Directorate of Distance and Continuing Education
Manonmaniam Sundaranar University
Tirunelveli – 627012, Tamil Nadu.

M.A.ECONOMICS
(First Year)

Statistics for Economists-I
(JMEC12)

Prepared by

Dr. G. Monikanda Prasad
Assistant Professor of Economics
Manonmaniam Sundaranar University
Tirunelveli – 627 012.

STATISTICS FOR ECONOMISTS-I

Unit	Contents
I	**Introduction and Collection of Data** Introduction – Nature and scope of Statistics – Uses and Limitations of Statistics – Data Collection – Primary and Secondary Data – Tools for collecting primary Data – Requisites of Good Questionnaire – Sources of Secondary Data.
II	**Classification and presentation of Data** Classification and Tabulation of Data – Types – Frequency Distribution – Cumulative Frequency Distribution – Class interval – Diagrams – Types – Graphical Representation – Histogram – Frequency Polygon – Ogive Curve – Lorenz Curve.
III	**Measures of Central Tendency** Measures of Central Tendency – Requisites of a Good Average – Arithmetic Mean, Median, and Mode – Relative Merits and Demerits.
IV	**Measures of Dispersion** Absolute and Relative Measures of Dispersion – Range – Quartile Deviation – Mean Deviation – Standard Deviation - Variance – Coefficient of Variation – Skewness and Kurtosis.
V	**Correlation and Regression** Correlation – Types of Correlation – Methods – Karl Pearson's Co-efficient of Correlation – Spearman's Rank Correlation – Regression Equations – Distinction between Correlation and Regression Analysis.

UNIT-I
INTRODUCTION AND COLLECTION OF DATA

Introduction

Statistical enquiry/survey is conducted only to collect information pertaining to the topic of study. The information may be quantitative or qualitative, but in statistical enquiry, we expert numerical information. Information may also be collected from sources other than enquiry. It means that there are two sources of information or data. Simply speaking, the two sources of data are:

Primary sources and Secondary sources.

To collect first-hand information, it is necessary to conduct enquiry or survey and for second hand information enquiry is not necessary, but there must be enough secondary sources. The primary data are original in character while secondary data have already been collected by someone for some other purpose and are now available for the present study. The census data are primary data to the census department, but to researchers and other people, they are secondary.

Nature and scope of statistics

Statistics is the study of collection, organization, analysis, interpretation and presentation of data with the use of quantified models. In short, it is a mathematical tool that is used to collect and summarize data. Scope of Statistics: It presents the facts in numerical figures

Nature of Statistics

Statistics is both science and art. Statistical methods are systematic and have a general application which makes it a science. Further, the successful application of these methods requires skills and experience of using the statistical tools. These aspects make it an art.

Scope of Statistics

The scope of Statistics is very immense, the application of statistics goes into diverse fields such as solving social problems, industrial and scientific problems. Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments.

Uses and limitations of Statistics

Its primary scope is descriptive and inferential statistics, which includes summarizing and analyzing data, making predictions, and generalizations about a population. However, statistics also has its limitations, which include being limited to numerical data, sampling bias, and the inability to establish causation.

Statistics provide the information to educate how things work. They're used to conduct research, evaluate outcomes, develop critical thinking, and make informed decisions.

Data Collection

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

Primary Data

As discussed about, the data collected for the first time by the investigator himself or by his agents are called primary data. They are original in nature. They have to be presented, analyzed and interpreted by the researcher. Generally the data collected are purpose oriented.

Merits of Primary data

1. If the investigation is good, the data collected will be accurate and reliable.
2. Adequate and topic specific data can be collected.

3. They are original in nature and so primary data make the knowledge world prosperous.
4. As there is personal contact between the informants and the investigator, misinterpretation can be avoided and accuracy can be enhanced.
5. Questionnaire method of data collection makes the coverage of wide areas easier.

Demerits

1. The collection of primary data is more expensive and time consuming.
2. If requires a lot of men and material.
3. If the agents appointed for collecting information are inefficient, the data collected may be inaccurate.
4. There are changes for personal bias are prejudice.
5. If there is non- response from informants, then there will be undue delay in completing the survey.

Tools for Collecting Primary Data

As is well known, gathering primary data is costly and time intensive. The main techniques for gathering data are observation, interviews, questionnaires, schedules, and surveys.

Primary data refers to the first hand data gathered by the researcher himself. Secondary data means data collected by someone else earlier. Surveys, observations, experiments, questionnaire, personal interview, etc. Government publications, websites, books, journal articles, internal records etc.

Requisites of Good Questionnaire

Primary data refers to the first hand data gathered by the researcher himself. Secondary data means data collected by someone else earlier. Surveys, observations, experiments,

questionnaire, personal interview, etc. Government publications, websites, books, journal articles, internal records etc.

Sources of Secondary data

Secondary data is research data that has previously been gathered and can be accessed by researchers. The term contrasts with primary data, which is data collected directly from its source.

Secondary data is usually gathered from the published (printed) sources. A few major sources of published information are as follows: Published articles of local bodies, and central and state governments. Statistical synopses, census records, and other reports issued by the different departments of the government.

Secondary data in research methodology is any information or statistics that researchers have already collected through their primary resources. Secondary data is readily available for other individuals to reference as they conduct their own primary research, allowing them to gain insights into different processes that contribute to a research process. So, what can be primary data for one researcher may be secondary for another, depending on how they sourced it. Secondary researchers can gather data from various sources and summarise it into a new document that is easier to understand.

Secondary data can be a direct by-product of someone else's research procedures, and likely took the initial researcher significant time to develop and publish before it became readily available for other people to use. While primary data tends to be more time-consuming to gather, secondary data often requires minimal research, especially when using resources from the Internet of other digital mediums. The use of search engines and online databases has reduced the level of effort that was previously necessary for gathering large amounts of secondary data.

Advantages of Secondary Data

Listed are a few advantages of secondary data.

- **Easy to access**: Data is available anywhere and anytime it can be in the form of periodicals, magazines, or can be accessed anytime through the internet. People generally use secondary data maximum nowadays to evaluate their studies. A very small example is the students who depend on books, internet sites, and teachers to access information and prepare for exams.
- **Low cost or cost-effective**: The secondary data is of low cost as data are available at cheap rates, for example, the internet access, newspaper, or periodicals are available at cheaper rates and available in large quantities, so there is no non-availability of data to its users. Thus it is cost-effective.
- **Less time taking**: Data is available quickly and readily while primary data need to be collected first and then only after summarization data are used. Time taken to collect and analyze data is less than secondary data that is quickly available. Therefore it takes less time to take the source of data.
- **Various sources are available to collect data:** Secondary data is not only available through one source, but there are multiple sources like books, magazines, the internet, periodicals, and many more. Therefore various sources are available to collect data for analysis for its users. These sources are easily accessible and readily available to their users.
- **Data can be collected by anyone**: Anyone can collect data whether he /she is specialized in collecting it or not, depending upon the use. Also, there is no ownership of data that can be claimed by its user as data has already been shared by its owner, who was a primary collector of data.

- **The study is based on longitudinal analysis:** Since the data has been collected over years, thus a longitudinal analysis is done by the researchers with the help of secondary data. The data collected is more reliable and valid for users.

Disadvantages of Secondary Data

Listed are a few disadvantages of secondary data.

1. **Inaccuracy**: It is a limitation of secondary data that the data collected over the past few years may be inaccurate. The basis of data collected may not be correct or the analysis or interpretation made may not be accurate or relevant.
2. **Data may be sometimes outdated:** The data provided through different sources may also be outdated as it has been stored and managed for many years. Therefore it may also sometimes be outdated and may not be relevant for today's scenario.
3. **Not compatible with the needs of the user:** Since data is related to past surveys and according to the needs of the researchers of that time. It may happen that the present user of this data may not need or not have topics relevant to his study or research. Therefore here instead of outdated data, the data becomes irrelevant for the user to be used in research.
4. **Anyone can access data**: There is no privatization of data by its owner, data can be accessed by anyone willing to research on that topic. There is no secrecy of data but the user of data cannot appeal their possession or ownership of the data they accessed.
5. **Data quality cannot be controlled:** The researchers have no control over the quality of data. As data is already surveyed by researchers according to their relevant basis and there may be changes in the surroundings and other factors that may lead to the change in the data provided thus no proper quality can be controlled.
6. **Data can be biased**: Since data collected by the researcher is based on his/her opinion, therefore data is biased. And it may also have an impact on the data collected by the user of the secondary data.

UNIT –II
CLASSIFICATION AND PRESENTATION OF DATA

Classification:

Classification is used to group similar data into categories, while tabulation is used to present data in a structured and organized manner. Method: Classification involves grouping data based on certain criteria, while tabulation involves organizing data into rows and columns.

Types

Classification of data is also used in tabular presentation and is of four types; viz., Geographical or Spatial Classification, Chronological or Temporal Classification, Qualitative Classification, and Quantitative Classification.

Frequency Distribution

A frequency distribution is a representation, either in a graphical or tabular format , that displays the number of observations within a given interval. The frequency is how often a value occurs in an interval while the distribution is the pattern of frequency of the variable.

Cumulative Frequency Distribution

What is Cumulative Frequency Distribution? Cumulative frequency distribution is a form of frequency distribution that represents the sum of a class and all classes below it.

Remember that frequency distribution is an overview of all distinct values (or classes of values) and their respective number of occurrence.

Example of cumulative frequency

Number of books read in a month	Frequency	Cumulative Frequency
2	1	1
3	3	1 + 3 = 4
4	5	4 + 5 = 9
5	2	9 + 2 = 11
6	1	11 + 1 = 12

Here is an example of a cumulative frequency table. The cumulative frequency table shows how many books a student read each month over a one-year period. Look at the table to see the student read 2 books in one month, 3 books in three of the months, 4 books in five of the months, and so on.

Class Interval

Class interval refers to the numerical width of any class in a particular distribution. In maths, class interval is defined as the difference between the upper class limit and the lower class limit. The size of the class into which a particular data is divided. Eg. divisions on a histogram or bar graph.

Diagram

(a) One – dimensional diagram
(b) Two – dimensional diagram
(c) Three – dimensional diagram
(d) Pictograms
(e) Cartograms

Graphical Representation

Graphical representation is a form of visually displaying data through various methods like graphs, diagrams, charts, and plots. It helps in sorting, visualizing, and presenting data in a clear manner through different types of graphs. Statistics mainly use graphical representation to show data.

The four different types of graphical representation

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot.

Graphical data display

Graphical displays communicate comparisons, relationships, and trends. They emphasize and clarify numbers. To choose the appropriate type of display, first define the purpose of the report, and then identify the most effective display to suit that purpose.

Histogram

Histogram is a graph that shows the frequency of numerical data using rectangles. The height of a rectangle (the vertical axis) represents the distribution frequency of a variable (the amount, or how often that variable appears).

Difference between Bar Graph and Histogram

A histogram is one of the most commonly used graphs to show the frequency distribution. As we know that the frequency distribution defines how often each different value occurs in the data set. The histogram looks more similar to the bar graph, but there is a difference between them. The list of differences between the bar graph and the histogram is given below:

Histogram	Bar Graph
It is a two-dimensional figure	It is a one-dimensional figure
The frequency is shown by the area of each rectangle	The height shows the frequency and the width has no significance.
It shows rectangles touching each other	It consists of rectangles separated from each other with equal spaces.

Types of Histogram

The histogram can be classified into different types based on the frequency distribution of the data. There are different types of distributions, such as normal distribution, skewed distribution, bimodal distribution, multimodal distribution, and so on. The histogram can be used to represent these different types of distributions. The different types of a histogram are:

- Uniform histogram
- Symmetric histogram
- Bimodal histogram
- Probability histogram

Uniform Histogram

A uniform distribution reveals that the number of classes is too small, and each class has the same number of elements. It may involve distribution that has several peaks.

Symmetric Histogram

A symmetric histogram is also called a bell-shaped histogram. When you draw the vertical line down the center of the histogram, and the two sides are identical in size and shape, the histogram is said to be symmetric. The diagram is perfectly symmetric if the right half portion of the image is similar to the left half. The histograms that are not symmetric are known as skewed

Bimodal Histogram

If a histogram has two peaks, it is said to be bimodal. Bimodality occurs when the data set has observations on two different kinds of individuals or combined groups if the centers of the two separate histograms are far enough to the variability in both the data sets.

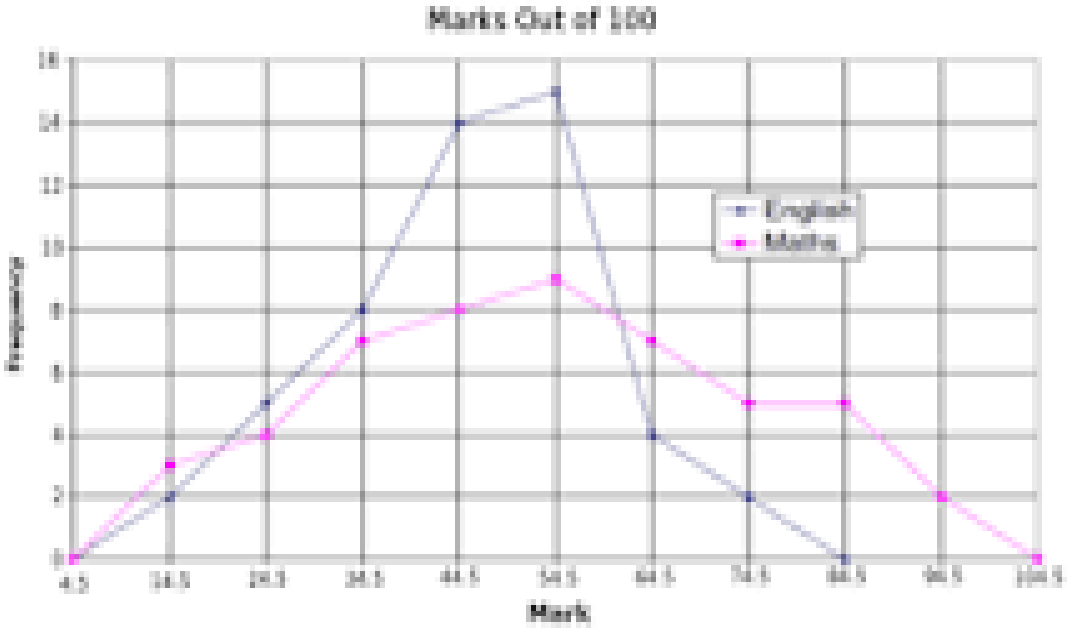
Probability Histogram

A Probability Histogram shows a pictorial representation of a discrete probability distribution. It consists of a rectangle centered on every value of x, and the area of each rectangle is proportional to the probability of the corresponding value. The probability histogram diagram is begun by selecting the classes. The probabilities of each outcome are the heights of the bars of the histogram.

Frequency Polygon

A frequency polygon is a line graph of class frequency plotted against class midpoint. It can be obtained by joining the midpoints of the tops of the rectangles in the histogram.

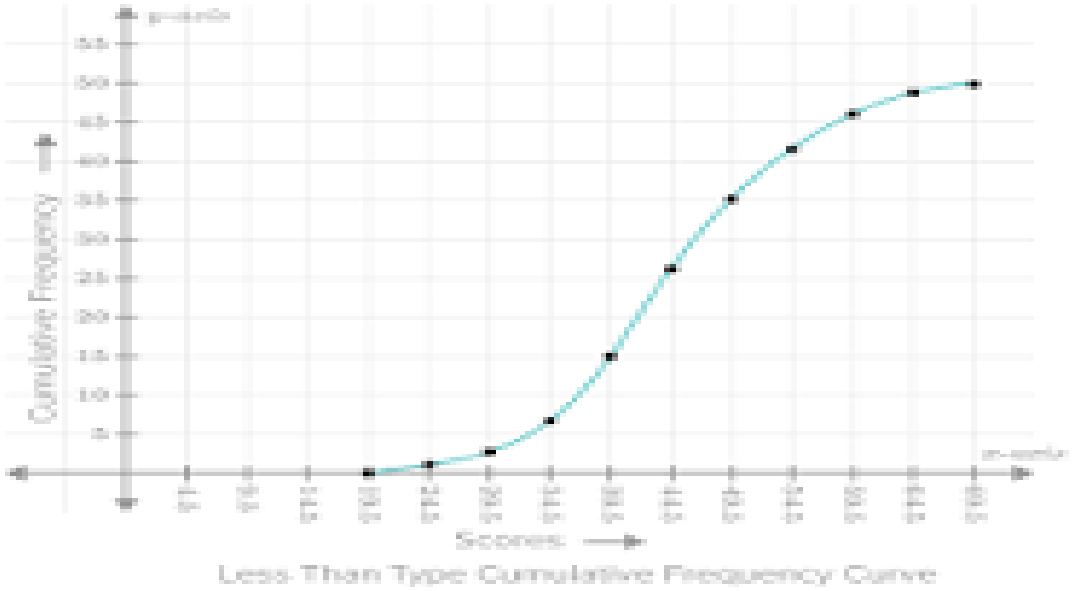
The percentage of a frequency polygon



To construct a relative frequency polygon:Sum the number of points in each interval, divide the sum of each interval by the total number of data points, and multiply by 100.

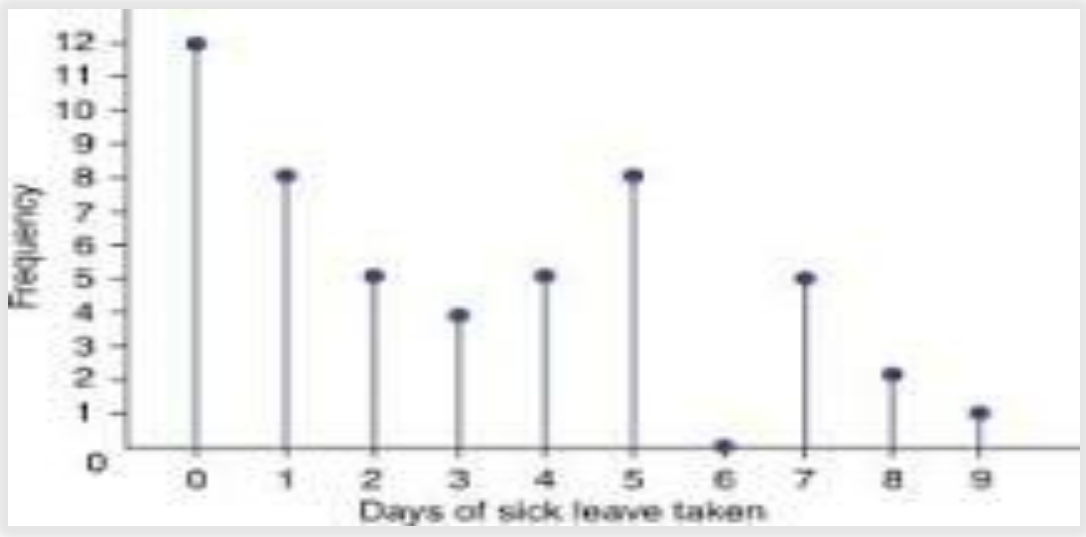
The result is the percentage of the total number of data points that is represented by each interval.

Community frequency polygon



The cumulative frequency polygon is essentially a line graph drawn on graph paper by plotting actual lower or upper limits of the class intervals on the -axis and the respective cumulative frequencies of these class intervals on the –axis.

Frequency table



A frequency table is simply a “t-chart” or two-column table which outlines the various possible outcomes and the associated frequencies observed in a sample.

Ogive Curve

The Ogive is defined as the frequency distribution graph of a series. The Ogive is a graph of a cumulative distribution, which explains data values on the horizontal plane axis and either the cumulative relative frequencies, the cumulative frequencies or cumulative per cent frequencies on the vertical axis.

Ogive Graph

The graphs of the frequency distribution are frequency graphs that are used to exhibit the characteristics of discrete and continuous data. Such figures are more appealing to the eye than the tabulated data. It helps us to facilitate the comparative study of two or more frequency distributions. We can relate the shape and pattern of the two frequency distributions.

The two methods of Ogives are:

- Less than Ogive
- Greater than or more than Ogive

Less than Ogive

The frequencies of all preceding classes are added to the frequency of a class. This series is called the less than cumulative series. It is constructed by adding the first-class frequency to the second-class frequency and then to the third class frequency and so on. The downward cumulation results in the less than cumulative series.

Greater than or More than Ogive

The frequencies of the succeeding classes are added to the frequency of a class. This series is called the more than or greater than cumulative series. It is constructed by subtracting the first class, second class frequency from the total, third class frequency from that and so on. The upward cumulation result is greater than or more than the cumulative series.

Uses of Ogive Curve

Ogive Graph or the cumulative frequency graphs are used to find the median of the given set of data. If both, less than and greater than, cumulative frequency curve is drawn on the same graph, we can easily find the median value. The point in which, both the curve intersects, corresponding to the x-axis, gives the median value. Apart from finding the medians, Ogives are used in computing the percentiles of the data set values.

Lorenz Curve

A **Lorenz curve**, developed by American economist Max Lorenz in 1905, is a graphical representation of income inequality or wealth inequality.

Lorenz curve

A Lorenz curve is a graphical representation of income inequality or wealth inequality. The graph plots percentiles of the population on the horizontal axis according to income or wealth.Lorenz curve represents the distribution of income in an economy. It is represented by a straight line that depicts the perfect distribution of income. Lorenz curve is beneath that line which shows the estimated distribution of income.

Merits of Lorenz curve

The Lorenz curve shows how income or wealth is distributed among a population. It plots the cumulative percentage of people on the x-axis and the cumulative percentage of income or wealth on the y-axis. The straight diagonal line represents perfect equality, where everyone has the same share of income or wealth.

Demerits of Lorenz curve

It is not possible to determine which distribution has more inequality. In the lifetime of an individual, there will be variation in income and this variation is not taken into consideration when inequality in the Lorenz Curves is analyzed. These are the limitations of the Lorenz Curves.

UNIT-III
MEASURES OF CENTRAL TENDENCY

Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

Requisites of a Good Average

To become good, a single value should satisfy the following conditions

1. Good Average should be based on all the observations: Only those averages, where all the data are used give best result, whereas the averages which use less data are not representative of the whole group.
2. Good Average should not be unduly affected by extreme value: No term should affect the average too much.
3. It should be rigidly defined. It means that there should be only one interpretation and only one value whoever may calculate it. In other words, a good average should be defined by an algebraic formula, which gives the same answer thought different people compute the average from the same data.

4. It should be based on all the items given. No value should not be omitted in the calculation of an average.
5. It should not be unduly affected by extreme values, here we mean is, very large values or very small values. Though every item influences the value of am average, any single value should not affected the value of an average unduly. For example in a series of 5,10,20,30,100, the last value 100 influences result unduly.
6. It should be capable of further algebraic treatment. It means that the average calculated should be used for further calculations or application of other statistical tools. For example, arithmetic mean is used in many further calculations such correlation and index number.
7. It should have sampling stability. The average calculated for different samples from a population should have minimum fluctuation or variation in value. It does not mean, all averages calculated should be the same, but with minimum difference.

Arithmetic Mean, Median, and Mode

Arithmetic Mean or simply mean is the value calculated by dividing the total sum of all values given by the number of values. There are three types of arithmetic mean. They are:

(a) Simple Arithmetic Mean

(b) Combined Arithmetic Mean

(c) Weighted Arithmetic Mean

Simple Arithmetic Mean

The arithmetic mean is the simplest and most widely used measure of a mean, or average. It simply involves taking the sum of a group of numbers, then dividing that sum by the count of the numbers used in the series.

Individual Observation:

Find the arithmetic mean for the following data under direct method:

S. No.	Marks
1	50
2	60
3	75
4	45
5	80

$\sum X = 310,\ n = 5$

$\overline{X} = \frac{310}{5} = 62$

$\overline{X} = 62$

$$\overline{X} = \frac{\sum X}{n}$$

Short-cut Method

Find the arithmetic mean for the following data under short-cut method:

S. No	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Income	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**

Solution

S. No	**Income**	$dx = x - 400$
1	100	-300
2	200	-200
3	300	-100
4	400	0
5	500	100
6	600	200
7	700	300
8	800	400
Total		400

A= Assumed Mean, $dx = X - A$

$A = 400, \sum d = 400, n = 8$

$$\overline{X} = A + \frac{\sum dX}{n}$$

$$\overline{X} = 400 + \frac{400}{8}$$

$\overline{X} = 400 + 50$

$\overline{X} = 450$

Calculation of Arithmetic Mean-Discrete Series

Calculate Arithmetic Mean

Income	**20**	**30**	**40**	**50**	**60**	**70**
f	**8**	**12**	**20**	**10**	**6**	**4**

Solution:

Income	**f**	*fx*
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280
Total	60	2460

$\overline{X} = 2460, n = 60$

$$\overline{X} = \frac{\sum fx}{n}$$

$$\overline{X} = \frac{2460}{60} = 41$$

Calculate mean from the following data using short cut method:

X	2	10	15	4	6	8
f	3	5	3	2	4	3

Solution:

X	f	$dx = x - A$	*fdx*
2	3	-3	-9
7	5	2	10
5	3	0	0
4	2	-1	-2
6	4	1	4
8	3	3	9
Total	20		$-11+23=12$

$A = 5, n = 20$

$$\overline{X} = A + \frac{\sum fdx}{n}$$

$$\overline{X} = 5 + \frac{12}{20}$$

$\overline{X} = 5 + 0.6,$

$\overline{X} = 5.6$

Continuous Series

There are three methods for calculating mean from continuous series.They are i. Direct method, ii. Short-cut method and Step-deviation method.

Class	0-2	2-4	4-6	6-8	8-10
Students	1	3	4	1	1

Class	f	m	fm
0-2	1	1	1
2-4	3	3	9
4-6	4	5	20
6-8	1	7	7
8-10	1	9	9
Total	10		46

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

$\sum fm = 46, N = 10$

$$\overline{X} = \frac{46}{10} = 4.6$$

$\overline{X} = 4.6$

Combined Arithmetic Mean

Combined arithmetic mean can be computed if we know the mean and number of items in each groups of the data. Geometric Mean (GM) is n th root of a number. GM is a measure of central tendency.

Weighted Arithmetic Mean

The weighted arithmetic mean is similar to an **ordinary** arithmetic mean (the most common type of average), except that instead of each of the data points.

Merits and Demerits of Mean

If the Arithmetic mean satisfies all the requisites of a good average, then there will be no demerits. Arithmetic mean is considered as the best average from the practical point of view. Though is the best average, it is not free from limitation. One can easily find out a particular measures merits and demerits if it is compared with the requisites of a good average.

Merits

1. It is easy to understand.
2. It is easy to calculate.
3. It is rigidly defined.
4. It is based on all the items of the series.
5. It is a more stable measure then other measures. It is less affected by sampling fluctuations.
6. It is amenable for further statistical calculation such as correlation and index number.
7. There is no necessity for arranging values in the calculation of mean whereas arrangement is mandatory for calculating median.

Demerits

1. It is unduly affected by extreme values.
2. It is necessary to convert inclusive classes into exclusive classes.
3. In the case of open-end classes, class adjustment has to be done.
4. It is not a good measure in extremely asymmetrical distribution.

Median

The median is the middle value in a set of data. First, organize and order the data from smallest to largest. To find the midpoint value, divide the number of observations by two. If there are an odd number of observations, round that number up, and the value in that position is the median.

Individual Observation:

Find the median for the following data.

X	12	10	15	5	8	18	20

Solution:

Arranged in ascending/ decending order
5
8
10
12
15
18
20

$$M = the\ size\ of \left(\frac{n+1}{2}\right)^{th} item$$

$$n = 7$$

$$M = the\ size\ of \left(\frac{7+1}{2}\right)^{th} item$$

$$M = the\ size\ of\ 4th\ item$$

$$M = the\ size\ of\ 4th\ item\ is\ 12$$

Calculation of Median-Discrete Series

Calculate Median using discrete series

x	**5**	**10**	**4**	**8**	**6**	**11**
f	**3**	**7**	**4**	**6**	**2**	**4**

$$M = the\ size\ of \left(\frac{n+1}{2}\right)^{th} item$$

Solution:

X	**F**	**c.f**
4	**4**	**4**
5	**3**	**7**
6	**2**	**9**
8	**6**	**15**
10	**7**	**22**
11	**4**	**26**

$$n = 26$$

$$M = the\ size\ of \left(\frac{26+1}{2}\right)^{th} item$$

$$M = the\ size\ of\ 13.5\ item\ is\ 8.$$

Continuous Series

Calculate Median for the following data:

Class	10-20	20-30	30-40	40-50	50-60
Students	7	6	4	2	6

Solution:

Class	**F**	**c.f**
10-20	7	7 (c.f)
20-30	6	13 (N/2)
30-40	4	17
40-40	2	19
50-60	6	25

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

$$N/2 = \frac{25}{2} = 12.5$$

$$M = 20 + \frac{25/2 - 7}{6} \times 10$$

$$M = 20 + \frac{12.5 - 7}{6} \times 10$$,,

$$M = 20 + \frac{5.5}{6} \times 10$$

$$M = 20 + 0.19 \times 10,$$
$$M = 20 + 1.9, M = 21.9$$

Merits

1. It is simply to calculate.
2. It is easy to understand.
3. Its is not affected by extreme values.
4. It can be located graphically.
5. It can be applied even in open-end classes with out adjustment.
6. It is the most appropriate average to deal with qualitative data.
7. It is very useful in markedly skewed distribution, particularly in income distribution.

Demerits

1 .Not based on all observations

2. Lack of representative character

3. Effect of sampling fluctuations

4. Lack of algebraic treatment.

Mode

The mode is the value that appears most frequently in a data set. A set of data may have one mode, more than one mode, or no mode at all. Other popular measures of central tendency include the mean, or the average of a set, and the median, the middle value in a set.

Calculate Mode for the following data:

Weight	**45-55**	**55-65**	**65-75**	**75-85**	**85-95**
Players	**5**	**8**	**10**	**15**	**8**

Solution:

Weight	**f**
35-45	**4**
45-55	**5**
55-65	**8**
65-75	**10 f_1**
75-85	**15 f_0**
85-95	**8 f2**

$$Z = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

Δ_1 = f_0-f_1, Δ_2 = f_0-f_2

L=75, Δ_1 = 15-10, Δ_2 = 15-8

$$Z = 75 + \frac{5}{5+7} \times 10$$

$$Z = 75 + \frac{50}{12}$$

$$Z = 75 + 4.17$$

$$Z = 79.17$$

Merits or Uses of Mode:

1. Mode is the term that occur most in the series hence it is not an isolated value like Median nor it is value like mean that may not be there in the series.

2. It is not affected by extreme values hence is a good representative of the series.

3. It can be found graphically also.

4. For open end intervals it is not necessary to know the length of open intervals.

5. It can also be used in case of Quantitative phenomenon.

6. With only just a single glance on data we can find its value. It is simplest.

7. It is the most used average in day today life, such as average marks of a class, average number of students in a section, average size of shoes, etc.

Demerits or Limitations of Mode:

1. Mode cannot be determined if the series is bimodal or multimodal.

2. Mode is based only on concentrated values; other values are not taken into account in-spite of their big difference with the mode. In continuous series only the lengths of class intervals are considered.

3. Mode is most affected by fluctuation of sampling.

4. Mode is not so rigidly defined. Solving the problem by different methods we won't get the same results as in case of mean.

5. It is not capable of further algebraic treatment. It is impossible to find the combined mode of some series as is in case of Mean

6. Also we can't find the total of whole series from value of mode as is in case of Mean.

7. If the number of terms is too large, only then we can call it as the representative value.

8. It is also said that sometimes mode is ill-defined, ill- definite and indeterminate.

Mode vs. Mean vs. Median

Mean, median, and mode are all different ways of noting the center of a data set.

Mode is the most common set of numbers, while mean is the average and median is the midpoint.

Mean

Mean is the average of a set of numbers. To calculate the mean, begin by adding up all of the data points and dividing by the total number of data points. For example, suppose you have the following series of numbers:

- 3, 3, 6, 9, 16, 16, 16, 27, 27, 37, 48

Added together, you get 208. Divide 208 by 11 (the number of data points) to get the mean, which is 18.9.

Median

The median is the data point in the middle of a set. To find the median, the numbers in the set must be arranged from smallest to largest. Let's use the numbers in the example above:

- 3, 3, 6, 9, 16, 16, 16, 27, 27, 37, 48

The median is 16, the data point in the exact middle of the set. This set has an odd number of data points, which makes it easier to find the middle. For a set with an even number of data points, you'd take the mean of the two middle numbers to find the median.

Mode

For example, in the following list of numbers, 16 is the mode since it appears more times in the set than any other number:

- 3, 3, 6, 9, **16, 16, 16**, 27, 27, 37, 48

A set of numbers can have more than one mode (this is known as *bimodal* if there are two modes) if there are multiple numbers that occur with equal frequency, and more times than the others in the set.

- **3, 3, 3**, 9, **16, 16, 16**, 27, 37, 48

In the above example, both the number 3 and the number 16 are modes as they each occur three times and no other number occurs more often.

UNIT-IV
MEASURES OF DISPERSION

Statistics refers to the study of data for a specific purpose. The data is analyzed, studied, and then interpreted. When the data is graphically represented, it gives us a clear and coherent idea about the salient features of the interpreted data. When the values are represented, they form the measure of central tendency. We may have come across them as mean, median, and mode. Through them, we get to know where the data is centered. However, it is equally important to know where the data is scattered or how much the data is bunched around the measure of central tendency. Whether it is grouped or ungrouped. In simple words, what the measure of central tendency might not tell us, the measure of dispersion can.

Definition of Measures of Dispersion

Dispersion in simple words means “scattered” or “spread”. In statistical data, dispersion refers to the extent to which the data is distributed. It can either be tightly clustered or widely scattered.

Absolute Measure of Dispersion

The absolute measure of dispersion consists of the same units as the original data set. This measure is used to identify the variations in terms of the average of deviations, such as standard deviation or mean deviation. It is inclusive of standard deviation, range, quartile deviation, etc.

There are different types of absolute measures of dispersion, such as:

Range: Range is the difference between the maximum value present in the data set and the minimum value present. Range = Maximum value – minimum value.

For example, the set range 1, 3, 5, 7, 9 is 9 – 1 = 8.

Variance: For finding variance, the first step is to subtract the mean from each value in the given set and then square each number. Add these squares and divide them by the total number of values present in the set to get the variance.

$\sigma 2$ (variance) = $\sum(X-\mu)2/ N$

Standard Deviation: Square root of the variance is simply known as the Standard deviation. S.D. (Standard Deviation) = $\sqrt{\sigma}$

Quartiles and quartile deviation: Quartile refers to the values which divide the set into quarters. Quartile deviation can be found by dividing the distance between the third and the first quartile by 2.

Mean and mean deviation: Mean is the average of all numbers in a given set. Further, calculating the average deviation from the mean value is known as the mean deviation.

Relative Measures of Dispersion

The relative measures of dispersion are employed for comparing the distribution of two data sets or more. They are used to compare the different unit sets. These include:

Coefficient of range

Coefficient of variation

Coefficient of standard deviation

Coefficient of quartile deviation

Coefficient of mean deviation

We calculate the dispersion coefficient when two series varying widely in their averages are to be compared. It is also used to compare two series having varied measurement units. The coefficient of dispersion is denoted as C.D.

The commonly used coefficients of dispersion have been listed below:

Coefficient of dispersion in terms of range

C.D. =$(Xmax–Xmin)/(Xmax+Xmin)$

Coefficient of dispersion in terms of Quartile deviation:

C.D. =(Q3–Q1)/(Q3+Q1)

Coefficient of dispersion in terms of Standard deviation:

C.D. = S.D./Mean

Coefficient of dispersion in terms of Mean deviation:

C.D. = Mean deviation/Average

The Coefficient of Variation

The Coefficient of Variation or CV is a measure of scattering/dispersion of given information details around the mean value. In mathematics, a coefficient is defined as an integer that is multiplied with the variable of a single element or the terms of a polynomial. It is usually a number, but sometimes may be followed by a letter in an expression.

For example: ax^2+bx+c.

Here

- x denotes the variable
- 'a' and 'b' are the coefficients of the equation.

CV is also known as relative standard deviation and in general, displays the size of a standard deviation to its mean.

Absolute and Relative Measures of Variation

Measures of dispersion may be either absolute or relative. Absolute measures of dispersion are expressed in the same statistical unit in which the original data are given such as rupees, kilograms, tonnes, etc. These values may be used to compare the variations in two distributions provided the variables are expressed in the same units and of the same average size.

A measures of relative dispersion is the ratio of a measures of absolute dispersion to an appropriate average. It is sometimes called a coefficient of dispersion, because "coefficient" means a pure number that is independent of the unit of measurement.

Range

Range is the simplest method of studying dispersion. It is defined as the difference between the value of the smallest item and the value of the largest item included in the distribution. Symbolically,

Where Range = L – S

L = Largest item, and

S = Smallest item.

The relative measure corresponding to range, called the coefficient of range, is obtained by applying the following formula:

$$\text{Coefficient of Range} = \frac{L-S}{L+S}$$

Calculate range and its coefficient.

The following are the prices of shares of AB co. Ltd. From Monday to Saturday:

Day	Price (Rs)	day	Price (Rs.)
Monday	200	Thursday	160
Tuesday	210	Friday	220
Wednesday	208	Saturday	250

Solution:

Range = L – S

Here L = 250 and S = 160

Range = 250 – 160 = Rs. 90

Coefficient of Range = $\frac{L-S}{L+S} = \frac{250-160}{250+160} = \frac{90}{410} = 0.22.$

Merits and Limitations: The merits and limitations of Range can be enumerated here.

Merits

- Amongst all the methods of studying dispersion range is the simplest to understand and the easiest to compute.
- It takes minimum time to calculate the value of range. Hence, if one is interested in getting a quick rather than a very accurate picture of variability one may compute range.

Limitations.

- Range is not based on each and every item of the distribution.
- It is subject to fluctuations of considerable magnitude from sample to sample.
- Range cannot tell us anything about the character of the distribution within the two extreme observations.

The Quartile Deviation

The range as a measure of dispersion discussed above has certain limitations. It is based on two extreme items and it fails to take account of the scatter within the range. From this there is reason to believe that if the dispersion of the extreme items is discarded, the limited range thus established might be more instructive. For this purpose there has been developed a measure called the interquartile range, the range which includes the middle 50 per cent of the distribution.

In other words, interquartile range represents the difference between the third quartile and the first quartile.

Symbolically,

Interquartile range = $Q_3 - Q_1$

Very often the interquartile range is reduced to the form of the Semi-interquartile range or quartile deviation by dividing it by 2.

Symbolically,

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

Coefficient of Q.D. = $\frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Individual Observations:

Coefficient of quartile deviation can be used to compare the degree of variation in different distributions.

Computation of Quartile Deviation

The process of computing quartile deviation is very simple. We have just to compute the values of the upper and lower quartiles. The following illustrations would clarify calculations.

Find out the value of quartile deviation and its coefficient from the following data:

Roll No	1	2	3	4	5	6	7
Marks	20	28	40	12	30	15	50

Solution:

Marks arranged in ascending order:

12	15	20	28	30	40	50

Q_1 = Size of $\frac{N+1}{4} th\, item = Size\, of\, \frac{7+1}{4} = 2nd\, item$

Size of 2nd item is 15. Thus $Q_1 = 15$

Q_3 = Size of 3 $\left(\frac{N+1}{4}\right)$ th item = Size of $\left(\frac{3*8}{4}\right)$ th item = 6^{th} item

Size of 6^{th} item is 40. Thus $Q_3 = 40$

$$\text{Q. D.} = \frac{Q_3 - Q_1}{2} = \frac{40-15}{2} = 12.5.$$

$$\text{Coefficient of Q. D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} == \frac{40-15}{40+15} = \frac{25}{55} = 0.455.$$

Discrete Series

Compute coefficient of quartile deviation from the following data:

Marks	10	20	30	40	50	60
No. of Student	4	7	15	8	7	2

Solution:

Calculation coefficient of quartile from the following data:

Marks	Frequency	c.f.	Marks	Frequency	c. f.
10	4	4	40	8	34
20	7	11	50	7	41
30	15	26	60	2	43

Q_1 = Size of $\frac{N+1}{4}\ th item\ \frac{43+1}{4} = 11^{th}$ item

Size of 11^{th} item is 20. Thus, $Q_1 = 20$

Q_3 = Size of $3\left(\frac{N+1}{4}\right)$ th item = $\frac{3*44}{4} = 33^{rd}$ item

Size of 33^{rd} item is 40. Thus, $Q_3 = 40$

$$\text{Q. D.} = \frac{Q_3 - Q_1}{2} = \frac{40-15}{2} = 10$$

$$\text{Coefficient of Q. D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40-20}{40+20} = 0.333.$$

Calculate quartile deviation and the coefficient of quartile deviation from the following data:

Wages in Rupees per week	Less than 35	35-37	38-40	41-43	Over 43
Number of wage earners	14	62	99	18	7

Solution:

Wages (Rs. Per week)	**f**	**c.f.**
Less then 35	14	14
35-37	62	76
38-40	99	175
41-43	18	193
Over 43	7	200

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

$$Q_1 = Size\,of\ \frac{N}{4}th\,item = \frac{200}{4} = 50th\,item$$

Q_1 lies in the class 35-37.

$$Q_1 = L + \frac{N/4 - c.f.}{f} * i$$

L = 35, N/4 = 50, c.f. = 14, f = 62, i = 2

$$Q_1 = 35 + \frac{50-14}{62} * 2 = 35 + 1.16 = 36.16$$

$$Q_1 = Size\,of\ \frac{3N}{4}th\,item = \frac{3*200}{4} = 150th\,item$$

Q_3 lies in the class 38-40.

$$Q_1 = L + \frac{N/4 - c.f.}{f} * i$$

L = 38, N/4 = 150, c.f. = 76, f = 99, i = 2

$$Q_3 = 38 + \frac{150-76}{99} * 2 = 38 + 1.49 = 39.49$$

$$Q.D. = \frac{39.49 - 36.16}{2} = 1.67$$

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

$$= \frac{39.49 - 36.16}{39.49 + 36.16} = \frac{3.33}{75.65} = 0.044$$

Merits and Limitations

Merits. In certain respects it is superior to range as a measure of dispersion.

- It has a special utility in measuring variation in case of open end distributions or one in which the data may be ranked but measured quantitatively.
- It is also useful in erratic or badly slewed distributions, where the other measures of dispersion would be warped by extreme values. The quartile deviation is not affected by the presence of extreme values.

Limitation.

Quartile deviation ignores 50% items. i,e., the first 25% and the last 25%. As the value of quartile deviation does not depend open every item of the series, it cannot be regarded as a good method of measuring dispersion.

- It is not capable of mathematical manipulation.
- Its value is very much affected by sampling fluctuations.
- It is fact not a measure of dispersion as it really does not show the scatter around an average but rather a distance on a scale, i.e., quartile deviation is not itself measured from an average, but it is a positional average.

The Mean Deviation

The mean deviation is also known as the average deviation. It is the average difference between the items in a distribution and the median or mean of that series.

Calculation of Mean Deviation

Discrete Series. In discrete series the formula for calculating mean deviation is

$$M.D. = \frac{\sum f \backslash D}{N}$$ (by the same logic as given before)

/D/ denotes deviation from median ignoring sings.

Steps

- Calculate the median of the series.
- Take the deviation of the item from median ignoring sings and denote them by /D/.
- Multiply these deviations of the respective frequencies and obtain the total $\sum f / D /$.
- Divide the total obtained in step (ii) by the number of observation. This gives us the value of mean deviation.

Calculate mean deviation from the following series.

X	10	11	12	13	14
F	3	12	18	12	3

Solution:

X	**F**	**/D/**	**f/D/**	**c.f.**
10	3	2	6	3
11	12	1	12	15
12	18	0	0	33
13	12	1	12	45
14	3	2	2	48

$$M.D. = \frac{\sum f \setminus D}{N}$$

Median = Size of $\frac{N+1}{4} th item = th item = \frac{48+1}{2} = 24.5 th item$

Size of 24.5^{th} item is 12, hence Median = 12

$$M.D. = \frac{36}{48} = 0.75.$$

Calculate the mean deviation from the mean for the following data:

Size	2	4	6	8	10	12	14	16
Frequency	2	2	4	5	3	2	1	1

Solution:

X	**F**	**fx**	**/x-8/ /D/**	**f/D/**
2	2	4	6	12
4	2	8	4	8
6	4	24	2	8
8	5	40	0	0
10	3	30	2	6
12	2	24	4	8
14	1	14	6	6
16	1	16	8	8
	N=20	**$\sum$ fX=160**		**$\sum$ f/D/=56**

$$\bar{X} \frac{\sum fX}{N} = \frac{160}{20} = 8$$

$$M.D. = \frac{\sum f \setminus D}{N} = \frac{56}{20} = 2.8 .$$

Calculation of Mean Deviation – Continuous Series

For calculation mean deviation in continuous series the procedure remains the same as discussed above. The only difference is that here we have to obtain the mid-point of the various classes and take deviations of these points from median. The formula is same, i.e.

$$M.D. = \frac{\sum f \backslash D}{N}$$

Find the median and mean deviation of the following data:

Size	Frequency	Size	Frequency
0-10	7	40-50	16
10-20	12	50-60	14
20-30	25	60-70	8
30-40	25		

Solution:

Calculation of median and mean deviation

Size	f	c.f.	m.p.	/m-35.2/ /D/	f/D/
0-10	7	7	5	30.2	211.4
10-20	12	19	15	20.2	242.4
20-30	18	37	25	10.2	183.6
30-40	25	62	35	0.2	5.0
40-50	16	78	45	9.8	156.8
50-60	14	92	55	19.8	277.2
60-70	8	100	65	29.8	238.4
	N=100				∑ f/D/=1314.8

$$\text{Median} = Size\,of\ \frac{N}{2}th\,item = \frac{100}{2} = 50th\,item$$

Median lies in the class 30-40

$$Med. = L + \frac{N/4 - c.f.}{f} * i$$

L = 30, N/4 = 50, c.f. = 37, f = 25, i = 10

$$\text{Med.} = 30\frac{50+37}{25} \times 10 = 30 + 5.2 = 35.2$$

$$\text{M.D .} = \frac{\sum f / D/}{N} = \frac{1314.8}{100} = 13.148$$

The Standard Deviation

The standard deviation concept was introduced by Karl Pearson in 1823. It is by far the most important and widely used measure of studying dispersion. Its significance lies in the fact that it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion. Standard deviation is also known as root mean square deviation for the reason that it is the square root of the mean of the squared deviation from that arithmetic mean.

Difference between Mean Deviation and Standard Deviation

Both these measure of dispersion are based on each and every item of the distribution.

But they differ in the following respects:

- Algebraic signs are ignored while calculating mean deviation whereas in the calculating of standard deviation signs are taken into account.
- Mean deviation can be computed either from median or mean. The standard deviation, on the other hand, is always computed from the arithmetic mean because the sum of the squares of the deviation of items from arithmetic mean is the least.

Calculating of Standard Deviation

Individual Observations: In case of individual observation standard deviation may be computed by applying any of the following two methods:

1. By taking deviation of the items from the actual mean.

2. By taking deviations of the items from an assumed mean.

Deviations taken from actual mean. When deviations are taken from actual mean the following formula is applied:

$$\sigma^* = \sqrt{\frac{\sum x^2}{N}}$$

$$x(X - \bar{X})$$

Steps:

- Calculate the actual mean of the series, i.e., $\bar{X}$.
- Take the deviations of the items from the mean, i.e. find $x(X - \bar{X})$. Denote these deviations by x.
- Square these deviation and obtain the total $\sum x^2$.
- Divide $\sum x^2$ by the total number of observations, i.e. N and extract the square-root. This gives us the value of standard deviation.

Deviations taken from Assumed Mean

When the actual mean is in fractions, say it is 123.674 it would be too cumbersome to take deviations from it and then obtain square of these deviations. In such a case either the mean may be approximated or else the deviations be taken from an assumed mean and the necessary adjustment made in the value of the standard deviation. The former method of approximation is less accurate and therefore, invariably in such a case deviations are taken from assumed mean.

When deviations are taken from assumed mean the following formula is applied:

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

Steps.

- Take the deviation of the items from an assumed mean. i.e. obtain (X-A). Denote these deviations by d. Take the total of these deviations, i..e., obtain $\sum d$.
- Square these deviations and obtain the total $\sum d^2$.
- Substitute the values of $\sum d^2$, $\sum d$ and N in the above formula.

Solution:

x	(x-264)* d	D^2
240	-24	576
260	-4	16
290	+26	676
245	-19	361
255	-9	81
288	+24	576
272	+8	64
263	-1	1
277	+13	169
251	-13	169
$\sum X = 2641$	$\sum d = +1$	$\sum d^2 = 2689$

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

$$\sum d^2 = 2689, \sum d = +1, N = 10$$

$$\sigma = \sqrt{\frac{2689}{10} - \left(\frac{1}{10}\right)^2}$$

$$= \sqrt{268.9 - 0.01} = 16.398.$$

Calculate the standard deviation from the following observations:

240.12	240.13	240.15	240.12	240.17
240.15	240.17	240.16	240.22	240.21

Solution:

x	(x - 240)	D^2
240.12	+0.12	.0144
240.13	+0.13	.0169
240.15	+0.15	.0225
240.12	+0.12	.0144
240.17	+0.17	.0289
240.15	+0.15	.0225
240.17	+0.17	.0289
240.16	+0.16	.0256
240.22	+0.22	.0484
240.21	+0.21	.0441
N=10	$\sum d = +1.60$	$\sum d^2 = 0.2666$

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

$$= \sqrt{\frac{0.2666}{10} - \left(\frac{1.6}{10}\right)^2} = \sqrt{0.02666 - 0.0256} = 0.033$$

Calculation of Standard Deviation – Discrete Series.

For calculating standard deviation in discrete series, any of the following methods may be applied:

1. Actual mean method.
2. Assumed mean method.
3. Step deviation method.

(a) Actual Mean Method - When this method is applied, deviations are taken from the actual mean, i.e. we find $(X - \bar{X})$ and denote these deviations by x. These deviations are then squared and multiplied by the respective frequencies. The following formula is applied:

$$\sigma = \sqrt{\frac{\sum fx^2}{N}}, where = (X - \bar{X})$$

However, in practice this method is rarely used because if the actual mean is in fractions the calculation take a lot of time.

(b) Assumed Mean Method - when this method is used, the following formula is applied:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$, where d =(X - A).

(b) Step Deviation Method - when this method is used, the following formula is applied:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$, where $d = \frac{(X - A)}{i}$

Calculate the standard deviation:

X	45	50	55	60	65	70	75	80
f	3	5	8	7	9	7	4	7

Solution:

X	f	d = (X-60)/5	fd	fd^2
45	3	-3	-9	27
50	5	-2	-10	20
55	8	-1	-8	8
60	7	0	0	0
65	9	1	9	9
70	7	2	14	28
75	4	3	12	36
80	7	4	28	112
	N=50		$\sum fd = 36$	$\sum fd^2 = 240$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

$$= \sqrt{\frac{240}{50} - \left(\frac{36}{50}\right)^2} \times 5$$

$$= \sqrt{4.8 - 0.5184} \times 5$$

$$= 10.35$$

Standard Deviation-Continuous Series

Find the Standard Deviation from the following data:

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	15	15	23	22	25	10	5	10

Solution:

Class	f	m = m.p	d = m-35/10	fd	fd^2
0-10	15	5	-3	-45	135
10-20	15	15	-2	-30	60
20-30	23	25	-1	-23	23
30-40	22	35	0	0	0
40-50	25	45	1	25	25
50-60	10	55	2	20	40
60-70	5	65	3	15	45
70-80	10	75	4	40	160
	N=125			$\sum fd = 2$	$\sum fd^2 = 488$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

$$= \sqrt{\frac{1310}{480} - \left(\frac{-22036}{480}\right)^2} \times 5$$
$$= \sqrt{2.729 - .21} \times 5$$
$$= 7.936$$

Variance

The term variance was used to describe the square of the standard deviation by R. A. Fisher in 1913. The concept of variance is highly important in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variation in their original series. Variance is defined as follows:

$$Variance = \frac{\sum\left(X - \overline{X}\right)^2}{N}$$

Advantages

- The biggest advantage of using variance is gaining information about a data set. Whether you're an investor mitigating risk or a statistician who wants to understand the spread of a sample, the variance is information that people can use to draw quick inferences.
- It's faster to use a variance than to plot each number on a spread and determine the approximate distance between the mean and each variable. This measure allows people who use statistics to make important estimations with a relatively quick calculation that provides information about the range of a sample.
- Variance treats all numbers in a set the same, regardless of whether they're positive or negative, which is another advantage of using this formula.

Disadvantages

- One disadvantage of using variance is that larger outlying values in the set can cause some skewing of data, so it isn't necessarily a calculation that offers perfect accuracy. That's because, once squared, outliers on either side of the population can have a

significant weight associated with them, depending on the values in the rest of the sample.

- Some researchers, who prefer to work with smaller numbers exacerbate this, so they might prefer to work in standard deviations, which take the square root of the variance and is less likely to skew heavily toward high numbers. Variance can also be difficult to interpret, which is another reason its square root might be preferable for data analysis and simpler calculations.

Coefficient of Variation

The Standard of deviation is an absolute measure of dispersion. The corresponding relative measure is known as the coefficient of variation. This measure is developed by Karl Pearson.

Coefficient of variation is denoted by C.V $= \frac{\sigma}{\overline{X}} \times 100$

Merits:

The CV is independent of the unit in which the measurement has been taken, but standard deviation depends on units of measurement. Hence one should use the coefficient of variation instead of the standard deviation.

Limitations:

If the value of mean approaches 0, the coefficient of variation approaches infinity. So the minute changes in the mean will make major changes.

Skewness and Kurtosis

Skewness is a statistical measure of the asymmetry of a probability distribution. It characterizes the extent to which the distribution of a set of values deviates from a normal distribution. Skewness between -0.5 and 0.5 is symmetrical.

Skewness can be categorized as:

1. **Positive Skewness:** When the right side of the distribution (tail) is longer or fatter, the data set is said to have positive skewness. The mean and median in this case are higher than the mode.
2. **Negative Skewness:** In this case, the left side of the distribution is longer or fatter. The mean and median are less than the mode.
3. **Zero Skewness**:If the data set follows a perfect symmetrical distribution, it has zero skewness. The mean, median, and mode are all equal.

Kurtosis:

Kurtosis measures whether data is heavily heavy-tailed or light-tailed.Kurtosis is a measure of the "tailedness" of the probability distribution. A standard normal distribution has kurtosis of 3 and is recognized as mesokurtic. An increased kurtosis (>3) can be visualized as a thin "bell" with a high peak whereas a decreased kurtosis corresponds to a broadening of the peak and "thickening" of the tails. Kurtosis >3 is recognized as leptokurtic and <3 as platykurtic (lepto=thin; platy=broad). There are four different formats of kurtosis, the simplest is the population kurtosis; the ratio between the fourth moment and the variance.

Types of kurtosis include:

1. **Leptokurtic:**This refers to distributions with kurtosis greater than 3. They have heavier tails, meaning more outliers, and a sharper peak than the normal distribution.
2. **Mesokurtic:** This is the classification of a standard normal distribution with a kurtosis of 3.
3. **Platykurtic:** Refers to distributions with kurtosis less than 3. They have a flatter peak and lighter tails, meaning fewer outliers than the normal distribution.

Key Difference between Skewness and Kurtosis

S.No	Skewness	Kurtosis
1.	Measures asymmetry in a data set	Measures the "tailedness" of the distribution
2.	Three types: Positive, Negative, and Zero Skewness	Three types: Leptokurtic, Mesokurtic, Platykurtic
3.	Describes the shape and direction of the skew (left or right)	Describes the shape of the distribution's peak and tails
4.	Heavily influenced by the size and direction of the tail	Heavily influenced by outliers and extreme values
5.	Can help identify potential outliers	Determines the probability of extreme values
6.	In a normal distribution, the skewness is zero	In a normal distribution, the kurtosis is 3
7.	Has implications on mean, median, and mode	Has implications on the peak and tails of the distribution
8.	Used in portfolio theory to analyze returns	Used to evaluate investment risks
9.	Skewness value can be positive, negative, or zero	Kurtosis value can be positive and is always greater than 1
10.	Changes with the transformation of each data point	Changes with the transformation of square of each data point

UNIT-V
CORRELATION AND REGRESSION

CORRELATION:

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

Correlation refers to the statistical relationship between two entities. In other words, it's how two variables move in relation to one another. Correlation can be used for various data sets, as well. In some cases, you might have predicted how things will correlate, while in others, the relationship will be a surprise to you. It's important to understand that correlation does not mean the relationship is causal.

To understand how correlation works, it's important to understand the following terms:

Positive Correlation

A positive correlation would be 1. This means the two variables moved either up or down in the same direction together.

Negative Correlation

A negative correlation is -1. This means the two variables moved in opposite directions.

Zero or no Correlation:

A correlation of zero means there is no relationship between the two variables. In other words, as one variable moves one way, the other moved in another unrelated direction.

Types of correlation coefficients

While correlation studies how two entities relate to one another, a correlation coefficient measures the strength of the relationship between the two variables. In statistics, there are three types of correlation coefficients. They are as follows:

Pearson correlation:

The Pearson correlation is the most commonly used measurement for a linear relationship between two variables. The stronger the correlation between these two datasets, the closer it'll be to +1 or -1.

Spearman correlation:

This type of correlation is used to determine the monotonic relationship or association between two datasets. Unlike the Pearson correlation coefficient, it's based on the ranked values for each dataset and uses skewed or ordinal variables rather than normally distributed ones.

Kendall Correlation:

This type of correlation measures the strength of dependence between two datasets.

KARL PEARSON'S COEFFICIENT OF CORRELATION

The value of the correlation is obtained by the below formula and the value always lie between ± 1.

To calculate the Karl Pearson's Coefficient of Correlation, the following formula is

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

$$x = (X - \overline{X}); y = (Y - \overline{Y})$$

$$\sigma_x = St\,\text{an}\,dard\ deviation\ of\ series\ X$$

$$\sigma_y = St\,\text{an}\,dard\ deviation\ of\ series\ Y$$

Calculate Karl Pearson's coefficient of correlation from the following data.

X	48	35	17	23	47
Y	45	20	40	25	45

Solution:

X	x = X-34	x^2	Y	y = Y-35	y^2	xy
48	14	196	45	10	100	140
35	1	1	20	-15	225	-15
17	-17	289	40	5	25	-85
23	-11	121	25	-10	100	110
47	13	169	45	10	100	130
$\sum X = 170$	$\sum x = 0$	$\sum x^2 = 776$	$\sum Y = 175$	$\sum y = 0$	$\sum y^2 = 550$	$\sum xy = 280$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$x = (X - \bar{X}),\ y = (Y - \bar{Y})$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{170}{5} = 34\ ;\ \bar{Y} = \frac{\Sigma Y}{N} = \frac{175}{5} = 35$$

$$\Sigma xy = 280,\ \Sigma x^2 = 776,\ \Sigma y^2 = 550$$

$$r = \frac{280}{\sqrt{776 \times 550}} = \frac{280}{653 \cdot 299} = 0 \cdot 429$$

DIRECT METHOD OF FINDING CORRELATION COEFFICIENT:

Coefficient of Correlation can also be calculated without taking deviations, the following formula is used

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Calculate Karl Pearson's coefficient of correlation using direct method.

X	9	8	7	6	5	4	3	2	1
Y	15	16	14	13	11	12	10	8	9

Solution:

X	**X^2**	**Y**	**Y^2**	**XY**
9	81	15	225	135
8	64	16	256	128
7	49	14	196	98
6	36	13	169	78
5	25	11	21	55
4	16	12	144	48
3	9	10	100	30
2	4	8	64	16
1	1	9	81	9
$\sum X = 45$	$\sum X^2 = 285$	$\sum Y = 108$	$\sum Y = 1356$	$\sum XY = 597$

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$N = 9,\ \Sigma XY = 597,\ \Sigma X = 45,\ \Sigma Y = 108,\ \Sigma X^2 = 285,\ \Sigma Y^2 = 1{,}356$

$$r = \frac{9 \times 597 - 45 \times 108}{\sqrt{9 \times 285 - (45)^2}\sqrt{9 \times 1356 - (108)^2}}$$

$$r = \frac{5373 - 4860}{\sqrt{2565 - 2025}\sqrt{12{,}204 - 11{,}664}}$$

$$r = \frac{513}{\sqrt{540 \times 540}} = \frac{513}{540} = +0{\cdot}95.$$

Deviation taken from Assumed Mean:

Coefficient of Correlation can also be calculated taking deviations, the following formula is used

$$r = \frac{N\sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{N\sum d_x^2 - (\sum d_x)^2}\sqrt{N\sum d_y^2 - (\sum d_y)^2}}$$

Σd_x = sum of the deviations of X series from an assumed mean
Σd_y = sum of the deviations of Y series from an assumed mean.
$\Sigma d_x d_y$ = sum of the product of the deviations of X and Y series from their assumed means
Σd_x^2 = sum of the squares of the deviations of X series from an assumed mean
Σd_y^2 = sum of the squares of the deviations of Y series from an assumed mean

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$\Sigma d_x d_y = 2248,\ \Sigma d_x = 41,\ \Sigma d_y = 108$

$\Sigma d_x^2 = 1727,\ \Sigma d_y^2 = 3468,\ N = 8$

$$r = \frac{(8)(2248) - (41)(108)}{\sqrt{(8)(1727) - (41)^2}\sqrt{(8)(3468) - (108)^2}}$$

$$= \frac{17984 - 4428}{\sqrt{13816 - 1681}\sqrt{27744 - 11664}}$$

$$= \frac{13556}{\sqrt{12135}\sqrt{16080}} = \frac{13556}{110{\cdot}16 \times 126{\cdot}806} = \frac{13556}{13968{\cdot}95} = +0{\cdot}97$$

Regression Equations

A regression equation is used in stats to find out what relationship, if any, exists between sets of data. For example, if you measure a child's height every year you might find that they grow about 3 inches a year. That trend (growing three inches a year) can be modeled with a regression equation. In fact, most things in the real world (from gas prices to hurricanes) can be modeled with some kind of equation; it allows us to predict future events.

Since there are two regression equations., i.e., the regression equation of X on Y is used to describe the variations in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the value of Y for given changes in X.

Regression equation of Y on X

The regression equation of Y on X is expressed as follows:

$$Y=a+bX$$

In this equation, Y is a dependent variable, X is independent variable, a is Y-intercept and b is the slope of line.

a and b in the equation are called numerical constants because for any given straight line, their value does not change. Such a line is known as the 'line of best fit'

To determine the values of a and b, the following two normal equations are to be solved simultaneously;

$$\sum X = Na + b\sum Y$$
$$\sum XY = a\sum Y + b\sum Y^2$$

Deviation Taken from Arithmetic Means of X and Y

The above method of finding out regression equation is tedious. The calculations can very much be simplified if instead of dealing with the actual values of X and Y we take the

deviations of X and Y series from their respective means. In such a case the two regression equations are written as follows:

(i) Regressive Equation of X on Y ; $X - \overline{X} = r\frac{\sigma x}{\sigma y}(Y - \overline{Y})$

$$r\frac{\sigma x}{\sigma y} = \frac{\sum xy}{\sum y^2}$$

(ii)) Regressive Equation of Y on X; $Y - \overline{Y} = r\frac{\sigma y}{\sigma x}(\mathrm{X} - \overline{X})$

$$r\frac{\sigma y}{\sigma x} = \frac{\sum xy}{\sum x^2}$$

From the following table calculate the regression equation taking deviation from the mean of X and Y series.

X	6	2	10	4	8
Y	9	11	5	8	7

Solution:

X	$x = X - \overline{X}$	x^2	Y	$y = Y - \overline{Y}$	y^2	xy
6	0	0	9	+1	1	0
2	-4	16	11	+3	9	-12
10	+4	16	5	-3	9	-12
4	-2	4	8	0	0	0
8	+2	4	7	-1	1	-2
$\sum X = 30$	$\sum x = 0$	$\sum x^2 = 40$	$\sum Y = 40$	$\sum y = 0$	$\sum y^2 = 20$	$\sum xy = -26$

Regression Equation of X on Y : $X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{20} = -1{\cdot}3$$

$$\overline{X} = \frac{30}{5} = 6\ ;\ \overline{Y} = \frac{40}{5} = 8$$

Hence, $X - 6 = -1{\cdot}3\,(Y - 8) = -1{\cdot}3\,Y + 10{\cdot}4$

$X = -1{\cdot}3\,Y + 16{\cdot}4$ or $X = 16{\cdot}4 - 1{\cdot}3\,Y$

Regression Equation of Y on X : $Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{-26}{40} = -0{\cdot}65$$

$Y - 8 = -0{\cdot}65\,(X - 6) = -0{\cdot}65\,X + 3{\cdot}9$

$Y = -0{\cdot}65\,X + 11{\cdot}9$ or $Y = 11{\cdot}9 - 0{\cdot}65\,X$

Deviation Taken from Assumed Mean:

The two regression equations are written as follows:

(i) Regressive Equation of X on Y ; $X-\overline{X}=r\frac{\sigma x}{\sigma y}(Y-\overline{Y})$

$$r\frac{\sigma x}{\sigma y}=\frac{N\sum dxdy-(\sum dx\times\sum dy)}{N\sum dy^2-(\sum dy)^2}$$

(ii)) Regressive Equation of Y on X; $Y-\overline{Y}=r\frac{\sigma y}{\sigma x}(\mathrm{X}-\overline{X})$

$$r\frac{\sigma y}{\sigma x}=\frac{N\sum dxdy-(\sum dx\times\sum dy)}{N\sum dx^2-(\sum dx)^2}$$

From the following table calculate the regression equation taking deviation from the mean of X and Y series.

X	6	2	10	4	8
Y	9	11	5	8	7

Solution:

X	$x=X-\overline{X}$	x^2	Y	$y=Y-\overline{Y}$	y^2	xy
6	0	0	9	+1	1	0
2	-4	16	11	+3	9	-12
10	+4	16	5	-3	9	-12
4	-2	4	8	0	0	0
8	+2	4	7	-1	1	-2
$\sum X=30$	$\sum x=0$	$\sum x^2=40$	$\sum Y=40$	$\sum y=0$	$\sum y^2=20$	$\sum xy=-26$

Regression equation of X on Y : $X-\overline{X}=r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$

$$\overline{X}=\frac{\Sigma X}{N}=\frac{30}{5}=6\ ;\ \overline{Y}=\frac{\Sigma Y}{N}=\frac{40}{5}=8$$

$$r\frac{\sigma_x}{\sigma_y}=\frac{N\Sigma d_x d_y-\Sigma d_x\Sigma d_y}{N\Sigma d_y^2-(\Sigma d_y)^2}$$

$$=\frac{5(-21)-(5)(5)}{(5)(25)-(5)^2}=-\frac{-105-25}{125-25}=-\frac{130}{100}=-1{\cdot}3$$

$$X-6=-1{\cdot}3\,(Y-8)$$

$$X-6=-1{\cdot}3\,Y+10{\cdot}4 \quad \text{or} \quad X=16{\cdot}4-1{\cdot}3\,Y$$

Regression Equation of Y on X : $Y-\overline{Y}=r\frac{\sigma_y}{\sigma_x}(X-\overline{X})$

$$r\frac{\sigma_y}{\sigma_x}=\frac{N\Sigma d_x d_y-\Sigma d_x\Sigma d_y}{N\Sigma d_x^2-(\Sigma d_x)^2}$$

$$=\frac{5(-21)-(5)(5)}{(5)(45)-(5)^2}=\frac{-105-25}{200}=-0{\cdot}65$$

$$Y-8=-0{\cdot}65\,(X-6)$$

$$Y-8=-0{\cdot}65\,X+3{\cdot}9 \quad \text{or} \quad Y=11{\cdot}9-0{\cdot}65\,X$$

Key differences and similarities between correlation and regression

Both statistical measures are helpful in pointing out simple relationships among data. They can also help gauge the strength of a relation between variables. Here is a list of differences and similarities between these two statistical measures:

i) External and internal indicators:

Correlation can help a business establish a relationship between two internal variables, such as the number of non-direct sales and the price of the company's shares. While regression can be useful to establish a non-interdependent relation between one or more internal and dependent variables, such as direct sales, and one external variable, such as gross domestic product (GDP).

ii) Curvilinear or Complex relations:

Correlation can only describe simple and linear relations between two variables, it cannot explain a more complex relationship. Regression can describe curvilinear associations, which are relations that depend on a pattern, such as a relationship between the inflation and the cost of raw materials and the cash available to borrow.

iii) Methods of Measuring

Correlation uses a formula to calculate the sample correlation coefficient, represented by the letter r, which measures the strength of the relationship between two interdependent variables. Regression also uses a formula to first determine the influence of one variable over another and then to describe the type of relationship, such as linear or inversely proportional.

The sample correlation coefficient ranges from -1 to +1 and the closer it is to 0, the weaker the relationship is between the two variables. To represent a regression is better to use a scatter plot, which is a mathematical diagram where you can use dots to represent different numerical values.