based on all the items of the data.

(v) The value of the mode is affected significantly by the size of the class-intervals.

(vi) Although a mode is the value of a variate that occurs most frequently, its frequency does not represent a majority of the total frequencies.

6.6.2. *Usefulness of Mode*

(i) Mode is used when most typical value of a distribution is desired.

(ii) It is the most meaningful measure of central tendency in case of highly skewed distributions.

(iii) It can be useful for qualitative data also.

(iv) Mode can be found even if the frequency distribution has class-intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude. It can be calculated for Open-end classes also.

(v) It can be found graphically also.

(vi) It is the most used average in day-to-day life, such as average number of students in a section, average size of shoes/garment.

6.7. Geometric Mean

The geometric mean is a mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the *n*th root of the product of *n* numbers, i.e., for a set of numbers x_1, x_2, ..., x_n, the geometric mean is defined as

$$GM = (x_1 * x_2 * x_3 * \ldots\ldots x_n)^{(1/n)}$$

$$= (\prod x_i)^{(1/n)}$$

Where $\prod x_i$ means multiplication of all X values and ^(1/n) means nth root.

For instance, the geometric mean of two numbers, say 2 and 8 is square root of their product = (2*8)^(1/2) = $\sqrt{16}$ = 4 and the geometric mean of numbers 4, 1, and 1/32 is the cube root of their product = $\sqrt[3]{4 * 1 * 1/32}$ = $(1/8)^{1/3}$ = ½.

GM is generally calculated with the help of log by using the formula

G = Antilog $\{\frac{\sum \log Xi}{n}\}$, for individual observations and

G = Antilog $\{\frac{\sum fi * \log Xi}{N}\}$ where N = ∑fi

6.7.1. *Usefulness of Geometric mean*

(i) It is useful to find out average growth rates and returns on portfolio of securities.

(ii) It is considered to be the best average in the construction of index numbers.

(iii) It is suitable when large weights have to be given to small items and small weights to large items in social and economic fields.

(iv) It is capable of algebraic manipulation.

(v) It is good for combining numbers expressed in different units.

Problem of Geometric Mean: It can't be calculated for the series having an item as zero or negative.

6.8. Relationship between Arithmetic Mean and Geometric Mean

For a list of non-negative real numbers, the arithmetic mean is greater than or equal to the geometric mean. Taking the formulas for both types of mean, for 2 numbers 'x' and 'y' we get the inequality:

$$\frac{x+y}{2} \geq \sqrt{x * y}$$

For example, for the numbers 9, 12, 54, the arithmetic mean 25, is greater than the geometric mean 18.

6.9. Comparison of the three measures of Central Tendency:

Of the 3 parameters, it will depend on the situation as to which one we would use. The following table helps in deciding which of these three measures would be appropriate to use in the various situations:

Mean	Median	Mode
Represents centre of gravity of data set	Represents middle of data set	Represents most common value
Sensitive to extreme values i.e. distorted by outliers/skewed data	Not sensitive to extreme values i.e. not distorted by outliers/ skewed data	Data set may have no mode, one or multiple mode
Most useful when data are normally distributed	Most useful when data set is skewed or has few extreme values in one direction	Not very popular
Uses all data values	Ignores most of the information; it is based on just 50% of the observations	Easily determined even for categorical data set

6.10. Which measure to use – Mean, Median or Mode?

i. Of the 3 measures, it will depend on the situation as to which one we would use.

ii. AM fulfills all the requisites of a good average, however, it is largely affected by the extreme values so it must not be used in the case when data are skewed (have outliers)

iii. For example, in calculating no. of days by which a project is delayed or an ordinance/order is issued wherein in one or two cases the number of days are extraordinary long (say one year or more), we should use mode or median and not mean.

iv. In general it is a sound rule in practice to calculate AM and use it along with median and mode

Example to understand which measure of Average to use?

Consider the duration (days) of absence from work of 21 labourers: 1, 1, 2, 2, 3, 3, 4, 4, 4, 4, 5, 6, 6, 6, 7, 8, 9, 10, 10, 59, 80.

- Mean = 11 days
  - Not typical of the series as 19 of the 21 labourers were absent for less than 11 days
  - Distorted by extreme values 59 and 80.
- Median = 5 days
  - Better measure as fifty percent of labourers were absent for less than 5 days and 50% were absent for more than 5 days
- Mode = 4 days
  - Better measure as maximum number of labourers were absent for 4 days.

6.11. Objectives and Functions of Averages:

(i) Representative of the group: An average represents all the features of a group; hence the results about the whole group can be deduced from it. For e.g. the average height of an Indian male is 5'6"

(ii) Brief description: An average gives us simple and brief description of the main features of the whole data.

(ii) Helpful in comparison: The measures of central tendency or averages reduce the data to a single value which is highly useful for making comparative studies. For example, the average marks of the students of Section 'A' are less than that of Section 'B'.

(iv) Helpful in formulation of policies: Averages help a business or the economy in formulation of policies for their development.

(v) Base of other statistical Analysis: Statistical measures such as mean deviation, co-efficient of variation, correlation, analysis of time series and index numbers are also based on the averages.

6.12. Limitations of Average/Central Tendency

(i) Since average is a single value representing a group of values, it must be properly interpreted otherwise there is every possibility of jumping to a wrong conclusion.

(ii) Average may not exist in the given data set; for e.g. 160, 170, 190, 210, 180 have the average (mean) as 182 which does not exist in the given data.

(iii) An Average may give absurd results e.g. a couple having average 2.5 children

(iv) An Average may fail to give idea about formation of series; for e.g. 150, 170, 190, 210, 180 and 300, 500, 90, 8 and 2 have same average.

6.13 Calculation of Mean, Median and Mode

Ex. 11: No. of Assesses (in lakhs) who paid Income Tax during 2015-16 in a State

Month	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.
No. of Assesses	38	39	40	36	40	37	38	41	42	40

The Arithmetic mean of the number of assesses is:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{38 + 39 + 40 + 36 + 40 + 37 + 38 + 41 + 42 + 40}{10} = 39.1 \text{ lakh}$$

To find median and mode of this distribution, **data have been arranged in an ascending order as**:

Sl. No.	1	2	3	4	5	6	7	8	9	10
No. of Assesses	36	37	38	38	39	40	40	40	41	42

n = 10, an even no. so median is the average of $10/2 = 5^{th}$ and $(10/2 + 1) = 6^{th}$ observation.

Median = (39 + 40) /2 = 39.5

Mode of the data set is 40, as the value 40 occurs maximum number (3) of times in this distribution.

Ex. 12: Calculation of Mean, Median and Mode for Freq. Distribution

Xi	**fi**	**(fi*Xi)**	**c.f.**	
1	4	4	4	1-4 items
2	6	12	10	5-10 items
3	5	15	15	11-15 items
4	4	16	19	16-19 items
5	4	20	23	20- 23 items
6	3	18	26	24-26 items
7	4	28	30	27 – 30 items
8	6	48	36	31-36 items
9	4	36	40	37-40 items
10	3	30	43	41- 43 items
11	3	33	46	44- 46 items
12	4	48	50	47 – 50 items
Total Σ	**50**	**308**		

Mean $= \frac{\Sigma(\mathbf{fi} * \mathbf{Xi})}{\Sigma \mathbf{fi}} = \frac{308}{50} = 6.16$

N = 50 (even); N/2 = 25

Median = $1/2*(25^{th} + 26^{th})$ item

= ½*(6+6) = 6

Highest freq. 6 occurs twice so it is a bimodal distribution. Thus we'll use Empirical Relation for finding out mode:

Mode = 3*median - 2*mean

= 3*6 –2*6.16 = 5.68

Ex. 13: Find the mean of the following marks

Class	0-10	10-20	20-30	30-40	40-50
f	4	6	13	6	1

Solution:

Class	**f**	**x**	**f*x**
0 - 10	4	5	20
10 - 20	6	15	90
20 - 30	13	25	325
30 - 40	6	35	210
40 - 50	1	45	45
Total	30		690

Here: $\sum(fi * Xi) = 690$ and $\sum f=30$

$$\text{Mean} = \frac{\sum(\mathbf{fi}*\mathbf{Xi})}{\sum \mathbf{fi}} = \frac{690}{30} = 23$$

Ex. 14: Find the mean of the following data: 13, 15, 17, 19, 10, 15, 25

Solution:

$$\text{Mean} = \frac{\sum(\mathbf{Xi})}{n} = \frac{\mathbf{13+15+17+19+10+15+25}}{\mathbf{7}}$$

$$= \frac{\mathbf{114}}{\mathbf{7}} = 16.29$$

Ex. 15: Find the mean of the following marks using shortcut method

Marks(x)	5	15	25	35	45	55	65
No. of Students (f)	5	4	3	5	12	8	1

Solution:

x	**f**	**d = x –a = x - 35**	**f*d**
5	5	-30	-150
15	4	-20	- 80
25	3	-10	- 30
35	5	00	00
45	12	+10	120
55	8	+20	160
65	1	+30	30
Total	38		50

Let assumed Mean (a) = 35

Mean is given by $\bar{\mathbf{x}} = \boldsymbol{a} + \frac{\sum(\mathbf{fi}*\mathbf{di})}{\mathbf{N}}$

$$= 35 + \frac{50}{38}$$

$$= 35 + 1.32$$

Mean = 36.32

Ex. 16: Find the mean of the following data using Step Deviation method

Items	0-10	10-20	20-30	30-40	40-50
Frequency	2	5	1	3	12

Class Interval	**Frequency (f)**	**x**	**$d' = \frac{x-a}{h} = \frac{x-25}{5}$**	**fd'**
0-10	2	5	$\frac{5-25}{10} = -2$	2*-2 = -4
10-20	5	15	$\frac{15-25}{10} = -1$	5*-1 = -5
20-30	1	25	$\frac{25-25}{10} = 0$	1*0 = 0
30-40	3	35	$\frac{35-25}{10} = 1$	3*1 = 3
40 - 50	12	45	$\frac{45-25}{10} = 2$	12*2 = 24
Sum	**N=23**			**18**

Let assumed mean = 25

Mean $(\bar{x}) = a + \frac{\Sigma(fi*di')}{N} * h$

$= 25 + \frac{18}{23} * 10$

$= 25 + 7.83$

$= 32.83$

Multiple choice questions: choose the correct answer

Q1. Which of the following can be easily determined from an Ogive?

(a) Mean

(b) Median

(c) Mode

(d) None of the above [Ans. (b)]

Q2. The sample mean, $\overline{X}$ is:

(a) Always equal to the population mean

(b) Never equal to the population mean

(c) A statistic

(d) A parameter [Ans. (c)]

Q3. All the following are measures of central tendency, except:

(a) Mean

(b) Median

(c) Mean deviation

(d) Geometric mean [Ans. (c)]

Q4. Central value of a set of data is termed as:

(a) Mean

(b) Median

(c) Mode

(d) Geometric Mean [Ans. (b)]

Q5. The median of a set of numbers 4, 5, 8, 6, 3, 4, 8, 10, 8 is:

(a) 3

(b) 6

(c) 4.5

(d) 28 [Ans. (b)]

Q6. The median of a set of numbers 5, 5, 12, 15, 18, 11, 7, 9 is:

(a) 4

(b) 33/2

(c) 41

(d) 10 [Ans. (d)]

Q7. For grouped frequency data the median is calculated by:

(a) Median = $l + \frac{\frac{N}{2} - cf_{m-1}}{fm} * h$

(b) Median = $l + \frac{\frac{N}{2} + cf_{m-1}}{fm} * h$

(c) Median = $l - \frac{\frac{N}{2} - cf_{m-1}}{fm} * h$

(d) Median = $l - \frac{\frac{N}{2} + cf_{m-1}}{fm} * h$ [Ans. (a)]

Q8. Median is equivalent to:

(a) 50^{th} percentile

(b) 100^{th} percentile

(c) 75^{th} percentile

(d) 25^{th} percentile [Ans. (a)]

Q9. Central value of a set of 240 values can be obtained by:

(a) 3^{rd} quartile

(b) 120 percentile

(c) 2^{nd} quartile

(d) 4^{th} quartile [Ans. (c)]

Q10. For a group of n = 120 subjects, 40^{th} percentile will be:

(a) 48^{th} value

(b) 40^{th} value

(c) 30^{th} value

(d) None of the above [Ans. (a)]

Q11. The most frequently occurring value in a data is:

(a) Mean

(b) Mode

(c) Median

(d) Standard Deviation [Ans. (b)]

Q12. The mode value of the set 3, 5, 7, 9 and 11 is:

(a) No Mode

(b) 0

(c) 11

(d) 5 [Ans. (a)]

Q13. The mode value(s) in the set 2, 2, 3, 3, 3, 5, 9, 10, 11, 10, 12, 10 is:

(a) 12

(b) 3

(c) 10

(d) 3 and 10 [Ans. (d)]

Q14. GM of the numbers 8, 4, 2 is:

(a) 8

(b) 4

(c) 14/3

(d) 84 [Ans. (b)]

Q15. Which of the following is not a measure of central tendency?

(a) P_{50}

(b) Mode

(c) Geometric Mean

(d) Mean deviation [Ans. (d)]

Q16. The highest point of a frequency curve is most closely associated with:

(a) The Mean

(b) The Mode

(c) The Median

(d) The Geometric Mean [Ans. (b)]

Q17. Median is preferred to mean when:

(a) Population is large enough

(b) Low variance is seen

(c) Skewed distribution is seen

(d) None of the above [Ans. (c)]

Q18. If the top most score in a distribution is tripled which measure will change?

(a) Mean

(b) Median

(c) Mode

(d) All of the above [Ans. (a)]

Q19. Which measure of central tendency can take two values for a given distribution?

(a) The median

(b) The mode

(c) The mean

(d) None of the above [Ans. (b)]

Q20. Cumulative frequency is most closely associated with:

(a) Mean

(b) Median

(c) Mode

(d) Standard Deviation [Ans. (b)]

Q21. The correct formula for calculation of mode value with grouped data is:

(a) $l + \frac{f_1 - f_o}{f_1 - f_0 - f_2} * h$

(b) $l + \frac{f_1 - f_o}{2f_1 + f_0 + f_2} * h$

(c) $l + \frac{f_1 - f_o}{2f_1 - 2f_0 - 2f_2} * h$

(d) $l + \frac{f_1 - f_o}{2f_1 - f_0 - f_2} * h$ [Ans. (d)]

Q22. Which of the following is crude formula for mode?

(a) Mode = 3 Mean- Median

(b) Mode = 3 Mean-2 Median

(c) Mode = 2(Mean + Median)

(d) Mode = 3 Median - 2 Mean [Ans. (d)]

Q23. If mean and median of a grouped data are 16 and 30 respectively then mode value is:

(a) 18

(b) 58

(c) 0

(d) 92 [Ans. (b)]

TRY

Q1. What do you mean by central tendency? **(b)** What is the importance of measuring central tendency? **(c)** What are the limitations of Central Tendency/Averages?

Q2. Find the geometric means of: 2, 4, 8, 32, 64 and 128

Q3. Find the mode for the following data:

Mid value	:	115	125	135	145	155	165	175	185	195
Frequency	:	6	25	48	72	116	60	38	22	3

Q4. From the data given below calculate mode. Also construct a histogram & locate mode.

Value	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	328	350	720	664	598	524	378	244

Q5. The figure of 2.2 children per adult female was felt to be in some respects absurd and the Royal Commission suggested that the middle class be paid money to increase the average to a rounder and more convenient number." Commenting on the above statement, discuss the limitations of the arithmetic average. Also point out the characteristics of a good measure of central tendency.

Q6. The mean age of a combined group of men and women is 30 years. If the mean age of the group of men is 32 and that of the group of women is 27 find out the percentage of the men and women in the group.

Q7. The mean weight of 150 students in a certain class is 60 kgms. The mean weight of boys in the class is 70 kgms and that of girls is 55 kgms. Find the number of boys and the number of girls in the class.

Q8. Mention the names of averages which can be located on graph.

Q9. Mention three merits and three demerits each of **(a)** median **(b)** mode **(c)** Mean.

Q10. Calculate mode of the following data graphically:

Marks	0-5	5-10	10-15	15-20	20-25	25-30
No. of Students	6	10	20	12	8	4

Q11. Calculate mean and median

CI	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	5	8	7	12	28	20	10	10

Q12. Find no. of shops for class interval 30-40 if the mean profit per shop is Rs. 28.

Profit per shop (Rs.)	0-10	10-20	20-30	30-40	40-50	50-60
No. of shops	12	18	27	?	17	6

Q13. There are 80 students in class XI. The average marks in maths are 65 for the class. There are 2 sections with 50 students in first section whose average marks are 60. Find average marks of the other section.

Chapter 7
Measures of Dispersion

Central tendency alone is not sufficient to analyse the data. For more meaningful analysis of data it is necessary to study Dispersion. Dispersion is also known as the **spread or variability** of data.

Measures of dispersion or variation measure the "spread" in the data around a measure of Central Tendency. It indicates the degree of heterogeneity or variability in the given data. It is an important characteristic indicating the extent to which observations vary among themselves. To appreciate the use of dispersion, let us consider three data sets showing prices of three items A, B and C as under:

Prices of item 'A': 60, 60, 60, 60, 60

Prices of item 'B': 58, 59, 60, 61, 62

Prices of item 'C': 0, 30, 60, 90, 120

Here, three data sets have the same average (mean, median and mode), however, they differ in terms of spread. Thus, different sets of data may have the same measure of central tendency, but may differ greatly in terms of spread or scattered-ness i.e. in terms of dispersion. Dispersion indicates, on an average, by how much amount different items of the data set differ from the central tendency.

7.1. Significance of Measuring Dispersion - Measures of variations (dispersion) are calculated to serve the following purposes:

(a) ***To know the reliability of an average***: Measuring variability determines the reliability of an average by pointing out to what extent the average is a representative of the entire data. In the above example, mean of set 'A' is 60; it is the perfect representative of data set. In case of data set 'B', the variation is low. Therefore, in this case, the mean can be considered as a good representative of the data set. But in case of data set C the variation in individual figures is very large so the average is hardly a representative of all high and low figures such as 0 and 120.

(b) ***To serve as a basis of the control of variability***: Another purpose of measuring variability is to determine the nature and cause of variation in order to control the variation itself.

(c) ***To compare two or more series with regard to their variability***: Measures of dispersion (Relative measures in particular) enable comparisons of two or more distributions with regard to their variability. It also gives an idea of the consistency of two data sets; the lower the variability of the data set, the higher is the consistency.

(d) ***Use of other statistical measures***: Dispersion also facilitates the use of other statistical measures like correlation, regression, statistical inference, etc.

(e) ***Useful in Estimation and Quality Control***: Dispersion is also helpful in theory of Estimation, testing of hypothesis and Statistical Quality Control, etc.

7.2. Properties OR Requisites of a Good Measure of Dispersion:

A measure of dispersion is the average of the deviations of items from its mean i.e., it is an average of the second order. Hence, it should possess all the qualities of a good measure of an average. According to Yule and Kendall the qualities of a good measure of dispersion are as follows:

1 ***Simple to understand and easy to calculate***: Measures of Dispersion are used even by layman. So they should be simple to understand and easy to calculate.

2 ***Rigidly defined***: It means, for the same data, all the methods of calculation should produce the same answer for a measure. In other words different methods of computation leading to different answers is not desirable.

3 ***Based on all items:*** When it is based on all items, it will produce a more representative value. Thus, good measure of dispersion should be based on the entire data.

4 ***Amenable to further algebraic treatment***: It means combining values of measures for different groups, calculations of missing values, adjustment for wrong entries, etc., is possible without the knowledge of actual values of all the items of the data set.

5. ***Sampling stability***: It means that the average difference in the values obtained from the sample and the corresponding values from the population should be the least. If it is so for a measure of dispersion, it is the best measure.

6. ***Not unduly affected by the extreme items***: Extreme items or outliers, many times, are not true representatives of the data. So their presence should not affect the calculation of the measure of dispersion to a large extent.

7.3. Different Measures of Dispersion:

There are two types of measures of dispersion. They are Absolute and Relative measures.

(a) ***Absolute Measures*****:** The measure of dispersion which are expressed in terms of the original units of data are termed as Absolute Measures. Such measures are not suitable for comparing the variability of the distributions or series expressed in different units of measurement.

(b) ***Relative Measures*****:** The measure of dispersion which are expressed in terms of ratios or percentages are called Relative Measures of dispersion. These measures are pure numbers independent of the unit of measurement. Normally, a measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average. Hence, it is also known as **Coefficient of Dispersion.**

(c) ***Measures in Common use***: The following measures of absolute dispersion are in common use:

(i) Based on selected items of the data: (i) Range - spread for entire data (ii) Semi Inter Quartile Range or Quartile Deviation - spread for middle 50% data.

(ii) Based on all items of the data: (i) Mean Deviation - mean of the absolute deviations from central tendency. (ii) Standard Deviation or Root Mean Square Deviation about arithmetic mean.

(ii) A Graphic Method - Lorenz Curve.

The relative Measures of Dispersion corresponding to the measures of absolute dispersion are:

Absolute Measures of Dispersion	Relative Measures of Dispersion
i) Range	Coefficient of Range
ii) Quartile Deviation	Coefficient of Quartile Deviation
iii) Mean Deviation	Coefficient of Mean Deviation
iv) Standard Deviation	Coefficient of Standard Deviation or Coefficient of Variation

7.4. Range

The range is defined as the difference between the highest and the lowest value in a set of data. Thus, Range = $X_{max.} - X_{min}$.

For the three data sets giving the prices of the different items the ranges are:

Prices of item 'A': 60, 60, 60, 60, 60; Range = 60 – 60 = 0

Prices of item 'B': 58, 59, 60, 61, 62; Range = 62 – 58 = 4

Prices of item 'C': 0, 30, 60, 90, 120; Range = 120 – 0 = 120

The interpretation of the value of range is very simple. In this case, the variation is zero in case of prices of item **A,** the variation is small in case of prices of item B, and the variation is very large in case of prices of item C.

For grouped data, the range may be approximated as the difference between the upper limit of the largest class and the lower limit of the smallest class. The relative measure corresponding to range, called the coefficient of range, is obtained as:

$$\text{Coefficient of Range} = \frac{X_{max.} - X_{min.}}{X_{max.} + X_{min.}}$$

Again for the three data sets on prices, the Coefficient of Range are: Series 1: $\frac{60-60}{60+60} = \frac{0}{120} = 0$

Series 2: $\frac{62-58}{62+58} = \frac{4}{120} = 0.033$ and Series 3: $\frac{120-0}{120+0} = \frac{120}{120} = 1$ **(Largest)**

7.4.1 Computation of Range and its Coefficient

Ex. 1: Calculate Range and Coefficient of Range

No. of Assesses (in lakhs) who paid Income Tax during 2015-16 in a State

Month	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.
No. of Assesses	38	39	40	36	40	37	38	41	42	40

Here Range = $X_{max.} - X_{min}$ = 42 – 36 = 6 and Coefficient of Range = $\frac{42-36}{42+36} = \frac{6}{78} = \frac{1}{13} = 0.077$

Ex. 2: Calculate Range and Coefficient of Range

Class	20-25	25-30	30-35	35-40	40-45	45-50	50-55
f	10	12	8	20	11	4	5

Range = Upper limit of the largest class - Lower limit of the smallest class = 55 – 20 = 35 and

Coefficient of Range = $\frac{55-20}{55+20} = \frac{35}{75} = \frac{7}{15}$

7.4.2. ***Merits and Demerits of Range*****:** The range is very easy to calculate and it gives us some idea about the variability of the data. However, as only two extreme values are used for computing range, it is a crude measure of variation and is not considered very reliable. Moreover, it largely increases with the increase in the sample size.

The concept of range is extensively used in statistical quality control. Range is helpful in studying variations in the, prices of shares, debentures and agricultural commodities which are very sensitive to price changes. The range is a good indicator for weather forecast.

7.5. Quartile Deviation

Quartile deviation is defined as half the difference between the upper quartile (Q_3) and lower quartile (Q_1). In other word; Quartile Deviation (QD) = $\frac{Q_3 - Q_1}{2}$; where Q_1 is the first quartile and Q_3 is the third quartile.

The difference $Q_3 - Q_1$, the distance between the two quartiles may be called **Inter Quartile Range** and half of this, **Semi-Inter Quartile Range** is called Quartile Deviation.

Quartile Deviation (QD) is dependent on the two quartiles, and does not take into account the variability of the largest 25% and the smallest 25% of observations. It is, therefore, unaffected by extreme values. Another advantage of quartile deviation is that it is the only measure of variability which can be used for open-end distribution.

The main limitation of quartile deviation is that it does not depend on the magnitudes of all the observations. It is based on the middle 50% of the observations.

The relative measure of dispersion based on quartile deviation is called the coefficient of quartile deviation. The coefficient of quartile deviation is defined as:

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

7.5.1 ***Computation of Quartile Deviation and its Coefficient*****:** Quartiles can be calculated similar to that of Median, once the quartiles are available the Quartile deviation and the coefficient of Quartile Deviations can be calculated easily.

Ex. 3: Calculate the Quartile Deviation for the following set of Data

Weight (in Kgs.):	60	61	62	63	65	70	75	80
No. of Workers:	1	3	5	7	10	3	2	3

Weight in Kgs. (x)	No. of Workers (f)	Cumulative Frequency (c.f.)
60	1	1
61	3	4
62	5	9
63	7	16
65	10	26
70	3	29
75	2	31
80	2	33

Here N = 33

Q_1 = Size of $\frac{(N+1)}{4} = \frac{(33+1)}{4}$ or 8.5^{th} observation = $\frac{8\text{th observation} + 9\text{th obs.}}{2} = \frac{62+62}{2}$ = 62 Kgs.

Q_3 = Size of $3.\frac{(N+1)}{4}$ = 3*8.5 = 25.5 observation = $\frac{25\text{th observation} + 26\text{th obs.}}{2} = \frac{65+65}{2}$ = 65 Kgs.

So Quartile Deviation = $\frac{Q_3 - Q_1}{2} = \frac{65-62}{2}$ = 3/2 = 1.5

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{65-62}{65+62} = \frac{3}{127}$ = 0.024

Ex. 4: Height of Grade 10 Boys in a government school is given below, calculate Quartile Deviation and its Coefficient

Height (cm)	No. of boys (f)	Cumulative frequency
150 - 155	4	4
155 - 160	7	11
160 - 165	18	29
165 - 170	11	40
170 - 175	6	46
175 - 180	4	50

Here: $\frac{N}{4} = \frac{50}{4}$ = 12.5 so Q_1 class is 160-165

Using the formula: $Q_1 = l + \frac{\frac{N}{4} - cf}{f} * h$

$Q_1 = 160 + \frac{12.5-11}{18} * 5$

$= 160 + \frac{1.5}{18} * 5 = 160 + \frac{7.5}{18} = 160 + 0.417 = 160.42$

Here: $\frac{3N}{4} = \frac{150}{4}$ = 37.5 so Q_3 class is 165-170

Using the formula: $Q_3 = l + \frac{\frac{3N}{4} - cf}{f} * h$

$Q_3 = 165 + \frac{37.5 - 29}{11} * 5$

$= 165 + \frac{8.5}{11} * 5 = 165 + \frac{42.5}{18} = 160 + 2.361 = 162.36$

So Quartile Deviation $= \frac{Q_3 - Q_1}{2} = \frac{162.36 - 160.42}{2}$

$= ½ * 1.94 = 0.97$

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{162.36 - 160.42}{162.36 + 160.42} = \frac{1.94}{322.78} = 0.006$

Ex. 5: Compute an appropriate measure of dispersion for the data given below:

Marks out of 200	**No. of boys (f)**	**Cumulative frequency**
Below 85	12	12
85 – 90	16	28
90 – 95	39	67
95 - 100	56	123
100- 105	62	185
105 - 110	75	260
110 - 115	30	290
115 and above	10	300

Since the frequency distribution has **open-end classes,** Quartile Deviation will be the most appropriate measure of Dispersion,

Here: $\frac{N}{4} = \frac{300}{4} = 75$ so

Q_1 class is 95-100

Using the formula: $Q_1 = l + \frac{\frac{N}{4} - cf}{f} * h$

$Q_1 = 95 + \frac{75 - 67}{56} * 5$

$= 95 + \frac{8}{56} * 5 = 95 + \frac{40}{56} = 95 + \frac{5}{7} = 95.71$

Here: $\frac{3N}{4} = \frac{900}{4} = 225$ so Q_3 class is 105-110

Using the formula: $Q_3 = l + \frac{\frac{3N}{4} - cf}{f} * h$

$Q_3 = 105 + \frac{225 - 185}{75} * 5$

$= 105 + \frac{40}{75} * 5 = 105 + \frac{200}{75} = 105 + 2.67 = 107.67$

So Quartile Deviation $= \frac{Q_3 - Q_1}{2}$

$= \frac{107.67 - 95.71}{2} = ½ * 11.96 = 5.98$

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{107.67 - 95.71}{107.67 + 95.71} = \frac{11.96}{203.38} = 0.059$

7.6. Mean Deviation

Range and quartile deviations are not ideal measures of Dispersion as they are not based on all the observations of the given data. But mean (or average) deviation is ideal in this sense that it is based on all the given observations. It is computed as the arithmetic mean of the absolute deviations of the individual observations from the average of the given data. The average which is frequently used in computing the mean deviation is mean or median, though sometimes mode can also be used.

Formulae:

(i) For raw data: Mean Deviation (MD) about Mean $= \frac{\sum|Xi-\bar{x}|}{n}$

Where n = no. of items in the data set and $\bar{x}$ is the mean of the data set

(ii) For data arranged in a frequency distribution

Mean Deviation (MD) about Mean $= \frac{\sum fi*|Xi-\bar{x}|}{\sum fi}$

(iii) Coefficient of MD $= \frac{\text{Mean Deviation about Mean(MD)}}{\text{Mean}}$

(iv) For raw data: Mean Deviation (MD) about Median (Me) $= \frac{\sum|Xi-Me|}{n}$

Where n = no. of items in the data set and Me is the median of the data set

(ii) For data arranged in a frequency distribution

Mean Deviation (MD) about Median (Me) $= \frac{\sum fi*|Xi-Me|}{\sum fi}$

(iii) Coefficient of MD $= \frac{\text{Mean Deviation about Median(Me)}}{\text{Median(Me)}}$

7.6.1 Calculation of Mean Deviation

Ex6. : Find Mean Deviation from mean and median and corresponding coefficient of Mean Deviation

[A] Mean Deviation About Mean

xi	fi	Xi*fi	xi - $\bar{x}$ = xi – 7.6	$\|Xi-\bar{x}\|$ = $\|Xi-7.6\|$	fi*$\|Xi-\bar{x}\|$
6	5	30	= 6 - 7.6 = -1.6	1.6	= 5*1.6 = 8.0
7	4	28	= 7 - 7.6 = -0.6	0.6	= 4*0.6 = 2.4
8	4	32	= 8 - 7.6 = +0.4	0.4	= 4*0.4 = 1.6
9	3	27	= 9 - 7.6 = +1.4	1.4	= 3*1.4 = 4.2
10	2	20	= 10 - 7.6 = +2.4	2.4	= 2*2.4 = 4.8
Total $\sum$	18	137			21.0

Mean $(\bar{x}) = \frac{\sum fi * xi}{\sum fi} = \frac{137}{18} = 7.6$

Mean Deviation (MD) about Mean = $\frac{\sum fi * |Xi - \bar{x}|}{\sum fi} = \frac{21}{18} = 1.167$

Coefficient of MD = $\frac{\text{Mean Deviation about Mean(MD)}}{\text{Mean}} = \frac{1.167}{7.6} = 0.153$

[B] Mean Deviation About Median

xi	fi	Cf	xi - Med. = xi – 7.5	|Xi – Med.| = |Xi – 7.5|	fi*|Xi – Med.|
6	5	5	= 6 - 7.5 = -1.5	1.5	= 5*1.7 = 7.5
7	4	= 5 + 4 = 9	= 7 - 7.5 = - 0.5	0.5	= 4*0.5 = 2.0
8	4	= 9 + 4 = 13	= 8 - 7.5 = + 0.5	0.5	= 4*0.5 = 2.0
9	3	= 13 + 3 = 16	= 9 - 7.5 = + 1.5	1.5	= 3*1.5 = 4.5
10	2	= 16 + 2 = 18	= 10 - 7.5 = + 2.5	2.5	= 2*2.5 = 5.0
Total $\sum$	18				21.0

$\frac{N+1}{2} = \frac{18+1}{2} = 9.5$

So median $= 9.5^{th}$ item $= \frac{9\text{th item} + 10\text{th item}}{2} = \frac{7+8}{2} = 7.5$

Mean Deviation (MD) about Median = $\frac{\sum fi * |Xi - Med|}{\sum fi} = \frac{21}{18} = 1.167$

Coefficient of MD = $\frac{\text{Mean Deviation about Median}}{\text{Median}} = \frac{1.167}{7.5} = 0.156$

Ex.7: Height of Grade 10 Boys in a government school is given below, calculate MD and its Coefficient

[A] Mean Deviation about Median

Height (cm)	**No. of boys (f)**	**xi**	**Cumulative frequency**	xi - Med. = xi – 163.9	|Xi – Med.| = |Xi – 7.5|	fi*|Xi – Med.|
150 - 155	4	152.5	4	152.5 – 163.9 = -11.4	11.4	= 4*11.4= 45.6
155 - 160	7	157.5	11	157.5 – 163.9 = -6.4	6.4	= 7*6.4 = 44.8
160 - 165	**18**	162.5	**29**	162.5 – 163.9 = -1.4	1.4	= 18*1.4 = 25.2
165 - 170	11	167.5	40	167.5 – 163.9 = 3.6	3.6	= 11*3.6 = 39.6
170 - 175	6	172.5	46	172.5 – 163.9 = 8.6	8.6	= 6*8.6 = 51.6
175 - 180	4	177.5	50	177.5 – 163.9 = 13.6	13.6	= 4*13.6 =54.4
Total	50					261.2

Here $\frac{N}{2} = \frac{50}{2} = 25$ so the median class is 160-165

Using the formula: Median = $l + \frac{\frac{N}{2} - cf}{f} * h$

Median = $160 + \frac{25 - 11}{18} * 5$

$= 160 + \frac{14}{18} * 5 = 160 + \frac{70}{18} = 163.9$

Mean Deviation (MD) about Median = $\frac{\sum fi * |Xi - Med|}{\sum fi} = \frac{261.2}{50} = 5.22$

Coefficient of MD = $\frac{\text{Mean Deviation about Median}}{\text{Median}} = \frac{5.22}{163.9} = 0.0319$

[B] Mean Deviation About Mean

Height (cm)	**No. of boys (f)**	xi	Xi * fi	xi - Mean. = xi – 164.5	|Xi – Mean| = |Xi – 7.5|	fi*|Xi – Mean|
150 - 155	4	152.5	610.0	152.5 – 164.5 = -12.0	12.0	= 4*12 = 48
155 - 160	7	157.5	1102.5	157.5 – 164.5 = -7.0	7.0	= 7*7 = 49
160 - 165	18	162.5	2925.0	162.5 – 164.5 = -2.0	2.0	= 18*2 = 36
165 - 170	11	167.5	1842.5	167.5 – 164.5 = 3.0	3.0	= 11*3 = 33
170 - 175	6	172.5	1035.0	172.5 – 164.5 = 8.0	8.0	= 6*8 = 48
175 - 180	4	177.5	710.0	177.5 – 164.5 = 13.0	13.0	= 4*13 = 52
Total	50		8225.0			266

Mean $(\bar{x}) = \frac{\sum fi * xi}{\sum fi} = \frac{8225.0}{50} = 164.50$

Mean Deviation (MD) about Mean = $\frac{\sum fi * |Xi - \bar{x}|}{\sum fi} = \frac{266}{50} = 5.32$

Coefficient of MD = $\frac{\text{Mean Deviation about Mean(MD)}}{\text{Mean}} = \frac{5.32}{164.50} = 0.0323$

7.6.2. Characteristics of Mean/Average Deviation: An important property of Mean Deviation is that it has the minimum value when deviations are taken from median, i.e., Mean Deviation about median is the least. The relative measure corresponding to the mean deviation, called the **coefficient of mean deviation,** is obtained by dividing mean deviation by the particular average used in computing the mean deviation. Thus, if mean deviation has been computed from median, the coefficient of mean deviation shall be obtained by dividing the mean deviation by the median.

Mean deviation is based on all observations and hence takes into account the variability of each of the items in the data set. However, the practice of neglecting signs and taking absolute deviations makes it difficult to be treated algebraically. Although the average deviation is a good measure of variability, its use is limited.

7.6.3. *Merits and Demerits of Mean/Average Deviation*

Merits

(i) It is simple to understand and easy to calculate.

(ii) It is based on all the observations of a series. Thus, it shows the dispersion or scatter of the various items of a series from its central value.

(iii) It is not very much affected by the values of extreme items of a series.

(iv) It truly represents the average of deviations of the items of a series.

(v) It has practical usefulness in the field of business and commerce.

Demerits

(i) It is not rigidly defined in the sense that it is computed from any central value viz. Mean, Median, Mode etc. and thereby it can produce different results.

(ii) It violates the algebraic principle by ignoring the + and – signs while calculating the deviations of the different items from the central value of a series.

(iii) It is not capable of further algebraic treatment.

(iv) It is difficult to calculate when the actual value of an average comes out in fraction.

7.7. Standard Deviation

While computing the mean deviation we ignore the negative signs of the deviations of the items from the central tendency. This ignoring of signs, introduces some limitations on the measure. A mathematical solution for ignoring signs is squaring and then taking the under root. Accordingly Root **Mean Square Deviation or Standard Deviation** is calculated**.**

Like mean deviation, root mean square deviation can also be calculated by subtracting arithmetic mean or median or mode from the given values. Out of these three values, in every data, root mean square deviation about arithmetic mean is the least. So it is called Standard Deviation.

7.7.1. ***Meaning and calculation of Standard Deviation*****:** Standard deviation may be defined as the square root of the arithmetic mean of the squares of deviations of given observations from their arithmetic mean. It is usually denoted by the Greek letter σ (sigma). The major steps involved in the computation of standard deviation are:

I) Compute the arithmetic mean of the given series.

2) Calculate the deviations of various items from the arithmetic mean.

3) Compute the squares of all the individual deviations.

4) Total the squared deviations and divide the sum by the number of items *[Actually the sum is divided by the Degrees of Freedom particularly for small samples]*, it gives **variance**.

5) Square root of the resultant figure i.e. variance gives the standard deviation of the series.

7.7.2. Formulae and Examples: The various formulae for calculation of variance and Standard Deviation (SD) are discussed as under:

7.7.2.1 Formulae based on definition:

(a) Variance and standard Deviation for Raw Data

(i) Variance $(\sigma^2) = \frac{\sum(xi-\bar{x})^2}{n}$; (ii) Standard Deviation $(\sigma) = \sqrt{\frac{\sum(xi-\bar{x})^2}{n}}$

(b) Variance and standard Deviation for Frequency Table

(i) Variance $(\sigma^2) = \frac{\sum fi*(xi-\bar{x})^2}{\sum fi}$; (ii) Standard Deviation $(\sigma) = \sqrt{\frac{\sum fi*(xi-\bar{x})^2}{\sum fi}}$

Ex. 8: Find out standard deviation for a set of items: 3, 4, 8, 7, 9, 11.

x	$x - \bar{x}$ = x - 7	$(x - \bar{x})^2$
3	-4	16
4	-3	9
8	1	1
7	0	0
9	2	4
11	4	16
Sum =42		46

Mean $(\bar{x}) = \frac{\sum xi}{n} = \frac{42.0}{6} = 7.0$

Variance $(\sigma^2) = \frac{\sum(xi-\bar{x})^2}{n} = \frac{46}{6} = 7.667$

Standard Deviation$(\sigma) = \sqrt{\frac{\sum(xi-\bar{x})^2}{n}} =$

$\sqrt{7.667} = 2.769$

Ex.9: Height of Grade 10 Boys in a government school is given below, calculate Standard Deviation and variance

Height (cm)	**No. of boys (f)**	xi	xi * fi	xi - Mean = xi – 164.5	$(xi - Mean)^2$ = |Xi – 7.5|	$fi(xi - Mean)^2$
150 - 155	4	152.5	610.0	152.5 – 164.5 = -12.0	144.0	= 4*144= 576
155 - 160	7	157.5	1102.5	157.5 – 164.5 = -7.0	49.0	= 7*49 = 343
160 - 165	18	162.5	2925.0	162.5 – 164.5 = -2.0	4.0	= 18*4 = 72
165 - 170	11	167.5	1842.5	167.5 – 164.5 = 3.0	9.0	= 11*9 = 99
170 - 175	6	172.5	1035.0	172.5 – 164.5 = 8.0	64.0	= 6*64 = 384
175 - 180	4	177.5	710.0	177.5 – 164.5 = 13.0	169.0	= 4*169 =676
Total	50		8225.0			2150

Mean $(\bar{x}) = \frac{\sum fi * xi}{\sum fi} = \frac{8225.0}{50} = 164.50$

Variance $(\sigma^2) = \frac{\sum fi*(xi-\bar{x})^2}{\sum fi} = \frac{2150}{50} = 43$

Standard Deviation $(\sigma) = \sqrt{\frac{\sum fi*(xi-\bar{x})^2}{\sum fi}} = \sqrt{43} = 6.557$

7.7.2.2 ***Simple Formulae of Variance and standard deviation*****:** To make calculations simple, the following formulae for variance and standard deviations may be used; these are particularly helpful when the figures are in decimals.

(a) ***Variance and standard Deviation for Raw Data***

(i) Variance $(\sigma^2) = \frac{\sum xi^2}{n} - (\frac{\sum xi}{n})^2 = \frac{\sum xi^2}{n} - (\bar{x})^2$

(ii) Standard Deviation $(\sigma) = \sqrt{\frac{\sum xi^2}{n} - (\frac{\sum xi}{n})^2} = \sqrt{\frac{\sum xi^2}{n} - (\bar{x})^2}$

(b) ***Variance and standard Deviation for Frequency Table***

(i) Variance $(\sigma^2) = \frac{\sum fixi^2}{\sum fi} - (\frac{\sum fixi}{\sum fi})^2 = \frac{\sum fixi^2}{\sum fi} - (\bar{x})^2$

(ii) Standard Deviation $(\sigma) = \sqrt{\frac{\sum fixi^2}{\sum fi} - (\frac{\sum fixi}{\sum fi})^2} = \sqrt{\frac{\sum fixi^2}{\sum fi} - (\bar{x})^2}$

Ex. 10: Find out standard deviation

x	x^2
3	9
4	16
8	64
7	49
9	81
11	121
Sum =42	340

Variance $(\sigma^2) = \frac{\sum xi^2}{n} - (\frac{\sum xi}{n})^2 = \frac{340}{6} - (\frac{42}{6})^2$

$= 56.67 - 49 = 7.67$

Standard Deviation$(\sigma) = \sqrt{\frac{\sum xi^2}{n} - (\frac{\sum xi}{n})^2}$

$= \sqrt{7.67} = 2.769$

Ex. 11: Height of Grade 10 Boys in a government school is given below; calculate Standard Deviation and variance

Height (cm)	**No. of boys (f)**	xi	xi * fi	$fi*xi^2 = fixi * xi$
150 - 155	4	152.5	610.0	610.0 * 152.5 = 93025.00
155 - 160	7	157.5	1102.5	1102.5 * 157.5 = 173643.75
160 - 165	18	162.5	2925.0	2925.0 * 162.5 = 475312.50
165 - 170	11	167.5	1842.5	1842.5 * 167.5 = 308618.75
170 - 175	6	172.5	1035.0	1035.0 * 172.5 = 178537.50
175 - 180	4	177.5	710.0	710.0 * 177.5 = 126025.00
Total	50		8225.0	1355162.5

Mean $(\bar{x}) = \frac{\sum fi*xi}{\sum fi} = \frac{8225.0}{50} = 164.50$

Variance $(\sigma^2) = \frac{\sum fixi^2}{\sum fi} - (\frac{\sum fixi}{\sum fi})^2 = \frac{1355162.5}{50} - (\frac{8225.0}{50})^2 = 27103.25 - (164.50)^2$

$= 27103.25 - 27060.25 = 43$

Standard Deviation $(\sigma) = \sqrt{\frac{\sum fixi^2}{\sum fi} - (\frac{\sum fixi}{\sum fi})^2} = \sqrt{43} = 6.557$

7.7.2.3 ***Calculation of Variance and standard deviation: Shortcut Method***

(i) Variance $(\sigma^2) = \frac{\sum fidi^2}{\sum fi} - (\frac{\sum fidi}{\sum fi})^2$ where di = xi – a; a is the assumed mean.

(ii) Standard Deviation $(\sigma) = \sqrt{\frac{\sum fidi^2}{\sum fi} - (\frac{\sum fidi}{\sum fi})^2}$

Ex.12 Height of Grade 10 Boys in a government school is given below; calculate Standard Deviation and variance by short cut method

Marks (xi)	No. of boys (fi)	di = xi – a = xi - 7	fi * di	$fi*di^2$ = fidi * di
5	4	5- 7 = -2	-8	16
6	7	6 - 7 = -1	-7	07
7	18	7- 7 = 0	0	00
8	11	8- 7 = 1	11	11
9	6	9- 7 = 2	12	24
10	4	10 - 7 = 3	12	36
Total	50		20	94

Let assumed mean (a) = 7

Variance (σ^2) = $\frac{\sum fidi^2}{\sum fi}$ - $(\frac{\sum fidi}{\sum fi})^2$ where di = xi – a; a is the assumed mean.

$= \frac{94}{50} - (\frac{20}{50})^2 = 1.88 - (0.40)^2$

$1.88 - 0.16 = 1.72$

Standard Deviation (σ)= $\sqrt{\frac{\sum fidi^2}{\sum fi} - (\frac{\sum fidi}{\sum fi})^2} = \sqrt{1.72} = 1.311$

7.7.2.4 ***Calculation of Variance and standard deviation: Step Deviation Method***

(i) Variance (σ^2) = $\left[\frac{\sum fidi^2}{\sum fi} - (\frac{\sum fidi}{\sum fi})^2\right] * h^2$; Where di = $\frac{xi - a}{h}$ and h = class height

(ii) Standard Deviation (σ) = $\left[\sqrt{\frac{\sum fidi^2}{\sum fi} - (\frac{\sum fidi}{\sum fi})^2}\right] * h$

Ex.13: Height of Grade 10 Boys in a government school is given below; calculate Standard Deviation and variance by Step Deviation method

Height (cm)	**No. of boys (f)**	xi	$di = \frac{xi - a}{h} = \frac{xi - 162.5}{5}$	= fi * di	$fi*di^2 =$ fidi * di
150 - 155	4	152.5	-2	-8	16
155 - 160	7	157.5	-1	-7	07
160 - 165	18	162.5	0	0	00
165 - 170	11	167.5	1	11	11
170 - 175	6	172.5	2	12	24
175 - 180	4	177.5	3	12	36
Total	50			20	94

Let assumed mean (a) = 162.5

Variance $(\sigma^2) = \left[\frac{\sum fidi^2}{\sum fi} - (\frac{\sum fidi}{\sum fi})^2\right] * h^2$ where di = xi – a; a is the assumed mean.

$= \frac{94}{50} - (\frac{20}{50})^2 = [1.88 - (0.40)^2] * 5^2$

$= [1.88 - 0.16] * 25 = 1.72 * 25 = 43$

Standard Deviation $(\sigma) = \left[\sqrt{\frac{\sum fidi^2}{\sum fi} - (\frac{\sum fidi}{\sum fi})^2}\right] * h$

$= \sqrt{43} = 6.557$

7.7.3. ***Properties of Standard Deviation*****:** Following are some important properties of Standard Deviation:

1) The value of standard deviation remains the same if each of the observations in a series is increased or decreased by a constant value say 'a'. Thus, if y = d + a, where **'a'** is a constant quantity, then standard deviation of Y is equal to standard deviation of X. In other words, standard deviation is independent of change of origin.

2) If each observation of given series is multiplied or divided by a constant value, standard deviation will also be similarly affected. Thus, if y = ax, where 'a' is a constant, then SD(Y) = a*SD(X). Thus, standard deviation is not independent of the change of scale.

3) For a given set of observations, standard deviation is never less than mean deviation about arithmetic mean and quartile deviation. In fact mean deviation is 4/5 σ and Q. D. is 2/3 σ for normal data.

4) Root mean square deviation calculated about a value other than arithmetic mean will always be higher than that of the standard deviation.

7.7.4. Merits and Demerits of Standard Deviation

MERITS: Among all the measures of dispersion, standard deviation is considered the best because it possesses almost all the requisites of a good measure of dispersion. Standard deviation had the following merits:

i) It is rigidly defined and is based on all the observations of the series.

ii) The unique property which makes standard deviation superior to other measures of dispersion is that it is amenable to algebraic treatment. Thus, if we are given the number of observations, mean and standard deviation for each of several groups, we can easily calculate the standard deviation of the composite group.

iii) Standard deviation is least affected by the fluctuations of sampling.

***DEMERITS*:** The main demerits of standard deviation as a measure of dispersion are:

i) The major limitation of SD is that it cannot be used for comparing the dispersion of two or more series of observations given in different units. A coefficient of variation has to be defined for this purpose.

ii) The process of squaring deviations from mean and then taking the square-root of the mean of these squared deviations is complicated.

In fact this gives rise to another limitation i.e., standard deviation is very much affected by the extreme values. The process of squaring deviations give undue importance to large deviations from arithmetic mean and less importance to items which are nearer to mean.

iii) The standard deviation cannot be computed for a distribution with open-end classes.

7.8. Coefficient of Variation

The coefficient of variation, also known as coefficient of standard deviation expressed in percentages, is based on the ratio of the standard deviation to the arithmetic mean of a series. Thus, coefficient of variation may be expressed as:

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation}}{Arithmetic\ Mean} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

The coefficient of variation is a relative measure of dispersion and is expressed in the form of percentages. So it can be conveniently used for comparing the variability or dispersion between the two sets of the observations given in different units or in the same units having wide variations in the average value. It may thus, be used to measure or compare the precision of two or more sets of observations.

Coefficient of Variation may be used to compare relative variability in the following conditions:

(i) Variation of the same character in two or more different series; for e.g. variation in the heights of the persons staying in Mumbai and staying in Delhi

(ii) Variation of the two or more different characters in the same series; for e.g. variation in the heights and weights in the same group of persons

The series for which C.V. is less is **more consistent or less variable.**

Ex. 14: In two series (i) of adults aged 21 years and (ii) children less than 1 year old, following measures were obtained for the height. Find which series shows greater variation or less consistency.

Persons	Mean Height	SD
Adults	160 cm	10 cm
Children	60 cm	5 cm

Solution:

$$\text{Coefficient of Variation for Adults} = \frac{\text{Standard Deviation}}{Arithmetic\ Mean} \times 100$$

$$= \frac{10}{160} \times 100 = 100/16 = 6.25\%$$

$$\text{Coefficient of Variation for Children} = \frac{\text{Standard Deviation}}{Arithmetic\ Mean} \times 100$$

$$= \frac{5}{60} \times 100 = 50/6 = 8.33\%$$

The coefficient of variation of heights of children is more. Thus the heights of children is more variable or less consistent.

Ex. 15: The following is the record of goals scored by Team A in a football season.

No. of goals scored in a match:	0	1	2	**3**	4
Number of matches:	1	9	7	5	3

For Team B, the average number of goals scored per match was 2.5 with a standard deviation of 1.25 goals. Find which team is more consistent. [TRY]

7.9. Comparison of Measures of Dispersion

To make a right choice, it is necessary to know the relative features of the various measures of dispersion. A comparison of these measures is given below:

1) ***Type*:** Range and quartile deviation are measures of dispersion which give the spread of the data, while mean deviation and standard deviation are measures of dispersions which give the average of the deviations of items from some central tendency measure.

2) ***Difficulty Level*:** Range is simple and easy to understand. Quartile deviation and mean deviation are to some extent understandable but standard deviation is comparatively complicated and abstract.

3**)** ***Items****:* Range and quartile deviation do not take into account all the items in a series. However, mean deviation and standard deviation take into account all the items of the data set.

4) ***Mathematical Treatment***: Range, quartile deviation and mean deviation are not capable of further mathematical treatment. Standard deviation is capable of further mathematical treatment.

5) ***Extreme Values***: Quartile deviation is not affected by the extreme or abnormal values of the items in a series. Between mean deviation and standard deviation, mean deviation is less affected by the extreme values. Range depends only on extreme items.

6) **Open-end Class**: Range, mean deviation and standard deviation cannot be calculated in case of a frequency distribution with open end classes. Quartile Deviation may be calculated for Open-end Class.

7) ***Reliability***: Standard deviation is considered to be the most reliable and dependable measure of dispersion. Range is least reliable while quartile deviation or mean deviation are not much reliable measures of dispersion. In fact standard deviation is least affected by sampling errors.

8) ***Use****:* Standard deviation is considered to be the best measure of dispersion. It possesses all the qualities and properties of a good and reliable measure of dispersion. Hence it is widely used in statistical analysis. Range, quartile deviation and mean deviation are not so popular and are used only in limited but appropriate cases.

7.10. Box plot (Box-and-whisker diagram or plot)

Box Plot is a convenient way of graphically depicting groups of numerical data through their five-number summaries viz. the smallest observation, lower quartile (Q_1), median (Q_2), upper quartile (Q_3) and the largest observation. A box plot may also indicate which observations, if any, might be considered outliers.

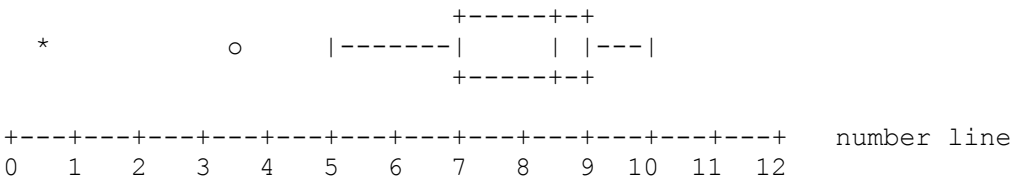
Box plots can be useful to display differences between populations without making any assumptions of the underlying statistical distribution. The spacing between the different parts of the box help indicate the degree of dispersion (spread) & skewness in the data and identify outliers. Box plots can be drawn either horizontally or vertically. For a data set, a horizontal box plot can be constructed as:

- Calculate the first quartile, median and third quartile
- Calculate the inter-quartile range (IQR) by subtracting the first quartile from the third quartile.
- Construct a box above the number line bounded on the left by the first quartile and on the right by the third quartile
- Indicate where the median lies inside the box with the presence of a symbol or a line dividing the box at the median value.
- The mean value of the data can also be labeled with a point.
- Any data observation which lies more than 1.5 * IQR lower than the first quartile or 1.5 * IQR higher than the third quartile is considered an outlier. Indicate the smallest value that is not an

outlier by connecting it to the box with a horizontal line or "whisker". Likewise, connect the largest value that is not an outlier to the box by a "whisker".

- Indicate outliers by open and closed dots. "Extreme" outliers, or those which lie more than three times the IQR to the left and right from the first and third quartiles respectively, are indicated by the presence of a closed dot. "
- Add an appropriate label to the number line and title the boxplot.

A Box Plot may look like:



For this data set:

- smallest non-outlier observation = 5 (left "whisker")
- lower (first) quartile ($Q1$) = 7
- median (second quartile) = 8.5
- upper (third) quartile ($Q3$) = 9
- largest non-outlier observation = 10 (right "whisker")
- interquartile range, IQR = $Q3 - Q1 = 2$
- the value 3.5 is a "mild" outlier, between 1.5*IQR and 3*IQR below $Q1$
- the value 0.5 is an "extreme" outlier, more than 3*IQR below $Q1$
- the data are skewed to the left (*negatively skewed*)

7.10.1. *Terms relevant to understanding box-plot*

Median (Q_2): The median (middle quartile) marks the mid-point of the data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less.

***Inter-quartile range*:** The middle “box” represents the middle 50% of scores for the group. The range of scores from lower to upper quartile is referred to as the inter-quartile range. The middle 50% of scores fall within the inter-quartile range.

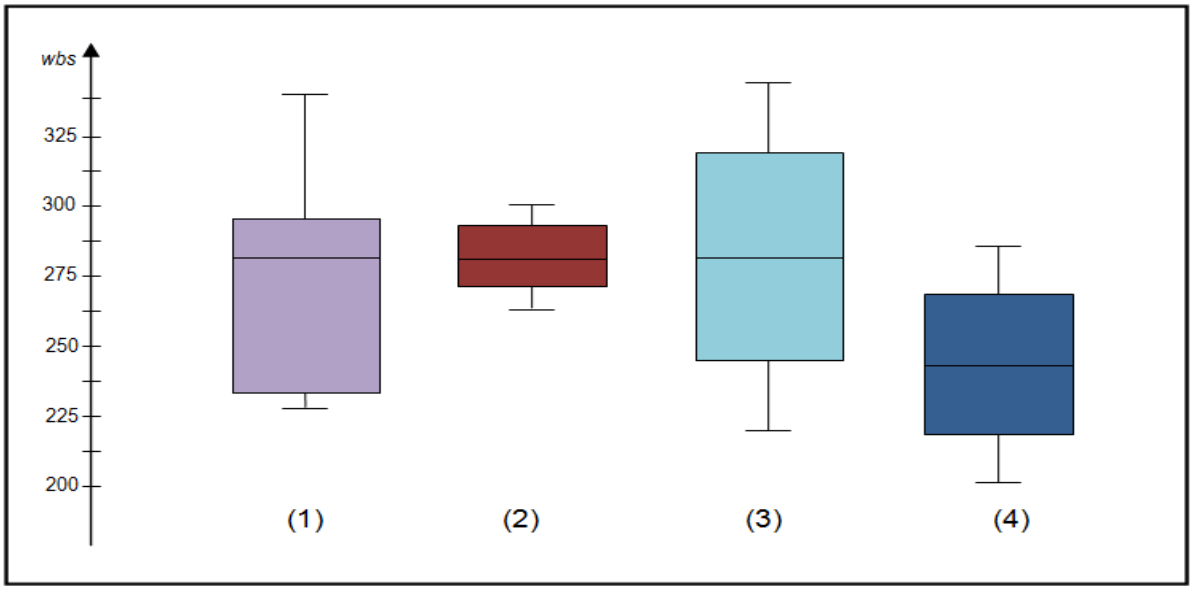
Upper quartile (Q_3): Seventy-five percent of the scores fall below the upper quartile.

Lower quartile (Q_1): Twenty-five percent of scores fall below the lower quartile.

***Whiskers*:** The upper and lower whiskers represent scores outside the middle 50%.

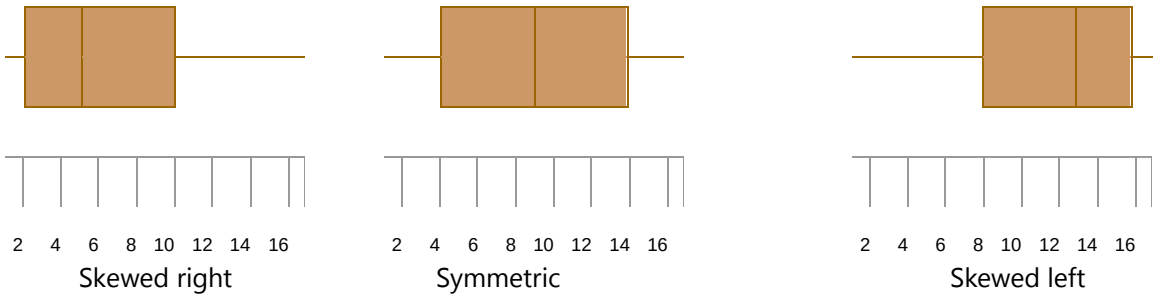
7.10.2 Interpreting Box plots

Box plots are used to show overall patterns of response for a group. They provide a useful way to visualise the range and other characteristics of responses for a large group. The diagram below shows 4 different box plot shapes and positions based on the responses of the students of four different sections of an institute:



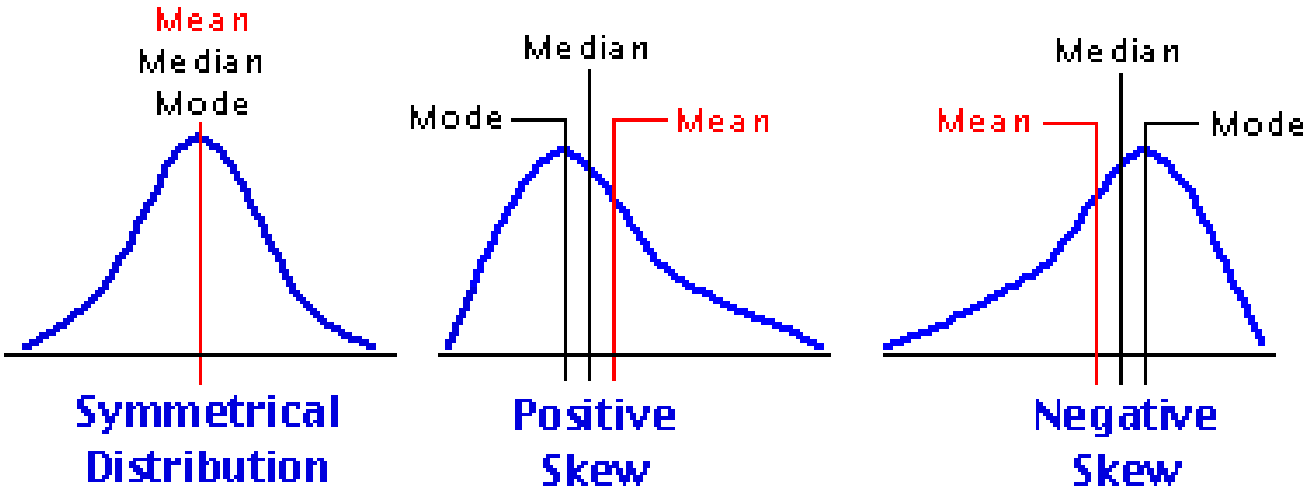
7.10.3. Interpretation of the four box-plots

- ***Box plot (2) is comparatively short*** suggesting that overall students have a high level of agreement with each other i.e. variability is much less.
- ***Box plots (1) and (3) are comparatively tall*** suggesting that students hold quite different opinions about this aspect with each other i.e. variability is high.
- ***One box plot is much higher or lower than another*** – compare (3) and (4) – This could suggest a difference between groups. The values in group 3 are higher (on an average) than the values in group 4, accordingly the median (average) in group 3 is higher than the median in group 4.
- ***The 4 sections of the box plot are uneven in size*** – See example (1). This shows that many students have similar views at certain parts of the scale, but in other parts of the scale students are more variable in their views. The long upper whisker in the example means that students' views are varied amongst the most positive quartile group, and very similar for the least positive quartile group.
- ***Same median, different distribution*** – See examples (1), (2), and (3). The medians (which is an average) are all at the same level. However the box plots in these examples show very different distributions (variations/skewness) of views.
- It is always important to consider the pattern of the whole distribution of responses in a box plot.



7.11. Measure of Skewness

Shape of data is measured by Skewness and Kurtosis. Skewness measures **lack of symmetry of data**. Positive or right skewed means longer right tail and Negative or left skewed means longer left tail.



7.11.1. Karl Pearson's skewness coefficient

Karl Pearson's skewness is defined as $S_k = \frac{mean - mode}{standard\ deviation}$ or $S_k = \frac{3(mean - median)}{standard\ deviation}$

Theoretically, the values of S_k vary between ± 3. However, normally these values lie between ± 1. Any threshold or rule of thumb is arbitrary, but if the skewness is greater than 1.0 (or less than -1.0), the skewness is substantial and the distribution is said to be far from symmetrical. In this case instead of Mean and Standard Deviation we should preferably use median and Quartile Deviation as measure of Central Tendency and Measure of Dispersion respectively. Alternately we may treat the outliers separately.

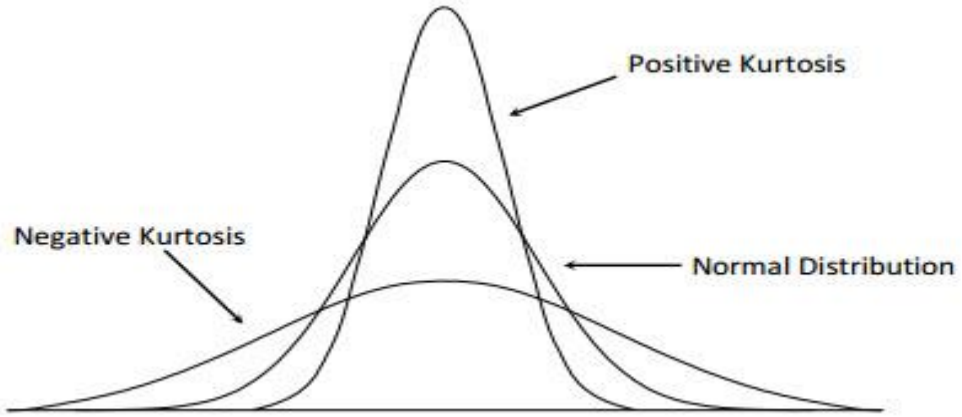
Examples of Skewness: Salary data is often positively skewed as many employees in a company make relatively little, while increasingly few people make very high salaries. Failure rate data is often left skewed. Consider light bulbs - very few will burn out right away, the vast majority lasting for quite a long time. Age of pensioners is positively skewed because there may be very less people with age 90-100 or above and a very large number of pensioners with age 60-70 years.

7.12. Kurtosis – measure of relative flatness or peakedness of data set

If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution as it is clear from the following figure in which all the three curves are symmetrical about the mean and have the same range. In addition to these measures, we should know one more measure which Prof. Karl Pearson calls as the 'Convexity of the Frequency Curve' or Kurtosis.

Kurtosis enables us to have an idea about the 'flatness or peakedness' of the frequency curve. It is measured by the coefficient β_2 (estimate b_2) or its derivation γ_2.

A standard normal distribution has mean $\mu = 0$; Std. Deviation $\sigma = 1$, skewness = 0, and kurtosis = 0.



Karl Pearson's coefficient of Kurtosis is given by the formula

$$b_2 = \frac{\Sigma(X_i - \bar{X})^4/n}{(\Sigma(X_i - \bar{X})^2/n)^2},$$

And $\gamma_2 = b_2 - 3$.

In probability theory and statistics, **kurtosis** is a measure of the "tailedness" of the probability distribution. Curve which is neither flat nor peaked is called the normal or **mesokurtic** curve; for such a curve $\beta_2 = 3$, and $\gamma_2 = 0$. Curve which is flatter than the normal curve is known as **platykurtic;** for such a curve $\beta_2 < 3$ and $\gamma_2 < 0$; they are said to have negative kurtosis. Curve which is more peaked than the normal curve is called **leplokurtic;** for such a curve $\beta_2 > 3$ and $\gamma_2 > 0$; they are said to have positive kurtosis.

Any threshold or **rule of thumb** is arbitrary, but if the kurtosis is greater than 2.0 (or less than -2.0), the kurtosis is substantial.

7.13. Comparison among dispersion, skewness and kurtosis: Dispersion, Skewness and Kurtosis are different characteristics of frequency distribution. Dispersion studies the scatter of the items round a central value. It does not show the extent to which deviations cluster below an average or above it. Skewness tells us about the clustering of the deviations above and below a measure of central tendency. Kurtosis studies the concentration of the items at the central part of a series. If items concentrate too much at the centre, the curve becomes 'LEPTOKURTIC – Positive Kurtosis' and if the concentration at the centre is comparatively less, the curve becomes 'PLATYKURTIC – Negative Kurtosis'.

Multiple choice questions: choose the correct answer

1. The measures of dispersion based on every item of the series is:

(a) range

(b) standard deviation

(c) quartile deviation

(d) None of these [Ans. (b)]

2. One of the measures of dispersion which is more useful in case of open-end distributions:

(a) range

(b) mean deviation

(c) standard deviation

(d) quartile deviation [Ans. (d)]

3. Standard deviation is always computed from:

(a) mean

(b) mode

(c) median

(d) Geometric Mean [Ans. (a)]

4. Which of the following measures is least affected by extreme items:

(a) quartile deviation

(b) range

(c) standard deviation

(d) mean deviation . [Ans. (a)]

5. Mean deviation is:

(a) less than S.D.

(b) more than S.D.

(c) not related to S.D.

(d) equal to Standard Deviation [Ans. (a)]

6. Coefficient of variation is given by:

(a) $\frac{\text{SD}}{Mean}$ x 100

(b) $\frac{Mean}{\text{SD}}$ x 100

(c) $\frac{\text{QD}}{Mean}$ x 100

(d) $\frac{\text{SD}}{Range}$ x 100 [Ans. (a)]

7. Which one of the following measurement does not divide a set of observations into equal parts?

(a) Quartiles

(b) Standard Deviations

(c) Percentiles

(d) Deciles [Ans. (b)]

8. Which one is the not measure of dispersion.

(a) The Range

(b) 50th Percentile

(c) Inter-Quartile Range

(d) Variance [Ans. (b)]

9. In a statistical analysis, dispersion of data is measured by:

(a) Geometric Mean

(b) Arithmetic Mean

(c) Mode

(d) Range [Ans. (d)]

10. Which of the following is the most commonly used as measure of dispersion?

(a) Mean deviation

(b) Standard deviation

(c) Range

(d) Quartile deviation [Ans. (b)]

11. Which measure is not dependent on the value of each score?

(a) Mean

(b) Variance

(c) Standard Deviation

(d) Range [Ans. (d)]

12. All are measures of dispersion except:

(a) Range

(b) Mean Deviation

(c) Standard Deviation

(d) Median [Ans. (d)]

13. Which of the following is called semi-interquartile range?

(a) $Q_1 - Q_3$

(b) $Q_3 - Q_1$

(c) $\frac{1}{2}(Q_3+Q_1)$

(d) $\frac{1}{2}(Q_3- Q_1)$ [Ans. (d)]

14. Which of the following is called Inter-quartile range?

(a) $\frac{1}{2}(Q_3- Q_1)$

(b) $Q_1 - Q_2$

(c) $Q_3 - Q_1$

(d) $Q_2 - Q_3$ [Ans. (c)]

15. The coefficient of Quartile Deviation is :

(a) $\frac{Q3 - Q1}{2}$

(b) $\frac{Q3+ Q1}{Q3- Q1}$

(c) $\frac{Q3 - Q1}{Q3+ Q1}$

(d) $\frac{Q3+ Q1}{2}$ [Ans. (c)]

16. Which of the following is an example of a relative measure of dispersion?

(a) Variance

(b) Mean Deviation

(c) Standard deviation

(d) Coefficient of Variation [Ans. (d)]

17. The square of the variance of a distribution is the:

(a) Standard deviation

(b) Mean Deviation

(c) Absolute dispersion

(d) None of the above [Ans. (d)]

18. Standard deviation (σ) is calculated by one of the following formulae:

(a) $\sqrt{\frac{\sum fixi^2}{\sum fi} - (\frac{\sum fixi}{\sum fi})^2}$

(b) $\sqrt{\frac{\sum fixi^2}{\sum fi} - \left(\frac{\sum fidi}{\sum fi}\right)^2}$

(c) $\sqrt{\frac{\sum fixi^2}{\sum fi} + \left(\frac{\sum fidi}{\sum fi}\right)^2}$

(d) $\sqrt{\frac{\sum fixi^2}{\sum pi} - (\frac{\sum fixi}{\sum fi})^2}$ [Ans. (a)]

19. 20 babies are born in a hospital on the same day. Each weighs 2.5 kg, the standard deviation is:

(a) 1

(b) 0

(c) 2.5

(d) 5 [Ans. (b)]

20. The general formula for an estimate of a variance is sum of squared deviations divided by:

(a) Population size

(b) Sample size

(c) Degrees of freedom

(d) Level of significance [Ans. (c)]

21. If a sample consists of three observations, 13, 14, 15 an estimate of the population standard deviation equals as:

(a) $\sqrt{\frac{2}{3}}$

(b) Zero

(c) $\sqrt{\frac{1}{3}}$

(d) 1 [Ans. (a)]

22. In a series of 300 boys, the mean systolic blood pressure was 120 mm of Hg and the standard deviation was found to be 20. The coefficient of variation is :

(a) 16.7%

(b) 8.3%

(c) 40%

(d) 30% [Ans. (a)]

23. Dispersion of a group of data can be graphically represented by:

(a) Normal curve

(b) Lorenz curve

(c) Curvilinear curve

(d) Cumulative frequency curve [Ans. (b)]

24. The semi- interquartile range is most closely related to the:

(a) Mean

(b) Median

(c) Mode

(d) Geometric Mean [Ans. (b)]

25. The properties of the standard deviation are most closely related to those of the :

(a) Mean

(b) Median

(c) Mode

(d) Range [Ans. (a)]

26. In one score in distribution is changed to another value, it is certain that:

(a) The range has changed

(b) The standard deviation has changed

(c) The semi-interquartile range has changed

(d) The Question does not make sense [Ans. (b)]

TRY

Q1. What are the requisites of a good measure of Dispersion?

Q2. What for a Measure of variation used? **(b)** What are various measures of dispersion? Name them.

Q3. Compare Mean Deviation and Std. Deviation.

Q4. Give the merits and demerits of Standard Deviation (SD)

Q5. Define coefficient of variation

Q6. What do you mean by range? Give its merits and demerits.

Q7. Calculate coefficient of variation from the following data:

Life in hrs.	0-50	50-100	100-150	150-200	200-250
No. of bulbs	2	8	60	25	5

Q8. Calculate standard deviation:

Variable	20-30	30-40	40-50	50-60	60-70	70-80	80-90
frequency	3	61	132	153	140	51	2

Q9. Calculate Range and Coefficient of range from the following data:

Marks	25	30	35	60	75
No. of Students	4	8	10	6	3

Q10. Calculate quartile deviation and its coefficient from the data given below:

Variable	20-30	30-40	40-50	50-60	60-70
frequency	4	6	8	6	4

Q11. Calculate the mean deviation about mean for the following distribution:

Classes	20-40	40-80	80-100	100-120	120-140
frequency	3	6	20	12	9

Q12. Find the standard deviation from the following data:

Marks	0-10	10-20	20-30	30-40	40-50
No. of Students	10	15	10	10	5

Q13. Compute third quartile from the given data: 11, 12, 14, 18, 22, 26, 30 (1)

Chapter 8
Correlation and Regression

The word correlation is used in everyday life to denote some form of association. We might say that there is a correlation between foggy days and attacks of breathlessness. However, in statistical terms correlation denotes association **between two quantitative variables**. It measures the degree/extent of the relationship between variables.

8.1. Types of Correlation

According to the direction of change in variables there are two types of correlation:

(a) *Positive Correlation*: Correlation between two variables is said to be positive if the values of one variable increase (or decrease) as the values of other variable also increase (or decrease). Some examples of positive correlation are correlation between: (i) Heights and weights of group of persons; (ii) House hold income and expenditure; (iii) Expenditure on advertising and sales revenue; it is because as the expenditure on advertising increases, sales revenue also increases. Thus, the change is in the same direction. Hence the correlation is positive.

(b) *Negative Correlation*: Correlation between two variables is said to be negative if the values of one variable increase (or decrease) as the values of other variable decrease (or increase). Some examples of negative correlations are correlation between (i) Volume and pressure of gas (ii) Price and demand of goods; (iii) Literacy and poverty in a country; it is because as the literacy level goes up in a country, the poverty in the country decreases. Thus the change in the values of two variables is in opposite direction. Therefore, the correlation between Literacy and poverty in a country is negative.

8.1.1 *Karl Pearson's Coefficient of Correlation*: It measures the degree/extent of linear relationship*/ association between two variables e.g. height and weight; income and income tax, etc.

The Karl Pearson's coefficient of correlation is denoted by 'r' and is described as $r = \frac{Cov(X,Y)}{\sigma_x * \sigma_y}$

Where: $Cov(x,y) = \frac{\sum (X - \bar{X}) * (Y - \bar{Y})}{N}$, called covariance between X & Y

σ_x = Standard deviation of series X = $\sqrt{\frac{\sum (X - \bar{X})^2}{N}}$

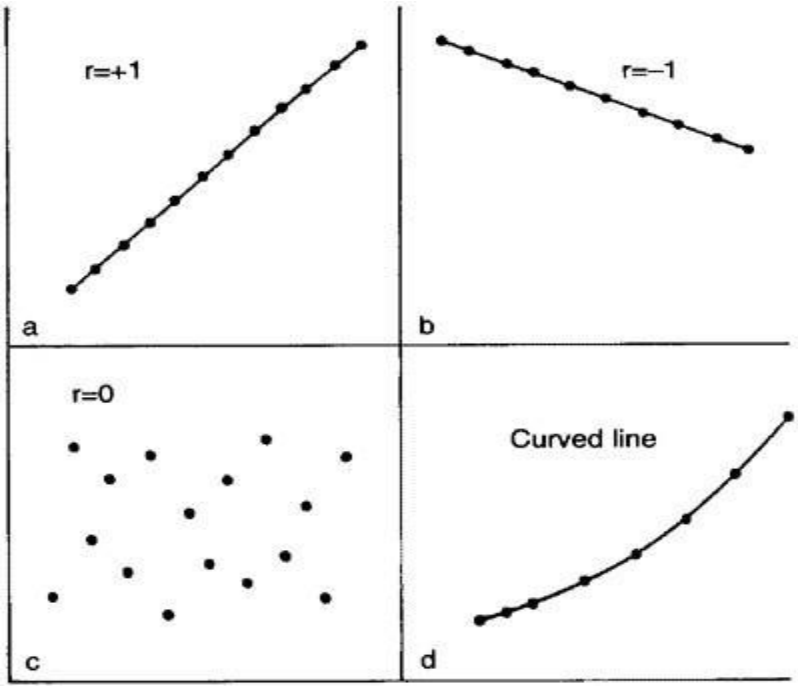
σ_y = Standard deviation of series Y = $\sqrt{\frac{\sum (Y - \bar{Y})^2}{N}}$

N = Number of pairs of observations

A simple formula for calculation: $r = \frac{N*\sum XY - \sum X * \sum Y}{\sqrt{[N*\sum X^2 - (\sum X)^2]} * \sqrt{[N*\sum Y^2 - (\sum Y)^2]}}$

Correlation lies between - 1 and + 1. When one variable increases as the other increases the correlation is **positive** and when one variable decreases as the other increases it is **negative**. Complete absence of correlation is represented by 0. Figure below gives graphical representations of some correlations.

*A linear relationship means a relationship that can be represented by a *line on a graph paper* (the word "linear" means "a line"). In case of linear relationship, the highest power of the variables x and y is 1; e.g. y = 3x + 7



(a) Perfect Positive Correlation
(b) Perfect Negative Correlation
(c) Absence of correlation
(d) Non-Linear (Curvilinear) Correlation

8.1.2 Calculation of the correlation coefficient: Table below indicates a set of 15 values [N = 15] of variables X and Y; calculate the correlation coefficient between the two variables and interpret the result.

X	Y	X^2	Y^2	XY
0.5	41	0.25	1681	20.50
0.7	55	0.49	3025	38.50
2.5	41	6.25	1681	102.50
4.1	39	16.81	1521	159.90
5.9	50	34.81	2500	295.00
6.1	32	37.21	1024	195.20
7.0	41	49.00	1681	287.00
8.2	42	67.24	1764	344.40
10.0	26	100.00	676	260.00
10.1	35	102.01	1225	353.50
10.9	25	118.81	625	272.50
11.5	31	132.25	961	356.50
12.1	31	146.41	961	375.10
14.1	29	198.81	841	408.90
15.0	23	225.00	529	345.00
118.70	541.00	1235.35	20695.00	3814.5

$$r = \frac{N*\sum XY - \sum X * \sum Y}{\sqrt{[N*\sum X^2 - (\sum X)^2]} * \sqrt{[N*\sum Y^2 - (\sum Y)^2]}}$$

$$= \frac{15*3814.5 - 118.7*541}{\sqrt{[15*1235.35 - (118.7)^2]} * \sqrt{[15*20695 - (541)^2]}}$$

$$= \frac{57217.5 - 64216.7}{\sqrt{18530.25 - 14089.69} * \sqrt{310425 - 292681}}$$

$$= \frac{-6999.2}{\sqrt{4440.56} * \sqrt{17744}}$$

$$= \frac{-6999.2}{66.637 * 133.207}$$

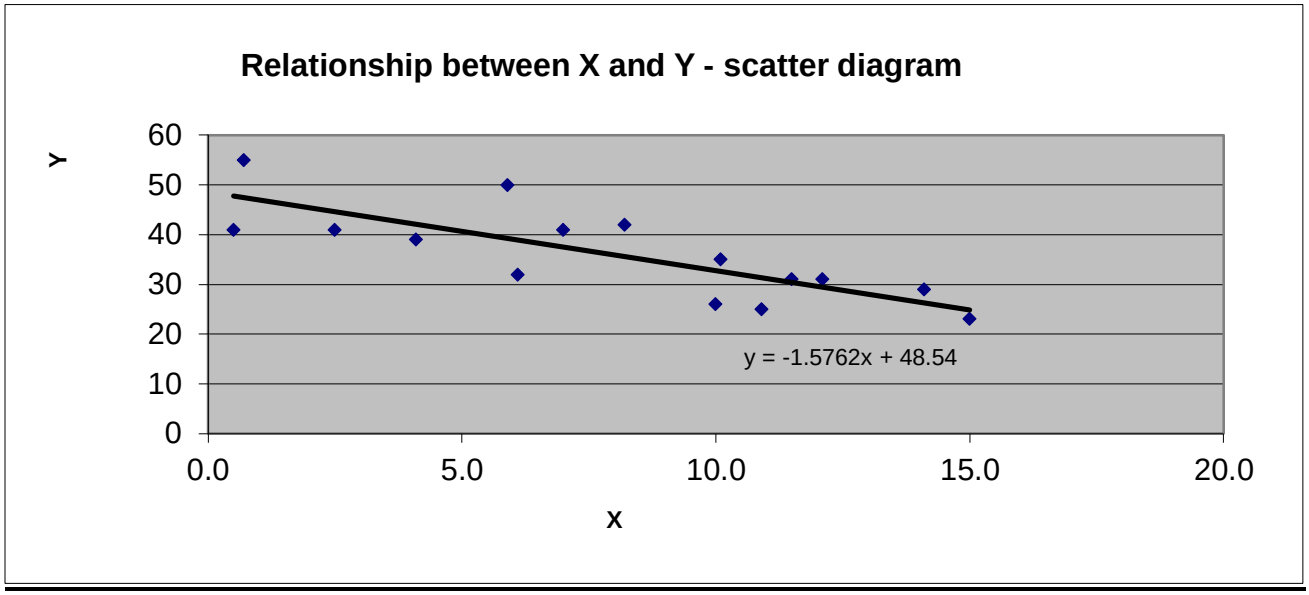
$$= \frac{-6999.2}{8876.489}$$

= - 0.7885

Interpretation: r = (-) 0.79; means a high negative correlation indicating that if the value of variable X increases, the value of variable Y will decrease and vice versa.

8.1.3. Scatter or dot diagram – Rough Measure of Correlation

Scatter or dot diagram is a graphic presentation to show the degree/extent of relationship (correlation) between two variables; it is also called Correlation Diagram. To plot a scatter diagram the given values of variables X and Y are marked in the XY plane in the form of dots. Normally independent variable is plotted along the X - axis and the dependent variable along the Y-axis.



Interpretation of the Scatter diagram: All the points lie around a down-sloping line so the correlation between the variables X and Y is high and negative.

8.1.4. *Short-Cut Method for Calculation of Correlation Coefficient*

When values of variables are big and actual means of variables X and Y are not whole number, calculation of correlation coefficient is somewhat cumbersome and we can use shortcut method in which deviations are taken from assumed mean for the variables X and Y. Formula for correlation coefficient by shortcut method:

$$r = \frac{n*\sum dxdy - \sum dx * \sum dy}{\sqrt{[n* \sum dx^2 - (\sum dx)^2]} * \sqrt{[n* \sum dy^2 - (\sum dy)^2]}};$$

where n = No. of pairs of observations,

a = Assumed mean of X, b = Assumed mean of Y,

dx = x – a : Sum of deviation from assumed mean a in x-series,

dy = y – b : Sum of deviation from assumed mean b in y-series.

Note: a and b are so chosen that they approximately lie in the middle of x and y series respectively.

Ex. 2: Calculate correlation coefficient from the following data by shortcut method:

x	10	12	14	18	20
y	5	6	7	10	12

Solution

x	y	dx= x - 14	dy = y - 8	dx*dy	dx^2	dy^2
10	5	-4	-3	12	16	9
12	6	-2	-2	4	4	4
14	7	0	-1	0	0	1
18	10	4	2	8	16	4
20	12	6	4	24	36	16
74	**40**	**4**	**0**	**48**	**72**	**34**

Using $r = \frac{n*\sum dxdy - \sum dx * \sum dy}{\sqrt{[n* \sum dx^2 - (\sum dx)^2]} * \sqrt{[n* \sum dy^2 - (\sum dy)^2]}}$; n = 5

$= \frac{5*48-4*0}{\sqrt{[5* 72 - (4)^2]} * \sqrt{[5* 34 - (0)^2]}}$

$= \frac{240 - 0}{\sqrt{360 -16} * \sqrt{170-0}}$

$= \frac{240}{\sqrt{344} * \sqrt{170}} = \frac{240}{18.55 * 13.04}$

$= \frac{240}{241.89} = 0.992$

8.1.5. *Interpretation of Coefficient of Correlation 'r'*

The coefficient of correlation describes not only the magnitude of correlation but also its direction. Thus **r = + 1** means Perfect Positive Correlation or complete agreement in the same direction; **r = - 1** means Perfect Negative Correlation i.e. complete agreement in the opposite direction and **r = 0** means No Linear relation.

As a rule of thumb, correlation coefficients between 0.00 and 0.25 are considered weak, between 0.25 and 0.75 moderate and between 0.75 and 1.00 high. Value of r = + 0.80 means correlation is strong and positive i.e. variables X and Y have **strong direct relationship**. Value - 0.26 means correlation is weak and negative i.e. variables X and Y have **weak inverse relationship**.

Value of correlation coefficient	**Correlation is**
+1	Perfect Positive Correlation
-1	Perfect Negative Correlation
0	No Correlation
0 to 0.25	Weak Positive Correlation
0.75 to (+1)	Strong Positive Correlation
(-) 0.25 to 0	Weak Negative Correlation
(-) 0.75 to (−1)	Strong Negative Correlation

8.1.*6. Properties of Correlation Coefficient*

(a) Correlation coefficient lies between −1 and +1.

(b) Correlation coefficient is independent of change of origin and scale; it means if a number is added to or subtracted from all the given values of the data, the correlation is not affected [Change of Origin]. Similarly if all the items of the data set are multiplied or divided by the same number, the correlation is not affected. [Change of Scale]

(c) If X and Y are two independent variables then correlation coefficient between X and Y is zero.

8.2. Spearman's Rank Correlation Coefficient(ρ)

Rank correlation is used when variables under consideration are not capable of quantitative measurement but can be arranged in serial order i.e. when we are dealing with qualitative characteristics (attributes) like honesty, beauty, morality, etc. It is based on the ranks of the items rather than actual values. However, it can be used even with the actual values after ordering/ranking them. Its examples are (i) correlation between honesty & wisdom (ii) finding out the degree of agreement between scores (ranks) given to the different Departments of an Organization based on their audit risks by two auditors. (iii) Correlation between Ranking of trainees at the beginning (x) and at the end (y) of a certain course.

Formula for Spearman Rank correlation coefficient is $(\rho) = 1 - \frac{6\sum D^2}{N(N^2-1)}$

D = Difference between the ranks of two items

N = The number of observations.

8.2.1. *Computation of Spearman's Rank Correlation Coefficient*

i) Give ranks to the values of items of both the series. Generally the item with the highest value is ranked 1 and then the others are given ranks 2, 3, 4, . . . etc. according to their values in the decreasing order.

ii) Find out the difference $D = R_1 - R_2$ where R_1 = Rank of x and R_2 = Rank of y [Note that $\Sigma D = 0$ (always)] for each pair of ranks.

iii) Calculate D^2 for each rank and then find ΣD^2

iv) Apply the formula.

8.2.2. *Spearman's rank correlation when there is a tie between two or more item*s: In case there is a tie i.e. same values are repeated for a variable; give the average rank to items getting the same rank. If m be the number of items of equal ranks, the factor $(m^3-m)/12$ is added to ΣD^2. If there are more than one such cases/items then this factor is added as many times as the number of such equal rank cases, so we have:

$$\rho = 1 - \frac{6\{\sum D^2 + \frac{1}{12}(m_1{}^3 - m_1) + \frac{1}{12}(m_2{}^3 - m_2) + \frac{1}{12}(m_3{}^3 - m_3) + -----\}}{N(N^2-1)}$$

Where N = total number of pairs of items and m_1, m_2, m_3 etc. indicate the number of times ranks are repeated.

Calculation of Rank Correlation [No repetition of Ranks]

Sl. No.	Rank X = R_1	Rank Y = R_2	$D = R_1 - R_2$	D^2
1	1	3	-2	4
2	3	1	2	4
3	7	4	3	9
4	5	5	0	0
5	4	6	-2	4
6	6	9	-3	9
7	2	7	-5	25
8	10	8	2	4
9	9	10	-1	1
10	8	2	6	36
Sum ∑			0	96

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6*96}{10(10^2-1)}$$

$$= 1 - \frac{6*96}{10*99} = 0.4181$$

8.2.3. *Merits and Demerits of Rank Correlation Coefficient*

Merits

(i) Spearman's rank correlation coefficient can be interpreted in the same way as the Karl Pearson's correlation coefficient;

(ii) It is easy to understand and easy to calculate;

(iii) For finding out the association between qualitative characteristics, rank correlation coefficient is the only method;

(iv) Rank correlation does not require the assumption of the normality of the population from which the sample observations are taken.

Demerits

1. Correlation coefficient can be calculated for bivariate frequency distribution but rank correlation coefficient cannot be calculated for bivariate frequency distribution; and
2. If the number of pairs of items n > 30, this formula is time consuming.

8.3. The Regression Analysis

Correlation Analysis is concerned with measuring the strength of the relationship between variables and measures the degree/extent of the relationship between variables. While Regression Analysis is used to ascertain the probable form of the relationship between variables with the ultimate objective to predict or estimate the value of one variable corresponding to a given value of other variable(s). The regression Analysis is often more useful than the correlation coefficient as it enables us to predict value of variable y for a given value of x and vice versa. For e.g. if we have regression equation between tax and income; for a given income we can find out the tax amount. Similarly if we have regression equation between no. of teachers and their salary, we can find out the salary bill of a school/Districts if we know the number of teachers.

There are two types of variables in regression analysis. The variable which is used for prediction is called independent variable. It is also known as regressor or predictor or explanatory variable. The variable whose value is predicted by the independent variable is called dependent variable. It is also known as regressed or explained variable.

8.3.1. ***Types of Regression*** - If scatter diagram shows some relationship between independent variable X and dependent variable Y, then the scatter diagram will be more or less concentrated round a curve, which may be called the curve of regression. When the curve is a straight line, it is known as line of regression and the regression is said to be linear regression. If the relationship between dependent and independent variables is not a straight line but curve of any other type then regression is known as nonlinear or curvilinear.

Regression can also be classified according to number of variables being used. If only two variables are used this is considered as simple regression whereas the involvement of more than two variables in regression is categorized as multiple regression.

8.3.2 Regression Lines: The relationship between two variables can be represented by a simple equation called the regression equation; in the simplest case this can be a straight line. For two variables X and Y, we have two regression lines i.e. regression line of X on Y and regression line of Y on X. The regression line of X on Y gives the most probable values/estimate of X for given values of Y whereas the regression line of Y on X gives the most probable values of Y for given values of X.

When the two sets of observations increase or decrease together (positive correlation) the line slopes upwards from left to right; when one set decreases as the other increases (negative correlation) the line slopes downwards from left to right.

(a) ***Regression Equation of Y on X*****:** Regression Equation of Y on X is expressed as: **$Y = a + byx*X$**

Where byx is regression coefficient of y on x; Y is a dependent variable; X independent variable and 'a' is Y intercept, the value of Y when X is zero.

Using **least square criteria*** the Regression Equation of Y on X becomes: $\mathbf{(Y - \bar{Y}) = r * \frac{\sigma_y}{\sigma_x} (X - \bar{X})}$

Where r is the correlation coefficient between variables X and Y, σ_x is the standard deviation (SD) of variable X; σ_y is the SD of variable Y while $\bar{x}$ and $\bar{y}$ are the means of variables X and Y respectively.

***The Least Square Criterion (the line of best fit):** "The sum of the squared deviations (Vertical for regression line of y on x / Horizontal for regression line of x on y) of the observed data points from the line of best fit is smaller than the sum of the squared deviations of the data points from any other line.
i.e. $\sum (X - Xe)^2$ or $\sum (Y - Ye)^2$ is minimum where Xe & Ye are the estimated values of variables X & Y based on regression lines.

(b) *Regression Equation of X on Y*: Regression Equation of X on Y is expressed as: **$X = a + bxy*Y;$**

Where bxy is regression coefficient of x on y; X is a dependent variable, Y independent variable and 'a' is X intercept, the value of X when Y is zero.

Using least square criteria the Regression Equation of X on Y becomes: $\mathbf{(X - \bar{X}) = r * \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})}$

Where r is the correlation coefficient between the variables X and Y, σ_x is the standard deviation of variable X, σ_y is the Standard Deviation of variable Y while $\bar{x}$ *and* $\bar{y}$ are the means of variables X and Y respectively.

(c*) Properties of Regression Lines:*

(i) When there is perfect correlation i.e. when $r = \pm 1$, the two regression lines coincide (become one line).

(ii) The farther the two regression lines from each other, the lesser is the degree of correlation.

(iii) The two regression lines always meet at point $(\bar{x}, \bar{y})$

(d) *Regression Coefficients:* $\boldsymbol{byx = r \frac{\sigma_y}{\sigma_x}}$ *and* $\boldsymbol{bxy = r \frac{\sigma_x}{\sigma_y}}$ *are* called the regression coefficients of Y on X and X on Y respectively. Properties of regression coefficients are:

(i) byx * bxy = r^2 (Coefficient of determination) i.e. Geometric mean of the regression coefficients is equal to the correlation between the variables.

(ii) If one of the regression coefficients is greater than one, then other must be less than one.

(ii) bxy, byx and r must have the same sign.

(e) *Normal Equation Method for Regression Lines*

(i*) Regression Equation of Y on X*: Let the Regression Equation of Y on X is: $y = a + b x$; normal equations to obtain the values of a and b are

$\Sigma y = n a + b \Sigma x$ ----------(1)

$\Sigma(x*y) = a \Sigma x + b \Sigma x^2$ ---------(2)

By solving equations (1) & (2) simultaneously, values of a & b can be determined to get regression line.

(ii) *Regression Equation of X on* Y: Let the Regression Equation of X on Y be: $x = a + b y$

Normal equations to get the values of a and b are

$\Sigma x = n a + b \Sigma y$ ----------(1)

$\Sigma(x*y) = a \Sigma y + b \Sigma y^2$ ---------(2)

By solving equations (1) & (2) simultaneously, values of a & b can be determined to get regression line.

8.4. Distinction between Correlation and Regression

Both correlation and regression have important role in the study of relationship between the variables but there are some distinctions between them as explained under:

(i) Correlation analysis studies the extent or degree of relationship between two or more variables while regression analysis tries to find out the type of relationship between two or more variables.

(ii) Correlation has limited application because it gives the strength of relationship while the purpose of regression is to "predict" the value of the dependent variable for the given values of one or more independent variables.

(iii) Correlation does not consider the concept of dependent and independent variables while in regression analysis one variable is considered as dependent variable and other(s) is/are as independent variable(s).

8.5. Coefficient of Determination

Ratio of explained variance to total variance gives the Coefficient of Determination, mathematically:

Coefficient of Determination (r^2) = Explained Variance / Total variance

8.5.1. Properties of Coefficient of Determination:

(i) Gives the percentage variation in the dependent variable accounted for by the independent variable.

(ii) Given by the square of the correlation coefficient, i.e., $r^{2.}$ For e.g. if r = 0.8, the coefficient of determination (r^2) = 0.64 so only 64% of the variation in the dependent variable has been explained by the independent variable and remaining 36% variation in the dependent variable is due to others factors.

(iii) Always non–negative and doesn't tell about direction of relationship.

(iv) It is more useful than Coefficient of Correlation.

8.6. Multiple and Partial Correlation and Regression

Regression analysis is a statistical technique which allows us to assess the relationship between one dependent variable (DV) and one or several independent variables (IVs). Multiple Regression is an extension of bi-variate regression in which several independent variables (IVs) are combined to predict the value of the dependent variable (DV). Regression may be assessed in a variety of manners, such as:

8.6.1. Partial regression and correlation - The relationship between two variables may be unclear because of the confounding (confusing) influence of another variable; for example, if we calculate the correlation between mental age and height in children 1 to 10 years of age; we may find a high correlation. Does that mean that height causes intelligence? The key factor is age, not height. Once we control for age, the relationship between height and mental age becomes trivial/insignificant.

One study was conducted to determine whether the number of hours spent studying was related to grades, the researchers found a negative correlation. This does not mean that studying fewer results in higher grades. Once they controlled for intelligence, the researchers found a significant positive relation between grades and hours of study.

Partial correlation may be written as $r_{12.3}$. This indicates that we are measuring the correlation between variables 1 and 2 with the effect of variable 3 removed from both the variables 1 and 2. Consider the example college grades (Variable-1), hours of study (Variable-2) and intelligence (variable 3). If we use partial correlation to measure correlation between hours of study and grades; the correlation between intelligence and grades (r_{13}) and the correlation between intelligence and hours of study (r_{23}) is removed. The confounding influence of intelligence is thus removed statistically, and the relationship between grades and hours of study can be measured accurately. This relationship would give the partial correlation.

8.6.2. Multiple regression and correlation*:* It studies the combined effect of all the variables acting on the dependent variable. The multivariate regression equation is of the form:

$Y = A + B_1X_1 + B_2X_2 + ... + B_nX_n + E$; where:

Y = The predicted value or the Dependent Variable (DV),

A = The Y intercept, the value of Y when all X's are zero,

X's = The values of the Independent Variables (IVs),

B = The coefficients of regression and E = An error term.

The goal of the regression is to derive the regression coefficients, or beta coefficients (B values). The beta coefficients allow the computation of reasonable Y values with the regression equation. When reporting multiple correlations, R^2 rather than R is often presented.

Although regression analysis reveals relationships between variables this does not imply that the relationships are causal. Demonstration of causality is not a statistical problem, but an experimental and logical problem.

8.7. Correlation and Causation

It is a common error to confuse correlation and causation. All that correlation shows is that the two variables are associated. There may be a third variable, **a confounding variable** that is related to both of them and indicates a correlation between the two variables. The relationship between two variables may be unclear because of the confounding influence of another variable; for example, if we calculate the correlation between mental age and height in children 1 to 10 years of age; we may find a high correlation. The key factor is age, not height. Once we control for age, the relationship between height and mental age becomes trivial/insignificant.

Another example is; as ice cream sales increase, the rate of drowning deaths increases sharply. Therefore, ice cream consumption causes drowning. This example fails to recognize the importance of time and temperature in ice cream sales and swimming. Ice cream is sold during the hot summer months at a much greater rate than during colder times and it is during these hot summer months that people are more likely to engage in activities involving water, such as swimming. The increased drowning deaths are simply caused by more exposure to water-based activities, not ice cream sales.

Ex.1: The following data pertain to the chlorine residual in a swimming pool at various times after it has been treated with chemicals:

No. of hours	2	4	6	8	10	12
Chlorine residual (parts per million)	1.8	1.5	1.4	1.1	1.1	0.9

(a) Calculate the Karl Pearson's coefficient of correlation between No. of hours and Chlorine residual.

(b) Fit a least squares line which will enable us to predict the chlorine residual content in terms of the numbers of hours since the pool has been treated with chemicals.

(c) Use the equation of the least squares line to estimate the chlorine residual in the pool 5 hours after it has been treated with chemicals.

Solution: Let No. of hours = x and Chlorine residual (parts per million) = y;

Sl. No.	x	Y	X^2	Y^2	XY
1	2	1.8	4	3.24	3.6
2	4	1.5	16	2.25	6.0
3	6	1.4	36	1.96	8.4
4	8	1.1	64	1.21	8.8
5	10	1.1	100	1.21	11.0
6	12	0.9	144	0.81	10.8
Total	**42**	**7.8**	**364**	**10.68**	**48.6**

$$(\bar{x}) = \frac{\sum xi}{n} = \frac{42.0}{6} = 7.0;$$

$$(\bar{y}) = \frac{\sum yi}{n} = \frac{7.8}{6} = 1.3$$

$$\sigma_x = \sqrt{\frac{\sum xi^2}{n} - (\frac{\sum xi}{n})^2} = \sqrt{\frac{364}{6} - (\frac{42}{6})^2}$$

$$= \sqrt{60.67 - 49} = 3.42$$

$$\sigma_y = \sqrt{\frac{10.68}{6} - (\frac{7.8}{6})^2}$$

$$= \sqrt{1.78 - 1.69} = 0.3$$

$$r = \frac{N*\sum XY - \sum X * \sum Y}{\sqrt{[N*\sum X^2 - (\sum X)^2]} * \sqrt{[N*\sum Y^2 - (\sum Y)^2]}}$$

$$= \frac{6*48.6-42*7.8}{\sqrt{[6*364 - (42)^2]} * \sqrt{[6*10.68 - (7.8)^2]}} = \frac{291.6-327.6}{\sqrt{2184-1764} * \sqrt{64.08-60.84}}$$

$$= \frac{-36}{\sqrt{420} * \sqrt{3.24}}$$

$$= \frac{-36}{20.49 * 1.8}$$

$$= \frac{-36}{36.88}$$

= -0.976

To predict the chlorine residual content we take y = Chlorine residual as dependent variable so the line of regression of y on X would be used:

$$(Y - \bar{Y}) = r * \frac{\sigma y}{\sigma x} (X - \bar{X})$$

Putting the values calculated above we have:

$$Y - 1.3 = -0.976 * \frac{0.3}{3.42} (X - 7)$$

$$Y = 1.3 - 0.0856 (X - 7)$$

$$Y = 1.3 - 0.0856*X + 0.599$$

$Y = 1.899 - 0.0856x$ is the required line of regression or line of best fit.

To estimate the chlorine residual in the pool 5 hours after it has been treated with chemicals, put x = 5

So we have $Y = 1.899 - 0.0856*5$

$= 1.899 - 0.428$

$= 1.471$ parts per million

Multiple choice questions: choose the correct answer

1. An association between two variables can be known by calculating:

(a) Coefficient of correlation

(b) Coefficient of regression

(c) Standard error of mean

(d) Standard deviation [Ans. (a)]

2. Coefficient of correlation lies from:

(a) 0 to +1

(b) -1 to 0

(c) -1 to +1

(d) -1.1 to +1.1 [Ans. (c)]

3. If r = -1 what does it mean?

(a) Weak correlation

(b) No correlation

(c) Perfect negative correlation

(d) Wrong calculation [Ans. (c)]

4. The correlation between Price and supply level of a commodity is 0.2. What does it indicate?

(a) Strong correlation

(b) Weak correlation

(c) Moderate correlation

(d) None of the above [Ans. (b)]

5. If r = 1.02 what does it indicate?

(a) Very strong +ve correlation

(b) Moderately +ve correlation

(c) Weak correlation

(d) Calculation of 'r' is wrong [Ans. (d)]

6. Not true about correlation coefficient 'r' is:

(a) Tells about Audit Risk of the Department

(b) -1 correlation shows perfect linear relationship

(c) Tells association between two variables

(d) Does not tell about causation [Ans. (a)]

7. If the correlation between height and weight is very strong. What will be the possible value of coefficient of correlation?

(a) 0.2

(b) 0.1

(c) 0.9

(d) -0.9 [Ans. (c)]

8. A correlation between two variables measures the degree to which they are:

(a) Mutually exclusive

(b) Casually related

(c) Positively skewed

(d) Associated [Ans. (d)]

9. Formula for Rank Correlation Coefficient 'ρ' is:

(a) $1 - \frac{6\sum D^2}{N(N^2-1)}$

(b) $1 - \frac{6\sum D^2}{N(N^2+1)}$

(c) $1 + \frac{6\sum D^2}{N(N^2-1)}$

(d) $1 + \frac{6\sum D^2}{N(N^2+1)}$ [Ans. (a)]

10. If $\sum D^2 = 0$, the value of rank correlation 'ρ' is

(a) 0

(b) + 1

(c) 2

(d) -1 [Ans. (b)]

11. The unit of correlation coefficient between height in feet and weight in kg is:

(a) Kg/feet

(b) Percentage

(c) Non-existent

(d) Feet/kg [Ans. (c)]

12. If r_{xy} is positive the relation between X and Y is of the type:

(a) When Y increases X increases

(b) When Y decreases X increases

(c) When Y increases X does not change

(d) Anything is possible [Ans. (a)]

13. Of the following three measures which can measure any type (linear and non-linear) of relationship?

(a) Karl Pearson's coefficient of correlation

(b) Spearman's rank correlation

(c) Scatter diagram

(d) Both (i) and (ii) are correct [Ans. (c)]

14. If precisely measured data are available the simple correlation coefficient is:

(a) More accurate than rank correlation coefficient

(b) Less accurate than rank correlation coefficient

(c) As accurate as the rank correlation coefficient

(d) Can be more or less accurate [Ans. (a)]

TRY

Q1. Raw material used in the production of a synthetic fiber is stored in a place which has no humidity control. Measurements of the relative humidity and the moisture content of samples of the raw material (both in percentages) on 12 days yielded the following results:

Humidity	46	53	37	42	34	29	60	44	41	48	33	40
Moisture Content	12	14	11	13	10	8	17	12	10	15	9	13

(a) Calculate the Karl Pearson's coefficient of correlation between Humidity and Moisture Content.

(b) Fit a least squares line which will enable us to predict the moisture content in terms of the relative humidity.

(c) Use the result to estimate (predict) the moisture content when the relative humidity is 38 percent.

Q2. The following data pertain to X, the amount of fertilizer (in kgs.) which a farmer applies to his soil, and Y, his yield of wheat (in quintals per acre):

X	112	92	72	66	112	88	42	126	72	52	28
Y	33	28	38	17	35	31	8	37	32	20	17

Assuming that the data can be looked upon as a random sample from a bivariate normal population, calculate r. Also, draw a scatter diagram of these paired data and judge whether the assumption seems reasonable.

Q3. Ranking of 10 trainees at the beginning (x) and at the end (y) of a certain course are given below:

Trainees	A	B	C	D	E	F	G	H	I	J
x	1	6	3	9	5	2	7	10	8	4
y	6	8	3	7	2	1	5	9	4	10

Calculate spearman's rank correlation coefficient. **[ρ=0.394]**

Q4. 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. **[ρ=0.95]**

Student 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

Q5. Comment: The correlation coefficient between the accidents in a particular year and the babies born in that year was found to be 0.8, which seems to be quite high.

Q6. The production manager of a company maintains that the flow time in days (y), depends on the number of operations (x) to be performed. The following data given the necessary information:

x	2	2	3	4	4	5	6	6	7	7
y	8	13	14	11	20	10	22	26	22	25

Plot a scatter diagram. Calculate the value the Karl Pearson's Correlation Coefficient. **[r (x, y) =0.78]**

Q7. From the following data of the age of husband and the age of wife, form two regression lines and calculate the husband's age when the wife's age is 16.

Husband's age	36	23	27	28	28	29	30	31	33	35
Wife's age	29	18	20	22	27	21	29	27	29	28

[Husband's age: x; Wife's age: y; y = 0.95x - 3.5; x = 0.8y + 10; 22.8yrs.]

Q8. The following table gives the ages and blood pressure of 10 women.

Age(x)	56	42	36	47	49	42	60	72	63	55
Blood Pressure (y)	147	125	118	128	145	140	155	160	149	150

(i) Find the correlation coefficient between X and Y.

(ii) Determine the least square regression equation of Y on X.

(iii) Estimate the blood pressure of a woman whose age is 45 years.

[(i) r = 0.89, (ii) y = 83.758 + 1.11x (iii) When x=45, y=134]

Q9. The following table gives the normal weight of a baby during the first six months of life:

Age in months	0	2	3	5	6
Weight in lbs.	5	7	8	10	12

Estimate the weight of a baby at the age of 4 months. **[9.2982 lbs]**

Q10. Does correlation imply causation? Explain with the help of an example.

[Hint: No. Correlation only implies co-variation.]

Q11. When rank correlation is more precise than simple correlation coefficient?

[Hint: When precise measurements of the variables are not possible.]

Chapter 9
Sampling Techniques and Estimation

9.1 Introduction

Process of Selection of a sample from a population to generate precise and valid estimates of population parameter(s) like mean or proportion is called Sampling while process of collecting relevant information/data in respect of each and every member/unit of the population is called Census.

Some obvious questions for sampling studies are what should be the ideal sample size, how to select the sample, what observational methods to use and what measurements to record. These issues are scientifically addressed in ***statistical sampling***.

In the basic statistical sampling setup, the population consists of a finite number of units –such as transactions, households, Gram Panchyats, vouchers, etc. The population size is normally denoted by N. With each unit of the population a value of variable of interest is associated, it may be referred to as x-value of the unit; for e.g. value of voucher/transaction, in the selected Districts number of schools without principal, value of tax obtained from a particular assesse. The units in the population are identifiable and may be labeled with numbers 1, 2, ….., N. A sample of the units from the population is selected and observed/Audited.

The data collected consist of the **x-values for each unit in the sample**, together with the unit's label/Identification. The procedure by which the sample of units is selected from the population is called ***sampling design***. The usual inference problem in sampling is to estimate some summary characteristics (Parameter) of interest of the population, such as ***mean, proportion*** **or total** after observing only a sample.

Sampling has broadly been categorized into two types namely Statistical (or Probability Sampling) and Non- statistical (or Non – Probability) Sampling.

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 possess the characteristic of the large group. This is the principal which governs sampling results.*

9.1.1 Advantages of sampling: Sampling has the following advantages:

(a) Less expensive: If data are collected for the entire population, the cost may be quite high. A sample is a small proportion of a population. Obviously the cost will be lower if the data are collected for a part of population.

(b) Saves Time: Use of sampling takes less time than census technique. Even data entry, tabulation, analysis etc., saves time in case of a sample than that of a population.

(c) Scope of sampling: To study the whole population would be impractical in a few cases. Some populations are so large that before the census is completed, the population would have changed. For e.g. to test blood, the entire blood can't be taken from the body.

(d) Accuracy of data is high: A sample permits a high degree of accuracy since a careful execution of field work is possible. Ultimately, the results of sampling studies turn out to be sufficiently accurate.

(e) Suitable in case of limited resources: The resources available within an organization are usually limited so studying the entire universe is not a viable proposition. So if limited resources exist, use of sampling is the only strategy for conducting a survey or audit.

(f) Other advantages: If we use sampling; it is possible to determine the extent of error. Further, the Non Sampling errors are likely to be less; even the Census Results are verified by sampling.

9.1.2 Disadvantages of sampling: Sampling has the following disadvantages:

(a) Chances of bias: The serious limitation of the sampling method is that it may have biased selection and thereby leads to wrong conclusions. Bias arises when the method of selection of sample is faulty. Small samples properly selected may be much more reliable than large samples poorly selected.

(b) Difficulties in selecting a truly representative sample: A sample produces reliable and accurate results only when it is representative of the whole group. Selection of a truly representative sample is difficult when the phenomena under study are of complex nature and there is considerable variability in the units.

(c) Lack of adequate subject knowledge: Use of sampling method requires adequate subject specific knowledge. When the researcher lacks specialized knowledge in sampling, he may commit serious mistakes. Consequently, the results of the study will be misleading.

(d) Changeability of units: Some of the units of sample may not cooperate with the researcher and some others may be inaccessible. Because of these problems, all the units of the sample may not be covered. The selected cases may have to be replaced by other cases; it may reduce the efficiency of the Sampling Results

(e) Impossibility of sampling: Deriving a representative sample is difficult, when the universe is too small or too heterogeneous. In this case, census study is the only alternative. Moreover, in studies requiring a very high standard of accuracy, the sampling method may be unsuitable.

Note: If time & money are not important factors and if population under consideration is not too large, census (Audit of all the records/units/GPs/Districts) is better than any sampling method

9.1.3 Various Terms used in Sampling: In order to appreciate the use of sampling, one must be familiar with the following terms:

(i) Population: Entire group of people/objects (vouchers, bills, audit entities) to which the researcher/auditor wishes to generalize the study/audit findings.

(ii) Sampling: Process of Selection of a sample from a population to generate precise and valid estimates of population parameter(s) like mean or proportion.

(iii) Census (100% enumeration): Process of collecting relevant information/data in respect of each and every member/unit of the population.

(iv) Statistical Inference: Drawing Conclusions (Inferences) about a population based on an examination/audit of sample(s) taken from the population.

(v) Sample: A sample is a part of the population, selected by the investigator/auditor as its representative to estimate information on certain characteristics of the original population

(vi) Sampling unit (Basic sampling unit): It is one of the units into which an aggregate (population) is divided for the purpose of selection of a sample; for e.g. Vouchers, cheques, bills, districts, schools, contracts, etc.

(vii) Sampling frame: List of all the sampling units **in the population** constitute Sampling Frame.

(viii) Sampling scheme: Method used to select sampling units from the sampling frame.

(ix) Parameter: Population characteristic like average, proportion based on all the units in the population; it is constant/fixed.

(x) Statistic: Sample characteristic like average, proportion based on sample values; it varies from sample to sample.

9.1.4 Sampling and Non Sampling Errors

(a) Sampling Errors: The errors due to use of sampling instead of 100% enumeration are called Sampling Errors; these errors are present to some extent in every sample design as no sample is a perfect mirror image of the population. Use of proper sample design and sufficient sample size reduce sampling error and increase the efficiency of the estimate(s) obtained by sampling. However, even the best sample can't eliminate the sampling errors entirely. It may be known that:

- The estimates of population parameter obtained vary from sample to sample.
- The sampling variance; like variance of average or proportion is the measure of variability of a sample estimator (a rule for calculating an estimate).
- The square root of the variance of the sample estimator is called the standard error (or the sampling error) of the estimator.
- The lesser the value of standard error of the estimator, the more efficient would be the estimator.

(b) Non Sampling Errors: Non-Sampling Errors in Audit include any misjudgement or mistake by the auditor that may lead to incorrect conclusion(s) based on audit. They occur even if the entire population is examined. By careful planning and supervision and by using appropriate audit technique non sampling errors can be reduced but they can't be eliminated. Some of the cases of non-sampling errors in audit are:

- Selecting inappropriate audit procedures to achieve specific objectives. For e.g. an auditor checks controller's signature on voucher and not the disbursement approval.
- Auditor may fail to recognize misstatements (errors) included in documents that (s) he examines; it may happen due to lack of experience or due to carelessness of the Auditor?
- Selecting inappropriate population for e.g. selecting only BPL households for audit of LPG subsidy scheme.
- Auditor makes an error in evaluation (say totalling mistake or skipping some vouchers containing error) of the results.

9.2 Sampling in Audit

In the early stages of the development of audit, it was not an uncommon practice for an auditor to perform a 100% examination of the entities and their records. However, as the economy grew, it became clear that 100 % examination of the tremendous volume of entries was unjustified and uneconomical. This led to the development of the test/sample check approach [Sampling Approach], which is both widely accepted and widely used in audit.

When sampling became a widely accepted tool in audit; another concept called '**Risk Assessment'** came almost simultaneously, which allowed auditor to focus on risky areas **through an objective analysis** of available information about the auditee unit. The main idea of risk analysis is to **identify risky areas in an objective way** so that the auditor can focus more on risky areas and optimally use available resources to meet overall audit objectives. Sampling based only on risk assessment may be **non-statistical** sampling.

9.3 Non-Statistical Sampling

A type of sampling in which, the units in the population do not have a known probability of being included in the sample is called Non-Statistical Sampling; it is also called non-probability Sampling. It is Subjective or Biased method of selecting a sample. It is used when (i) the number of elements in the population is either unknown or units in the population cannot be identified or (ii) there are time/ resource constraints and the statistical sample can't be drawn.

9.3.1 Why use Non-Statistical Sampling in Audit: Non-statistical sampling is less rigorous than Statistical Sampling then why would anybody specifically the Auditor use non-Statistical Sampling**.** The non- Statistical Sampling is used primarily because it is often less costly and time consuming still it can be as effective in achieving the audit objectives as the Statistical Sampling. It has been explained as under:

(i) Lower Training Costs: It usually takes less time to learn non-statistical Sampling approaches and this results in lower training costs.

(ii) Ease of Implementation: Because non-statistical approaches are less complex, they are generally easier and quicker to apply in the field. Also, the reduced complexity makes it less likely that the methods will be wrongly applied by the audit staff.

(iii) Impracticality of Random-Based Selection: In some cases it is not practical or not economical to apply random-based selection. A population of source documents may be large and unnumbered i.e. a list of units in the population (Sampling Frame) may not be available. For e.g. it may be possible that in a large village, the list of households is not available so to carry out the statistical sampling, the list of households needs to be prepared first which may be very much time consuming.

(iv) Proposed Adjustment Based on Qualitative Analysis: The increased precision of a statistical estimate is often not needed because **the proposed audit adjustment is based on the auditor's qualitative analysis of sample results rather than mathematical calculations**.

The use of qualitative analysis rather than quantification of sample results is often the most efficient and effective approach in Audit. Even when statistical sampling is used, the auditor may use sample sizes that are not large enough to reliably estimate the amount of misstatement. The sample results are used to identify situations in which the risk of material misstatement is unacceptably large. Once the situation is identified, the auditor relies on qualitative analysis to determine a proposed adjustment in sample results.

9.3.2 Dis-advantages of non-Statistical Sampling: Not representative of the population and it is not possible to (i) assess the validity of estimates (ii) Determine sample size objectively.

9.3.3 Methods of Non-Statistical Sampling for Audit: Some non-statistical sampling plans which can be used in Audit are as under:

(a) Block Selection: A block sample includes all the items of a selected time period. For example the selection of all the vouchers received in a treasury during the months of March and June. In this case, sample includes only two sampled items out of 12. A sample with so few items is generally not adequate to reach a reasonable audit conclusion for the entire year.

Block selection should be used with caution because valid references cannot be made beyond the period or block examined. If block sampling is used, many blocks should be selected to help minimize sampling risk.

(b) Haphazard Selection: A haphazard sample consists of sampling units without any conscious bias, that is, without any special reason for including or omitting items from the sample. It does not consist of sampling units in a careless manner; rather, units are selected in a manner that they can be expected to be representative of the population. The key to haphazard selection is to avoid being biased by the nature, size, appearance, or location of items. For example, if the auditor selects vouchers/files from a cabinet of drawers, the auditor should not select items only from the middle of the drawers.

With haphazard selection, there is no means of calculating the probability that a particular item will be selected for inclusion in the sample. As a result the method is not acceptable for statistical sampling. It does not mean that auditors should not use it for selecting items. Haphazard selection may be used when the auditor believes that it would produce a fairly representative sample and estimation of the parameter is not required.

(c) Judgmental Selection: Judgmental sample selection is based on the auditor's sound and seasoned judgment. Note that if the auditor only selects large or unusual items from the population or uses some other judgmental criterion for selection, the selection method has a **conscious bias** and cannot be considered a representative selection method. Three basic issues determine which items are selected:

(i) Period of items: A sufficient number of older accounts are included to provide adequate audit coverage.

(ii) Relative risk: Items prone to error due to their nature are given special attention.

(iii) Representativeness: Besides value and risk considerations, the auditor should be satisfied that the sample provides coverage over all types of items in the population.

(iv) Value of item: Generally it is believed that, the larger the value of an item; in terms of money value, the larger will be the risk. So a larger value item is given more importance/weight while selecting a sample.

9.4. Statistical Sampling

The essential features of statistical sampling are: (i) The sample items should have a known probability of selection which may not be equal (ii) The sample results should be evaluated in accordance with probability theory. Statistical Sampling is also called **Probability Sampling**.

Just because one of these requirements is met does not mean that the application is statistical. For example, if random number method is employed to select the sample, it will not lead to statistical sampling if no attempt is made to evaluate sample findings mathematically [condition no. (ii)].

Statistical sampling allows the auditors to optimize the sample size given the risk they are willing to accept. In this way both - over auditing and under auditing can be avoided. Moreover, it enables auditors to estimate/project the results about the population with a known reliability.

The users of the audit report expect fairness in the selection procedure and transparency in reporting while the auditors are interested in the audit findings. The **statistical sampling**, which provides ***estimates*** along with the ***reliability*** of the estimates, provides scientific solution to these problems. The audit reports based on this scientific approach are defensible. This enhances the acceptability and effectiveness of audit reports.

9.4.1 Advantages and Dis-advantages of statistical sampling:

Advantages:

- ❖ It provides estimates free from personal bias
- ❖ It permits application of objective methods of minimizing error under the resource constraints.
- ❖ It allows to draw valid conclusions about the population.
- ❖ It helps the auditor to design an efficient sampling plan.
- ❖ It helps the auditor to measure the sufficiency of evidence obtained
- ❖ It helps the auditor to quantify sampling risk

Dis-advantages:

- ❖ Needs sampling frame i.e. the list of units in the population.
- ❖ Compared to Non Probability Sampling it is difficult to apply
- ❖ Cost of designing and conducting the sampling survey is high.
- ❖ Cost of training is also high.

9.5 Various Statistical Sampling methods

9.5.1 Simple Random Sampling (SRS): It is the simplest form of random/statistical sampling; it consists of selecting the sample units ensuring equal probability of selection to every unit of the population. It is of two types**:**

(a) Simple Random Sampling with Replacement (SRSWR): In SRSWR, a unit is selected from the ***sampling frame*** (list of units in the population); the unit is replaced back and the next unit is selected; the process is repeated till a sample of the desired size is selected. As a result it is possible for a unit to be included in the sample more than once. In practice SRSWR is not attractive; same units can be selected more than once which may not add any value/additional information to audit. But in mathematical terms, it is simpler to relate the sample to population by SRSWR.

(b) Simple Random Sampling without Replacement (SRSWOR): In SRSWOR, a unit is selected for inclusion in the sample, it is removed from the sampling frame and the next unit is selected.

Thus, in this type of sampling a unit cannot be selected again. If a sample is selected with the help of random number table; a random number selected more than once is ignored in SRSWOR.

SRSWOR has two advantages:

- Elements are not repeated so resources are not wasted.
- Variance estimation is smaller (efficiency is higher) than SRSWR with same sample size.

We select a slip from 10 slips and replace the slip before selecting the next slip, then it is sampling with replacement; if we do not replace the slip before selecting the second slip, it is sampling without replacement.

(c) Advantages and disadvantages of Simple random sampling:

Advantages

(i) One of the greatest advantages of simple random sampling method is that it needs only a minimum knowledge of the study group of population in advance.

(ii) It is free from errors in classification/stratification.

(iii) It is totally free from bias and prejudice.

(iv) The method is simple to use.

(v) It is very easy to assess the sampling error in this method.

Disadvantages

(i) As compared to stratified sampling, it gives less efficient results for the same sample size.

(ii) The study of sample becomes time consuming if the units or items are widely dispersed.

(iii) It cannot be employed if the units of the population are heterogeneous in nature.

(iv) This method does not use available knowledge about the population.

(v) It needs complete list of units in the population (sampling frame)

(vi) It does not always produce a representative sample.

9.5.2 Systematic sampling: In Systematic Sampling the sample is chosen by selecting a random starting point and then picking every k^{th} (k is sampling interval) unit in succession from the sampling frame. The sampling interval 'k' is the ratio of population size (N) to sample size (n), rounded to the nearest integer. Systematic sampling is of 2 types (a) Linear and (b) Circular Systematic Sampling.

(a) Example of Systematic Sampling:

(i) Say, Target Population N= 54000 vouchers and sample size n = 6000

(ii) Sample Fraction (k) = Target Population / Sample size = 54000/6000 = 9

(iii) Number all vouchers of the population from 1 to 54000

(iv) Select a number between 1 to 9 (**here k = 9**) randomly

(v) Say, number 5 is selected then 5^{th} voucher is selected

(vi) Next 5+9 = 14^{th}, 14 + 9 = 23^{rd}, 32^{nd}, 41^{st} vouchers are selected and so on …

(b) Advantages and disadvantages of Systematic sampling

Advantages

1. Systematic sampling is less costly and easier to implement than SRS. It is because in this method random selection is done only for the first unit.
2. It ensures representativeness across the list (population) and is easy to implement.
3. It can help eliminate cluster selection i.e. selection of nearby units of the population.

Disadvantages

1. It works well only if the complete and up-to-date frame is available and if the units are randomly arranged in the frame; for this reason the units are arranged in some order say alphabetical or in increasing/decreasing order of value before selecting a sample.
2. If the population has a periodicity of the trait, this sampling technique may not give a representativeness sample. So it should not be used if the population already has some pattern.

9.5.3 Stratified sampling: It is a two-step process in which the entire population is partitioned into sub-populations, or strata. The strata should **be mutually exclusive and collectively exhaustive;** it means, every population unit should be assigned to one and only one stratum (singular for strata) and no population unit should be omitted. From each stratum, units are selected randomly, usually by SRS. The population units in each stratum should be **as homogeneous as possible**. A major objective of stratified sampling is to increase reliability without increasing cost. Stratification may be done on the basis of income, age, rural-urban, states, Revenue-Capital, Treasuries, major heads, etc.

(b) Allocation of Sample size in Stratified sampling

(i) Proportional Allocation: In this type of allocation, the sample size to be selected from the various strata is given by:

$n_i = (n/N)*Ni$; where n_i is the size of sample from the i^{th} strata, Ni is population size of the i^{th} strata; n is total sample size and N is the population size.

(ii) Optimum allocation ni's are chosen so as to

- Maximise the precision for fixed sample size n; **Neyman's Allocation**
- Maximise the precision for fixed cost or
- Minimise the total cost for fixed desired precision;

For each of these three allocations, specific mathematical formulae are available.

(iii) Disproportionate Allocation - No. of items selected from a stratum is independent of stratum size.

Note: Remember that in stratified sampling, a large sample would be required from a stratum if Stratum size Ni is large or Stratum variability Si (Standard Deviation) is large.

(c) Exercise: Proportional Allocation: Number of Vouchers coming from 3 treasuries are 300, 200 and 500 respectively. Find out the size of sample to be selected from each treasury if a proportional stratified sample of size 60 is required.

Here $N_1 = 300$, $N_2 = 200$ and $N_3 = 500$; $N = 1000$

using $n_i = (n/N)*N_i$; $i = 1, 2, 3$

$n_1 = (60/1000)* 300 = 18$, $n_2 = (60/1000)* 200 = 12$ and $n_3 = (60/1000)* 500 = 30$.

Thus a sample of 18, 12 and 30 vouchers will be selected from these strata.

Actually, the principal followed here is that the larger the Size of the Strata, the larger should be the sample from that Strata.

(d) Examples of stratified sampling in Audit:

(i) To select BPL households for a social audit; divide the population of BPL into three categories (strata) say top 25%, Middle 50% and Bottom 25% and select separate samples from 3 categories/strata; the sample from the three strata may or may not be equal.

(ii) Dividing contracts into value ranges and then selecting separate samples from each value range; say 100% contracts from highest value range, 50% contracts from next highest and so on.

(iii) Dividing the entire state into 3-4 geographical regions and then selecting required sample of districts from each of these regions.

(e) Advantages and disadvantages of Stratified sampling:

Advantages

1. Stratified sampling is more precise if variable of interest is associated with strata.
2. In this kind of sampling all the subgroups are represented, allowing separate conclusion about each of them; say separate conclusion for each state/District/treasury.
3. The stratified random sample also improves the representation of the various groups within the population, as well as ensures that strata are not over/under-represented. It helps the researcher/auditor to compare strata, as well as make more valid inferences about the population from the sample.

Disadvantages

1. In this method sampling error is difficult to measure
2. There is a loss of precision if a small number of units are sampled in individual heterogeneous strata.
3. This method is useful when there is a sufficient knowledge about the spread/variability of the population so that the strata are homogeneous.

9.5.4 Cluster sampling: In this type of sampling, the target population is first divided into mutually exclusive and collectively exhaustive sub-populations or clusters. It means, each unit should belong to one and only one subgroups and that none of the units of the population is left out. Then a **random sample**

of clusters is selected, based on a probability sampling technique such as SRS or Systematic Sampling. For each selected cluster, either all the units or a sample of units is drawn and audited. Heterogeneity within the cluster should be the same as that in population, ideally each cluster should be a small-scale representation of the population.

(b) Advantages and disadvantages of Cluster sampling:

Advantages

1. It is simple as complete list of units (sampling frame) is required only for the clusters selected in the sample which reduces efforts of finding out the complete sampling frame for all the clusters.
2. For this kind of sampling less travel/resources are required.

Disadvantages

1. It is imprecise if clusters are homogeneous (Large sample as compared to SRS is required for the same precision)
2. In this kind of sampling; Sampling Error is difficult to measure.

9.5.5 Stratified Sampling versus Cluster Sampling

- In both stratified and cluster sampling, the population is divided into well-defined groups.
- Stratified sampling is used when each group has small variation (more homogeneity) within itself but wide variation among the groups.
- Cluster Sampling in used in the opposite case, when there is considerable variation within each group but the groups are essentially similar to each other.
- In Stratified sampling estimate of each and every strata is available but not in cluster sampling.

Suppose in a state there are 20 Districts:

- We take a sample of 15 villages in each of the 20 Districts of a state to audit/study the implementation of MGNREGA. In all 300 villages are selected and audited/studied. This is an example of stratified sampling when estimates of the desired characteristics for each of the 20 Districts (Strata) would also be available
- On the other hand let us select 5 districts out of 20 and take a sample of 60 villages in the selected Districts only. In all 300 villages are selected and studied. This is an example of Cluster Sampling. In this case estimates of the desired characteristics for each of the Districts (Clusters) would not be available.

9.5.6 Multi Stage Sampling: Multistage sampling refers to a sampling plan where the sampling is carried out in stages using smaller and smaller sampling units at each stage. In a two-stage sampling design, a sample of primary units is selected and then a sample of secondary (Ultimate) units is selected from the selected primary units. The simplest form of two-stage sampling is to use Simple Random Sampling (SRS) at each stage – an SRS of primary units and then an SRS of secondary units within each selected primary unit. The primary units do not need to be of the same size and it is not necessary to select the same number of secondary units within each primary unit.

Stratified random sampling and cluster sampling can be viewed as special cases of two stage sampling. A stratified random sample is a census of the primary units (the strata) followed by an SRS of the secondary units within each primary unit. A cluster sample is an SRS of the primary units (the clusters) followed by a census of the secondary units within each selected primary unit.

We can use any probability sampling plan at each stage of a multistage plan and the plans can be different at each stage.

(a) Examples of Multi Stage Sampling: (i) Two Stage Sampling: In order to estimate the condition of highways under its jurisdiction and the cost of urgent repairs, the state Department of Transportation selected a number of "highway miles" in two stages. In the first stage, a number of highways were selected by SRS without replacement from the list of all highways maintained by the Department. In the second stage, a number of one-mile segments were selected by SRS without replacement from the total length of each selected highway; for example, if the length of highway 101 is 73 miles, it is seen as consisting of 73 one-mile segments ("highway miles"), from which a number of segments are selected at random. Highway engineers then visit the selected segments, inspect the pavement condition, rate the condition of the segment and estimate the cost of required repairs. For the purpose of this problem, assume there are 352 highways in the state, with a total length of 28,950 miles. A simple random sample of 30 highways is selected without replacement. From each selected highway, 10% of one-mile segments are then selected at the Second Stage.

(ii) Another example of Multi Stage Sampling is selection of households for the entire country. Suppose, we are interested in obtaining a sample of 'n' households in the country - the first stage units may be states, the second stage units may be districts in selected states, third stage units will be villages/urban blocks in the selected districts and ultimate stage units are households in the selected villages/urban blocks.

(b) Advantages and Disadvantages of Multi Stage Sampling: Multistage samples are used primarily for cost or feasibility (practicality) reasons. For example, to select an SRS of households in a state would be extremely difficult because no list of all households exists. However, we could proceed in stages: an SRS of villages/Urban blocks within the state, and then an SRS of households within each village/block. We now need to have a list of households for selected villages/blocks only. If, it is not available, it can be prepared on the spot by the investigator. This leads to great saving in operational time and cost. Two-stage sampling also has the flexibility to take a larger sample from the primary units which are larger or more variable.

The disadvantage of multi-stage sampling is that it is likely to be less efficient compared to a suitable single stage sampling of the same size.

9.6 Statistical Sampling Plans used in Audit

Mainly two types of sampling plans are used in Audit; they are (a) Attribute Sampling Plans or Attribute Sampling and (b) Variable Sampling Plans or Variable Sampling

9.6.1 Attribute Sampling Plans or Attribute Sampling: An attribute is a qualitative characteristic which cannot be measured quantitatively. However, the population may be classified into various classes w. r. t. the attribute. Attribute Sampling is used in Tests of Controls (TOC) i.e. to find out if the controls are effective or not; it is done by finding no. or percentage of deviations, etc. – it deals with the situation 'How Many'.

An attribute sampling situation is one dealing with the rate of occurrence or frequency of items in a population having a certain attribute. The attribute either exists or does not. A voucher either has an error or no error. There is no in-between situation. Examples of attributes sampling situations are estimating the number or percentage of schools without toilet or without principal; the frequency or ratio of obsolete items in an inventory or the frequency of errors in a file or document or in a group of receivable balances.

In attribute sampling, the results of a random sample are expressed as a sample frequency or, in auditing, as a sample error rate (e.g., one unbilled shipment in a sample of 100 shipping orders would be a 1 percent sample error rate). The sample error rate is also the most likely (estimated or projected) error rate in the population because a random sample is likely to be a representative of the population. An Attribute Sampling has the following types:

(a) Fixed sample size: The sampling techniques like Simple Random Sampling and systematic sampling which have been discussed in the previous paragraphs constitute fixed size attribute sampling if they are performed for 'Test of Control' to estimate the deviation/error rate of a population.

(b) Stop or Go Sampling (also called Sequential Sampling): Stop-or-go sampling is used for not so common cases; it prevents oversampling. We take a sample decide whether controls are effective or not; if sample is inconclusive, we take another sample and so on. This method aims at concluding that the error rate of the population is below a predefined level at a given confidence level by examining as few sample items as possible – the sampling stops as soon as the expected result is reached. This method is not well-suited for projecting the results to the population, though it can be useful for assessing system audit conclusions. This sampling

(i) Involves sampling a universe in increments and examining each incremental sample before deciding when to stop.

(ii) Allows auditors to determine from the smallest possible sample size if an error rate exceeds a predetermined level.

(iii) Provides assurance, within a fixed degree of confidence, that the error rate in a population is less than a predetermined acceptable error rate.

(iv) Does not provide an estimate of actual error rate; however, it can readily be converted into attribute sampling, which can be used to estimate error rate.

(v) Used when Auditor expects very small rate of deviation

(c) Discovery Sampling: Discovery sampling aims at auditing cases where a single error would be critical; it is therefore particularly useful for the detection of cases of fraud or avoidance of controls. Based on attribute sampling, this method assumes a zero (or at least very low) rate of error. It is not well suited for projecting the results to the population. Discovery sampling allows the auditor to conclude, based on a sample, whether the assumed very low or zero error rate in the population is a valid assumption. Discovery Sampling consist of the following Steps:

(i) Selects a sample of a given size; accepts population if the sample is error free & rejects it otherwise.

(ii) Used when chances of error are negligibly small and auditor is not interested in estimating/projecting no. of errors in the population.
(iii) For e.g. reliability = 95% & acceptable occurrence rate = 1% gives sample size of 300 from relevant table.
(iv) If there is no error in this sample size of 300, it is 95% certain that the deviation/error rate does not exceed 1%;
(v) An auditor may use discovery sampling if he has reasons to suspect fraudulent activity.

9.6.2 Variable Sampling Plans or Variable Sampling: Variable (or quantitative) sampling is used when the objective is to estimate a quantity (like amount of loss to government, average loss per transaction, etc.); it deals with "How Much". It is used primarily for substantive testing. Most commonly used variable sampling plan in Audit is Probability Proportional to Size (PPS) or Monetary Unit Sampling (MUS). A variable sample drawn from a city can estimate the average height of its citizens based on the average height in the sample. Examples of variables sampling situations in Audit may be the average invoice value of a group of purchase invoices or the average invoice gross profit of a group of sales invoices.

In variables sampling, the results of a random sample may be expressed as a sample average (e.g., the average invoice value in a sample of 100 invoices might be $500). The sample average is also the most likely (estimated or projected) population average. The auditor can, if the sample size has been chosen properly, have a set level of confidence that the true population average falls within the range of this most likely average plus or minus a specified precision. The Auditor can estimate the population totals also by multiplying the sample average with the population size.

(a) Probability Proportional to Size (PPS): PPS sampling assigns higher probability of selection for population units with higher sizes (size may be total expenditure, total population, total no. of BPL households in a GP, total number of vouchers in a treasury, **total risk based on one or more criteria**, etc.). In other words, the entities with higher sizes, based on some characteristics, will have higher chances of selection. **Monetary Unit Sampling (MUS)** in audit is an example of PPS sampling with money value of transactions as size measure. MUS is actually **PPS – Systematic Sampling**.

Selection of Systematic PPS Sample

Let's take treasuries as clusters, the objective is to select 15 clusters i.e. 15 treasuries using PPS; size being no. of vouchers in a treasury. We may also take the total value of vouchers of treasury as the size measure.

STEP I: List all Treasuries with number of vouchers in them; find the cumulative totals of number of vouchers by successive addition of number of vouchers in the different treasuries; as shown:

Treasury	No. of vouchers	Cumulative total
1	34	34
2	60	94
3	30	124
4	76	200
5	315	515 and so on
	….	
Grand Total		**2,356**

(ii) Divide the cumulative total = 2,356 by 15 [no. of clusters to select]

We have Sampling Interval = cumulative total/no. of clusters to be selected

= 2,356 / 15= 157.07; **Sampling Interval 'k' is 157**

Find a three digit random number [less than or equal to 157] say 123

Select in sample the cluster corresponding to 123 (123 or next number appearing in the Cumulative Total column).

Select remaining clusters from the cumulative distribution by adding 157 (sampling interval) each time.

Treasury	No. of vouchers	Cum. Total	Cluster Selected
3	30	124 *	selected
4	76	200	
5	315	515 **	**selected twice**

(2nd 123+157=280; 3rd 280+157=437)

Note that the fifth cluster is selected again; this means it becomes a **PPS with replacement**. If all the units of the selected clusters/treasuries are being examined (one stage cluster sampling), selection of a cluster again may not add to any value and would mean wastage of resources. If, however, a two stage sampling is being used then the cluster may be selected more than once and a different sample of the second stage units may be selected (without replacement) and audited. This would not lead to wastage of resources.

Selection of a Random PPS Without Replacement (PPSWOR):

Let us consider selection of 3 Districts from the 17 Districts of Punjab by PPSWOR, with size measure as expenditure on the scheme under review. List of Districts in Punjab along with expenditure for the scheme is obtained first. Cumulative totals are then obtained by successive addition of Expenditure of the Districts.

Sr. No.	Name of the Districts	Expenditure under the scheme (Rs. 000)	Cumulative Total
1	Amritsar	368	368
2	Bathinda	1095	1463
3	Faridkot	1009	2472
4	Fategarh Sahib	1536	4008
5	Firozpur	3419	7427
6	Gurdaspur	534	7961
7	Hoshiarpur	621	8582
8	Jalandhar	534	9116
9	Kapurthala	323	9439
10	Ludhiana	223	9662
11	Mansa	278	9940
12	Moga	660	10600
13	Muktsar	1474	12074
14	Nawanshahr	1613	13687
15	Patiala	1038	14725
16	Rupnagar	527	15252
17	Sangrur	2131	17383
Total		**17383**	

Here the cumulative total is 17383, a five digit number. So we obtain 5 digit random numbers from the random number table. Following is a page from the random number table which may be used for selection of a sample. We start with first row and move rightwards to select random numbers.

Row Number	Random numbers									
00000	**10097**	**32533**	**76520**	**13586**	**34673**	**54876**	**80959**	**09117**	**39292**	**74945**
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062

Table for sample selection

Random Number Selected	Decision	District selected	Reason
10097	Selected	Moga	Moga is selected as the selected random number, 10097 is greater than the cumulative total 9940 but less than (or equal to) 10600.
32533	No selection	X	As the selected random number is > 17383; **cumulative total of all the strata**
76520	No selection	X	As the selected random number is > 17383
13586	Selected	Nawanshahr	Nawabnshahr is selected as the selected random number 13586 is greater than the cumulative total 12074 but less than (or equal to) 13687
34673	No selection	X	As the selected random number is > 17383
54876	No selection	X	As the selected random number is > 17383
80959	No selection	X	As the selected random number is > 17383
09117	Selected	Kapurthala	Kapurthala is selected as the selected random number 9117 is greater than the cumulative total 9116 but less than (or equal to) 9439

Hence the selected districts are Moga, Nawanshahr and Kapurthala.

(b) Monetary Unit Sampling (MUS): Monetary Unit Sampling (MUS) is a sampling method in which the sampling unit is not an invoice or any other physical unit, but an individual rupee. However, when the individual rupee is selected, the auditor does not verify just that particular rupee, but the rupee acts as a hook and drags the whole invoice with it. For example, if as a result of sample selection, Rs.365 is selected for testing and if that rupee falls in voucher number 14, then that voucher will be audited.

Let us assume that there are 6 items out of which 2 items are to be selected. The values of the 6 items are 100, 200, 300, 400, 500 and 1000. If Simple Random Sampling is used to select 2 items, then all the items have equal chance of selection, as the sampling unit would be individual item. On the other hand, if MUS is used, then the total value of 6 items works out to Rs.2500, i.e., there are 2500 sampling units. As 2 items are to be selected, the sampling interval would be 2500/2 (Rs.1250). This means that one rupee out of every 1250 rupees would be selected. In such a case, the chances of 1000 rupee item getting picked up is 10 times more than the 100 rupee item getting picked up. Thus MUS has a bias towards high value items.

Advantages of MUS: The important advantages of MUS are:

(i) It normally produces smaller sample sizes than other substantive sampling plans.

(ii) There is no difficulty in expressing a conclusion in monetary terms.

(iii) No rupee stratification is necessary, as this will be accomplished automatically, thus avoiding problems of determining optimum strata boundaries and allocation of sample size among strata.

(iv) It is relatively easy to apply compared to other sampling plans.

(v) The problem of detecting the large but infrequent errors is solved, since all items greater than the sampling interval will be selected in the sample by themselves.

Disadvantages of MUS: The main disadvantages of MUS are the following:

(i) Accounts/items with 'nil' balances will have no chance of selection.

(ii) The more an item is understated; the less likely the item has a chance of selection. Hence, MUS is less useful for finding understatements.

(iii) A large percentage error in a small transaction can significantly increase the computed error limit.

(iv) It is very difficult to use MUS in a non-computerized environment as totaling the sample items in the population may be a difficult task.

9.6.3 Sampling Risk in Audit: Sampling risk arises from the possibility that, when a test of controls or a substantive test is restricted to a sample, the auditor's conclusions may be different from the conclusions he would reach if the tests were applied to all the items (100% enumeration). That is, a particular sample may contain proportionately more or less monetary misstatements or deviations from prescribed controls than exist in the population as a whole. The sampling risk varies inversely with sample size: the smaller the sample size, the greater is the sampling risk.

In performing substantive tests [Variable Sampling] the auditor faces two aspects of sampling risk:

(a) The risk of incorrect acceptance is the risk that the sample supports the conclusion that the recorded account balance is not materially misstated when it is materially misstated. This is called **risk of assessing control risk too low –β risk.**

(b) The risk of incorrect rejection is the risk that the sample supports the conclusion that the recorded account balance is materially misstated when it is not materially misstated. This is called **risk of assessing control risk too high –α risk.**

The auditor faces similar risk in performing tests of controls [Attribute Sampling] when sampling is used:

(a) The risk of assessing control risk too low is the risk that based on the sample the auditor concludes that the controls are operating effectively when they are not.

(b) The risk of assessing control risk too high is the risk that based on the sample the auditor concludes that the controls are not operating effectively while they are.

The risk of incorrect rejection or the risk of assessing control risk too high relates to the efficiency of the audit while the risk of incorrect acceptance or the risk of assessing control risk too low relates to the effectiveness of an audit in detecting an existing material misstatement.

9.7 Determining Sample Sizes for Conducting Audits/Surveys:

9.7.1 Sample Size for Un-stratified Mean Per Unit (MPU): The un-stratified MPU is a statistical plan whereby a sample mean is calculated and **projected as an estimated total**. Here the sample is selected with SRS, the sample mean $\bar{x}$ of sample values multiplied by the number of items in the population N produce an estimate of population total. Assuming normality, the optimum sample size under MPU is

$$n = \left(\frac{Z_r \cdot SD}{d}\right)^2$$

Where Z_r = confidence level coefficient [**Refer table given below**], d = precision or average margin of error and SD = ***Standard Deviation***.

Z Score - Table

Confidence Level	Z -value
75 %	1.15
80%	1.28
85%	1.44
90%	1.65
92%	1.75
94%	1.88
95%	1.96
96%	2.05
99%	2.58

Example: ABC Limited desires 95 per cent reliability and plans to use unrestricted (un-stratified) random sampling without replacement to estimate the value of inventory of a subsidiary. To estimate mean and standard deviation (SD) of the inventory population, a pilot sample of 30 items from the total population of 2000 items was selected. The pilot sample produced an arithmetic mean of Rs. 4000 and a (SD) of Rs. 150.

- 95 per cent reliability means firm is willing to tolerate 5 per cent chance of sampling error.

- It means that 5 % of the times, projection plus/minus precision may not include true population total.
- Based on 95 % reliability, reliability coefficient (Z_r) is 1.96 – based on normal tables.
- Precision (d) is judgmentally set equal to Rs. 30 (termed as materiality) – the amount considered material for this application; in terms of total deviation in the population it comes out to be Rs 60,000/-

Using the Formula, we have the sample size as:

$$n_o = (\frac{Z_r \,.\, SD}{d})^2 = (\frac{1.96 * 150}{30})^2 = (1.96*5)^2 = 9.8 * 9.8 = 96.04$$

= 96 (approx.)

Sixty- six additional sample items are added to the pilot sample of 30 to yield the total sample of 96. The 66 additional sample items are selected using SRSWOR.

9.7.2 Sample Size for Un-stratified Proportion of *audit objections* (*errors*): The projected number of audit objections in the population is the sample proportion of errors multiplied by number of items in the population. The optimum sample size under SRS is $n = \frac{Z_r{}^2 * P*(1-p)}{d^2}$, Where Z_r = confidence level coefficient, d= margin of error the Auditor is willing to tolerate and p = Proportion of errors that is expected in the population.

Example: An Auditor is trying to determine the proportion of bills that contain error. She asks, "How large a sample size do I need?" To answer a question like this we need to know:

1. How accurately does she need the answer i.e. guessing 'd'?
2. What level of confidence does she intend to use?
3. What is her current estimate of the proportion of errors?

Possible answers might be:

1. "We need a margin of error < 2.5%". [Margin of error normally ranges b/w 1% to 5% — we can choose any margin of error we like but must specify it.]. It may be noted that, as the margin of error reduces, the sample size increases. For e.g. if the margin of error is reduced to 2% from 4%, the sample size would increase 4 times.

2. 95% confidence intervals are typical but not mandatory — we could choose 90%, 99%, etc. For this example, we assume 95% confidence level. Here again; as the confidence level increases, the sample size also increases.

3. May be guided by past audits or general knowledge. Alternately, we may conduct a pilot with sample size of 30 to estimate expected error. Let's suppose the answer is 10%.

Thus we have Z_r = 1.96 (for 95% confidence coefficient); P= Proportion of errors that is expected in the population = 10% = 0.1 and d = margin of error = 2.5% =0.025.

Using the formula

$$n = \frac{Z_r{}^2 * p*(1-p)}{d^2} = \frac{1.96^2 * 0.1*(1-0.1)}{0.025^2} = 553.19$$

So a sample size of 554 would be required.

9.8 Statistical Inference/Estimation of Parameters

The task of statistical inference is to draw a conclusion about a population parameter (a characteristic of a population) like mean, proportion, etc. from the study of the sample statistic (the characteristic of sample). The real question in estimation is not whether our estimate is correct or not - but how close is it to the true value? The Closeness of estimate to the true value is measured by the Standard Error – which is the Standard Deviation of the Sampling Distribution.

9.8.1 Point Estimate vs. Interval Estimate: Estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample. Statisticians use sample statistics to estimate population parameters. An estimate of a population parameter may be expressed in two ways:

(a) Point estimate. A point estimate of a population parameter is a single value of a statistic. We may draw a sample of size n from the population and compute proportion p, or mean $\bar{x}$ and use them as point estimates of μ (Population mean) or P (Population Proportion). However, these can't be expected to be equal to μ or P. It would, therefore, be much more meaningful to estimate these parameters by interval estimates.

(b) Interval estimate. An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, $a < x < b$ is an interval estimate of the population parameter x. It indicates that the population parameter is greater than '*a'* but less than '*b'*. An interval estimate may be expressed as follows: Point estimator $\pm$ (reliability coefficient) x (std. error). Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

The interval estimate of a confidence interval is defined as: sample statistic $\pm$ margin of error; where margin of error depends upon the confidence level say 95% and the variability in the population/sample measured by the Standard Deviation.

(c) Confidence Level: The confidence level describes the likelihood/probability that the confidence interval includes the true population parameter. Suppose we collect all possible samples from a given population, and compute confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

(d) Margin of Error: In a confidence interval, the range of values above and below the sample statistic is called the margin of error. For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the votes. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. It means, "We are 95% confident that the independent candidate will receive votes between 25% and 35%".

9.8.2 Confidence Intervals for single mean (Population Standard Deviation σ Known - this is hardly ever true)

Assumptions

- Population Standard Deviation is Known
- Population is Normally Distributed
- If population is not Normal, we may use large samples

(a) Formula for Confidence Interval Estimate of single mean

$$\bar{x} - Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

where $Z_{\frac{\alpha}{2}}$ is the reliability coefficient; its value comes from normal table for given level of confidence α. σ is Population Standard Deviation, n is the sample size, $\bar{x}$ is the sample mean and μ is the population mean.

For Large Samples (n ≥ 30), 95 % confidence interval of sampling mean is: $\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$ and

99% confidence interval is $\bar{x} \pm 2.58 * \frac{\sigma}{\sqrt{n}}$

(b) Confidence Intervals for single mean- An Example:

If sample mean is 103, and σ = 20; sample size n = 100; we would claim that the population mean μ_x lies within the interval: $103 \pm (1.96) * \frac{20}{\sqrt{100}} = 103 \pm (1.96) * (2.00)$ i.e. b/w 99.08 and 106.92 in 95% of cases.

And between $103 \pm (2.58) * (2.00)$ i.e. b/w 97.84 and 108.16 in 99% of the cases.

(c) Calculation of Confidence Interval requires

(i) Confidence Level: it is generally fixed at 95%; in some cases it may be fixed at 90% or 99%. As the confidence level increases, the width of confidence interval will also increase.

(ii) Standard Deviation: Standard Deviation of the population is rarely known so it is estimated using the sample values; the larger the Standard Deviation (variation in the sample), the larger would be the width of the confidence interval.

(iii) Sample value: The value of sample mean $\bar{x}$ or sample Proportion p as the case may be; these values are calculated from the sample.

9.8.3 Confidence Interval Estimate of Proportion

Assumptions

- Two Categorical outcomes like error or no error; schools with or without principal.
- Population Follows Binomial Distribution
- Normal approximation can be used for larger samples

(b) Formula for Confidence Interval Estimate of single proportion

$$p - Z_{\frac{\alpha}{2}} * \sqrt{\frac{p*(1-p)}{n}} \leq P \leq p + Z_{\frac{\alpha}{2}} * \sqrt{\frac{p*(1-p)}{n}}$$

Where p is sample proportion and $Z_{\frac{\alpha}{2}}$ is reliability coefficient; its value comes from normal table for given level of confidence interval. P is the population proportion and n is the sample size.

(c) Example for Estimating Population Proportion: A random sample of 400 Voters showed 32 preferred Candidate A. Set up a 95% confidence interval estimate for P, the population proportion. Here sample proportion = 32/400 = 0.08

Using $p - Z_{\frac{\alpha}{2}} * \sqrt{\frac{p*(1-p)}{n}} \leq P \leq p + Z_{\frac{\alpha}{2}} * \sqrt{\frac{p*(1-p)}{n}}$

95% confidence interval estimate for P, the population proportion would be

$$0.08 - 1.96 * \sqrt{\frac{0.08*(1-0.08)}{400}} \leq P \leq 0.08 + 1.96 * \sqrt{\frac{0.08*(1-0.08)}{400}}$$

So 95 % confidence interval for sample proportion is:

$$\mathbf{0.053 < p < 0.107}$$

9.8.4 Where can we use Estimation in Audit: If it is decided to estimate/project the value of a parameter, then it is always advisable to have an interval estimate. It is because, we may draw a sample of size n from the population and compute Proportion p, or mean $\bar{x}$ and use them as point estimates of μ (Population mean) or P (Population Proportion). However, these can't be expected to be equal to μ or P. It would, therefore, be much more meaningful to estimate these parameters by interval estimates. Moreover, confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

9.9 Method of selection of Simple Random Sample With or Without Replacement Using Random Number Table

Let there be 'N' number of auditable units in the population/stratum from which 'n' number of units are to be selected.

Step 1: Prepare or get the list of units and associate serial numbers with each of the units.

Step 2: Open any page of the random number table randomly.

Step 3: Select a random number 'r' of appropriate number of digits, starting from the left most top corner of the table and proceed sequentially from left to right. If there are 100-999 units in the population, a 3 digit random number is selected and if there are 1000 to 9999 units in the population, a four digit random number.

Step 4: If 'r' is between 1 and N, then the unit corresponding to the r^{th} serial number is selected. If not, select the next random number in the sequence and proceed sequentially.

Repeat Step 3 and Step 4 until we select 'n' distinct units (note that if a random number is selected more than once it is ignored in case of sampling without replacement and the corresponding unit is selected again in case of sampling with replacement.)

EXERCISE:

Answer the following questions:

1) What is sampling?
2) What are sampling errors and non-sampling errors?
3) What is the difference between Simple Random sampling with replacement and Simple random sampling without replacement? Which of them would you prefer to use in Audit?
4) Give a few examples of use of Probability Proportional to Size sampling in Audit?
5) What is Step by Step sampling?
6) What are the advantages of sampling?
7) What is Non statistical sampling? Why is it used frequently in Audit in spite of it having many disadvantages?
8) What is Systematic sampling and what are its Advantages?
9) What is stratified sampling? How is it different from cluster sampling?
10) Explain Monetary Unit sampling with one example?

MCQs on Sampling Theory and Statistical Estimation

Q1. Standard Error is due to:

a) Observer error
b) Instrumental error
c) Sampling error
d) Conceptual error [Ans. (c)]

Q2. The correct formula for the standard error of the mean is:

a) σ^2/n
b) σ/n
c) $\sigma/\sqrt{n}$
d) $\sqrt{\sigma}/n$ [Ans. (c)]

Q3. Mean of 400 observations is 100 and standard deviation is 8. What will be standard error of mean?

a) 0.4
b) 1.0
c) 2.0
d) 4.0 [Ans. (a)]

Q4. If the sample size is multiplied by 100, then standard error of the sample mean is:

a) Multiplied by 100
b) Divided by 100
c) Divided by 10
d) Unaffected [Ans. (c)]

Q5. The standard error of prediction is a kind of:

a) Mean
b) Median
c) Standard deviation
d) None of the above [Ans. (c)]

Q6. In Random Sampling, chance of being selected is:

a) Not same and not known
b) Same and known
c) Same and not known
d) Not same but known [Ans. (b)]

Q7. About Random Sampling not true is:

a) Simple
b) Biased
c) Stratified
d) Systematic [Ans. (b)]

Q8. True about Simple Random Sampling is:

a) Technique provides least number of possible samples.
b) Every fixed unit is taken for selection
c) All units have equal chance to be selected
d) Only selected units have right to be selected [Ans. (c)]

Q9. For a survey, a village is divided into 5 lanes then each lane is sampled randomly. It is example of:

a) Simple Random Sampling
b) Stratified Random Sampling
c) Systematic Random Sampling
d) Multi stage Sampling [Ans. (b)]

Q10. Which is true of cluster sampling:

a) Every n^{th} case is chosen for study
b) Involves the use of Random Numbers
c) A natural group is taken as sampling unit
d) Stratification of population is done [Ans. (c)]

Q11. A selection of every unit from a normal distributed population is utilizing:

a) Simple Random Sampling
b) Systematic Sampling
c) Stratified Random Sampling
d) Complete enumeration [Ans. (d)]

Q12. In systematic random sampling, the sample interval is determined using:

a) Random numbers
b) $\frac{\text{Total population}}{\text{Sample size desired}}$
c) $\frac{\text{Total population}}{\text{Sample size desired}} * 100$
d) None of the above [Ans. (b)]

Q13. As sample size increases, standard deviation of sample means

a) Decrease
b) Increase
c) Remains the same
d) Approaches to infinity [Ans. (a)]

Q14. For qualitative data, the sample size can be calculated by:

a) $n= \frac{{Z_r}^2 * p^2 * (1-p)}{d^2}$, b) $n= \frac{{Z_r}^2 * p * (1-p)}{d^2}$

c) $n= \frac{{Z_r}^2 * p * (1-q)}{d^2}$ d) $n= \frac{{Z_r}^2 * q * (1-p)}{d^2}$ [Ans. (b)]

Where Z_r = confidence level coefficient, d= margin of error the Auditor is willing to tolerate and p = Proportion of errors that is expected in the population.

Q15. The term used to define the smallest element in sample selection is:

a) Sample frame
b) Statistical Unit
c) Population Unit
d) Sampling unit [Ans. (d)]

Q16. Which of the following is a probability based sample selection method?

a) Accidental sampling
b) Multistage stratified
c) Convenience sampling
d) Judgmental sampling [Ans. (b)]

Q17. Interviewing hockey players as they exit the arena is an example of what type of sampling?

a) Quota
b) Convenience
c) Simple random
d) Cluster [Ans. (b)]

Q18. Which of the following is NOT true of probability sampling

a) The number of elements to be included in the sample set can be pre-specified
b) Estimates are statistically projectable to the population
c) It is possible to specify the probability of selecting any particular sample of a given size
d) The results will always be more accurate than non –probability sampling. [Ans. (d)]

Q19. A sampling frame is

a) A summary of the various stages involved in designing a survey
b) An outline view of all the main clusters of units in a sample
c) A list of all the units in the population from which a sample will be selected
d) A wooden frame used to display tables of random numbers [Ans. (c)]

Q20. The sample mean is

a) always equal to the population mean
b) Never equal to the population mean
c) A statistic
d) A parameter [Ans. (c)]

Q21. Which one of the following is a non-sampling error?

a) Selecting inappropriate audit procedure to achieve audit objectives
b) Taking a very small sample due to resource constraints
c) Considering inappropriate population while selecting a sample
d) Both (a) and (c) above [Ans. (d)]

Q22. In Probability Proportional to Size (PPS) Sampling, chance of being selected is:

a) Not same and not known
b) Same and known
c) Same and not known
d) Not same but known [Ans. (d)]

Q23. Which of the following would lead to the smallest confidence interval?

a) Small sample size and confidence coefficient of 0.95
b) Large sample size and confidence coefficient of 0.95
c) Small sample size and confidence coefficient of 0.90
d) Large sample size and confidence coefficient of 0.90 [Ans. (d)]

Q24. Which of the following is an advantage of sampling?

a) Census results are verified by sampling
b) Non – sampling errors are likely to be less
c) The Quality of information is maintained
d) All of these [Ans. (d)]

Q25. Which of the following are examples of Attribute Sampling

a) Probability Proportional to Size (PPS) Sampling
b) Monetary Unit Sampling (MUS)
c) Sequential Sampling
d) None of these [Ans. (c)]

Q26. The sampling method in which a unit is not returned to the population before selecting the next unit is called:

a) Sampling With replacement
b) Sampling without replacement
c) There is no such sampling method
d) None of these [Ans. (b)]

Q27. Which of the following is an example of non-statistical sampling?

a) Sequential Sampling
b) Cluster Sampling
c) Haphazard Sampling
d) Simple Random Sampling with replacement [Ans. (c)]

Q28. Which of the following statements is correct concerning statistical sampling of attributes?

a) For large Populations, population size has little or no effect on determining the sample size.
b) As the population size doubles, the sample size should also be doubled.
c) The likely population error rate (also called expected deviation rate) has no effect on the sample size.
d) The tolerable error is not required in this kind of sampling. [Ans. (a)]

Q29. In cases where a single error could not be tolerated and the management expects 100% compliance; assuming that 100% check is not feasible, which type of sampling approaches would be appropriate?

a) Sequential Sampling

b) Cluster Sampling

c) Haphazard Sampling

d) Discovery Sampling [Ans. (d)]

Q30. Name the sampling method in which the sampling unit is not an invoice or any other physical unit, but an individual rupee. However, when the individual rupee is selected, the auditor does not verify just that particular rupee, but the rupee acts as a hook and drags the whole invoice with it.

a) Rupee Sampling

b) Monetary Unit Sampling

c) Haphazard Sampling

d) Discovery Sampling [Ans. (b)]

Q31. An important difference between Judgmental and Statistical Sampling is that in Statistical Sampling:

a) No judgment is required as formula are available for everything

b) Efficiency is better

c) Population estimate with reliability of estimate can be obtained

d) All of these [Ans. (c)]

Q32. The risk that the sample supports the conclusion that the recorded account balance is not materially misstated when it is materially misstated is called:

(a) Risk of assessing control risk too low.

(b) Risk of assessing control risk too high

(c) Neither of these two

(d) Either of these two [Ans. (a)]

Q33. Say, Target Population N= 54000 vouchers and sample size n = 6000; following the systematic sampling approach if 4^{th} unit is selected as the first unit in the sample then the next unit selected in the sample would be

a) 8^{th}

b) 10^{th}

c) 12^{th}

d) 13^{th} [Ans. (c)]

Q34. Three strata has population sizes 500, 300 and 200. If it is decided to use proportional allocation with a total sample size of 90; the sample size from the second stratum would be:

a) 21

b) 30

c) 27

d) The question is incomplete [Ans. (c)]

Portion of Random Number Table

Row Sl. number	Random Numbers									
00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
00014	99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
00015	65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
00016	80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
00018	69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
00019	09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
00020	91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
00021	80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
00022	44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
00023	12550	73742	11100	02040	12860	74697	96644	89439	28707	25815
00024	63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
00025	61196	90446	26457	47774	51924	33729	65394	59593	42582	60527
00026	15474	45266	95270	79953	59367	83848	82396	10118	33211	59466
00027	94557	28573	67897	54387	54622	44431	91190	42592	92927	45973
00028	42481	16213	97344	08721	16868	48767	03071	12059	25701	46670
00029	23523	78317	73208	89837	68935	91416	26252	29663	05522	82562
00030	04493	52494	75246	33824	45862	51025	61962	79335	65337	12472
00031	00549	97654	64051	88159	96119	63896	54692	82391	23287	29529
00032	35963	15307	26898	09354	33351	35462	77974	50024	90103	39333
00033	59808	08391	45427	26842	83609	49700	13021	24892	78565	20106
00034	46058	85236	01390	92286	77281	44077	93910	83647	70617	42941
00035	32179	00597	87379	25241	05567	07007	86743	17157	85394	11838

00036	69234	61406	20117	45204	15956	60000	18743	92423	97118	96338
00037	19565	41430	01758	75379	40419	21585	66674	36806	84962	85207
00038	45155	14938	19476	07246	43667	94543	59047	90033	20826	69541
00039	94864	31994	36168	10851	34888	81553	01540	35456	05014	51176
00040	98086	24826	45240	28404	44999	08896	39094	73407	35441	31880
00041	33185	16232	41941	50949	89435	48581	88695	41994	37548	73043
00042	80951	00406	96382	70774	20151	23387	25016	25298	94624	61171
00043	79752	49140	71961	28296	69861	02591	74852	20539	00387	59579
00044	18633	32537	98145	06571	31010	24674	05455	61427	77938	91936
00045	74029	43902	77557	32270	97790	17119	52527	58021	80814	51748
00046	54178	45611	80993	37143	05335	12969	56127	19255	36040	90324
00047	11664	49883	52079	84827	59381	71539	09973	33440	88461	23356
00048	48324	77928	31249	64710	02295	36870	32307	57546	15020	09994
00049	69074	94138	87637	91976	35584	04401	10518	21615	01848	76938
00050	09188	20097	32825	39527	04220	86304	83389	87374	64278	58044
00051	90045	85497	51981	50654	94938	81997	91870	76150	68476	64659
00052	73189	50207	47677	26269	62290	64464	27124	67018	41361	82760
00053	75768	76490	20971	87749	90429	12272	95375	05871	93823	43178
00054	54016	44056	66281	31003	00682	27398	20714	53295	07706	17813
00055	08358	69910	78542	42785	13661	58873	04618	97553	31223	08420
00056	28306	03264	81333	10591	40510	07893	32604	60475	94119	01840
00057	53840	86233	81594	13628	51215	90290	28466	68795	77762	20791
00058	91757	53741	61613	62269	50263	90212	55781	76514	83483	47055
00059	89415	92694	00397	58391	12607	17646	48949	72306	94541	37408
00060	77513	03820	86864	29901	68414	82774	51908	13980	72893	55507
00061	19502	37174	69979	20288	55210	29773	74287	75251	65344	67415
00062	21818	59313	93278	81757	05686	73156	07082	85046	31853	38452
00063	51474	66499	68107	23621	94049	91345	42836	09191	08007	45449
00064	99559	68331	62535	24170	69777	12830	74819	78142	43860	72834
00065	33713	48007	93584	72869	51926	64721	58303	29822	93174	93972
00066	85274	86893	11303	22970	28834	34137	73515	90400	71148	43643
00067	84133	89640	44035	52166	73852	70091	61222	60561	62327	18423
00068	56732	16234	17395	96131	10123	91622	85496	57560	81604	18880
00069	65138	56806	87648	85261	34313	65861	45875	21069	85644	47277
00070	38001	02176	81719	11711	71602	92937	74219	64049	65584	49698
00071	37402	96397	01304	77586	56271	10086	47324	62605	40030	37438
00072	97125	40348	87083	31417	21815	39250	75237	62047	15501	29578
00073	21826	41134	47143	34072	64638	85902	49139	06441	03856	54552

00074	73135	42742	95719	09035	85794	74296	08789	88156	64691	19202
00075	07638	77929	03061	18072	96207	44156	23821	99538	04713	66994
00076	60528	83441	07954	19814	59175	20695	05533	52139	61212	06455
00077	83596	35655	06958	92983	05128	09719	77433	53783	92301	50498
00078	10850	62746	99599	10507	13499	06319	53075	71839	06410	19362
00079	39820	98952	43622	63147	64421	80814	43800	09351	31024	73167
00080	59580	06478	75569	78800	88835	54486	23768	06156	04111	08408
00081	38508	07341	23793	48763	90822	97022	17719	04207	95954	49953
00082	30692	70668	94688	16127	56196	80091	82067	63400	05462	69200
00083	65443	95659	18288	27437	49632	24041	08337	65676	96299	90836
00084	27267	50264	13192	72294	07477	44606	17985	48911	97341	30358
00085	91307	06991	19072	24210	36699	53728	28825	35793	28976	66252
00086	68434	94688	84473	13622	62126	98408	12843	82590	09815	93146
00087	48908	15877	54745	24591	35700	04754	83824	52692	54130	55160
00088	06913	45197	42672	78601	11883	09528	63011	98901	14974	40344
00089	10455	16019	14210	33712	91342	37821	88325	80851	43667	70883
00090	12883	97343	65027	61184	04285	01392	17974	15077	90712	26769
00091	21778	30976	38807	36961	31649	42096	63281	02023	08816	47449
00092	19523	59515	65122	59659	86283	68258	69572	13798	16435	91529
00093	67245	52670	35583	16563	79246	86686	76463	34222	26655	90802
00094	60584	47377	07500	37992	45134	26529	26760	83637	41326	44344
00095	53853	41377	36066	94850	58838	73859	49364	73331	96240	43642
00096	24637	38736	74384	89342	52623	07992	12369	18601	03742	83873
00097	83080	12451	38992	22815	07759	51777	97377	27585	51972	37867
00098	16444	24334	36151	99073	27493	70939	85130	32552	54846	54759
00099	60790	18157	57178	65762	11161	78576	45819	52979	65130	04860
00100	03991	10461	93716	16894	66083	24653	84609	58232	88618	19161
00101	38555	95554	32886	59780	08355	60860	29735	47762	71299	23853
00102	17546	73704	92052	46215	55121	29281	59076	07936	27954	58909
00103	32643	52861	95819	06831	00911	98936	76355	93779	80863	00514
00104	69572	68777	39510	35905	14060	40619	29549	69616	33564	60780
00105	24122	66591	27699	06494	14845	46672	61958	77100	90899	75754
00106	61196	30231	92962	61773	41839	55382	17267	70943	78038	70267
00107	30532	21704	10274	12202	39685	23309	10061	68829	55986	66485
00108	03788	97599	75867	20717	74416	53166	35208	33374	87539	08823
00109	48228	63379	85783	47619	53152	67433	35663	52972	16818	60311
00110	60365	94653	35075	33949	42614	29297	01918	28316	98953	73231
00111	83799	42402	56623	34442	34994	41374	70071	14736	09958	18065

00112	32960	07405	36409	83232	99385	41600	11133	07586	15917	06253
00113	19322	53845	57620	52606	66497	68646	78138	66559	19640	99413
00114	11220	94747	07399	37408	48509	23929	27482	45476	85244	35159
00115	31751	57260	68980	05339	15470	48355	88651	22596	03152	19121
00116	88492	99382	14454	04504	20094	98977	74843	93413	22109	78508
00117	30934	47744	07481	83828	73788	06533	28597	20405	94205	20380
00118	22888	48893	27499	98748	60530	45128	74022	84617	82037	10268
00119	78212	16993	35902	91386	44372	15486	65741	14014	87481	37220
00120	41849	84547	46850	52326	34677	58300	74910	64345	19325	81549
00121	46352	33049	69248	93460	45305	07521	61318	31855	14413	70951
00122	11087	96294	14013	31792	59747	67277	76503	34513	39663	77544
00123	52701	08337	56303	87315	16520	69676	11654	99893	02181	68161
00124	57275	36898	81304	48585	68652	27376	92852	55866	88448	03584
00125	20857	73156	70284	24326	79375	95220	01159	63267	10622	48391
00126	15633	84924	90415	93614	33521	26665	55823	47641	86225	31704
00127	92694	48297	39904	02115	59589	49067	66821	41575	49767	04037
00128	77613	19019	88152	00080	20554	91409	96277	48257	50816	97616
00129	38688	32486	45134	63545	59404	72059	43947	51680	43852	59693
00130	25163	01889	70014	15021	41290	67312	71857	15957	68971	11403
00131	65251	07629	37239	33295	05870	01119	92784	26340	18477	65622
00132	36815	43625	18637	37509	82444	99005	04921	73701	14707	93997
00133	64397	11692	05327	82162	20247	81759	45197	25332	83745	22567
00134	04515	25624	95096	67946	48460	85558	15191	18782	16930	33361
00135	83761	60873	43253	84145	60833	25983	01291	41349	20368	07126
00136	14387	06345	80854	09279	43529	06318	38384	74761	41196	37480
00137	51321	92246	80088	77074	88722	56736	66164	49431	66919	31678
00138	72472	00008	80890	18002	94813	31900	54155	83436	35352	54131
00139	05466	55306	93128	18464	74457	90561	72848	11834	79982	68416
00140	39528	72484	82474	25593	48545	35247	18619	13674	18611	19241
00141	81616	18711	53342	44276	75122	11724	74627	73707	58319	15997
00142	07586	16120	82641	22820	92904	13141	32392	19763	61199	67940
00143	90767	04235	13574	17200	69902	63742	78464	22501	18627	90872
00144	40188	28193	29593	88627	94972	11598	62095	36787	00441	58997
00145	34414	82157	86887	55087	19152	00023	12302	80783	32624	68691
00146	63439	75363	44989	16822	36024	00867	76378	41605	65961	73488
00147	67049	09070	93399	45547	94458	74284	05041	49807	20288	34060
00148	79495	04146	52162	90286	54158	34243	46978	35482	59362	95938
00149	91704	30552	04737	21031	75051	93029	47665	64382	99782	93478

00150	94015	46874	32444	48277	59820	96163	64654	25843	41145	42820
00151	74108	88222	88570	74015	25704	91035	01755	14750	48968	38603
00152	62880	87873	95160	59221	22304	90314	72877	17334	39283	04149
00153	11748	12102	80580	41867	17710	59621	06554	07850	73950	79552
00154	17944	05600	60478	03343	25852	58905	57216	39618	49856	99326
00155	66067	42792	95043	52680	46780	56487	09971	59481	37006	22186
00156	54244	91030	45547	70818	59849	96169	61459	21647	87417	17198
00157	30945	57589	31732	57260	47670	07654	46376	25366	94746	49580
00158	69170	37403	86995	90307	94304	71803	26825	05511	12459	91314
00159	08345	88975	35841	85771	08105	59987	87112	21476	14713	71181
00160	27767	43584	85301	88977	29490	69714	73035	41207	74699	09310
00161	13025	14338	54066	15243	47724	66733	47431	43905	31048	56699
00162	80217	36292	98525	24335	24432	24896	43277	58874	11466	16082
00163	10875	62004	90391	61105	57411	06368	53856	30743	08670	84741
00164	54127	57326	26629	19087	24472	88779	30540	27886	61732	75454
00165	60311	42824	37301	42678	45990	43242	17374	52003	70707	70214
00166	49739	71484	92003	98086	76668	73209	59202	11973	02902	33250
00167	78626	51594	16453	94614	39014	97066	83012	09832	25571	77628
00168	66692	13986	99837	00582	81232	44987	09504	96412	90193	79568
00169	44071	28091	07362	97703	76447	42537	98524	97831	65704	09514
00170	41468	85149	49554	17994	14924	39650	95294	00556	70481	06905
00171	94559	37559	49678	53119	70312	05682	66986	34099	74474	20740
00172	41615	70360	64114	58660	90850	64618	80620	51790	11436	38072
00173	50273	93113	41794	86861	24781	89683	55411	85667	77535	99892
00174	41396	80504	90670	08289	40902	05069	95083	06783	28102	57816
00175	25807	24260	71529	78920	72682	07385	90726	57166	98884	08583
00176	06170	97965	88302	98041	21443	41808	68984	83620	89747	98882
00177	60808	54444	74412	81105	01176	28838	36421	16489	18059	51061
00178	80940	44893	10408	36222	80582	71944	92638	40333	67054	16067
00179	19516	90120	46759	71643	13177	55292	21036	82808	77501	97427
00180	49386	54480	23604	23554	21785	41101	91178	10174	29420	90438
00181	06312	88940	15995	69321	47458	64809	98189	81851	29651	84215
00182	60942	00307	11897	92674	40405	68032	96717	54244	10701	41393
00183	92329	98932	78284	46347	71209	92061	39448	93136	25722	08564
00184	77936	63574	31384	51924	85561	29671	58137	17820	22751	36518
00185	38101	77756	11657	13897	95889	57067	47648	13885	70669	93406
00186	39641	69457	91339	22502	92613	89719	11947	56203	19324	20504
00187	84054	40455	99396	63680	67667	60631	69181	96845	38525	11600

00188	47468	03577	57649	63266	24700	71594	14004	23153	69249	05747
00189	43321	31370	28977	23896	76479	68562	62342	07589	08899	05985
00190	64281	61826	18555	64937	13173	33365	78851	16499	87064	13075
00191	66847	70495	32350	02985	86716	38746	26313	77463	55387	72681
00192	72461	33230	21529	53424	92581	02262	78438	66276	18396	73538
00193	21032	91050	13058	16218	12470	56500	15292	76139	59526	52113
00194	95362	67011	06651	16136	01016	00857	55018	56374	35824	71708
00195	49712	97380	10404	55452	34030	60726	75211	10271	36633	68424
00196	58275	61764	97586	54716	50259	46345	87195	46092	26787	60939
00197	89514	11788	68224	23417	73959	76145	30342	40277	11049	72049
00198	15472	50669	48139	36732	46874	37088	73465	09819	58869	35220
00199	12120	86124	51247	44302	60883	52109	21437	36786	49226	77837
00200	19612	78430	11661	94770	77603	65669	86868	12665	30012	75989
00201	39141	77400	28000	64238	73258	71794	31340	26256	66453	37016
00202	64756	80457	08747	12836	03469	50678	03274	43423	66677	82556
00203	92901	51878	56441	22998	29718	38447	06453	25311	07565	53771
00204	03551	90070	09483	94050	45938	18135	36908	43321	11073	51803
00205	98884	66209	06830	53656	14663	56346	71430	04909	19818	05707
00206	27369	86882	53473	07541	53633	70863	03748	12822	19360	49088
00207	59066	75974	63335	20483	43514	37481	58278	26967	49325	43951
00208	91647	93783	64169	49022	98588	09495	49829	59068	38831	04838
00209	83605	92419	39542	07772	71568	75673	35185	89759	44901	74291
00210	24895	88530	70774	35439	46758	70472	70207	92675	91623	61275
00211	35720	26556	95596	20094	73750	85788	34264	01703	46833	65248
00212	14141	53410	38649	06343	57256	61342	72709	75318	90379	37562
00213	27416	75670	92176	72535	93119	56077	06886	18244	92344	31374
00214	82071	07429	81007	47749	40744	56974	23336	88821	53841	10536
00215	21445	82793	24831	93241	14199	76268	70883	68002	03829	17443
00216	72513	76400	52225	92348	62308	98481	29744	33165	33141	61020
00217	71479	45027	76160	57411	13780	13632	52308	77762	88874	33697
00218	83210	51466	09088	50395	26743	05306	21706	70001	99439	80767
00219	68749	95148	94897	78636	96750	09024	94538	91143	96693	61886
00220	05184	75763	47075	88158	05313	53439	14908	08830	60096	21551
00221	13651	62546	96892	25240	47511	58483	87342	78818	07855	39269
00222	00566	21220	00292	24069	25072	29519	52548	54091	21282	21296

Source:222 rows of random numbers from the random number table of RAND