1st SEM BSc Psychology

UNIVERSITY OF CALICUT

STA 1C 02- DESCRIPTIVE STATISTICS

Prepared by
Shafeeka Chungath
Assistant Professor
Dept. Of Commerce

CPA College of Global Studies
Puthanathani

SEMESTER I
STA 1C 02- DESCRIPTIVE STATISTICS
Contract Hours per week: 4
Number of credits: 3
Number of Contact Hours: 72
Course Evaluation: External 60 Marks+ Internal 15 Marks
Duration of Exam: 2 Hours
Objectives
1. To generate interest in Statistics
2. To equip the students with the concepts of basic Statistics
3. To provide basic knowledge about Statistical methods

Module 1: A basic idea about data- collection of data, primary and secondary data, organization, planning of survey and diagrammatic representation of data
10 Hours
Module 2: Classification and tabulation- Classification of data, frequency distribution, formation of a frequency distribution, Graphic representation viz. Histogram, Frequency Curve, Polygon, Ogives, Bar diagram and Pie diagram
10 Hours
Module 3: Measure of central tendency- Arithmetic Mean, Median, Mode, Geometric Mean, Harmonic Mean, Combined Mean, Advantages and disadvantages of each average
20 Hours

Module 4: Measures of dispersion- Range, Quartile Deviation, Mean Deviation, Standard Deviation, Combined Standard Deviation, Percentiles, Deciles, Relative Measures of Dispersion, Coefficient of variation
16 Hours

Module 5: Skewness and Kurtosis- Pearson's and Bowley's coefficient of skewness, Percentile Measure of Kurtosis
16 Hours

References
1. Gupta, S.P. Statistical Methods. Sultan Chand and Sons: New Delhi.
2. Gupta, S.C., & Kapoor, V.K. Fundamentals of Applied Statistics. New Delhi: Sultan Chand and Sons.

3. Garret, H.E., &Woodworth, R.S. Statistics in Psychology and Education. Bombay: Vakila, Feffex and Simens Ltd.
4. Mood, A.M., Graybill, F.A and Boes, D.C. Introduction to Theory of Statistics. 3rd Edition Paperback – International Edition.
5. Mukhopadhyay,P. Mathematical Statistics. New central Book Agency (P) Ltd: Calcutta.

MODULE 1

A BASIC IDEA ABOUT DATA

"**Statistics** as the aggregate of facts affected to mark extent by the multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for the predetermined purpose and placed in relation to each other"

- Horace Secrist –

Functions of statistics

- Simplifies complexity
- Definiteness
- Comparison
- Formulation of policies
- Formulating and testing hypothesis
- Prediction
- Tests and laws of other science
- Studies relationship

Stages in a statistical investigation

1. Collection of data
2. Organization or presentation of data
3. Analysis of data
4. Interpretation of data

COLLECTION OF DATA

- Collection of data is the process of enumeration together with the proper recording of results.
- Statistical survey or enquiry means a search for the collection of facts of given problem.
- A statistician begins the work with the collection of data, ie, numerical facts or raw data.
- From these raw data a statistician can analyses the data and come to a final conclusion.

Planning and designing of enquiry

- Planning is the most important element of a survey.
- It examines the preliminaries of data collection such as:
    - Objective of an enquiry
    - Scope
    - Statistical unit
    - Source of data
    - Method of data collection
    - The frames
    - Standard of data
    - Type of enquiry

Classification of data

Statistical data may be classified as two:

1. Primary data
2. Secondary data

Primary data

Primary data is data that is collected by a researcher from first-hand sources, using methods like surveys, interviews, or experiments. It is collected with the research project in mind, directly from primary sources. The term is used in contrast with the term secondary data.

Methods of collecting primary data

- Direct personal observation
  - Under this method the investigator collects the data personally.
  - He has to go to the spot for conducting interview
- Indirect oral interview
  - This method adopts the opposite technique of data collection.
  - In this case the data collection does not take place directly from the source but taking interviews of the persons who have close relation with the source.
- Information through agencies
  - In this method, the investigator appoints local agents or enumerators in different parts of the area.
  - These agents or enumerators are asked to collect information and transmit it to the investigator.
- Mailed questionnaires
  - In this method, a list of questions is prepared relating to the problem under investigation, is printed and then sent out to the informants through post.
  - It is requested that it may be returned to the investigator properly filled up.
  - A covering letter is also sent with the questionnaire.
  - A stamped self- addressed envelope is also attached.
- Schedules sent through enumerator
  - In this method, list of questions or schedules are sent to the informants through the enumerators.
  - They read the questions to the informants and record their answers on the same schedules.
  - At first, enumerator explains the aims and objectives of the enquiry and asks them for co-operation

Secondary data

- Secondary data refers to data that is collected by someone other than the primary user. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data that was originally collected for other research purposes.

Precautions in the use of secondary data

- ❖ Suitability for the purpose
- ❖ Adequacy of data
- ❖ Reliability or dependability
- ❖ Security

Diagrammatic Presentation of Data

Diagrams play an important role in statistical data presentation. Diagrams are nothing but geometrical figures like lines, bars, circles, squares, etc. Diagrammatic data presentation allows us to understand the data in an easier manner.

Advantages of Diagrammatic Data Presentation:-

- **Easy to understand** – Diagrammatic data presentation makes it easier for a common man to understand the data.
- **Simplified Presentation** – You can represent large volumes of complex data in a simplified and intelligible form using diagrams.
- **Reveals hidden facts** – When you classify and tabulate data, some facts are not revealed. Diagrammatic data presentation helps in bringing out these facts and also relations.
- **Quick to grasp** – Usually, when the data is represented using diagrams, people can grasp it quickly.
- **Easy to compare** – Diagrams make it easier to compare data.
- **Universally accepted** – Almost all fields of study like Business, economics, social institutions, administration, etc. use diagrams. Therefore, they have universal acceptability.

Planning of survey

A **survey** plan begins with objectives that describe why and for whom the **survey** is being done. The **survey** objectives tell a lot about the data that need to be collected. The objectives also help determine the population to be targeted.

a sample survey consists of the following steps:-

1. Define the target population. ...
2. Select the **sampling** scheme and **sample** size. ...
3. Develop the **questionnaire**. ...
4. Recruit and train the field investigators. ...
5. Obtain information as per the **questionnaire**. ...
6. Scrutinize the information gathered. ...
7. Analyze and interpret the information.

MODULE 2

CLASSIFICATION AND TABULATION

Classification is the process of grouping **data** into different categories, on the basis of nature, behavior, or common characteristics.

Objects of Classification:

1. It condenses the mass of data in an easily assailable form.

2. It eliminates unnecessary details.

3. It facilitates comparison and highlights the significant aspect of data.

4. It enables one to get a mental picture of the information and helps in drawing inferences.

5. It helps in the statistical treatment of the information collected.

Types of classification:

Statistical data are classified in respect of their characteristics.

a) Chronological classification

b) Geographical classification

c) Qualitative classification

d) Quantitative classification

a) Chronological classification: In chronological classification the collected data are arranged according to the order of time expressed in years, months, weeks, etc., The data is generally classified in ascending order of time.

b) Geographical classification: In this type of classification the data are classified according to geographical region or place. For instance, the production of paddy in different states in Iraq, production of wheat in different countries etc.,

c) Qualitative classification: In this type of classification data are classified on the basis of same attributes or quality like sex, literacy, religion, employment etc., Such attributes cannot be measured along with a scale.

d) Quantitative classification: Quantitative classification refers to the classification of data according to some characteristics that can be measured such as height, weight, etc.,

Formation of frequency distribution

3 categories of frequency distribution are:

- Uni-variate frequency distribution
- Bi-variate frequency distribution
- Multi-variate frequency distribution

Terms used in frequency distribution

- Frequency:-it is the number of times an observation repeats in the data. It denoted as f.
- Lower and upper limits:-all the left end limits of a frequency distribution are called lower limit and all the right end limits are called upper limits.
- Class boundaries:- Class boundaries are the data values which separate classes. They are not part of the classes or the dataset. The lower class boundary of a class is defined as the average of the lower limit of the class in question and the upper limit of the previous class.
- Class width or class interval:- it is the difference between the corresponding upper and lower limits.
- Mid points or class mark:- arithmetic mean of the upper and lower boundaries are called mid point.

$$\frac{\text{Upper boundary} + \text{lower boundary}}{2}$$

Steps in forming a frequency distribution

- Find the largest and lowest observation in the data find out range, width and number of classes.
- After fixing class width and number of classes, write down various classes.
- Represent the classes according to nature(whether discrete or continuous)

Types of frequency tables

- Less than cumulative frequency table
- Greater than cumulative frequency table

Weight in kg (CB)	Cumulative Frequency	
	Less than	More than
43.50	0	33 + 3 or 36
48.50	0 + 3 or 3	29 + 4 or 33
53.50	3 + 4 or 7	24 + 5 or 29
58.50	7 + 5 or 12	17 + 7 or 24
63.50	12 + 7 or 19	8 + 9 or 17
68.50	19 + 9 or 28	0 + 8 or 8
73.50	28 + 8 or 36	0

- Relative frequency table

Chili	Frequency Chosen-Number	Relative Frequency
1	13	13/50=.26
2	7	7/50=.14
3	5	5/50=.1
4	10	10/50=.2
5	7	7/50=.14
6	5	5/50=.1
7	3	3/50=.06

TABULATION

Tabulation is a process of summarizing data and presenting it in a compact form, by putting data into statistical table.

statistical table has at least four major parts and some other minor parts.

- (1) The Title.
- (2) Table number
- (3) The Box Head (column captions)
- (4) The Stub (row captions)
- (5) The Body.
- (6) Foot Notes.
- (7) Source Notes. The general sketch of **table** indicating its **necessary parts** is shown below

GRAPHS

a graph can be defined as a pictorial representation or a diagram that represents data or values in an organized manner. The points on the graph often represent the relationship between two or more things.

Difference between graphs and diagrams

a) Diagrams are very attractive to eye, but graphs are not attractive to eyes.
b) Diagrams do not add anything to the meaning of the data, and hence that cannot be used by statisticians, but graphs make data more meaningful.
c) Diagrams create effective and long lasting impressions in he minds of on lookers, but graphs does not create such impression.
d) Diagrams help us in making comparison between data, but graphs help us in studying the cause and effect relationship between two variables.

e) Diagrams are rarely used to present frequency distribution, but graphs are mostly used for such presentation.

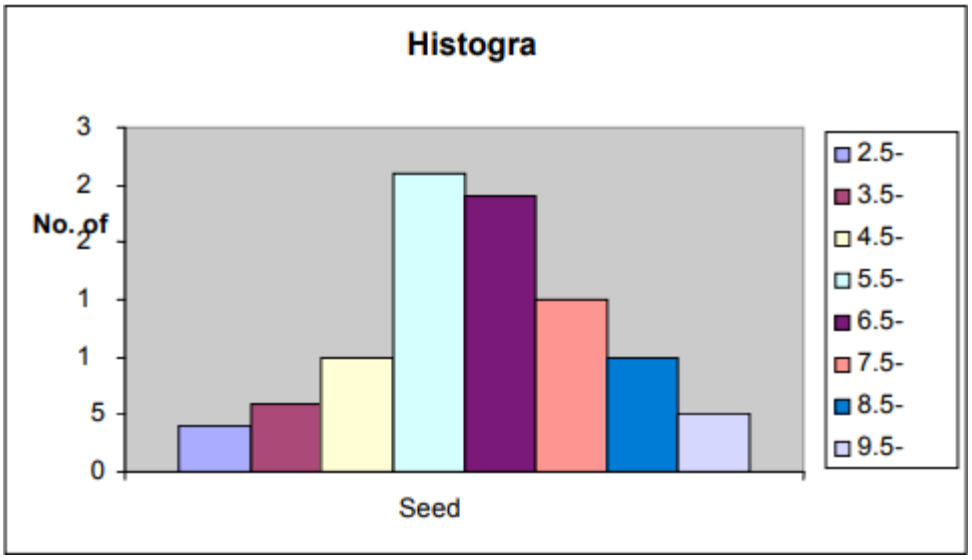
<u>Most commonly used graphical methods are:</u>

- Histogram

  When the data are classified based on the class intervals it can be represented by a histogram. Histogram is just like a simple bar diagram with minor differences. There is no gap between the bars, since the classes are continuous.

  Draw a histogram for the following data

Seed Yield (gms)	No. of Plants
2.5-3.5	4
3.5-4.5	6
4.5-5.5	10
5.5-6.5	26
6.5-7.5	24
7.5-8.5	15
8.5-9.5	10
9.5-10.5	5

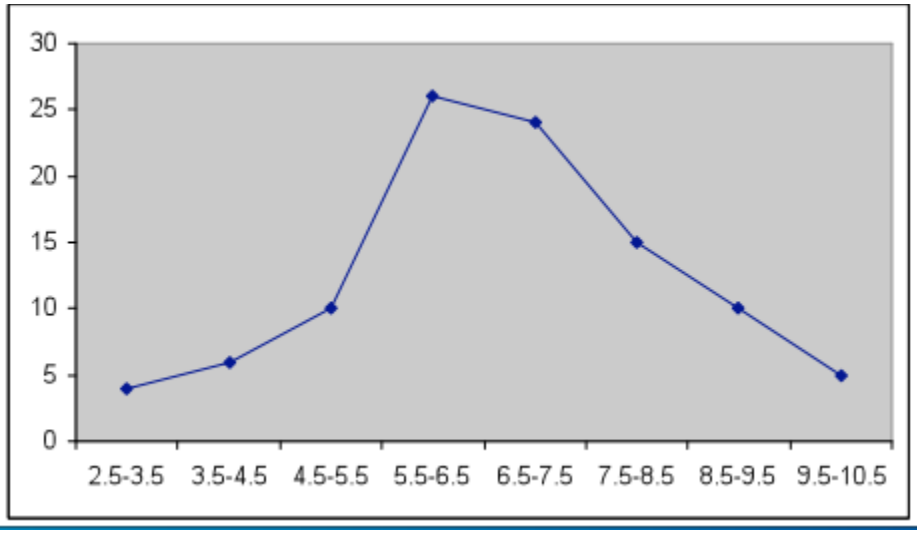


- Frequency polygon

  The frequencies of the classes are plotted by dots against the mid-points of each class. The adjacent dots are then joined by straight lines. The resulting graph is known as frequency polygon

Draw frequency polygon for the following data

Seed Yield (gms)	No. of Plants
2.5-3.5	4
3.5-4.5	6
4.5-5.5	10
5.5-6.5	26
6.5-7.5	24
7.5-8.5	15
8.5-9.5	10
9.5-10.5	5

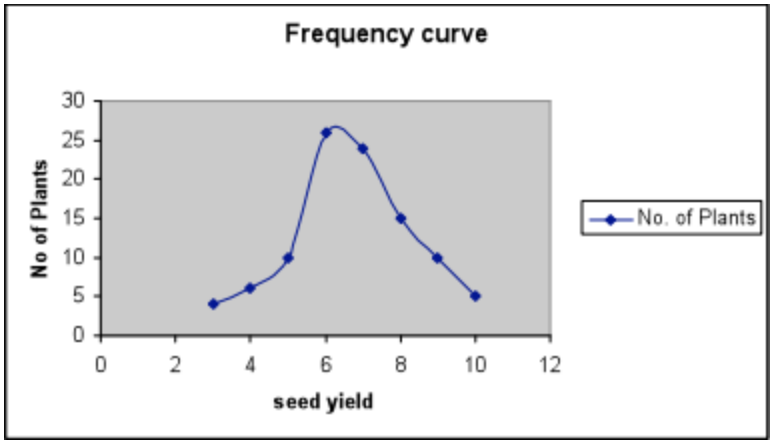


- Frequency curve

  The procedure for drawing a frequency curve is same as for frequency polygon. But the points are joined by smooth or free hand curve.

  Draw frequency curve for the following data

Seed Yield (gms)	No. of Plants
2.5-3.5	4
3.5-4.5	6
4.5-5.5	10
5.5-6.5	26
6.5-7.5	24
7.5-8.5	15
8.5-9.5	10
9.5-10.5	5



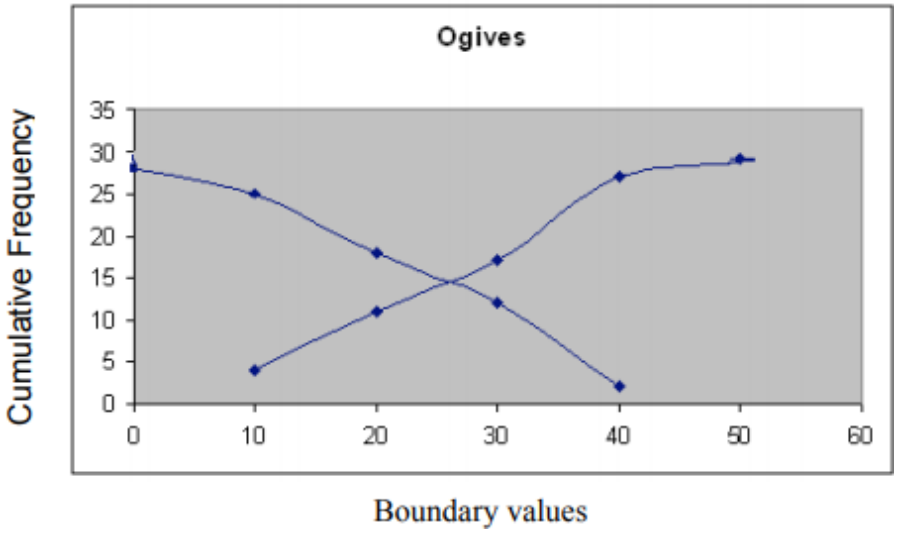
- Cumulative frequency curves or ogives

  Ogives are known also as cumulative frequency curves and there are two kinds of ogives. One is less than ogive and the other is more than ogive.

  Less than ogive: Here the cumulative frequencies are plotted against the upper boundary of respective class interval.

  Greater than ogive: Here the cumulative frequencies are plotted against the lower boundaries of respective class intervals.

Continuous Interval	Mid Point	Frequency	< cumulative Frequency	> cumulative frequency
0-10	5	4	4	29
10-20	15	7	11	25
20-30	25	6	17	18
30-40	35	10	27	12
40-50	45	2	29	2



Module 3
Measure of central tendency

Measures Of Central Tendency.

- ✓ It also Know as Averages.
- ✓ It is a single significant Figure.
- ✓ Which sum up the characteristics of a group of figures.
- ✓ It is conveys General Idea of Whole group.
- ✓ It is Generally located at the Centre or middle of the distribution

Type of averages

1) Arithmetic Mean
2) Median
3) Mode
4) Harmonic mean
5) Geometric mean

1. Arithmetic mean

The average of a set of numerical values, as calculated by adding them together and dividing by the number of terms in the set.

Find the arithmetic mean of the first 7 natural numbers. Solution:

The first 7 natural numbers are 1, 2, 3, 4, 5, 6 and 7. Let

x denote their arithmetic mean.

Then mean = Sum of the first 7 natural numbers/number of natural numbers

$x = (1 + 2 + 3 + 4 + 5 + 6 + 7)/7$

$= 28/7$

$= 4$

Hence, their mean is 4.

2. Median

The **median** is the middle number in a sorted, ascending or descending,

Find the median of the following set of points in a game:

15, 14, 10, 8, 12, 8, 16

Solution:

First arrange the point values in an ascending order (or descending

order). 8, 8, 10, 12, 14, 15, 16

8, 8, 10, 12, 14, 15, 16

Middle position

The number of point values is 7, an odd number. Hence, the median is the value in the middle position.

Median = 12

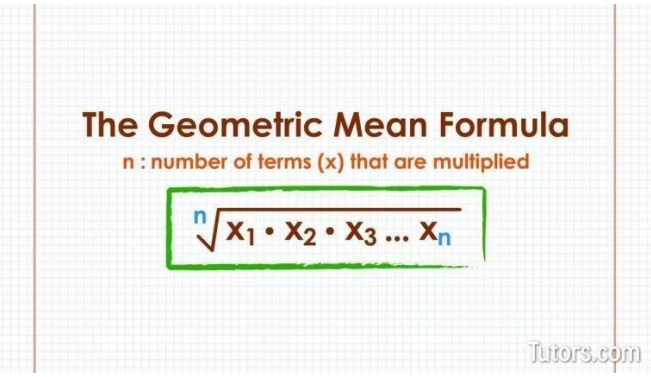
3. **Mode**

The **mode** is the number that appears most frequently in a data set

Example: in {6, 3, 9, 6, 6, 5, 9, 3} the Mode is 6 (it occurs most often).

4. **Geometric mean**

The geometric mean is a mean or average, which indicates the central tendency or typical



value of a set of numbers by using the product of their values.

Example: What is the Geometric Mean of **1, 3, 9, 27 and 81**?

- First we multiply them: $1 \times 3 \times 9 \times 27 \times 81 = 59049$
- Then (as there are 5 numbers) take the 5th root: $\sqrt[5]{59049} = \mathbf{9}$

5. **Harmonic mean**

The harmonic mean is one of several kinds of average, and in particular, one of the Pythagorean means. Typically, it is appropriate for situations when the average of rates is desired

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

The reciprocals of 1, 2 and 4 are:

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

Now add them up:

$$1 + 0.5 + 0.25 = 1.75$$

Divide by how many:

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

The reciprocal of that average is our answer: Harmonic Mean = $\frac{3}{\mathbf{1.75}} =$

1.714 (to 3 places)

MODULE 4

MEASURE OF DISPERSION

1. Range
2. Quartile deviation
3. Mean deviation
4. Standard deviation

1. Range

The **range** is the difference between the largest and the smallest observation in the data.

Example: 1, 3, 5, 6, 7 => Range = 7 -1= 6

2. Quartile deviation

The **Quartile Deviation** is a simple way to estimate the spread of a distribution about a measure of its central tendency

Quartile Deviation = (Q3 – Q1) / 2

Coefficient of Quartile Deviation = (Q3 – Q1) / (Q3 + Q1)

Example:-

Consider a data set of following numbers: 22, 12, 14, 7, 18, 16, 11, 15, 12. You are required to calculate the Quartile Dev Solution:

First, we need to arrange data in ascending order to find Q3 and Q1 and avoid any duplicates

7, 11, 12, 13, 14, 15, 16, 18, 22

Calculation of Q1 can be done as follows,

Q1 = ¼ (9 + 1)

=¼ (10)

Q1=2.5 Term

Calculation of Q3 can be done as follows,

Q3=¾ (9 + 1)

=¾ (10)

Q3= 7.5 Term

Mean deviation

The mean of the absolute values of the numerical differences between the numbers of a set (such as statistical data) and their mean or median

Calculation of MD

Individual Series	**Discrete Series**	**Continuous Series**
$\frac{\sum /d/}{n}$	$\frac{\sum f/d/}{N}$	$\frac{\sum f/d/}{N}$

Example: the Mean Deviation of 3, 6, 6, 7, 8, 11, 15, 16

Step 1: Find the **mean**:

$$\text{Mean} = 3 + 6 + 6 + 7 + 8 + 11 + 15 + 16\mathbf{8} = 728 = 9$$

Step 2: Find the **distance** of each value from that mean:

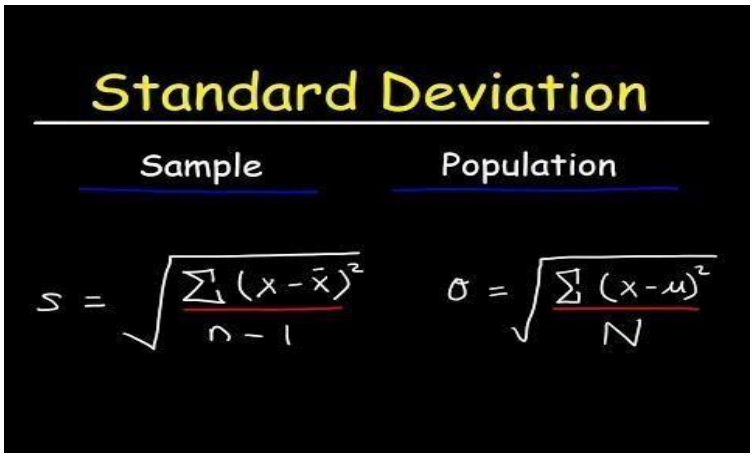
Value	Distance from 9
3	6
6	3
6	3
7	2
8	1
11	2
15	6
16	7

Step 3. Find the **mean of those distances**:

$$\textbf{Mean Deviation} = 6 + 3 + 3 + 2 + 1 + 2 + 6 + 7\mathbf{8} = 308 = \mathbf{3.75}$$

1. <u>Standard deviation</u>

It is the square root of the mean of the square of the deviations of all values of a series from their arithmetic mean



A. IritliviAual Series:

Deviation can be taken from Acti:at Mean and following foriiiiila is used.

P‘rom AetMe£ be-m

BW.- Where X x' is the squared deviation from Actual Mean ;

$$\int^2$$

Here alsn the dexñatic›ns car tre taken from Wctvtal c›r Assured Mf ear.

From Actual Mean:

S.D.=L /M

Where x‘ is the sqii.ai’e of deviations from actual inean,fdenotes corresponding frequent ; N = *f*

$$\sqrt{\qquad \int^2 \over N}$$

ezB i8 tot O0, @ tez as E.a. k/,
/ is J›e 6equency of eacb variable
dX ia tke deviation of' variable 6•oot Assumed roean.

C. CONINUOUS SERIES

Here we take the deviations from Actual or Assumed Mean as desired from the Mid Point of Class-Intervals.

From Actual Mean

$$\text{S.D.} = \sqrt{\frac{\Sigma fx^2}{N}}$$

Where x^2 is the square of deviations from actual mean and f is the corresponding frequency; $N = \Sigma f$

From Assumed Mean

$$\text{S.D.} = \sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2}$$

Where dx^2 is the squared deviation from Assumed Mean and dx is the deviation from Assumed Mean ; f is the corresponding frequency and $N = \Sigma f$.

Important Note :- In case of Continuous Series, if Class Intervals are of Equal Size or any common factor can be taken from dx, then we proceed for ***Step-Deviation Method*** as per the following formula.

$$\text{S.D.} = \sqrt{\frac{\Sigma fd'x^2}{N} - \left(\frac{\Sigma fd'x}{N}\right)^2} \times i$$

Where dx' is the step-deviation such that $\therefore \quad d'x = \frac{dx}{i}$

Here i is the class-interval.

Note :-When class intervals are equal we take step deviations directly as dx instead of dx'.

MODULE 5
SKEWNESS AND KURTOSIS

Skewness

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

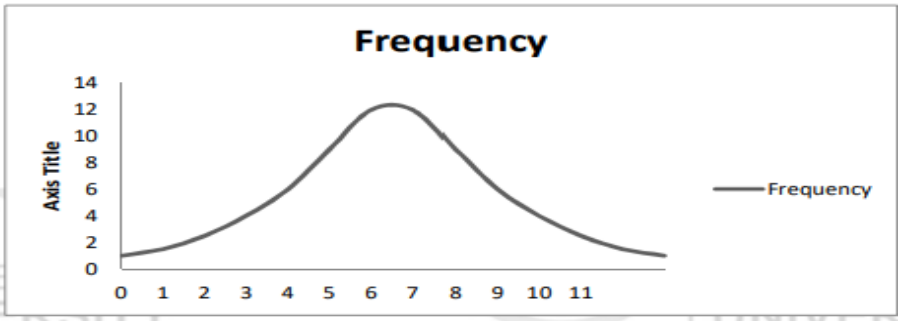


Fig. 4.1: Symmetrical Curve

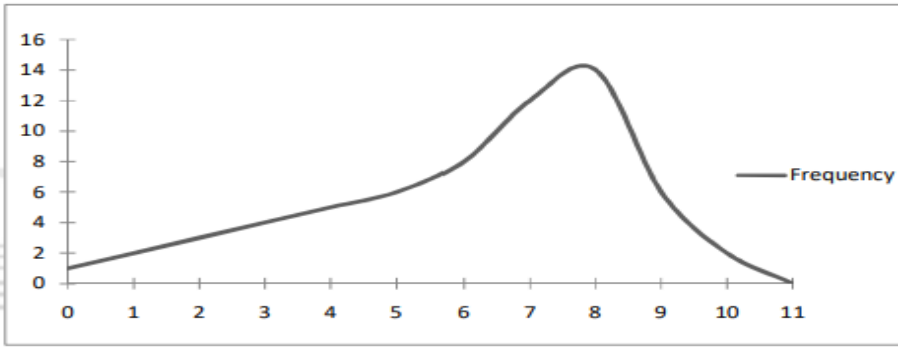


Fig. 4.2: Negative Skewed Curve

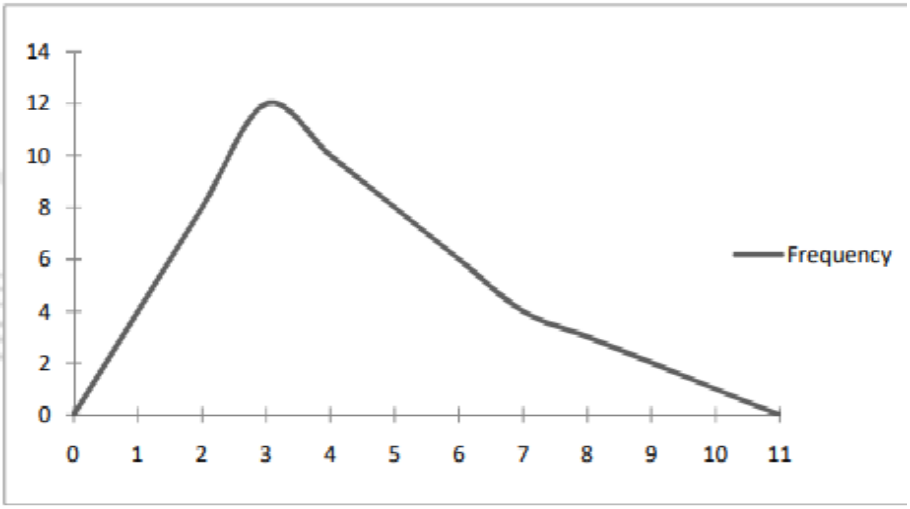


Fig. 4.3: Positive Skewed Curve

1. β and γ Coefficient of Skewness

Karl Pearson defined the following β and γ coefficients of skewness, based upon the second and third central moments:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

It is used as measure of skewness. For a symmetrical distribution, β_1 shall be zero. β_1 as a measure of skewness does not tell about the direction of skewness, i.e. positive or negative. Because μ_3 being the sum of cubes of the deviations from mean may be positive or negative but μ_3^2 is always positive. Also, μ_2 being the variance always positive. Hence, β_1 would be always positive. This drawback is removed if we calculate Karl Pearson's Gamma coefficient $\gamma_1$which is the square root of β_1 i. e.

$$\gamma_1 = \pm\sqrt{\beta_1} = \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{\mu_3}{\sigma^3}$$

Then the sign of skewness would depend upon the value of μ_3 whether it is positive or negative. It is advisable to use γ_1 as measure of skewness.

2. Karl Pearson's Coefficient of Skewness

This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

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

The value of this coefficient would be zero in a symmetrical distribution. If mean is greater than mode, coefficient of skewness would be positive otherwise negative. The value of the Karl Pearson's coefficient of skewness usually lies between ± 1 for moderately skewed distubution. If mode is not well defined, we use the formula

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

By using the relationship

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

Here, $-3 \le S_k \le 3$. In practice it is rarely obtained.

3. Bowleys's Coefficien of Skewness

This method is based on quartiles. The formula for calculating coefficient of skewness is given by

$$S_k = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$$

$$= \frac{(Q_3 - 2Q_2 + Q_1)}{(Q_3 - Q_1)}$$

The value of Sk would be zero if it is a symmetrical distribution. If the value is greater than zero, it is positively skewed and if the value is less than zero it is negatively skewed distribution. It will take value between +1 and -1.

2. **Kelly's Coefficient of Skewness**

The coefficient of skewness proposed by Kelly is based on percentiles and deciles. The formula for calculating the coefficient of skewness is given by Based on Percentiles

$$S_k = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{10})}$$

$$= \frac{(P_{90} - 2P_{50} + P_{10})}{(P_{90} - P_{10})}$$

where, P_{90}, P_{50} and P_{10} are 90^{th}, 50^{th} and 10^{th} Percentiles.

Based on Deciles

$$S_k = \frac{(D_9 - 2D_5 + D_1)}{D_9 - D_1}$$

where, D_9, D_5 and D_1 are 9^{th}, 5^{th} and 1^{st} Decile.

Remarks about Skewness

1. If the value of mean, median and mode are same in any distribution, then the skewness does not exist in that distribution. Larger the difference in these values, larger the skewness;

2. If sum of the frequencies are equal on the both sides of mode then skewness does not exist;

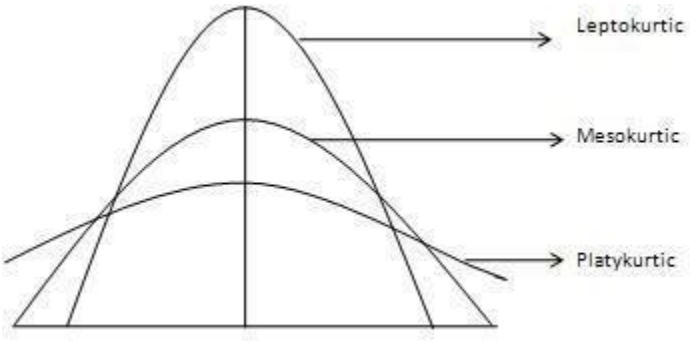
3. If the distance of first quartile and third quartile are same from the median then a skewness does not exist. Similarly if deciles (first and ninth) and percentiles (first and ninety nine) are at equal distance from the median. Then there is no asymmetry;

4. If the sums of positive and negative deviations obtained from mean, median or mode are equal then there is no asymmetry; and

5. If a graph of a data become a normal curve and when it is folded at middle and one part overlap fully on the other one then there is no asymmetry

Measure of kurtosis

Prof. Karl Pearson has called it the "Convexity of a Curve". Kurtosis gives a measure of flatness of distribution. The degree of kurtosis of a distribution is measured relative to that of a normal curve. The curves with greater peakedness than the normal curve are called "Leptokurtic". The curves which are more flat than the normal curve are called "Platykurtic". The normal curve is called "Mesokurtic."



Measure of kurtosis

For calculating the kurtosis, the second and fourth central moments of variable are used. For this, following formula given by Karl Pearson is used:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$