BUSSINESS STATISTICS

Scope and Limitations of Statistics

Dr. T N KAVITHA

Assistant Professor of Mathematics

SCSVMV

Scope and Limitations of Statistics

Here we discuss the Importance, Scope, and Limitations of Statistics.

The scope of Statistics was limited in ancient times as the government used statistics for the purpose of administration alone.

Gradually, the subject became more and more popular and its application has become more extensive. Now is hardly any field of human activity where statistics are not used.

Now is used by economists, businessmen, scientists, administrators, etc.

What is the Scope of Statistics?

1. Statistics help in economic planning

All economic plans of a formulated on the basis of statistical data. The success of the plan is also evaluated with the help of statistics.

Economic problems such as production, consumption, wages, price profits, unemployment, poverty, etc. can be expressed numerically.

2. Statistics in business and management

Statistics are very useful to businessmen. It helps businessmen in formulating policies regarding business and forecasting future trends.

3. Statistics in administration

Efficient administration cannot be perceived without statistics. Statistics have been used from the time of origin of statistics to collect information regarding the military and fiscal policies.

4. Statistics in research

Statistical methods are extensively used in every type of research work. Whether it is agriculture, health, or social science, the statistics help in carrying out different types of researches.

What is the Importance of Statistics?

1. Statistics and planning
2. Statistics and economics
3. Statistics and business
4. Statistics and industry
5. Statistics and modern science
6. Statistics, psychology and education
7. Statistics and war
8. Statistics and mathematics

What are the Limitations of Statistics?

Limitations of Statistics

1. Statistics is unable to explain individual items
2. Statistics are unable to study qualitative characters
3. Statistical results are not accurately correct
4. Statistics deal with average
5. Statistics is only one of the methods of studying a given problem
6. Statistics is liable to be misused
7. Qualitative Aspect Ignored
8. To Many methods to study problems
9. Results are true only on average
10. Statistical laws are not exact

THANK YOU

BUSSINESS STATISTICS

Data Collection Methods

Dr. T N KAVITHA

Assistant Professor of Mathematics

SCSVMV

Data Collection Methods

In Statistics, the data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to evaluate the outcome of the problem. The data collection methods allow a person to conclude an answer to the relevant question. Most of the organizations use data collection methods to make assumptions about future probabilities and trends.

Data collection method

A data can be classified into two types, namely

- Primary Data Collection methods
- Secondary Data Collection methods

Primary Data Collection Methods

Primary data or raw data is a type of information that is obtained directly from the first-hand source through experiments, surveys, or observations. The primary data collection method is further classified into two types.

They are

- Quantitative Data Collection Methods
- Qualitative Data Collection Methods

Methods performed to collect the data

Quantitative Data Collection Methods

It is based on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures. This method is cheaper than qualitative data collection methods, and it can be applied in a short duration of time.

Qualitative Data Collection Methods

It does not involve any mathematical calculations. This method is closely associated with elements that are not quantifiable. This qualitative data collection method includes interviews, questionnaires, observations, case studies etc. There are several methods to collect this type of data. They are

- Observation Method
- Questionnaire Method
- Interview Method

Observation Method

The different types of observations are:

- Structured and unstructured observation
- Controlled and uncontrolled observation
- Participant, non-participant and disguised observation

Interview Method

The method of collecting data in terms of oral or verbal responses. It is achieved in two ways, such as

- **Personal Interview** – In this method, a person known as an interviewer is required to ask questions face to face to the other person. The personal interview can be structured or unstructured, direct investigation, focused conversation etc.
- **Telephonic Interview** – In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views orally.

Questionnaire Method

In this method, the set of questions are mailed to the respondent.

They should read, reply and subsequently return the questionnaire.

The questions are printed in the definite order on the form.

A good Questionnaire

A good survey should have the following features:

- Short and simple
- Should follow a logical sequence
- Provide adequate space for answers
- Avoid technical terms
- Should have good physical appearance such as colour, quality of the paper to attract the attention of the respondent

Schedules

This method is similar to the questionnaire method with a slight difference. The enumerations are specially appointed for the purpose of filling the schedules. It explains the aims and objects of the investigation and may remove misunderstandings if any have come up. Enumerations should be trained to perform their job with hard work and patience.

Secondary Data Collection Methods

Secondary data is data collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals etc. It may be either published data or unpublished data.

Published data are available in various resources

- Government publications
- Public records
- Historical and statistical documents
- Business documents
- Technical and trade journals

Unpublished data includes

- Diaries
- Letters
- Unpublished biographies etc.

Thank you

BUSSINESS STATISTICS

Types of Classifications&Tabulation

Dr. T N KAVITHA

Assistant Professor of Mathematics

SCSVMV

DATA CLASSIFICATION:

The grouping of related facts/data into different classes according to certain common characteristic.

Basis of data Classification: Broadly 4 broad basis

1. Geographical
2. Chronological or Temporal
3. Qualitative
4. Quantitative

1. Geographical

Geographical classifications i.e. area wise

- Total Population of india by states and by districts
- No. of death due to covid-19 by countries.
- Deaths in Tamilnadu by districts

so area wise classification i.e world , countries, states, districts, region, ...

2. Chronological or Temporal

Chronological classification or Temporal classification i.e. on the basis of time Table:

by year wise , month wise, week wise, daywise, time wise,.....

3. Qualitative classification

Qualitative i.e. on the basis of some attributes or characters

Example: People by place of residence, Rural- Urban, Male-Female, Illiterate-Literate .....

Sex	Urban	Rural
Boys	200	390
Girls	167	100

4.Quantitative:

Quantitative: On the basis of quantitative class intervals

For example students of a college may be classified according to weight as follows

Weight of students of a college Wt. In (LBS)	90-100	100-110	110-120	120-130	130-140	140-150	Total
No. of students	50	200	260	360	90	40	1000

Presentation of data

Presentation of data refers to an exhibition or putting up data in an attractive and useful manner such that it can be easily interpreted. The three main forms of presentation of data are:

1.Textual presentation

2.Data tables OR Tabulation

3.Diagrammatic presentation

Tabulation

Tabulation is the systematic arrangement of the statistical data in columns or rows.

It involves the orderly and systematic presentation of numerical data in a form designed to explain the problem under consideration.

Tabulation helps in drawing the inference from the statistical figures.

Types of Tabulation

In general, the tabulation is classified in two parts, that is a simple tabulation, and a complex tabulation.

Simple tabulation, gives information regarding one or more independent questions.

Complex tabulation gives information regarding two mutually dependent questions.

Two-Way Table

These types of table give information regarding two mutually dependent questions.

For example,

How many millions of the persons are in the Divisions?

The Two-Way Tables will answer the question by giving the column for female and male.

Three-Way Table

Three-Way Table gives information regarding three mutually dependent and inter-related questions.

For example, from one-way table, we get information about population, and from two-way table, we get information about the number of male and female available in various divisions.

Now we can extend the same table to a three way table, by putting a question, “How many male and female are literate?”

Thus the collected statistical data will show the following, three mutually dependent and inter-related questions:

1.Population in various division.

2.Their sex-wise distribution.

3.Their position of literacy.

Components of Data Tables

- Table Number:
- Title:
- Headnotes:
- Stubs:
- Caption:
- Body or field:
- Footnotes:
- Source:

Table Number

Table Number: Each table should have a specific table number for ease of access and locating. This number can be readily mentioned anywhere which serves as a reference and leads us directly to the data mentioned in that particular table.

Title

Title: A table must contain a title that clearly tells the readers about the data it contains, time period of study, place of study and the nature of classification of data.

Headnotes

Headnotes: A headnote further aids in the purpose of a title and displays more information about the table. Generally, headnotes present the units of data in brackets at the end of a table title.

Stubs

Stubs: These are titles of the rows in a table. Thus a stub display information about the data contained in a particular row.

Caption

Caption: A caption is the title of a column in the data table. In fact, it is a counterpart if a stub and indicates the information contained in a column.

Body or field

Body or field: The body of a table is the content of a table in its entirety. Each item in a body is known as a ‘cell’.

Footnotes

Footnotes: Footnotes are rarely used. In effect, they supplement the title of a table if required.

Source

Source: When using data obtained from a secondary source, this source has to be mentioned below the footnote.

Construction of Data Tables

There are many ways for construction of a good table. However, some basic ideas are:

- The title should be in accordance with the objective of study:
- Comparison:
- Alternative location of stubs:
- Headings:
- Footnote:
- Size of columns:
- Use of abbreviations:
- Units:

The title should be in accordance with the objective of study

The title should be in accordance with the objective of study: The title of a table should provide a quick insight into the table.

Comparison

Comparison: If there might arise a need to compare any two rows or columns then these might be kept close to each other.

Alternative location of stubs

Alternative location of stubs: If the rows in a data table are lengthy, then the stubs can be placed on the right-hand side of the table.

Headings

Headings: Headings should be written in a singular form. For example, ‘good’ must be used instead of ‘goods’.

Footnote

Footnote: A footnote should be given only if needed.

Size of columns

Size of columns: Size of columns must be uniform and symmetrical.

Use of abbreviations

Use of abbreviations: Headings and sub-headings should be free of abbreviations.

Units

Units: There should be a clear specification of units above the columns.

Advantages of Tabulation

1.The large mass of confusing data is easily reduced to reasonable form that is understandable to kind.

2.The data once arranged in a suitable form, gives the condition of the situation at a glance, or gives a bird eye view.

3.From the table it is easy to draw some reasonable conclusion or inferences.

4.Tables gave grounds for analysis of the data.

5.Errors, and omission if any are always detected in tabulation.

Thank you

BUSSINESS STATISTICS

Graphical Representation

Dr. T N KAVITHA

Assistant Professor of Mathematics

SCSVMV

Graphical Representation

Graphical Representation is a way of analysing numerical data. It exhibits the relation between data, ideas, information and concepts in a diagram. It is easy to understand and it is one of the most important learning strategies. It always depends on the type of information in a particular domain.

Types of graphical representation

There are different types of graphical representation. Some of them are as follows:

Line Graphs

Bar Graphs

Histograms

Line Plot

Frequency Table

Circle Graph

Stem and Leaf Plot

Line Graphs,Bar Graphs &Histograms

- **Line Graphs** – Line graph or the linear graph is used to display the continuous data and it is useful for predicting future events over time.
- **Bar Graphs** – Bar Graph is used to display the category of data and it compares the data using solid bars to represent the quantities.
- **Histograms** – The graph that uses bars to represent the frequency of numerical data that are organised into intervals. Since all the intervals are equal and continuous, all the bars have the same width.

Line Plot & Frequency Table

🕐**Line Plot –** It shows the frequency of data on a given number line. ‘ x ‘ is placed above a number line each time when that data occurs again.

🕐**Frequency Table –** The table shows the number of pieces of data that falls within the given interval.

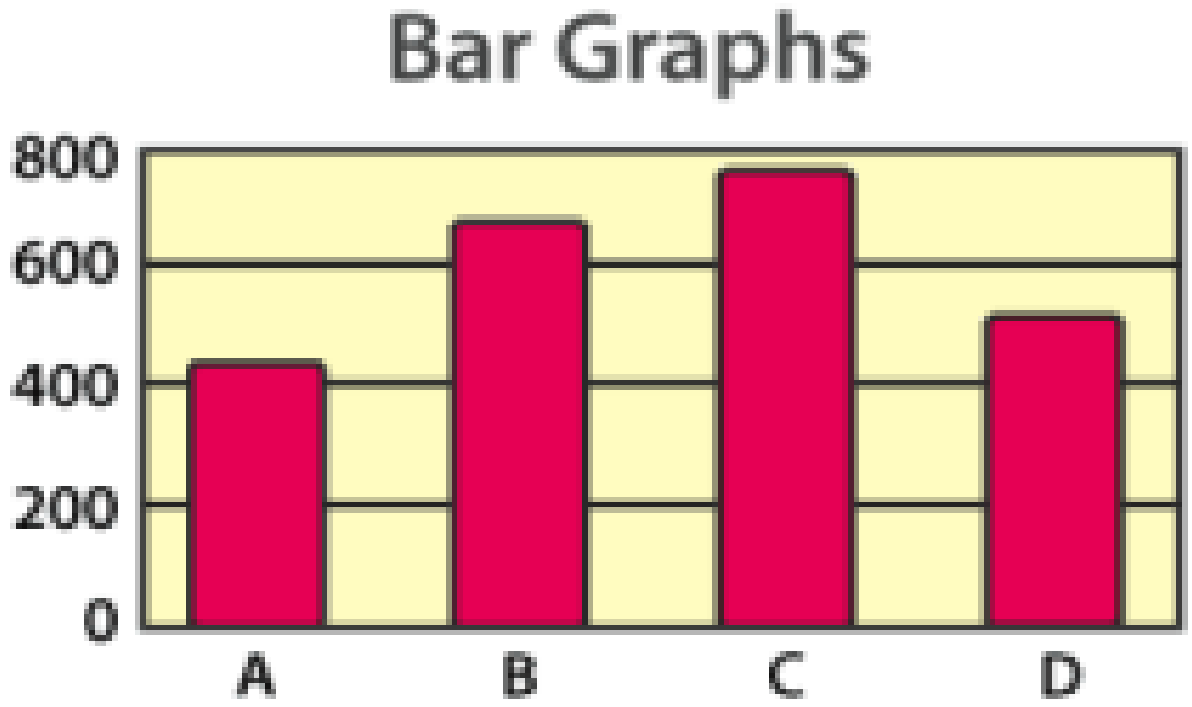
Circle Graph &Stem and Leaf Plot

- **Circle Graph** – Also known as the pie chart that shows the relationships of the parts of the whole. The circle is considered with 100% and the categories occupied is represented with that specific percentage like 15%, 56%, etc.
- **Stem and Leaf Plot** – In the stem and leaf plot, the data are organised from least value to the greatest value. The digits of the least place values from the leaves and the next place value digit forms the stems.

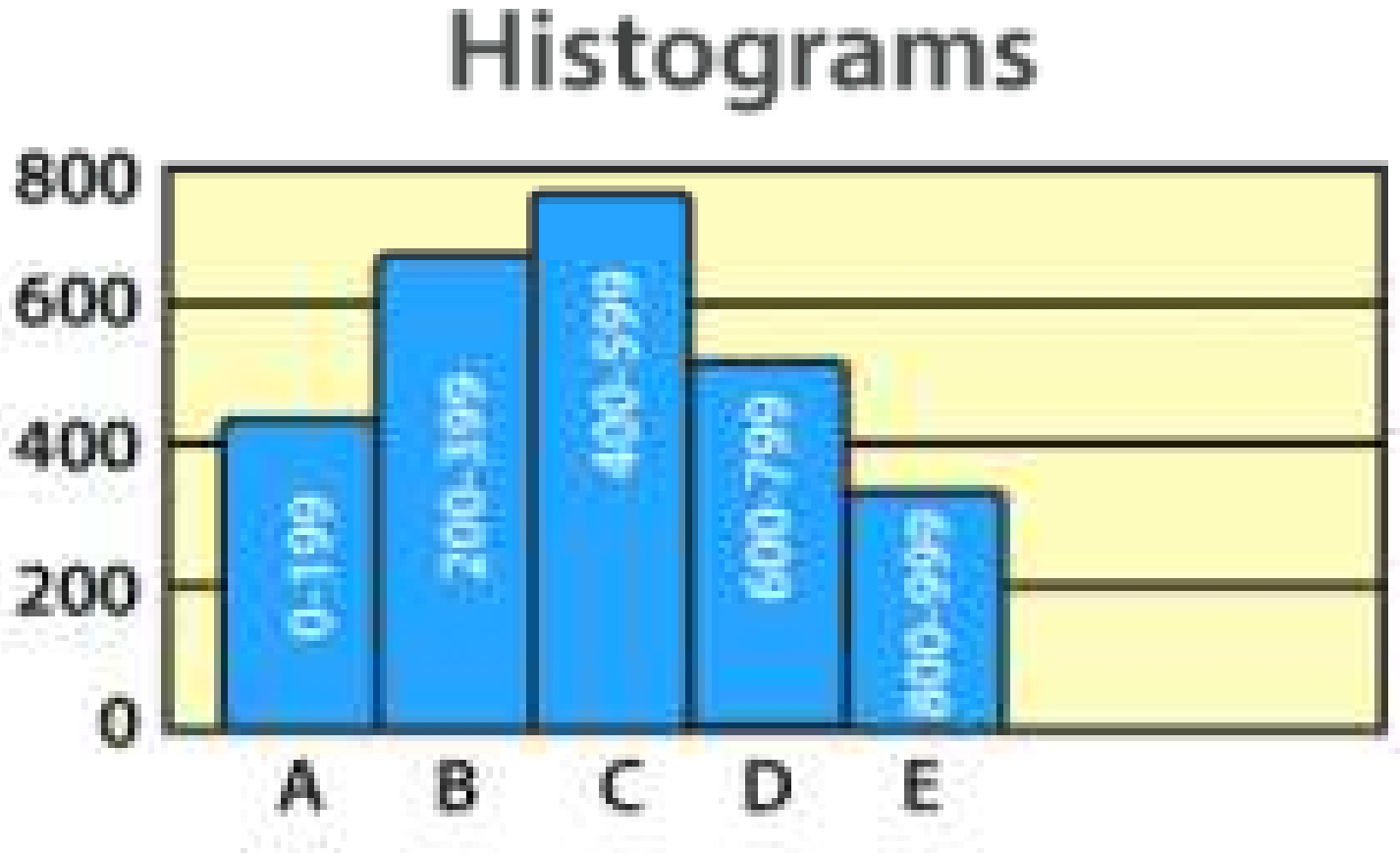
Box and Whisker Plot

🕐**Box and Whisker Plot** – The plot diagram summarises the data by dividing into four parts. Box and whisker show the range (spread) and the middle (median) of the data.

TYPES OF GRAPHICAL REPRESENTATION



TYPES OF GRAPHICAL REPRESENTATION

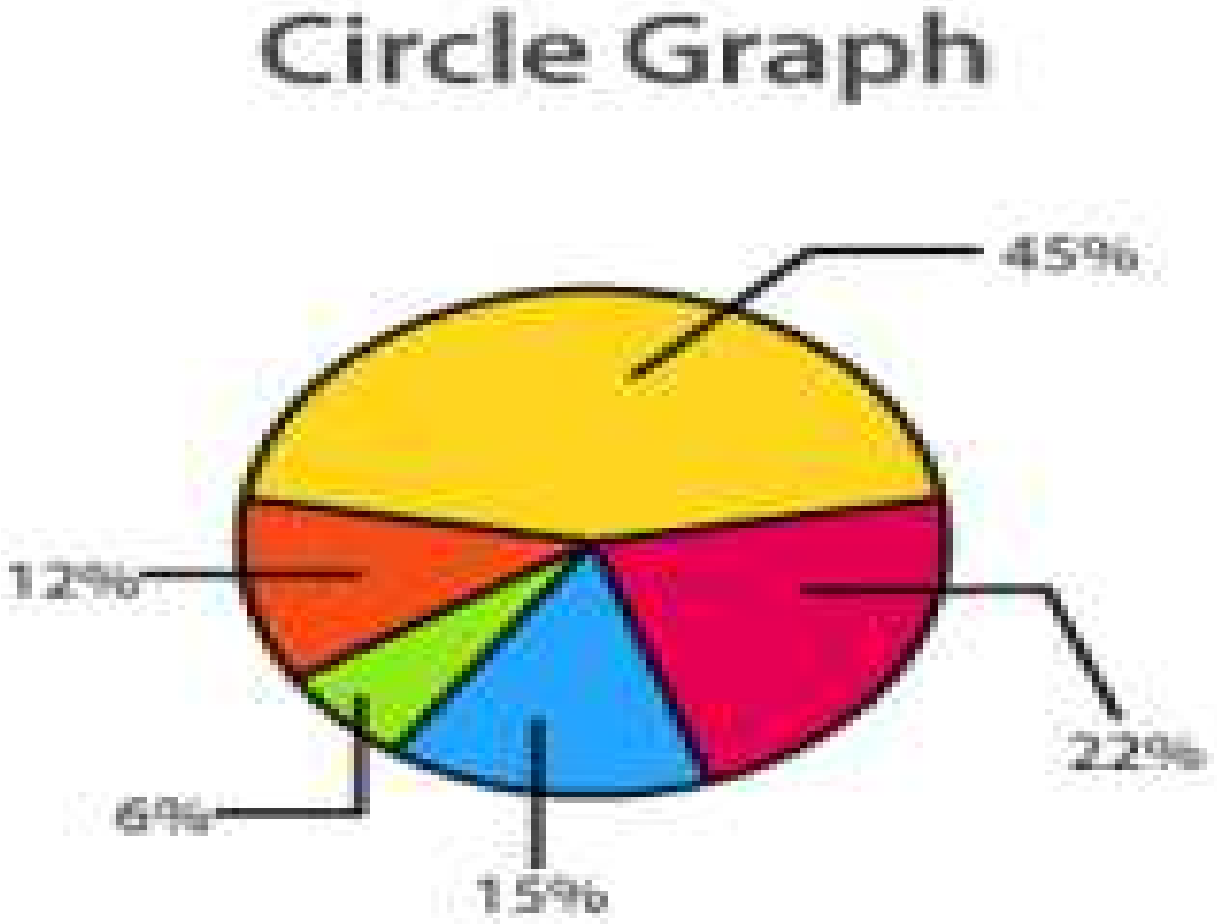


TYPES OF GRAPHICAL REPRESENTATION

Frequency Table

Rulers of France		
Reign (Years)	Tally	Frequency
1-15	𝍸 𝍸 𝍸 III	18
16-30	𝍸 𝍸 I	11
31-45	𝍸 I	6
46-60	IIII	4
61-75	I	1

TYPES OF GRAPHICAL REPRESENTATION



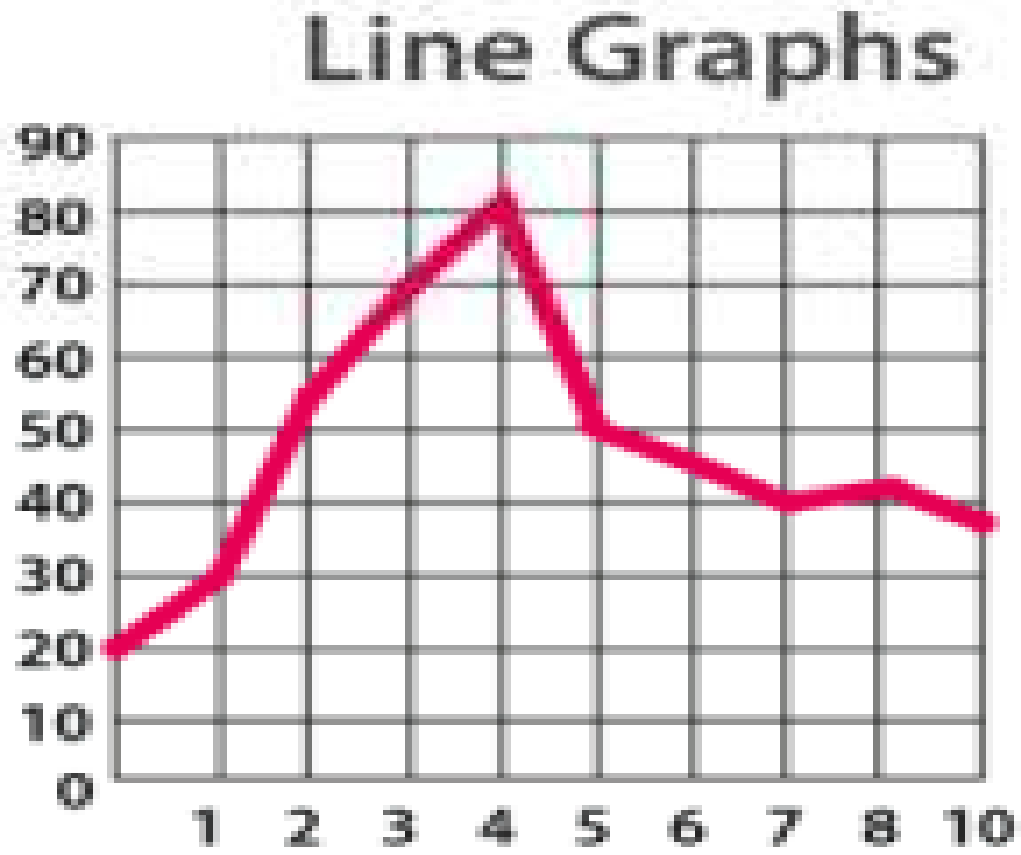
TYPES OF GRAPHICAL REPRESENTATION

Stem and Leaf Plot

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

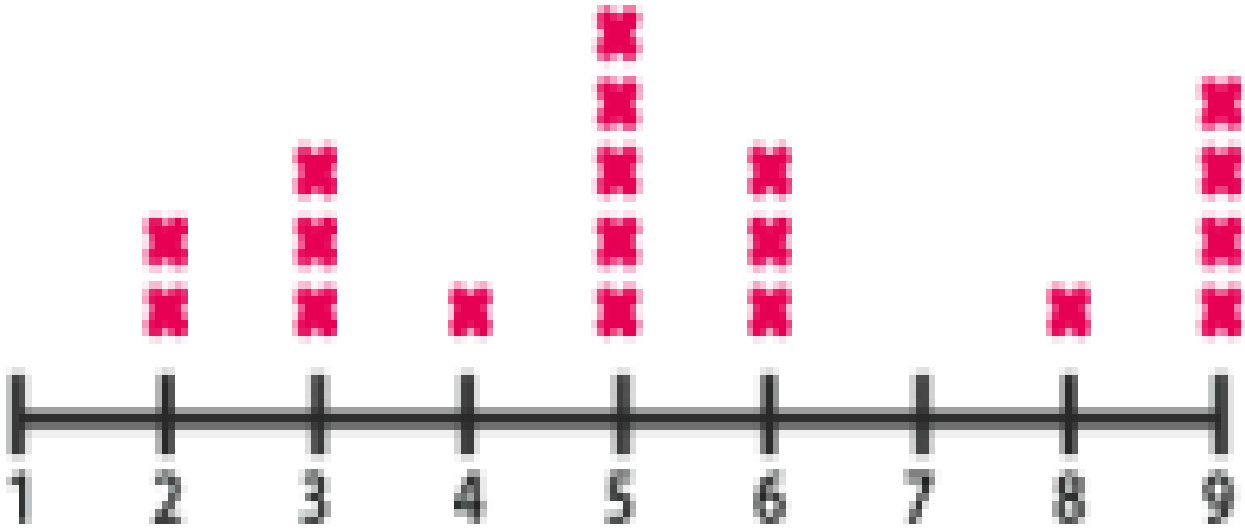
Key : 6 | 3 = 63 Year

TYPES OF GRAPHICAL REPRESENTATION



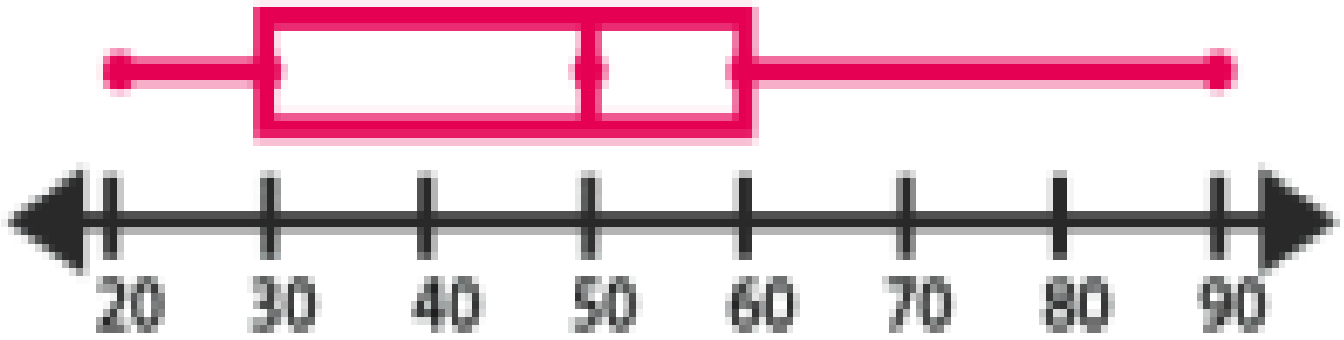
TYPES OF GRAPHICAL REPRESENTATION

Line Plot



TYPES OF GRAPHICAL REPRESENTATION

Box and Whisker Plot



General Rules for Graphical Representation of Data

There are certain rules to effectively present the information in the graphical representation. They are:

- **Suitable Title:** Make sure that the appropriate title is given to the graph which indicates the subject of the presentation.
- **Measurement Unit:** Mention the measurement unit in the graph.

Rules for Graphical Representation

Proper Scale: To represent the data in an accurate manner, choose a proper scale.

Index: Index the appropriate colours, shades, lines, design in the graphs for better understanding.

Data Sources: Include the source of information wherever it is necessary at the bottom of the graph.

Keep it Simple: Construct a graph in an easy way that everyone can understand.

Neat: Choose the correct size, fonts, colours etc in such a way that the graph should be a visual aid for the presentation of information.

Principles of Graphical Representation

Algebraic principles are applied to all types of graphical representation of data.

In graphs, it is represented using **two lines called coordinate axes**.

The **horizontal axis is denoted as the x-axis** and the **vertical axis is denoted as the y-axis.**

The point at which two lines intersect is called an **origin ‘O’.**

Principles of GraphicalRepresentation

Consider x-axis,

the distance from the origin to the **right side will take a positive value** and the distance from the origin to the **left side will take a negative value.**

Similarly, for the y-axis,

the points above the origin will take a positive value, and the points below the origin will take a negative value.

frequency distribution

- Generally, the frequency distribution is represented in four methods, namely
- 🕛Histogram
- 🕛Smoothed frequency graph
- 🕛Pie diagram
- 🕛Cumulative or ogive frequency graph
- 🕛Frequency Polygon

Merits of Using Graphs

Some of the merits of using graphs are as follows:

- The graph is easily understood by everyone without any prior knowledge.
- It saves time
- It allows us to relate and compare the data for different time periods
- It is used in statistics to determine the mean, median and mode for different data, as well as in the interpolation and the extrapolation of data.

What are the Advantages of Graphical Method?

Some of the advantages of graphical representation are:

- It makes data more easily understandable.
- It saves time.
- It makes the comparison of data more efficient.

THANK YOU

BUSSINESS STATISTICS

Measures of central Tendency

Dr. T N KAVITHA

Assistant Professor of Mathematics

SCSVMV

Measures of Central Tendency

- Mean - Average
- Median - Middle most value
- Mode - Most repeated value

Mean (Individual Observations)

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

$$Mean = \frac{\sum x}{n} = \frac{x1 + x2 + x3 + x4 + ...}{n}$$

$$Mean = A + \frac{\sum d}{n}$$

where d=x-A,

A-Assumed mean(Individual Observations)

Mean(Discrete Series)

$$\text{Mean} = \frac{\sum fx}{N} \text{ where } N = \sum f$$

$$\text{Mean} = A + \frac{\sum fd}{N} \text{ where } d=x\text{-}A,\ A\text{-Assumed mean}$$

$$\text{Mean} = A + \frac{\sum fd}{N} \times i \text{ where } d = \frac{x - A}{i},\ A\text{-Assumed mean},\ i = \text{class interval}$$

Mean(Continuous Observations)

$$\text{Mean} = \frac{\sum fm}{N} \text{ where } N = \sum f \text{ , m- mid value}$$

$$\text{Mean} = A + \frac{\sum fd}{N} \text{ where d=m-A, A-Assumed mean}$$

$$\text{Mean} = A + \frac{\sum fd}{N} \times i \text{ where } d = \frac{m - A}{i} \text{ , A-Assumed mean , i= class interval}$$

Median (Individual Observations)

$$\text{Median} = \frac{n+1}{2}\text{th item(Individual Observations)(EVEN DATA)}$$

Median(Discrete Series)

$$\text{Median} = \frac{N+1}{2} \text{ th item}$$

Median(Continuous Observations)

$$\text{Median} = \frac{N}{2}\text{th item}$$

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

Where L- Lower limit of the median class

$$N = \sum f$$

Cf-cumulative frequency preceding the median class

f- frequency of the median class

i-class interval

Mode-(Individual Observations)

Mode = the item which is occurred more number of times.

Mode(Discrete Series)

Mode = the item which is occurred more number of times.

Mode(Continuous Observations)

$$\text{Mode} = M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Where L – lower limit of the modal class

f_0 -frequency of the class preceding the modal class

f_1 - frequency of the modal class

f_2 - frequency of the class succeeding the modal class

i-class interval

Empirical relation:

Relation Between Mean, Median and Mode

If the value of the n=mode is equal to the value of the median and the mean then we call it as symmetrical data set. For such data sets, there is a simple relationship between the three M's (mean, median and mode):

Mean - Mode =3 (Mean – Median)

(OR)

Mode = Mean - 3 Mean + 3 Median

(OR)

Mode = 3 Median - 2 Mean

Find the AM , median and mode of the following set of observations:
25,32,28,34,24,31,36; 27,29,30.

$$\text{Mean} = \frac{\sum x}{n} = \frac{296}{10} = 29.6$$

Assending order 24, 25, 27, 28, 29, 30, 31, 32, 34, 36

Median $= [(n+1)/2]^{th}$ item $= (10+1)/2$ th item $= 5.5^{th}$ item

5.5^{th} item $= (5^{th}$ item $+ 6^{th}$ item$)/2 = (29+30)/2 = 29.5$

There is no mode.

Find the mean, median and mode for the following data:

Class(x)	**frequency(f)**
0-10	20
10-20	5
20-30	3
30-40	8
40-50	10
50-60	35
60-70	10
70-80	4
80-90	3
90-100	2

Class(x)	m	frequency(f)	fm	cf
0-10	5	20	100	20
10-20	15	5	75	25
20-30	25	3	75	28
30-40	35	8	280	36
40-50	45	10 f_0	450	46
50-60(median, Mode)	**55**	**35 f_1**	**1925**	**81**
60-70	65	10 f_2	650	91
70-80	75	4	300	95
80-90	85	3	255	98
90-100	95	2	190	100
		$N = \sum f = 100$	$\sum fm = 4300$	

$$N = \sum f = 100$$

$$\text{Mean} = \frac{\sum fm}{N} \text{ where } N = \sum f$$

$$= 4300/100 = 43$$

Median = (N/2)th item = (100/2)th item = 50^{th} item

50^{th} item is 50-60 (the median class) = 55

$$\textbf{Median} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

$$= 50 + \frac{50 - 46}{35} \times 10 = 50 + (0.1142 \text{x} 10) = 50 + 1.142$$

$$= 50 + 1.1428 = 51.1428$$

Modal class is 50-60

$$M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 50 + \frac{35 - 10}{2(35) - 10 - 10} \times 10$$

$$= 50 + 5 = 55$$

Thank you

BUSSINESS STATISTICS

Measures of Dispersion

Dr. T N KAVITHA

Assistant Professor of Mathematics

SCSVMV

MEASURES OF DISPERSION

- Range
- Quartile Deviation
- Mean Deviation:
- Standard Deviation

Range

Range = L - S

L- Largest values

S- Smallest value

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

Quartile Deviation

Quartile Deviation :

$$QD = \frac{Q_3 - Q_1}{2}$$

Q_1- First quartile

Q_3-3rd quartile

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Mean Deviation(Individual Observations)

Mean Deviation about mean:

$$MD = \frac{\sum |D|}{N} \text{ where } D = x - \bar{x} \text{ (Individual Observations)}$$

(OR)

Mean Deviation about median:

$$MD = \frac{\sum |D|}{N} \text{ where } D = x - \text{Median} \text{ (Individual Observations)}$$

Mean Deviation(Discrete series)

Mean Deviation about mean:

$$MD = \frac{\sum f|D|}{N} \text{ where } D = x - \bar{x} \text{ (Discrete series)}$$

(OR)

Mean Deviation about median:

$$MD = \frac{\sum f|D|}{N} \text{ where } D = x - \text{Median} \text{ (Discrete series)}$$

Mean Deviation(Continuous series)

Mean Deviation about mean:

$$MD = \frac{\sum f|D|}{N} \text{ where } D = m - \bar{x} \text{ (Continuous series)}$$

(OR)

Mean Deviation about median:

$$MD = \frac{\sum f|D|}{N} \text{ where } D = m - \text{Median} \text{ (Continuous series)}$$

Coefficient of Mean Deviation = MD/mean

Coefficient of Mean Deviation = MD/median

Standard Deviation(Individual Observations)

$$\sigma = \sqrt{\frac{\sum x^2}{n}} \quad \text{Where } x = X - \bar{X} \text{ (Individual Observations)}$$

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} \quad \text{where } d = X - A \text{ ; A- Assumed mean(Individual Observations)}$$

Standard Deviation(Discrete series)

$$\sigma = \sqrt{\frac{\sum fx^2}{n}} \quad \text{where } x = X - \bar{X} \text{ (Discrete series)}$$

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \quad \text{where } d = X - A \text{ ; A- Assumed mean(Discrete series)}$$

Standard Deviation(Continuous series)

$$\sigma = \sqrt{\frac{\sum fx^2}{n}} \text{ where } d = m - \bar{X} \text{ (Continuous series)}$$

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times i \text{ where } d = \frac{m - A}{i}; \text{ A- Assumed mean(Continuous series)}$$

Coefficient of Variation

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

Range

Example 1: Find the range for the following three sets of data:

Set 1:	05	15	15	05	15	05	15	15	15	15
Set 2:	8	7	15	11	12	5	13	11	15	9
Set 3:	5	5	5	5	5	5	5	5	5	5

Solution: In each of these three sets, the highest number is 15 and the lowest number is 5.

Since the range is the difference between the maximum value and the minimum value of the data, it is 10 in each case. L = 15 , S = 5

Range = 15 - 5 = 10

Example 2: Calculate the coefficient of range separately for the two sets of data given below:

Set 1	8	10	20	9	15	10	13	28
Set 2	30	35	42	50	32	49	39	33

Solution:

Set 1

L = 28 , S= 8 Range= 28-8=20

Set2

L= 50 , S= 30 Range= 50-30=20

It can be seen that the range in both the sets of data is the same:

Set 1 28-8=20

Set 2 50-30=20

Coefficient of range in Set 1 is: $\frac{28-8}{28+8} = 0.55$

Coefficient of range in set 2 is: $\frac{50-30}{50+30} = 0.25$

Example 3: Find the range for the following frequency distribution:

Size of Item	20-40	40-60	60-80	80-100	100-120	Total
Frequency	7	11	30	17	20	70

Solution

Here, the upper limit of the highest class is 120

and the lower limit of the lowest class is 20.

Hence, the range is 120 - 20 = 100.

The coefficient of range is calculated by the formula: (L-S)/ (L+S) = 120-20 / 120+20

= 100/ 20 = 0.71428

THANK YOU

BUSSINESS STATISTICS

Geometric Mean, weighted Arithmetic mean & Harmonic Mean

Dr. T N KAVITHA

Assistant Professor of Mathematics

SCSVMV

Harmonic mean

A simple way to define a harmonic mean is to call it the reciprocal of the arithmetic mean of the reciprocals of the observations. The most important criteria for it is that none of the observations should be zero.

Harmonic mean

The formula is:

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \ldots}$$

Where **a,b,c,...** are the values, and **n** is how many values.

Steps:

- Calculate the reciprocal (1/value) for every value.
- Find the average of those reciprocals (just add them and divide by how many there are)
- Then do the reciprocal of that average (=1/average)

Properties of Harmonic Mean

If all the observation taken by a variable are constants, say k, then the harmonic mean of the observations is also k

The harmonic mean has the least value when compared to the geometric mean and the arithmetic mean

A.M <= G.M <= H.M

What is the harmonic mean of 1, 2 and 4?

The reciprocals of 1, 2 and 4 are:

$$\frac{1}{1} = 1, \quad \frac{1}{2} = 0.5, \quad \frac{1}{4} = 0.25$$

Now add them up:

$$1 + 0.5 + 0.25 = 1.75$$

Divide by how many:

$$\text{Average} = \frac{1.75}{3}$$

The reciprocal of that average is our answer:

$$\text{Harmonic Mean} = \frac{3}{1.75} = \mathbf{1.714} \text{ (to 3 places)}$$

Advantages of Harmonic Mean

- A harmonic mean is rigidly defined
- It is based upon all the observations
- The fluctuations of the observations do not affect the harmonic mean
- More weight is given to smaller items

Disadvantages of Harmonic Mean

- Not easily understandable
- 🕐Difficult to compute

We travel 10 km at 60 km/h, than another 10 km at 20 km/h, what is our average speed?

$$\text{Harmonic mean} = 2/\left(\frac{1}{60} + \frac{1}{20}\right) = 30 \text{ km/h}$$

Check: the 10 km at 60 km/h takes 10 minutes, the 10 km at 20 km/h takes 30 minutes, so the total 20 km takes 40 minutes, which is 30 km per hour

Geometric Mean

The Geometric Mean is a special type of average where we multiply the numbers together and then take a square root (for two numbers), cube root (for three numbers) etc.

Geometric mean = G.M.

$$= (x_1 f_1 . x_2 f_2 \ldots x_n f_n)^{(1/N)}$$

Geometric Mean

A geometric mean is a mean or average which shows the central tendency of a set of numbers by using the product of their values. For a set of n observations, a geometric mean is the nth root of their product. The geometric mean G.M., for a set of numbers $x_1, x_2, \ldots, x_n$ is given as

$$G.M. = (x_1 . x_2 \ldots x_n)^{1/n}$$

or,

$$G.M. = (\pi_{i=1}^{n} x_i)^{1/n}$$

$$= \sqrt[n]{(x_1, x_2, \ldots, x_n)}.$$

Geometric Mean

The geometric mean of two numbers, say x, and y is the square root of their product x × y. For three numbers, it will be the cube root of their products i.e., $(x\ y\ z)^{1/3}$.

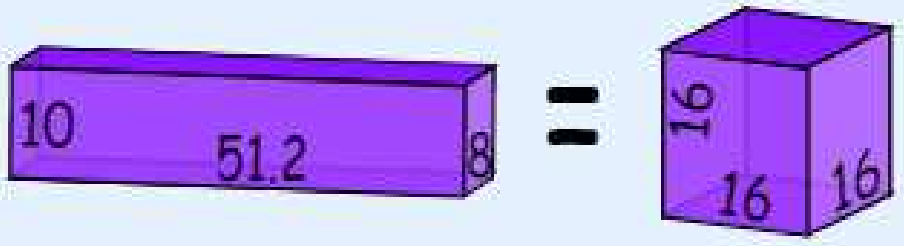
What is the Geometric Mean of 10, 51.2 and 8?

- First we multiply them: $10 \times 51.2 \times 8 = 4096$
- Then (as there are three numbers) take the cube root: $\sqrt[3]{4096} = \mathbf{16}$

In one line:

$$\textbf{Geometric Mean} = \sqrt[3]{(10 \times 51.2 \times 8)} = 16$$

It is like the volume is the same:



$10 \times 51.2 \times 8 = 16 \times 16 \times 16$

Relation Between Geometric Mean and Logarithms

In order to make our calculation easy and less time consuming we use the concept of logarithms in the calculation of geometric means.

Since, G.M. = $(x_1 . x_2 \ldots x_n)^{1/n}$

Taking log on both sides, we have

log G.M. = $(1/n)$ (log $((x_1 . x_2 \ldots x_n))$

or, log G.M. = $(1/n)(\log x_1 + \log x_2 + \ldots + \log x_n)$

or, log G.M. = $(1/n) \sum I = 1n \log x_i$

or, G.M. = Antilog$(1/n) \sum I = 1n \log x_i))$.

Geometric Mean of Frequency Distribution

For a grouped frequency distribution, the geometric mean G.M. is

$$G.M. = (x_1 f_1 . \ x_2 f_2 \ldots \ x_n f_n)^{1/N},$$

where $N = \sum i = 1n\ f_i$

Taking logarithms on both sides, we get

$$\log G.M. = 1/N \ (f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n)$$

$$= 1/N \ [\sum i = 1n\ f_i \log x_i \ ].$$

Calculate the geometric and harmonic mean of the given data

x	2	4	5	8
f	3	3	2	2

$$\text{Geometric mean} = \text{G.M.} = (x_1 f_1 . x_2 f_2 \ldots x_n f_n)^{1/N)}$$

$$\text{Here, } N = 3 + 3 + 2 + 2 = 10$$

$$\text{G.M.} = (2.3 \times 4.3 \times 5.2 \times 8.2)^{1/10} = (11520)^{(1/10)} = 0.406$$

$$\text{H.M.} = \frac{1}{\frac{1}{N}\sum \text{i}} = \frac{1}{\frac{1}{N}\sum f_i x_i}$$

$$= \frac{1}{\frac{1}{10}(3/2 + ¾ + 2/5 + 2/8)} . = \frac{1}{\frac{1}{10}(1.5+0.75+0.4+0.25)} = \frac{1}{\frac{1}{10}(2.9)}$$

$$= \frac{10}{(2.9)} = \frac{10\times 10}{(2.9)\times 10} = \frac{100}{(29)} = 3.44827$$

Properties of Geometric Means

- The logarithm of geometric mean is the arithmetic mean of the logarithms of given values
- If all the observations assumed by a variable are constants, say K >0, then the G.M. of the observation is also K
- The geometric mean of the ratio of two variables is the ratio of the geometric means of the two variables
- The geometric mean of the product of two variables is the product of their geometric means

Geometric Mean of a Combined Group

Suppose G1, and G2 are the geometric means of two series of sizes n1, and n2 respectively. The geometric mean G, of the combined groups, is:

$$\log G = (n_1 \log G_1 + n_2 \log G_2) / (n_1 + n_2)$$

or, $$G = \text{antilog } [(n_1 \log G_1 + n_2 \log G_2) / (n_1 + n_2)]$$

In general for n_i geometric means, i = 1 to k,

we have

$$G = \text{antilog } [(n_1 \log G_1 + n_2 \log G_2 + \ldots + n_k \log G_k) / (n_1 + n_2 + \ldots + n_k)]$$

Advantages of Geometric Mean

- A geometric mean is based upon all the observations
- It is rigidly defined
- The fluctuations of the observations do not affect the geometric mean
- It gives more weight to small items

Disadvantages of Geometric Mean

- A geometric mean is not easily understandable by a non-mathematical person
- If any of the observations is zero, the geometric mean becomes zero
- If any of the observation is negative, the geometric mean becomes imaginary

Weighted Mean

Weighted Mean is an average computed by giving different weights to some of the individual values. If all the weights are equal, then the weighted mean is the same as the arithmetic mean.

Formula of weighted Mean

The Weighted mean for given set of non-negative data $x_1, x_2, x_3, \ldots x_n$ with non-negative weights $w_1, w_2, w_3, \ldots w_n$ can be derived from the formula given below.

$$\bar{x} = \frac{(w_1x_1 + w_2x_2 + \ldots + w_nx_n)}{(w_1 + w_2 + \ldots + w_n)}$$

Where,

x is the repeating value

w is the number of occurrences of x weight

$\bar{x}$ is the weighted mean

Example of Weighted Mean

Question: Suppose that a marketing firm conducts a survey of 1,000 households to determine the average number of TVs each household owns. The data show a large number of households with two or three TVs and a smaller number with one or four. Every household in the sample has at least one TV and no household has more than four. Find the mean number of TVs per household.

Number of TVs per Household	1	2	3	4
Number of Households	73	378	459	90

Solution:

As many of the values in this data set are repeated multiple times, you can easily compute the sample mean as a weighted mean. Follow these steps to calculate the weighted arithmetic mean:

- Step 1: Assign a weight to each value in the dataset:
- $x_1 = 1$, $w_1 = 73$
- $x_2 = 2$, $w_2 = 378$
- $x_3 = 3$, $w_3 = 459$
- $x_4 = 4$, $w_4 = 90$

Step 2:

Compute the numerator of the weighted mean formula.

Multiply each sample by its weight and then add the products together:

$$\sum 4i = 1 w_i x_i = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$$

$$= (1)(73) + (2)(378) + (3)(459) + (4)(90)$$

$$= 73 + 756 + 1377 + 360$$

$$= 2566$$

Step 3:

Now, compute the denominator of the weighted mean formula by adding the weights together.

$$\sum 4i = 1wi = w1+w2+w3+w4$$

$$= 73 + 378 + 459 + 90$$

$$=1000$$

Step 4:

Divide the numerator by the denominator

$\sum 4i=1 w_i x_i / \sum 4i=1 wi$

$=2566 / 1000$

$=2.566$

The mean number of TVs per household in this sample is 2.566.

THANK YOU

UNIT III
CORRELATION & REGRESSION

BY
DR. T N KAVITHA
ASSISTANT PROFESSOR OF MATHEMATICS
SCSVMV

Correlation and Regression analysis

Correlation Coefficient :

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$$

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

Where $\mathrm{x} = \mathrm{X} - \overline{\mathrm{X}}$ and $\mathrm{y}=\mathrm{Y}-\overline{\mathrm{Y}}$

$$\mathbf{r} = \frac{\mathrm{N}\sum \mathrm{dxdy} - \left(\sum \mathrm{dx}\right)\left(\sum \mathrm{dy}\right)}{\sqrt{\mathrm{N}\sum \mathrm{dx}^{2-} \left(\sum \mathrm{dx}\right)^2}\sqrt{\mathrm{N}\sum \mathrm{dy}^2 - \left(\sum \mathrm{dy}\right)^2}}$$

Where dx=X-A

dy=Y-B

MISCELLANEOUS PROBLEMS

1. Find the Coefficient of correlation for the following data:

X	35	40	60	79	83	95
Y	17	28	30	32	38	49

Solution :

X	Y	$dx = X - 65$	$dy = Y - 32$	dx^2	dy^2	$dx\,dy$
35	17	− 30	− 15	900	225	450
40	28	− 25	− 4	625	16	100
60	30	− 5	− 2	25	4	10
79	32	14	0	196	0	0
83	38	18	6	324	36	108
95	49	30	17	900	289	510
392	194	2	2	2970	570	1178

$$\bar{X} = \frac{392}{6} = 65.33$$

$$\bar{Y} = \frac{194}{6} = 32.33$$

Co-efficient of correlation

$$r(X, Y) = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2} \cdot \sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6(1178)-2(2)}{\sqrt{6(2970)-4}\cdot\sqrt{6(570)-4}}$$

$$= \frac{7064}{\sqrt{17816}\cdot\sqrt{3416}} = \frac{7064}{(133.48)(58.45)}$$

$$= \frac{7064}{7801.91} = 0.906$$

2. Calculate the coefficient of correlation from the following data: $\Sigma X = 50$, $\Sigma Y = -30$, $\Sigma X^2 = 290$, $\Sigma Y^2 = 300$, $\Sigma XY = -115$, $N = 10$

Solution : $\Sigma X = 50$, $\Sigma Y = -30$, $\Sigma X^2 = 290$, $\Sigma Y^2 = 300$, $\Sigma XY = -115$, $N = 10$

Co-efficient of correlation

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

$$= \frac{10(-115) - (50)(-30)}{\sqrt{10(290) - (50)^2}\sqrt{10(300) - (-30)^2}}$$

$$= \frac{-1150 + 1500}{(20)\left(\sqrt{2100}\right)} = \frac{350}{(20)(45.83)}$$

$$= \frac{350}{916.6} = 0.382$$

$$r = 0.382$$

3. Calculate the correlation coefficient from the data given below:

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

$$r(x, y) = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}} = \frac{57}{\sqrt{60} \cdot \sqrt{60}} = \frac{57}{60}$$

$$r = 0.95$$

Regression:

(i) Regression equation of X on Y

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

(ii) Regression equation of Y on X

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

(i) Regression equation of X on Y

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

(ii) Regression Equation of Y on X

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

Regression coefficient of X on Y

$$b_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\,\Sigma Y^2 - (\Sigma Y)^2}$$

(ii) Regression coefficient *of* Y on X

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\,\Sigma X^2 - (\Sigma X)^2}$$

$$b_{yx} = \frac{N\Sigma dx\,dy - (\Sigma dx)(\Sigma dy)}{N\,\Sigma dx^2 - (\Sigma dx)^2}$$

$$b_{xy} = r\,\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} \qquad b_{yx} = r\,\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2}$$

The Correlation coefficient $r = \pm \sqrt{b_{xy} \times b_{yx}}$

Calculate the two regression equations of X on Y and Y on X from the data given below, taking deviations from a actual means of X and Y.

Price(Rs.)	10	12	13	12	16	15
Amount demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs.20.

Solution:

Calculation of Regression equation

X	$x=(X-13)$	x^2	Y	$y=(Y-41)$	y^2	xy
10	−3	9	40	−1	1	3
12	−1	1	38	−3	9	3
13	0	0	43	2	4	0
12	−1	1	45	4	16	−4
16	3	9	37	−4	16	−12
15	2	4	43	2	4	4
$\sum X=78$	$\sum x=0$	$\sum x^2=24$	$\sum Y=246$	$\sum y=0$	$\sum y^2=50$	$\sum xy=-6$

(i) Regression equation of X on Y

$$X-\overline{X} = r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

$$\overline{X} = \frac{78}{6}=13\ ,\ \overline{Y} = \frac{246}{6}=41$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-6}{50} = -0.12$$

$$X-13 = -0.12\,(Y-41)$$

$$X-13 = -0.12Y+4.92$$

$$X = -0.12Y+17.92$$

(ii) Regression Equation of Y on X

$$Y-\overline{Y} = r\frac{\sigma_y}{\sigma_x}(X-\overline{X})$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = -\frac{6}{24} = -0.25$$

$$Y-41 = -0.25\,(X-13)$$

$$Y-41 = -0.25\,X+3.25$$

$$Y = -0.25\,X+44.25$$

When X is 20, Y will be

$= -0.25\ (20)+44.25$

$= -5+44.25$

$= 39.25$ (when the price is Rs. 20, the likely demand is 39.25)

Obtain regression equation of Y on X and estimate Y when X=55 from the following

X	40	50	38	60	65	50	35
Y	38	60	55	70	60	48	30

Solution:

X	Y	$dx=(X-48)$	dx^2	$dy=(Y-50)$	dy^2	$dxdy$
40	38	−8	64	−12	144	96
50	60	2	4	10	100	20
38	55	−10	100	5	25	−50
60	70	12	144	20	400	240
65	60	17	289	10	100	170
50	48	2	4	-2	4	-4
35	30	-13	169	-20	400	260
$\sum X=338$	$\sum Y=361$	$\sum dx=2$	$\sum dx^2=774$	$\sum dy=11$	$\sum dy^2=1172$	$\sum dxdy=732$

Table 9.9

$$\overline{X} = \frac{\Sigma X}{N} = \frac{338}{7} = 48.29$$

$$\overline{Y} = \frac{\Sigma Y}{361}$$

(i) Regression coefficients of Y on X

$$b_{yx} = \frac{N\Sigma dx\, dy - (\Sigma dx)(\Sigma dy)}{N\,\Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{7(732) - (2)(11)}{7(774) - (2)^2}$$

$$= \frac{5124 - 22}{5418 - 4}$$

$$= \frac{5102}{5414}$$

$$= 0.942$$

$$b_{yx} = 0.942$$

(ii) Regression equation of Y on X

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

Y–51.57 = 0.942(X–48.29)

Y = 0.942X–45.49+51.57=0.942 #–45.49+51.57

Y = 0.942X+6.08

The regression equation of Y on X is Y= 0.942X+6.08 Estimation of Y when X= 55

Y= 0.942(55)+6.08=57.89

Find the means of X and Y variables and the coefficient of correlation between them from the following two regression equations:

$2Y - X - 50 = 0$

$3Y - 2X - 10 = 0.$

Solution:

We are given

$2Y - X - 50 = 0 \quad \ldots (1)$

$3Y - 2X - 10 = 0 \quad \ldots (2)$

Solving equation (1) and (2)

We get $Y = 90$

Putting the value of Y in equation (1)

We get $X = 130$

Hence $\overline{X} = 130$ and $\overline{Y} = 90$

Calculating correlation coefficient

Let us assume equation (1) be the regression equation of Y on X

$2Y = X+50$

$$2Y = X+50$$

$$Y = \frac{1}{2}X+25 \text{ therefore } b_{yx} = \frac{1}{2}$$

Clearly equation (2) would be treated as regression equation of X on Y

$$3Y-2X-10 = 0$$

$$2X = 3Y-10$$

$$X = \frac{3}{2}Y-5 \text{ therefore } b_{xy} = \frac{3}{2}$$

The Correlation coefficient $r = \pm \sqrt{b_{xy} \times b_{yx}}$

$$r = \sqrt{\frac{1}{2} \times \frac{3}{2}} = 0.866$$

Example

Find the means of X and Y variables and the coefficient of correlation between them from the following two regression equations:

$4X-5Y+33=0$

$20X-9Y-107=0$

Solution:

We are given

$4X-5Y+33=0 \quad \ldots (1)$

$20X-9Y-107=0 \quad \ldots (2)$

Solving equation (1) and (2)

We get $Y = 17$

Putting the value of Y in equation (1)

We get $X = 13$

Hence $\overline{X} = 13$ and $\overline{Y} = 17$

Calculating correlation coefficient

Let us assume equation (1) be the regression equation of X on Y

$$4X = 5Y - 33$$

$$X = \frac{1}{4}(5Y - 33)$$

$$X = \frac{5}{4}Y - \frac{33}{4}$$

$$b_{xy} = \frac{5}{4} = 1.25$$

Let us assume equation (2) be the regression equation of Y on X

$$9Y = 20X - 107$$

$$Y = \frac{1}{9}(20X - 107)$$

$$Y = \frac{20}{9}X - \frac{107}{9}$$

$$b_{yx} = \frac{20}{9} = 2.22$$

But this is not possible because both the regression coefficient are greater than

So our above assumption is wrong. Therefore treating equation (1) has regression equation of Y on X and equation (2) has regression equation of X on Y. So we get

$$b_{yx} = 5/4 = 1.25$$

$$b_{xy} = 9/20 = 0.45$$

$$r = \pm\sqrt{1.25 \times 0.45} = \pm\sqrt{0.5625}$$

$$= \pm 0.75$$

The following table shows the sales and advertisement expenditure of a form

	Sales	Advertisement expenditure (Rs. Crores)
Mean	40	6
SD	10	1.5

Coefficient of correlation r= 0.9. Estimate the likely sales for a proposed advertisement expenditure of Rs. 10 crores.

Solution:

Given $\overline{X}=40, \overline{Y}=6, \sigma_x=10, \sigma_y=1.5$ and $r=0.9$

Equation of line of regression x on y is

$$X-\overline{X} = r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

$$X-40 = (0.9)\frac{10}{1.5}(Y-6)$$

$$X-40 = 6Y-36$$

$$X = 6Y+4$$

When advertisement expenditure is 10 crores i.e., Y=10 then sales X=6(10)+4=64 which implies sales is 64.

There are two series of index numbers P for price index and S for stock of the commodity. The mean and standard deviation of P are 100 and 8 and of S are 103 and 4 respectively. The correlation coefficient between the two series is 0.4. With these data obtain the regression lines of P on S and S on P.

Solution:

Let us consider X for price P and Y for stock S. Then the mean and SD for P is considered as X-Bar = 100 and σ_x=8. respectively and the mean and SD of S is considered as Y-Bar =103 and σ_y=4. The correlation coefficient between the series is $r(X,Y)$=0.4

Let the regression line X on Y be

$$X-\overline{X} = r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

$$X-100 = (0.4)\frac{8}{4}(Y-103)$$

$$X-100 = 0.8(Y-103)$$

$$X-0.8Y-17.6 = 0 \quad \text{or} \quad X=0.8Y+17.6$$

The regression line Y on X be $Y-\overline{Y} = r\frac{\sigma_y}{\sigma_x}(X-\overline{X})$

$$Y-103 = (0.4)\frac{4}{8}(X-100)$$

$$Y-103 = 0.2\,(X-100)$$

$$Y-103 = 0.2\,X-20$$

$$Y = 0.2\,X+83 \quad \text{or} \quad 0.2\,X-Y+83=0$$

In a laboratory experiment on correlation research study the equation of the two regression lines were found to be $2X-Y+1=0$ and $3X-2Y+7=0$. Find the means of X and Y. Also work out the values of the regression coefficient and correlation between the two variables X and Y.

Solution:

Solving the two regression equations we get mean values of X and Y

$$2X-Y = -1 \qquad \ldots (1)$$

$$3X-2Y = -7 \qquad \ldots (2)$$

Solving equation (1) and equation (2) We get $X=5$ and $Y=11$

Therefore the regression line passing through the means $\overline{X}=5$ and $\overline{Y}=11$

The regression equation of Y on X is $3X-2Y=-7$

$$2Y = 3X+7$$

$$Y = \frac{1}{2}(3X+7)$$

$$Y = \frac{3}{2}X+\frac{7}{2}$$

$$\therefore \quad b_{yx} = \frac{3}{2}(>1)$$

The regression equation of X on Y is

$$2X - Y = -1$$

$$2X = Y - 1$$

$$X = \frac{1}{2}(Y - 1)$$

$$X = \frac{1}{2}Y - \frac{1}{2}$$

$$\therefore \quad b_{xy} = \frac{1}{2}$$

The regression coefficients are positive

$$r = \pm \sqrt{b_{xy} \cdot b_{yx}} = \pm \sqrt{\frac{3}{2} \times \frac{1}{2}}$$

$$= \sqrt{\frac{3}{2} \times \frac{1}{2}}$$

$$= \sqrt{\frac{3}{4}}$$

$$= 0.866$$

$$r = 0.866$$

Rank Correlation:

Spearman's Rank Correlation Coefficient

$$\rho = 1 - \frac{6\Sigma d^2}{N\left(N^2 - 1\right)}$$

The following are the ranks obtained by 10 students in commerce and accountancy are given below

Commerce	6	4	3	1	2	7	9	8	10	5
Accountancy	4	1	6	7	5	8	10	9	3	2

To what extent is the knowledge of students in the two subjects related?

Solution :

Rank in Commerce (R_X)	Rank in Accountancy (R_Y)	$d = R_X - R_Y$	d^2
6	4	2	4
4	1	3	9
3	6	−3	9
1	7	−6	36
2	5	−3	9
7	8	−1	1
9	10	−1	1
8	9	−1	1
10	3	7	49
5	2	3	9
			$\Sigma d^2 = 128$

Rank Correlation is given by

$$\rho = 1 - \frac{6\Sigma d^2}{N\left(N^2 - 1\right)}$$

$$= 1 - \frac{6(128)}{10(100-1)} = 1 - \frac{768}{990} = 1 - 0.7758$$

$$\rho = 0.2242$$

Estiamted Cost (X)	Actual Cost (Y)	R_x	R_y	$d = R_x - R_y$	d^2
300	273	2	1	1	1
450	486	4	5	−1	1
800	734	7	7	0	0
250	297	1	2	−1	1
500	631	6	6	0	0
975	872	8	8	0	0
475	396	5	3	2	4
400	457	3	4	−1	1
					$\Sigma d^2 = 8$

Rank Correlation is given by

$$\rho = 1 - \frac{6\Sigma d^2}{N\left(N^2 - 1\right)} = 1 - \frac{6(8)}{8(64-1)}$$

$$= 1 - \frac{6}{63} = 1 - 0.095$$

$$\rho = 0.905$$

Rank correlation for Repeated ranks:

$$\rho = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^2 - m_3) + \ldots\ldots\ldots\ldots\right]}{n(n^2-1)}$$

Where m1,m2,…. are the number of times a value repeated

Example 4: Calculate rank correlation coefficient from the following data:

Expenditure on advertisement	10	15	14	25	14	14	20	22
Profit	6	25	12	18	25	40	10	7

Solution: Let us denote the expenditure on advertisement by x and profit by y

x	Rank of x (R_x)	y	Rank of y (R_y)	$d = R_x - R_y$	d^2
10	8	6	8	0	0
15	4	25	2.5	1.5	2.25
14	6	12	5	1	1
25	1	18	4	−3	9
14	6	25	2.5	3.5	12.25
14	6	40	1	5	25
20	3	10	6	−3	9
22	2	7	7	−5	25
					$\sum d^2 = 83.50$

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{m(m^2-1)}{12} + \ldots\right\}}{n(n^2-1)}$$

Here rank 6 is repeated three times in rank of x and rank 2.5 is repeated twice in rank of y, so the correction factor is

$$\frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}$$

Hence rank correlation coefficient is

$$r_s = 1 - \frac{6\left\{83.50 + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}\right\}}{8(64-1)}$$

$$r_s = 1 - \frac{6\left\{83.50 + \frac{3 \times 8}{12} + \frac{2 \times 3}{12}\right\}}{8X\ 63}$$

$$r_s = 1 - \frac{6(83.50 + 2.50)}{504}$$

$$r_s = 1 - \frac{516}{504}$$

$$r_s = 1 - 1.024 = -\ 0.024$$

There is a negative association between expenditure on advertisement and profit.

SKEWNESS

The measures of central tendency and variation do not reveal all the characteristics of a given set of data. For example, two distributions may have the same mean and

standard deviation but may differ widely in the shape of their distribution. Either the distribution of data is symmetrical or it is not. If the distribution of data is not symmetrical, it is called *asymmetrical* or *skewed*. Thus skewness refers to the lack of symmetry in distribution.

THANK YOU

BUSSINESS STATISTICS

Measures of Skewness

Dr. T N KAVITHA

Assistant Professor of Mathematics

SCSVMV

Skewness

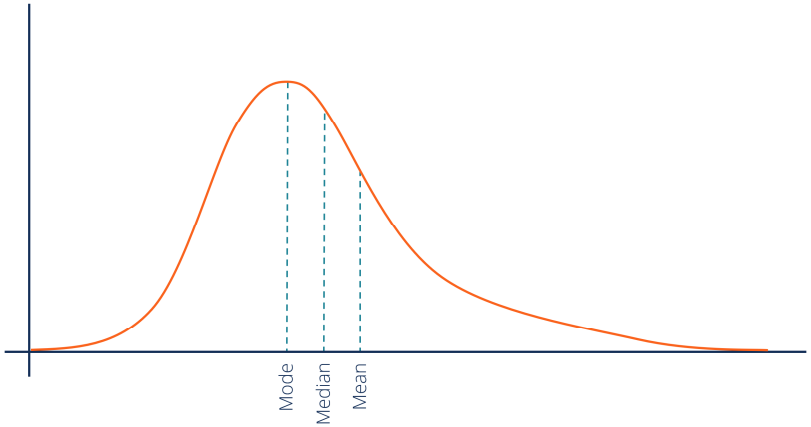
The measures of central tendency and variation do not reveal all the characteristics of a given set of data.

For example, two distributions may have the same mean and standard deviation but may differ widely in the shape of their distribution.

Either the distribution of data is symmetrical or it is not. If the distribution of data is not symmetrical, it is called asymmetrical or skewed. Thus skewness refers to the lack of symmetry in distribution.

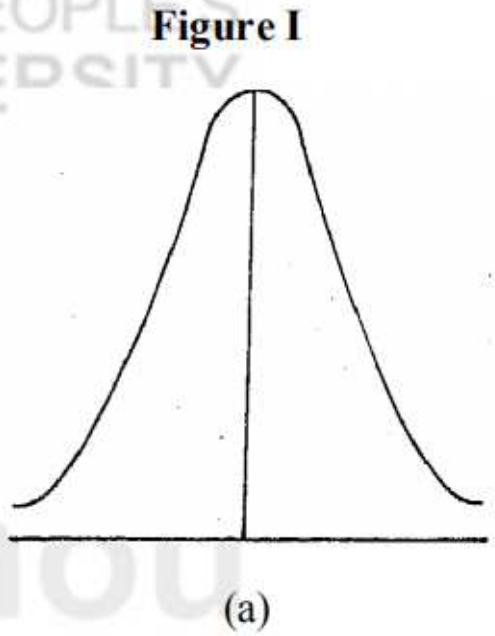
Skewness

Asymmetrical or skewed distribution



symmetrical distribution

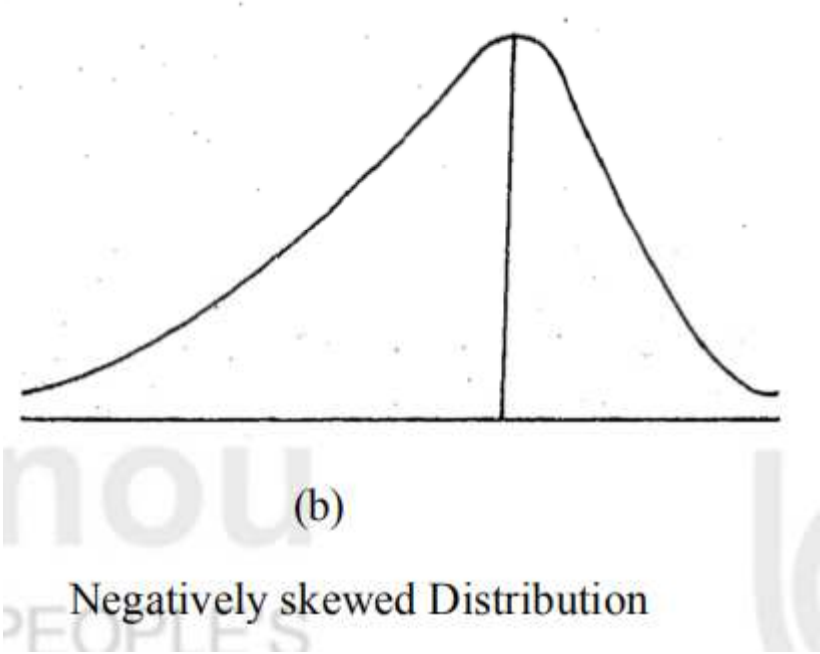
mean = median = mode.



Negative skewness

If the longer tail is towards the lower value or left hand side, the skewness is negative. Negative skewness arises when the mean is decreased by some extremely low values, thus making

mean < median < mode.



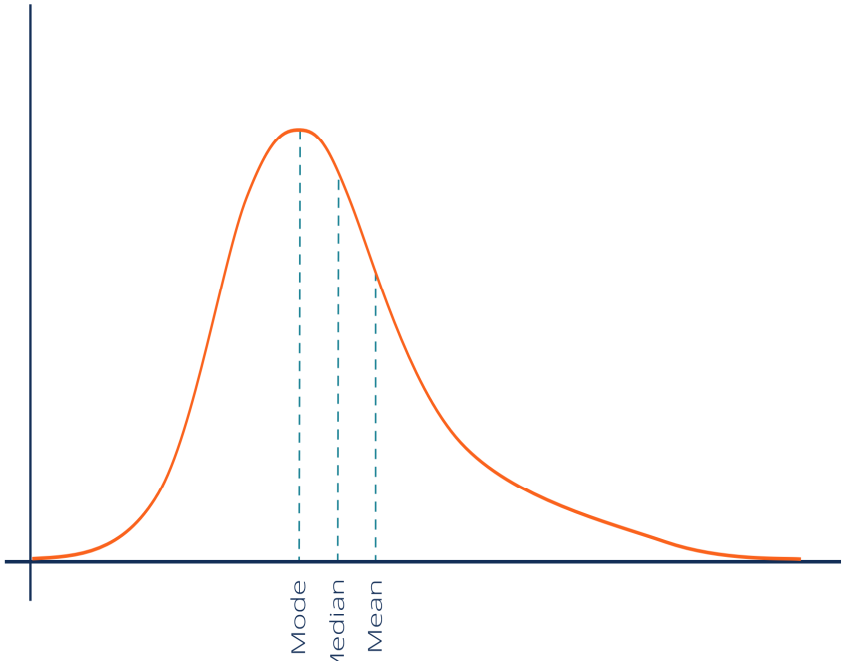
(b)

Negatively skewed Distribution

Positive skewness

If the longer tail of the distribution is towards the higher values or right hand side, the skewness is positive. Positive skewness occurs when mean is increased by some unusually high values, thereby making

mean > median > mode.



RELATIVE SKEWNESS

In order to make comparisons between the skewness in two or more distributions, the coefficient of skewness (given by Karl Pearson) can be defined as:

Karl Pearson coefficient of skewness

$$SK. = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$$

If the mode cannot he determined, then using the approximate relationship,
Mode = 3Median - 2 Mean, the above formula reduces to

$$SK. = \frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$$

Note

If the value of this coefficient is zero, the distribution is symmetrical; (mean = median=mode)

If the value of the coefficient is positive, it is positively skewed distribution,(mean>median>mode)

If the value of the coefficient is negative, it is negatively skewed distribution. (mean<median<mode)

The value of this coefficient usually lies between ± 1

Bowley's coefficient of skewness

Bowley's coefficient of skewness

$$SK = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

When we are given open-end distributions where extreme values are present in thedata or positional measures such as median and quartiles, the formula for coefficient of skewness (given by Bowley) is more appropriate.

Absolute Measures of Skewness

Following are the absolute measures of skewness:

1. Skewness (Sk) = Mean – Median
2. Skewness (Sk) = Mean – Mode
3. Skewness (Sk) = (Q3 - Q2) - (Q2 - Q1)

For comparing to series, we do not calculate these absolute mearues we calculate the relative measures which are called coefficient of skewness.

Coefficient of skewness are pure numbers independent of units of measurements.

For a distribution Karl Pearson's coefficient of skewness is 0.64,S.D is 13 and mean is 59.2 Find mode and median.

Solution:

We have given $Sk = 0.64$, $\sigma = 13$ and $Mean = 59.2$

Therefore by using formulae

$$Sk = \frac{\text{Mean -Mode}}{\sigma}$$

$$0.64 = \frac{59.2 \text{ -Mode}}{13}$$

$$\text{Mode} = 59.20 - 8.32 = 50.88$$

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

$$50.88 = 3 \text{ Median - } 2\ (59.2)$$

$$Median = \frac{50.88 \text{ -}118.4}{3} = \frac{169.28}{3} = 56.42$$

Karl Pearson’s coefficient of skewness is 1.28, its mean is164and mode 100, find the standard deviation.

Using the formulae, we have

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$1.28 = \frac{164 - 100}{\sigma}$$

$$\sigma = \frac{64}{1.28} = 50$$

The following are the marks of 150 students in an examination.Calculate Karl Pearson's coefficient of skewness.

0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
10	40	20	0	10	40	16	14

Class	f	x	CF	$d' = \frac{x-35}{10}$	fd'	fd'^2
0-10	10	5	10	−3	−30	90
10-20	40	15	50	−2	−80	160
20-30	20	25	70	−1	−20	20
30-40	0	35	70	0	0	0
40-50	10	45	80	+1	10	10
50-60	40	55	120	+2	80	160
60-70	16	65	136	+3	48	144
70-80	14	75	150	+4	56	244
					$\sum fd' = 64$	$\sum fd'^2 = 828$

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

$$= 40 + \frac{75 - 70}{10} \times 10 = 45$$

$$\text{Mean } (\bar{x}) = A + \frac{\sum_{i=1}^{k} fd'}{N} \times h$$

$$= 35 + \frac{64}{150} \times 10 = 39.27$$

$$\text{Standard Deviation } (\sigma) = h \times \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2}$$

$$= 10 \times \sqrt{\frac{828}{150} - \left(\frac{64}{150}\right)^2}$$

$$= 10 \times \sqrt{5.33} = 23.1$$

Therefore, coefficient of skewness:

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

$$= \frac{3(39.27 - 45)}{23.1} = -0.744$$

THANK YOU