SARDAR PATEL UNIVERSITY
THIRD SEMESTER
(EFFECTIVE FROM JUNE, 2019)
SUBJECT: STATISTICS
COURSE CODE: US03CSTA21
(DESCRIPTIVE STATISTICS)

Course credit: 4
No. of lectures per week: 4
All units carry equal Weightage
Weightage: Internal – 30%, External – 70%
Objectives:

The main objective of this course is to acquaint students with some basic concepts in Statistics. They will be introduced to some elementary statistical methods of analysis of data. At the end of this course students are expected to be able to analyze the data.

1. **To tabulate statistical information given in descriptive form,**
2. **To use graphical techniques and interpret,**
3. **To compute various measures of central tendency, dispersion, skewness,**
4. **To analyze data pertaining to attributes and to interpret the results,**
5. **To apply statistics in the various fields.**

Unit - I Analysis of Quantitative data - I

- ➢ **Types of data**
  - ➢ **Quantitative data : Discrete and Continuous**
  - ➢ **Qualitative data : Nominal and Ordinal**
- ➢ **Measures of central tendency**
  - ➢ **Mean, Median, Mode**
  - ➢ **Geometric mean**
  - ➢ **Harmonic mean**
  - ➢ **Weighted mean**
  - ➢ **Combined mean**
  - ➢ **Merits & demerits**
  - ➢ **Properties (with proof)**
  - ➢ **Examples**

Unit - II Analysis of Quantitative data - II

- ➢ **Partition values and their graphical representation**
- ➢ **Measures of Dispersion : Range, Quartile derivation, Mean Derivation, Standard derivation**

- Coefficient of variation(C.V)
- Merits & Demerits
- Properties (with proof)
- Box – and – whisker plot
- Lorenz curve
- Stem – and – Leaf diagram
- Raw moments
- Central moments
- Relationship between raw and central moments
- Skewness
- Kurtosis
- Examples

Unit - III Index numbers

- Introduction
- Uses of index numbers
- Steps for construction of index numbers
- Problems in the construction of index numbers
- Methods of constructing index numbers
  - Simple (Unweighted) Aggregate method
  - Weighted Aggregate method
    - Laspeyre’s Price Index
    - Paasche’s Price Index
    - Fisher’s Price Index
    - Marshall Edgeworth Price Index
- Tests of consistency of Index number
  - Time reversal test
  - Factor reversal test

Unit - IV Vital Statistics

- Introduction
- Uses of Vital statistics and methods of collecting vital statistics
- Measurement of Mortality:
  - Crude Death Rate (CDR)
  - Specific Death Rate (SDR)
  - Standardized Death Rate (STDR)
- Measurement of Fertility:
  - Crude Birth Rate (CBR)
  - General Fertility Rate (GFR)
  - Specific Fertility Rate (SFR)

    - **Total Fertility Rate (TFR)**
- **Measurement of population growth**
  - **Methods of measuring population growth**
    - **Crude rate of natural increase**
    - **Vital index**
    - **Gross Reproduction Rate (GRR)**

References:

1. **Gupta S.C. : Fundamentals of Statistics**
2. **Gupta S.C. : Applied Statistics**
3. **Gupta S.C. and V.K.Kapoor : Fundamentals of Mathematical Statistics**
4. **Agarwal B.L. : Basic statistics**
4. **Ken Black : Business Statistics**
5. **Gupta S.C. : Fundamentals of Applied Statistics**

Unit – I

Analysis of Quantitative data

Meaning/Definition:

(i) Statistics is a science which deals with collection, presentation, analysis and interpretation of numerical data.

(ii) Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and at calculated risk.

Types of Data:

(a) Qualitative or Categorical Data:

When the characteristic under study concerns qualitative phenomena that is only classified in categories, the data are called categorical data. The qualitative phenomenon under study is called an attribute. For example, Literacy, Honesty, blood type, Sex, Nationality etc.

OR

The objects being studied are grouped into categories based on some qualitative trait. The resulting data are merely labels or categories.

(i) Nominal Data: A set of data is said to be nominal if the observations (data items) belonging to it can be classified into categories and it can be assigned a code in the form of a number when the numbers are simply labels. For example, in a data set males could be coded as 1, females as 0; marital status of an individuals could be coded as 1 if married, 0 if unmarried.

(ii) Ordinal Data: A set of data is said to be ordinal if the observations (data items) belonging to it can be classified into categories and it can be assigned a code in the form of a number when the numbers are also important OR A set of data is said to be ordinal if the observations (data items) belonging to it can be classified into categories that can be ranked (put in order or where order is also important) For example, you might ask patients to express the amount of pain they are feeling on a scale of 1 to 5. A score of 5 means severe pain and 1 means low pain. Another example would be movie ratings * to *****.

Examples on Nominal Data:

(1) Sex: 1 = Male, 0 = Female

(2) Smoking status:1 = Smoker, 0 = Non-smoker
(3) Symptoms of respiratory disease: 1 = Present, 0 = Absent
(4) Blood group: 1 = O group, 2 = A group, 3 = B group, 4 = AB group
(5) Disease of diabetes: 1 = Yes, 0 = No
(6) Digestibility of Iron: 1 = Yes, 0 = No
(7) Type of Birth: 1 = With complication, 0 = No complication
(8) Prominent wrinkles: 1 = Yes, 0 = No
(9) Birth defect: 1 = Present, 0 = Absent
(10) Citizen: 1 = Indian, 0 = Non - Indian
(11) Medium of Instruction: 1 = Gujarati, 0 = English
(12) Religion: 1 = Hindu, 0 = Non-Hindu

Examples on Ordinal Data:

(1) Test of juice: 1 = Very tasty, 2 = Tasty, 3 = Bad
(2) I.Q. of children: 1 = Above average, 2 = Average, 3 = below average
(3) "Cigarette should be ban in public places?": 1 = strongly agree, 2 = agree, 3 = Neutral, 4 = Disagree,
5 = strongly disagree.
(4) Drinking level: 0 = Non-drinker, 1 = Light to moderate drinker, 2 = heavy drinker
(5) Quality of High school: 1 = Superior, 2 = Average, 3 = Poor
(6) How often do you visit the zoo?
1 = Never 2 = Rarely 3 = Sometimes 4 = Often 5 = Always
(7) How do you feel about the Principal's performance this year?
1 = Strongly approve 2 = Somewhat 3 = Neutral/No opinion 4 = Somewhat disapprove 5 = Strongly disapprove

Remark: **Likert scale** is commonly used in survey research. It is often used to measures respondent's attitudes by asking they agree or disagree with a particular question or statement.

(b) Quantitative or Numerical data:

When the characteristic under study is measured on a numerical scale (quantitative phenomena), the resulting data consists of set of numbers. The quantitative phenomenon under study is called variable.

For example, Blood cholesterol level, Amount of weight loss, birth weight, Intensity of earthquake etc.

OR

The objects being studied are measured based on some quantitative trait. The resulting data are set of numbers.

(i) Numerical data classified as Discrete or Continuous data.

(a) Discrete data: Only certain values are possible or Numeric data that have a finite no. of possible values.

(b) Continuous data: Any value within an interval is possible or Continuous data have infinite possibilities.

Variable:

The word variable means something that can vary i.e. Change. A variable takes on different numerical values. OR

A quantity which can vary from one individual to another is called variable.

OR

A quantitative characteristic under study is called variable.

For example, age, height, weight of a person, birth weight, temperature, income, consumption of electricity etc.

There are two types of variable (i) Discrete and (ii) Continuous

(i) Discrete variable:

A variable which can take only an integer value in the specified range is called discrete variable. For example, No. of children, No. of accidents, Marks, Blood cholesterol level etc.

(ii) Continuous variable:

A variable which can take any (integer as well as real) value in the specified range is called continuous variable. For example, Percentage of marks, Body temperature, age, height, temperature etc.

Attribute:

A qualitative characteristic under study is called an attribute. For example, Sex, Religion, Blood group etc.

Classify the following as Variable/Attribute:

1 Nutritional value of crops
2 Drinking level
3 Blood group
4 I.Q. of Children
5 No. of diseased plants
6 Digestibility of Iron
7 Fatty acid in vegetable oil
8 Protein level in milk
9 State of seed after sowing
10 No. of fruit consumption per day
11 Product quality of biotech food
12 Packaging material for biotech food
13 Deficiency of vitamin
14 Birth weight
17 Blood pressure (mmHg)
18 Smoking status
19 Citizen
20 Medium of instruction
21 Pregnancy duration(in days)
22 Eye colour
23 Religion
24 Sex
25 % of Attendance
26 Economical condition
27 Sex ratio
28 Body mass Index (BMI)
29 Type of birth
30 Smoking should be ban in public places?

Measures of Central Tendency

To understand the concept of the above let us consider the following example:

A study is conducted to determine if dieting plus exercise is more effective in producing weight loss than dieting alone. Twelve pairs of matched subjects are run in the study. Subjects are matched on initial weight, initial level of exercise, age, and sex. One member of each pair is put one diet for 3 months. The other member receives the same diet but in addition is put on a moderate exercise regime. The following scores indicate the weight loss in pounds over the 3-monthe period for each subject:

Pair	1	2	3	4	5	6	7	8	9	10	11	12
Diet+ Exercise	24	20	22	15	23	21	16	17	19	25	24	13
Diet alone	16	18	19	16	18	18	17	19	13	18	19	14

(i) Identify the objective of the above problem.
(ii) Which statistical measure do you calculate? Why?

Objective: To compare two different methods of producing weight loss.

To achieve the said objective one such measure is to calculate an average (mean).

Averages are the measures which condense a huge set of numerical data into single numerical values which are representative of the entire data set (distribution). They give us an idea about the concentration of the values in the central part of the distribution. In brief, average of a statistical data is the value of variable which is representative of the entire data set (distribution).

Two series of observations are not comparable because of the unsystematic variations generally present in the series (sets of numbers) but constants make it possible to compare the series easily.

Averages are very much useful for

(i) Describing the distribution in concise manner.

(ii) Comparative studies of different distributions.

(iii) Computing various other statistical measures such as dispersion (variation), skewness (lack of symmetry), kurtosis etc.

The various measures of central tendency are

(i) Mean or Arithmetic mean (A.M)
(ii) Median
(iii) Mode
(iv) Geometric mean (G.M)
(v) Harmonic mean (H.M)

Requisites of a good (ideal) measure of central tendency:

There are various measures of central tendency. The difficulties lies in choosing the measures as no hard and fast rules have been made to select anyone. However, some norms have been set which work as a guideline for choosing a particular measure of central tendency.

A measure of central tendency is good or satisfactory if it possesses the following characteristics:

(1) It should be rigidly defined. It means that the definition should be clear and unambiguous so that it leads to one and only one interpretation by the different persons.
(2) It should be easy to calculate and understand.
(3) It should be based on all the observations.
(4) It should be least affected by extreme observations.
(5) It should be stable with regarding to sampling. It means that if a no. of samples of same size is drawn from a population, the measures of central tendency having the minimum variation among the different calculated values.

(1) <u>Mean or Arithmetic mean</u>:

Mean of a given set of observations is their sum divided by the number of observations. It is the most common and useful measure of central tendency.

For ungrouped (raw) data:

Let $Xi, i = 1,2 \dots n$ be the given n observations then their mean is denoted by $\bar{X}$ and is defined as

$$\bar{X} = \frac{Sum\ of\ all\ observations}{no.of observations} = \frac{1}{n}\sum_{i=1}^{n} Xi$$

For Grouped data:

For Simple (Discrete) frequency distribution:

Let $(Xi, fi), i = 1,2 \dots n$ be the given frequency distribution then their mean is denoted by $\bar{X}$ and is defined as

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} fiXi\ \ Where\ N = \sum_{i=1}^{n} fi$$

For grouped frequency distribution:

In case of grouped frequency distribution, $Xi's$ are the mid-values of respective classes.

Merits and Demerits of Mean

Merits:

(1) It is rigidly defined.

(2) It is easy to calculate and understand.

(3) It is based on all the observations.

(4) Of all the averages, mean is stable regarding sampling.

Demerits:

(1) It is very much affected by extreme observations.

(2) It cannot be used in case of open-end classes.

(3) It cannot be determined graphically.

(4) It may lead to wrong conclusions if the details of the data from which it is calculated are not available.

Deviation about any arbitrary value A:

If $Xi, i = 1, 2 \dots n$ be n observations and A is any arbitrary value. Then $Xi - A$, $i = 1, 2 \dots n$ is called deviation of ith observation about any value A.

Deviation about any mean:

If $Xi, i = 1, 2 \dots n$ be n observations and $\bar{X}$ is the mean then $Xi - \bar{X}, i = 1,2 \dots n$ is called deviation of ith observation about mean.

Properties of Mean:

(1) The algebraic sum of the deviations of the observations from their mean is always zero.

Mathematically,

$$\sum_{i=1}^{n}(Xi - \bar{X}) = 0\ or\ \sum_{i=1}^{n} fi(Xi - \bar{X}) = 0$$

Proof:

$$\sum_{i=1}^{n}(Xi - \bar{X}) = \sum_{i=1}^{n} Xi - \bar{X}\sum_{i=1}^{n} 1 = n\bar{X} - n\bar{X} = 0 \quad \because \bar{X} = \frac{1}{n}\sum_{i=1}^{n} Xi$$

OR

$$\sum_{i=1}^{n}(Xi - A) = 0\ if\ A = \bar{X}$$

Proof:

$$\sum_{i=1}^{n}(Xi - A) = \sum_{i=1}^{n} Xi - A\sum_{i=1}^{n} 1 = 0$$

$$\Rightarrow \sum_{i=1}^{n} Xi = nA$$

$$\Rightarrow A = \frac{1}{n}\sum_{i=1}^{n} Xi = \bar{X}$$

(2) The sum of the squares of deviations of the given set of observations is minimum when taken from mean.

Mathematically,

$$S = \sum_{i=1}^{n}(Xi - A)^2\ is\ minimum\ when\ A = \bar{X}$$

Or

For a frequency distribution,

$$S = \sum_{i=1}^{n} fi(Xi - A)^2\ is\ minimum\ when\ A = \bar{X}\ where\ \bar{X} = \frac{1}{N}\sum_{i=1}^{n} fiXi, N = \sum_{i=1}^{n} fi$$

$i.e$

$$\sum_{i=1}^{n}(Xi - \bar{X})^2\ or\ \sum_{i=1}^{n} fi(Xi - \bar{X})^2\ is\ minimum$$

Proof:

Here we apply the principle of maxima and minima from differential calculus. For S to be minimum, we should have

$$\frac{\partial S}{\partial A} = 0\ and\ \frac{\partial^2 S}{\partial A^2} > 0$$

We have

$$S = \sum_{i=1}^{n} fi(Xi - A)^2\ ----------(1)$$

Differentiating (1) w.r.to A and equating to zero, we get

$$\frac{\partial S}{\partial A} = \sum_{i=1}^{n} 2fi(Xi - A)(-1)$$

$$= -2\sum_{i=1}^{n} fi(Xi - A)\ ----------(2)$$

Now

$$\frac{\partial S}{\partial A} = 0 \Rightarrow -2\sum_{i=1}^{n} fi(Xi - A) = 0$$

$$\Rightarrow \sum_{i=1}^{n} fi(Xi - A) = 0$$

$$\Rightarrow \sum_{i=1}^{n} fiXi - A\sum_{i=1}^{n} fi = 0$$

$$\Rightarrow \sum_{i=1}^{n} fiXi - NA = 0 \because N = \sum_{i=1}^{n} fi$$

$$\Rightarrow A = \frac{1}{N}\sum_{i=1}^{n} fiXi = \bar{X}$$

Differentiating (2) w.r. to A, we get

$$\frac{\partial^2 S}{\partial A^2} = -2\sum_{i=1}^{n} fi(-1) = 2\sum_{i=1}^{n} fi = 2N > 0$$

Hence S is minimum at the point $A = \bar{X}$

(3) Mean depends on change of origin as well as scale.

Proof:

Let $Xi, i = 1, 2 \ldots n$ be n observations then their mean is denoted by $\bar{X}$ and is given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Xi$$

Let us define a new variable ui as $ui = \frac{Xi-A}{C}, i =$ $1,2 \ldots n$ *where A is new origin and C be the new scale*

From the above, we have

$Xi = A + Cui, i = 1,2 \ldots n$

Taking summation over i from 1 to n we get,

$$\sum_{i=1}^{n} Xi = \sum_{i=1}^{n} A + C\sum_{i=1}^{n} ui = nA + C\sum_{i=1}^{n} ui$$

Dividing both the sides by n, we get

$\bar{X} = A + C\bar{u}$

Which shows that mean depends on change of origin and scale.

(4) Combined Mean:

If $\bar{X}_1, \bar{X}_2, \ldots \bar{X}_k$ be the means of k groups (series) with $n_1, n_2 \ldots n_k$ no. of observations resp. then the mean of combined group (all the observations) with $n = n_1 + n_2 + \cdots + n_i + \cdots + n_k$ observations is given by

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots n_i\bar{X}_i + \cdots n_k\bar{X}_k}{n_1 + n_1 + \cdots + n_i + \cdots n_k}$$

Proof:
Let $\left(X_{i1}, X_{i2}, \dots X_{ij}, \dots X_{ik}\right), i = 1{,}2 \dots k, j = 1{,}2, \dots ni$ be the observations in k groups respectively.

Now

$$\bar{X}_i = \frac{1}{ni}\sum_{j=1}^{ni} X_{ij}, i = 1{,}2 \dots k$$

$$\therefore \sum_{j=1}^{ni} X_{ij} = n_i\bar{X}_i\,, i = 1{,}2 \dots k$$

Now

$$\bar{X} = Combined\ Mean = Mean\ of\ all\ (n)\ observations$$

$$= \frac{Sum\ of\ all\ (n) observations}{Total\ no.of\ observations}$$

$$= \frac{1}{n}\left[\sum_{j=1}^{n1} X_{1j} + \sum_{j=1}^{n2} X_{2j} + \cdots + \sum_{j=1}^{ni} X_{ij} + \cdots + \sum_{j=1}^{nk} X_{kj}\right]$$

or

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{ni} Xij = \frac{1}{n}\sum_{i=1}^{k} ni\overline{Xi} \quad \because \bar{X}_i = \frac{1}{ni}\sum_{j=1}^{ni} X_{ij}$$

Where $n = n_1 + n_2 + \cdots + n_i + \cdots + n_k$

$$\therefore \bar{X} = \frac{1}{n}\left[n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots + n_i\bar{X}_i + \cdots + n_k\bar{X}_k\right]$$

Hence the result.

In particular, if $\bar{X}_1\ and\ \bar{X}_2$ be the means of two groups with n_1, n_2 no. of observations respectively; then the mean $\bar{X}$ of combined group with $n_1 + n_2$ observations is given by

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

If $n_i = n\ \forall i = 1{,}2 \dots k$

i.e. no. of observations in each group is same then

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \cdots \bar{X}_i + \cdots \bar{X}_k}{k}$$

i.e. mean of combined group is mean of all means.

(5) <u>Weighted Mean:</u>

Let $X_1, X_2, \dots, X_n$ be n observations and $W_1, W_2, \dots, W_n$ be the corresponding weights then the weighted mean is given by

$$\bar{X}_w = \frac{W_1X_1 + W_2X_2 + \cdots + W_iX_i + \cdots W_nX_n}{W_1 + W_2 + \cdots + W_i + \cdots + W_n} = \frac{\sum W_iX_i}{\sum W_i}$$

If $W_i = W \; \forall i = 1,2 \ldots n$ then

If $W_i = W \; \forall \; i = 1, 2 \ldots \; n$ then

$$\bar{X}_w = \frac{W \sum X_i}{W \sum 1} = \frac{1}{n} \sum Xi = \bar{X}$$

i.e. when each observations has equal Weightage then weighted mean is same as mean.

(ii) Median:

Median is that value of the variable which divides the data (set of observations) into two equal parts so that the no. of observations below median and above median is equal. Thus, we see that against mean which is based on all the observations the median is the only positional average i.e. its value depends on the middle position (term).

For ungrouped (raw) data:

Let $Xi, i = 1,2 \ldots n$ be the given n observations.

Steps:

(1) Arrange the data either in ascending or descending order.

(2) Median is the middle term or mean of two middle terms according as the no. of observations is odd or even.

$$Median = \begin{cases} Value\ of\ \left(\frac{n+1}{2}\right)^{th}\ observations, if\ n\ is\ odd \\ Mean\ of\ \left(\frac{n}{2}\right)^{th}\ and\ \left(\frac{n}{2}+1\right)^{th}\ observations, if\ n\ is\ even \end{cases}$$

For Grouped data:

For simple frequency distribution:

Let $(Xi, fi), i = 1,2 \ldots n$ be the given frequency distribution

Steps:

(1) Calculate the cumulative frequency of less than type.

(2) Calculate $\left(\frac{N+1}{2}\right)$

(3) Select the cumulative frequency just greater than (or equal to) $\left(\frac{N+1}{2}\right)$

(4) The value of the variable corresponding to selected cumulative frequency is median.

For grouped frequency distribution:

Let $(Xi - Xi + 1, fi), i = 1,2 \ldots n$ be the given grouped frequency distribution.

Steps:

(1) Calculate the cumulative frequency of less than type.

(2) Calculate $\left(\frac{N}{2}\right)$

(3) Select the cumulative frequency just greater than (or equal to) $\left(\frac{N}{2}\right)$

(4) The class corresponding to selected cumulative frequency is called median class and median is calculated by the following formula

$$Median = l + \left(\frac{\frac{N}{2} - F_{<}}{f}\right) \times c$$

Where l = lower limit of a median class

$F_{<}$ = cumulative frequency of the class previous to median class

f = frequency of a median class

c = class-width of a median class

Remark: classes must be continuous.

Merits and Demerits of Median:

Merits:

(1) It is rigidly defined.

(2) It is easy to calculate and understand.

(3) It is not affected by extreme observations and hence it is very much useful in case of open-end classes.

(4) It can be determined graphically.

Demerits:

(1) It is not based on all the observations.

Remark:

The sum of the absolute deviations of a given set of observations is minimum when taken from median.

(iii) Mode:

Mode is the value of variable which occurs most frequently (maximum no. of times) in the given data (set of observations). Mode is a measure which representing the common or typical value of the data.

Uses:

(i) Average size of shoe sold in a shop is 8.

(ii) Average size of shirt sold in a readymade garment shop is 90 (XL).

(iii) Average student in a hostel spends Rs. 1500 per month.

In all the above cases, the average referred to as mode.

For ungrouped (raw) data:

Let $Xi, i = 1,2 \dots n$ be the given n observations.

From the given data select that value which occur maximum no. of times (most often).

For simple frequency distribution:

Let $(Xi, fi), i = 1,2 \dots n$ be the given frequency distribution

Steps:

(1) Select the maximum frequency.

(2) The value of the variable corresponding to selected frequency is mode.

For grouped frequency distribution:

Let $(Xi - Xi + 1, fi), i = 1,2 \dots n$ be the given grouped frequency distribution.

Steps:

(1) Select the maximum frequency.

(2) The class corresponding to selected frequency is called the modal class.

(3) Mode is determined by the following formula

$$Mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times c$$

Where

l = lower limit of modal class.

f_1 = frequency of modal class.

f_0 = frequency previous to modal class.

f_2 = frequency next to modal class.

c = class-width of modal class.

Remark: classes must be continuous.

Merits and Demerits of Mode:

Merits:

(1) It is easy to calculate.

(2) It can be determined graphically.

(3) It is not at all affected by extreme observations.

Demerits:

(1) Mode is not rigidly defined. It is ill-defined iff

(a) Maximum frequency is repeated.

(b) Maximum frequency occurs either in the very beginning or at the end.

(c) The given distribution is irregular.

(2) It is not based on all the observations.

(iv) Geometric Mean (G.M):

Geometric mean of a set of observations is the n^{th} root of their product.

For ungrouped (raw) data:

Let $Xi, i = 1, 2 \ldots n$ be the given n observations. Then their Geometric mean is defined as

$$G.M = \left(\prod_{i=1}^{n} Xi\right)^{\frac{1}{n}} = nth\ root\ of\ their\ products.$$

In particular, if $n = 2$ (i.e. with two observations X_1 and X_2 then geometric mean can be computed by taking the square root of their product.

If $n > 2$, the no. of observations is greater than 2, then computation of n th root is very tedious. In such case the calculations are facilitated by making the use of logarithms.

Taking the logarithm on both sides, we get

$$\log(G.M) = \frac{1}{n} log\left(\prod_{i=1}^{n} Xi\right)$$

$$\therefore G.M = Antilog\left(\frac{1}{n}\sum_{i=1}^{n} log(Xi)\right)$$

Thus we see that logarithm of G.M is the mean of their logarithms.

For Grouped data:

For simple frequency distribution:

Let $(Xi, fi), i = 1, 2 \ldots\ n$ be the given frequency distribution. Then the Geometric mean is given by

$$\therefore G.M = Antilog\left(\frac{1}{N}\sum_{i=1}^{n} filog(Xi)\right), where\ N = \sum_{i=1}^{n} fi$$

For grouped frequency distribution:

In case of grouped frequency distribution, Xi's are the mid-values of respective classes.

Remark: If one of the numbers (observation) is zero, G.M is zero.

(v) Harmonic Mean (H.M):

Harmonic mean is the reciprocal of the mean of the reciprocals of the given observations.

For ungrouped (raw) data:

Let $Xi, i = 1, 2 \ldots n$ be the given n observations. Then their Harmonic mean is denoted by

$$H.M = \frac{1}{\frac{1}{n}\sum\frac{1}{Xi}} = reciprocal\ of\ the\ mean\ of\ their\ reciprocals.$$

i.e. H.M of n observations is reciprocal of the mean of their reciprocals.

For Grouped data:

For simple frequency distribution:

Let $(Xi, fi), i = 1, 2 \ldots\ n$ be the given frequency distribution. Then the Harmonic mean is given by

$$H.M = \frac{1}{\frac{1}{N}\sum\frac{fi}{Xi}} = \frac{N}{\sum\frac{fi}{Xi}}\ , where\ N = \sum fi$$

For grouped frequency distribution:

In case of grouped frequency distribution, Xi's are the mid-values of respective classes.

Remark: H.M cannot be calculated if one of the numbers (observation) is zero.

Relationship between A.M, G.M and H.M

$$\boldsymbol{A.M \geq G.M \geq H.M}$$

The sign of equality holds if and only if all the n numbers (observations) are equal.

Proof:

We shall establish the result for two numbers only, although the result holds true for n observations.

Let a and b be two real positive numbers i.e. $a > 0, b > 0$ then

$$A.M = \frac{a+b}{2}$$

$$G.M = \sqrt{ab}$$

$$H.M = \frac{2}{\frac{1}{a}+\frac{1}{b}} = \frac{2ab}{a+b}$$

Consider $A.M - G.M = \frac{a+b}{2} - \sqrt{ab} = \frac{a+b-2\sqrt{ab}}{2} = \frac{\left(\sqrt{a}-\sqrt{b}\right)^2}{2} \geq 0$

$\therefore A.M \geq G.M - - - (1)$

The sign of equality holds only if $\sqrt{a} - \sqrt{b} = 0$

$$\Rightarrow \sqrt{a} = \sqrt{b}$$

$$\Rightarrow a = b$$

i.e. if and only if the two numbers are equal.

Also consider $G.M - H.M = \sqrt{ab} - \frac{2ab}{a+b}$

$$= \sqrt{ab} - \frac{2\sqrt{ab}\sqrt{ab}}{a+b}$$
$$= \sqrt{ab}\left(1 - \frac{2\sqrt{ab}}{a+b}\right) = \frac{\sqrt{ab}}{a+b}\left(a + b - 2\sqrt{ab}\right)$$
$$= \frac{\sqrt{ab}}{a+b}\left(\sqrt{a} - \sqrt{b}\right)^2 \geq 0$$

$\therefore G.M - H.M \geq 0 - - - (2)$

The sign of equality holds only if $\sqrt{a} - \sqrt{b} = 0$

$$\Rightarrow \sqrt{a} = \sqrt{b}$$
$$\Rightarrow a = b$$

i.e. if and only if the two numbers are equal.

From (1) and (2)

$A.M \geq G.M \geq H.M$

The sign of equality holds if and only if the two numbers (observations) are equal.

Remark: (i) For two numbers $G^2 = AH$ Where A, G, H are A.M, G.M, H.M respectively.

Proof: Let $a > 0$ and $b > 0$ are two positive numbers. Then

$$A \times H = \frac{a+b}{2} \times \frac{2ab}{a+b} = ab = G^2$$

For more than two observations, the result $G^2 = AH$ holds only if the numbers (observations) are in G.P

Quantiles (Partition values):

Quantiles are the values which divide the entire data (set of numbers or observations) into some number of equal parts. The number of parts may be two, four, eight, ten or hundred.

Quartiles:

Quartiles are the values which divide the entire data (set of numbers or observations) into four equal parts. They are 3 in numbers namely $Q1, Q2, Q3$.

The ith quartile Qi is the value of X (variable) corresponding to the cumulative frequency just greater than (or equal to) $\frac{i\times N}{4}, i = 1,2,3$.

For continuous frequency distribution, the class corresponding to the cumulative frequency just greater than (or equal to) $\frac{i\times N}{4}$ is called ith quartile class and is given by

$$Qi = l + \frac{\frac{iN}{4} - F_<}{f} \times C, i = 1,2,3$$

Octiles:

Octiles are the values which divide the entire data (set of numbers or observations) into eight equal parts. They are 7 in numbers namely $O1, O2, \ldots O7$.

The jth octile Oj is the value of X (variable) corresponding to the cumulative frequency just greater than (or equal to), $\frac{j\times N}{8}, i = 1,2, \ldots 7$

For continuous frequency distribution, the class corresponding to the cumulative frequency just greater than (or equal to) $\frac{j\times N}{8}$ is called jth octile class and is given by

$$Oj = l + \frac{\frac{jN}{8} - F_<}{f} \times C, j = 1,2, \ldots 7$$

Deciles:

Deciles are the values which divide the entire data (set of numbers or observations) into ten equal parts. They are 9 in numbers namely $D1, D2, \dots, D9$.

The kth decile Dk is the value of X (variable) corresponding to the cumulative frequency just greater than (or equal to) $\frac{k\times N}{10}, k = 1,2, \dots 9$

For continuous frequency distribution, the class corresponding to the cumulative frequency just greater than (or equal to) $\frac{k\times N}{10}$ is called kth decile class and is given by

$$Dk = l + \frac{\frac{kN}{10} - F_{<}}{f} \times C, k = 1,2, \dots 9$$

Percentiles:

Percentiles are the values which divide the entire data (set of numbers or observations) into hundred equal parts. They are 99 in numbers namely P1, P2,…P99.

The mth percentile Pm is the value of X (variable) corresponding to the cumulative frequency just greater than (or equal to) $\frac{m\times N}{100}, m = 1,2, \dots 99$

For continuous frequency distribution, the class corresponding to the cumulative frequency just greater than (or equal to) $\frac{m\times N}{100}$ is called $mth\ o$ctile class and is given by

$$Pm = l + \frac{\frac{mN}{100} - F_{<}}{f} \times C, m = 1,2, \dots 99$$

WHAT HAVE WE DISCUSSED?

- **Average is a number that represents or shows the central tendency of a set of observations.**
- **Mean is one of the representative values of the data.**
- **Median is also a form of representative value. It refers to the value which lies in the middle of the data with half of the observations below it and other half above it.**
- **Mode is another form of central tendency or representative value. The mode of a set of observations is the observation that occurs most often (very often).**

Unit - II

Measures of Dispersion, Skewness & Kurtosis

DISPERSION (VARIATION):

Averages or measures of central tendency gives only the value around which the other observations concentrated or clustered. If we are given only the average of a series of observations, we cannot form complete idea about the distribution since there may exist a no. of distributions (sets of observations) may have same measure of central tendency (averages) but may differ widely from each other. To clear this point, consider the **% of marks** of 3 students in 5 examinations as follows:

Student	Examination					Average
	1	2	s3	4	5	
A	60	60	60	60	60	60
B	40	50	60	50	40	60
C	20	40	60	80	100	60

i.e. in the above example, all the three students have same averages i.e. 60 but the actual observations are different. Hence averages are not adequate measure to describe the data (distribution) completely. Thus the measure of central tendency must be supported by some other measure. One such measure is dispersion.

Dispersion (Variation) means scatteredness. We study dispersion to have an idea of the homogeneity (compactness) or heterogeneity (scatter) of the distribution.

The commonly used measures of dispersion are:

(i) Range

(ii) Quartile deviation ($Q.D$)

(iii) Mean deviation ($M.D$)

(iv) Standard deviation ($S.D$)

(i) Range:

Range is the simplest measure of dispersion. It is defined as the difference between the two extreme observations.

$$Range = L - S$$
$$= Xmax - Xmin$$

Where $Xmax$ is the largest and $Xmin$ is the smallest observations.

(ii) Quartile Deviation ($Q.D$):

Quartile deviation is based on upper quartile Q_3 and lower quartile Q_1 .It is defined as

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

(iii) Mean Deviation (M.D):

Let $Xi, i = 1, 2 \dots n$ be the given n observations then mean deviation about any arbitrary value A is defined as

$$M.D(about\ A) = \frac{1}{n}\sum_{i=1}^{n}|Xi - A|$$

where $A =$ mean or median or mode

For simple frequency distribution:

Let $(Xi, fi), i = 1, 2 \dots\ n$ be the given frequency distribution then the M.D about any arbitrary value A is defined as

$$M.D(about\ A) = \frac{1}{N}\sum_{i=1}^{n} fi|Xi - A|$$

For grouped frequency distribution:

In case of grouped frequency distribution, $Xi's$ are the mid-values of respective classes.

(iv) Standard Deviation (S.D):

Standard deviation is the superior measure of dispersion compared to any other measure of dispersion. It is denoted by S or Sx or σ and is defined as the positive square root of the mean of the squares of deviations of the given observations from their mean.

For ungrouped (raw) data:

Let $Xi, i = 1{,}2 \dots n$ be the given n observations.

$$S.D = S\ or\ \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Xi - \bar{X})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Xi^2 - (\bar{X})^2}$$

For Simple frequency distribution:

Let $(Xi, fi), i = 1{,}2 \dots n$ be the given frequency distribution.

$$S.D = S\ or\ \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{n} fiXi^2 - (\bar{X})^2}$$

$$where\ N = \sum_{i=1}^{n} fi$$

For Grouped frequency distribution:

In case of grouped frequency distribution, Xi's are the mid-values of the respective classes.

Variance:

A square of standard deviation is called variance.

Properties:

(1) Standard deviation (Variance) is independent of change of origin but depend on scale.

Proof: Let $(Xi, fi), i\ = 1, 2 \dots, n$ be the given frequency distribution.

$$S.D = S\ or\ \sigma\ or\ S_x = \sqrt{\frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^2}$$

$$where\ N = \sum_{i=1}^{n} fi\ and\ \bar{X} = \frac{1}{N}\sum_{i=1}^{n} fiXi$$

Let us define a new variable ui as

$ui = \frac{Xi-A}{C}, i = 1{,}2 \dots n\ \ where\ A\ is\ new\ origin\ and\ C\ be\ the\ new\ scale$

From the above, we have

$Xi = A + Cui, i = 1,2 \dots n$

$\therefore \bar{X} = A + C\bar{u}$
$\therefore Xi - \bar{X} = C(ui - \bar{u}), i = 1,2 \dots n$
Squaring both the sides and multiplying by fi, we get

$$\sum_{i=1}^{n} fi(Xi - \bar{X})^2 = \sum_{i=1}^{n} C^2 fi(ui - \bar{u})^2 = C^2 \sum_{i=1}^{n} fi\,(ui - \bar{u})^2 N = \sum_{i=1}^{n} fi$$

Dividing both the sides by $N = \sum fi$ and taking square root, we get

$$S \; or \; \sigma \; or \; S_x = \sqrt{\frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^2} = C\sqrt{\frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^2} = CS_u$$

Which shows that standard deviation (variance) independent of change of origin A but depend on scale C.

(2) Relationship between variance and mean square deviation.

OR

In usual notation, prove that $S^2 \geq \sigma^2$
Where

$$S^2 = Mean\ square\ deviation = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - A)^2 , N = \sum_{i=1}^{n} fi$$

$$\sigma^2 = Variance = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^2$$

Proof:
Let $(Xi, fi), i = 1, 2 \dots, n$ be the given frequency distribution.
Now

$$S^2 = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - A)^2$$

$$= \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X} + \bar{X} - A)^2$$

$$= \frac{1}{N}\sum_{i=1}^{n} fi[(Xi - \bar{X})^2 + (\bar{X} - A)^2 + 2(Xi - \bar{X})(\bar{X} - A)]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{n} fi(Xi - \bar{X})^2 + (\bar{X} - A)^2 \sum_{i=1}^{n} fi + 2(\bar{X} - A)\sum_{i=1}^{n} fi(Xi - \bar{X})\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^2 + (\bar{X} - A)^2 \quad \because \sum_{i=1}^{n} fi(Xi - \bar{X}) = 0$$

$\therefore S^2 = \sigma^2 + A\ non\ negative\ quantity$
$\therefore S^2 \geq \sigma^2 - - - (I)$

In other words, mean square deviation is greater than (or equal) the variance.
The sign of equality will holds in (I) iff
$(\bar{X}-A)^2=0 \Longrightarrow \bar{X}-A=0 \Longrightarrow \bar{X}=A$

(3) Standard deviation is superior than other measure of dispersion.

Standard deviation is most important and widely used measures of dispersion. It is rigidly defined and based on all the observations. The squaring the deviations $(Xi-\overline{X})^2$ removes the drawback of ignoring the signs of deviations in computing the mean deviation. Moreover, of all measures of dispersion, standard deviation is stable regarding sampling. Thus, we see that, standard deviation satisfies all most all properties laid down for an ideal measure of dispersion.

Coefficient of variation (C.V):

Coefficient of variation $(C.V)$ is the relative measure of dispersion. It is defined as

$$C.V=\frac{S.D}{Mean}\times 100$$

For comparing variability of two (or more) distributions, we compute *C.V* for each distribution. A distribution with smaller *C.V.* is said to be more homogeneous or uniform or less variable (steady) or consistent than the other and the distribution with greater *C.V* is said to be more heterogeneous or more variable than the other.

Combined Variance:

Let $\left(X_{i1},X_{i2},\dots X_{ij},\dots X_{ik}\right), i=1,2\dots k, j=1,2,\dots ni$ be the observations in k groups respectively.

Now

$$\bar{X}_i = Mean\ of\ ith\ group = \frac{1}{ni}\sum_{j=1}^{ni} X_{ij}, i=1,2\dots k$$

$$\therefore \sum_{j=1}^{ni} X_{ij} = n_i\bar{X}_i\,, i=1,2\dots k$$

$$Si^2 = Variance\ of\ ith\ group = \frac{1}{ni}\sum_{j=1}^{ni}(Xij-\overline{Xi})^2$$

Now

$$S^2 = Combined\ variance = Variance\ of\ all\ (n) observations$$

$$=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{ni}(Xij-\bar{X})^2$$

Where

$$\bar{X} = Combined\ mean = Mean\ of\ all\ (n) observations$$

$$n=\sum_{i=1}^{k} ni = n_1+n_2+\cdots+n_i+\cdots+n_k$$

$$\therefore\ S^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{ni}(Xij - \overline{Xi} + \overline{Xi} - \bar{X})^2$$

$$= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{ni}(Xij - \overline{Xi})^2 + \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{ni}(\overline{Xi} - \bar{X})^2 + Product\ term\ be\ zero$$

$$= \frac{1}{n}\sum_{i=1}^{k} niSi^2 + \frac{1}{n}\sum_{i=1}^{k} ni(\overline{Xi} - \bar{X})^2$$

$$= \frac{1}{n}\sum_{i=1}^{k}[niSi^2 + ni(Xi - \bar{X})^2]$$

Let $di = \overline{Xi} - \bar{X}, i = 1,2 \ldots k$

$$\therefore\ S^2 = \frac{1}{n}\sum_{i=1}^{k}[niSi^2 + nidi^2]$$

Where

$$n = n_1 + n_2 + \cdots + n_i + \cdots + n_k = \sum_{i=1}^{k} ni$$

In particular , for $k\ = 2$

$$S^2 = \frac{1}{n}\sum_{i=1}^{2}[niSi^2 + nidi^2]$$

$$= \frac{1}{n}\left[n_1{S_1}^2 + n_1{d_1}^2 + n_2{S_2}^2 + n_2{d_2}^2\right]$$

$$= \frac{1}{n}\left[n_1\left({S_1}^2 + {d_1}^2\right) + n_2\left({S_2}^2 + {d_2}^2\right)\right]$$

Statement:

If $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_i, \ldots \bar{X}_k$ be the means and ${S_1}^2, {S_2}^2, \ldots, {S_i}^2, \ldots {S_k}^2$ be the variances of k groups with $n1, n2, \ldots, ni, \ldots, nk$ no.of observations respectively; then the variance S^2 of combined group (all the observations) with $n1 + n2 + \cdots + ni + \cdots + nk$ observations is given by

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}[niSi^2 + nidi^2]$$

Box - and - Whisker Plot:

Box plots graphically display the variation in the given data. Box plots are particularly effective for displaying sets of data alongside each other for the purpose of visual comparisons.

The five-number summary consists of the Median (Q_2), the quartiles $(Q_1\ and\ Q_3)$ and the smallest and largest values in the data set (distribution). Immediate visuals of a Box-and-Whisker plot are the centre, the spread, and the overall range of the distribution.

An outlier is any data point that is more than 1.5 times IQR (IQR – Inter Quartile Range = $Q_3 - Q_1$) from either end of the box. To find an outlier, calculate
$Lower\ Limit\ (LL) = Q_1 - 1.5(IQR)$
$Upper\ Limit\ (UL) = Q_3 + 1.5(IQR)$
Any values less than LL and above UL are called outliers.

Moments:

Raw Moments (Moments about any arbitrary value A):

Let $(Xi, fi), i = 1,2 \ldots n$ be the given frequency distribution and A be any arbitrary value then rth raw moment (moment about any value A) is denoted by $m_r{}'$ and is defined as

$$m_r{}' = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - A)^r, r = 0,1, \ldots \qquad N = \sum_{i=1}^{n} fi$$

In particular,
For

$$r = 0, m_0{}' = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - A)^0 = \frac{1}{N}\sum_{i=1}^{n} fi = 1$$

$$r = 1, m_1{}' = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - A)^1 = \frac{1}{N}\left(\sum_{i=1}^{n} fiXi - A\sum_{i=1}^{n} fi\right) = \frac{1}{N}\sum_{i=1}^{n} fiXi - A = \bar{X} - A$$

$\therefore \bar{X} = A + m_1{}'$

If $A = 0$ then

$$m_r{}' = \frac{1}{N}\sum_{i=1}^{n} fiXi^r, r = 0,1, \ldots$$

Central Moments or Moments about Mean:

Let $(Xi, fi), i = 1, 2 \ldots, n$ be the given frequency distribution and $\bar{X}$ be the mean then rth central moment (moment about mean) is denoted by m_r and is defined as

$$m_r = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^r, r = 0,1, \ldots \qquad N = \sum_{i=1}^{n} fi$$

In particular,
For

$$r = 0, m_0 = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^0 = \frac{1}{N}\sum_{i=1}^{n} fi = 1$$

$$r = 1, m_1 = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^1 = \frac{1}{N}\sum_{i=1}^{n} fiXi - \bar{X}\frac{1}{N}\sum_{i=1}^{n} fi = \bar{X} - \bar{X} = 0$$

$$r = 2, m_2 = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^2 = Variance\ and\ so\ on.$$

Remark: If $A = \bar{X}\ then\ m_r = m_r{}'$

Central moments in terms of raw moments

OR

Express central moments in terms of raw moments

The rth central moment is denoted by m_r and is defined as

$$m_r = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^r, r = 0,1, \ldots \qquad N = \sum_{i=1}^{n} fi$$

$$= \frac{1}{N}\sum_{i=1}^{n} fi(Xi - A + A - \bar{X})^r$$

We know that

$\bar{X} = A + m_1{}'$

$\therefore A = \bar{X} - m_1{}'$

$$\therefore m_r = \frac{1}{N}\sum_{i=1}^{n} fi\big(Xi - A + (-m_1{}')\big)^r$$

We know that

$$(a+b)^n = \sum_{j=0}^{n}\binom{n}{j} a^{n-j} b^j \quad (Using\ Binomial\ expansion)$$

$$\therefore m_r = \frac{1}{N}\sum_{i=1}^{n} fi\left[\sum_{j=0}^{r}\binom{r}{j}(Xi - A)^{r-j}(-m_1{}')^j\right]$$

$$= \frac{1}{N}\left[\sum_{j=0}^{r}\binom{r}{j}(-m_1{}')^j\left\{\sum_{i=1}^{n} fi(Xi - A)^{r-j}\right\}\right]$$

$$= \sum_{j=0}^{r}\binom{r}{j}(-m_1{}')^j\left\{\frac{1}{N}\sum_{i=1}^{n} fi(Xi - A)^{r-j}\right\}$$

$$= \sum_{j=0}^{r}\binom{r}{j}(-m_1{}')^j m_{r-j}{}'$$

$provided\ r > j$, r=1, 2…, $r = 0,1, \ldots$

$$\because m_r{}' = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - A)^r$$

In particular,

For

$$r = 1, m_1 = \sum_{j=0}^{1}\binom{1}{j}(-m_1{}')^j m_{1-j}{}' = \binom{1}{0}(-m_1{}')^0 m_1{}' + \binom{1}{1}(-m_1{}')^1 m_0{}'$$

$$= m_1{}' - m_1{}' = 0$$

$\because m_0{}' = 1$

$$r = 2, m_2 = \sum_{j=0}^{2} \binom{2}{j} (-m_1{}')^j m_{2-j}{}' = m_2{}' - (m_1{}')^2$$

$$r = 3,, m_3 = \sum_{j=0}^{3} \binom{3}{j} (-m_1{}')^j m_{3-j}{}' = m_3{}' - 3m_2{}'(m_1{}') + 2(m_1{}')^3$$

$$r = 4,, m_4 = \sum_{j=0}^{4} \binom{4}{j} (-m_1{}')^j m_{4-j}{}' = m_4{}' - 4m_3{}'(m_1{}') + 6m_2{}'(m_1{}')^2 - 3(m_1{}')^4$$

Raw moments in terms of central moments
OR
Express raw moments in terms of central moments

The rth raw moment is denoted by $m_r{}'$ and is defined as

$$m_r{}' = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - A)^r, r = 0,1, \dots \qquad N = \sum_{i=1}^{n} fi$$

$$= \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X} + \bar{X} - A)^r$$

We know that

$\bar{X} = A + m_1{}'$

$\therefore \bar{X} - A = m_1{}'$

$$\therefore m_r{}' = \frac{1}{N}\sum_{i=1}^{n} fi\big(Xi - \bar{X} + (m_1{}')\big)^r$$

We know that

$$(a + b)^n = \sum_{j=0}^{n} \binom{n}{j} a^{n-j} b^j \quad (Using\ Binomial\ expansion)$$

$$\therefore m_r{}' = \frac{1}{N}\sum_{i=1}^{n} fi \left[\sum_{j=0}^{r} \binom{r}{j} (Xi - \bar{X})^{r-j} (m_1{}')^j\right]^r$$

$$= \frac{1}{N}\left[\sum_{j=0}^{r} \binom{r}{j} (m_1{}')^j \left\{\sum_{i=1}^{n} fi\, (Xi - \bar{X})^{r-j}\right\}\right]$$

$$= \sum_{j=0}^{r} \binom{r}{j} (m_1{}')^j \left\{\frac{1}{N}\sum_{i=1}^{n} fi\, (Xi - \bar{X})^{r-j}\right\}$$

$$\therefore m_r{}' = \sum_{j=0}^{r} \binom{r}{j} m_{r-j} (m_1{}')^j$$

$$provided\ r > j, r = 2, 3 \dots \qquad \because m_r = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^r$$

In particular,
For

$$r = 2, m_2{}' = \sum_{j=0}^{2} \binom{2}{j} m_{2-j}(m_1{}')^j = \binom{2}{0} m_2(m_1{}')^0 + \binom{2}{1} m_1(m_1{}')^1 + \binom{2}{2} m_0(m_1{}')^2$$
$$= m_2 + (m_1{}')^2$$
$$r = 3, m_3{}' = \sum_{j=0}^{3} \binom{3}{j} m_{3-j}(m_1{}')^j = m_3 + 3m_2m_1{}' + (m_1{}')^3$$
$$r = 4, m_4{}' = \sum_{j=0}^{4} \binom{4}{j} m_{4-j}(m_1{}')^j = m_4 + 4m_3m_1{}' + 6m_2(m_1{}')^3 + (m_1{}')^4$$

Effect of change of origin and scale on central moments:

OR

Moments (raw or central) depends upon change of scale but independent of change of origin.

Let $(Xi, fi), I = 1,2,\dots,n$ be the given frequency distribution. Then rth central moment is defined as

$$m_r = \frac{1}{N}\sum_{i=1}^{n} fi(Xi - \bar{X})^r - - - (I)$$

Let us define a new variable ui as $= \frac{Xi-A}{C}, i = 1,2 \dots n\ where\ A\ is\ new\ origin\ and\ C\ be\ the\ new\ \ scale.$

From the above, we have

$Xi = A + Cui, i = 1,2 \dots n$ and

$\bar{X} = A + C\bar{u}$

$\therefore Xi - \bar{X} = C(ui - \bar{u}), i = 1,2 \dots n$

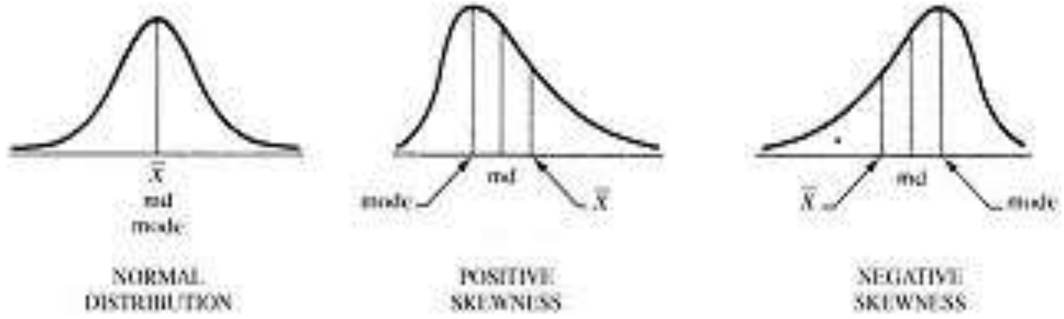
$$\therefore m_{r(x)} = \frac{1}{N}\sum_{i=1}^{n} fi\{C(ui - \bar{u})\}^r = C^r \frac{1}{N}\sum_{i=1}^{n} fi(ui - \bar{u})^2 = C^r\, m_{r(u)}$$

i.e. moments about mean (central moments) are independent of change of origin but depend on scale.

- **SKEWNESS:**

A frequency distribution is said to be skewed if it is not symmetric. The literal meaning of skewness is lack of symmetry. A frequency distribution is said to be positively (negatively) skewed if it has longer tail towards right (left). The degree of skewness is measured by coefficient.

Dispersion studies the degree of variation in the given distribution while skewness attempts at studying the direction of variation.



- **Karl-Pearson's coefficients of skewness:**

We know that for symmetrical distribution, mean, median, and mode coincide. Hence, if they are at different places then the distribution is skewed. By keeping this point in mind, Karl-Pearson gave the coefficient for the measurement of skewness as

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

$$= \frac{3(Mean - Median)}{S.d}, \quad \text{if mode is ill defined}$$

If $S_K = 0$ then frequency distribution is not skewed (symmetric).
If $S_K < 0$ then frequency distribution is negatively skewed.
If $S_K > 0$ then frequency distribution is positively skewed.

In some situations mode is not defined or difficult to find, the Karl-Pearson's coeff. of skewness can be defined by using empirical relation as

- **Bowley's coefficient of skewness:**

For a symmetrical distribution, the two quartiles namely Q_1 and Q_3 are equidistance from the median i.e.Q_2. The coefficient of skewness based on quartile is defined as

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

Remark:
(1) Bowley's coefficient of skewness ranges from -1 to 1.
(2) Bowley's coefficient of skewness is useful in the following situations:
(a) When mode is ill defined
(b) When the distribution has open-end classes.

- **Moments Coefficient of Skewness:**

Moments coefficient of skewness is denoted by β_1 and is defined as

$$\beta_1 = \frac{{m_3}^2}{{m_2}^3}$$

Where m_2 and m_3 are second and third central moments. The gamma coefficient of skewness is defined as

$$\gamma_1 = \sqrt{\beta_1} = \frac{m_3}{{m_2}^{\frac{3}{2}}}$$

If $\gamma_1 < 0$ $then\ the\ frequency\ distribution\ is\ negatively\ skewed.$
If $\gamma_1 = 0$ $then\ the\ frequency\ distribution\ is\ symmetric.$

If $\gamma_1 > 0$ *then the frequency distribution is positively skewed*

Result: The moment coefficient of skewness is independent of change of origin and scale.

- **Kurtosis**

Kurtosis is the peakedness of a frequency curve. Even if two distributions has same average, dispersion and skewness, one may have higher (lower) concentration of values near the mode, and in this case, its frequency curve will show a sharper peak (or flatter peak) than the other. This characteristic of a frequency distribution is known as kurtosis. The literal meaning of kurtosis is ‘peakedness’ or ‘flatness’ of a frequency curve. A frequency curve is said to be leptokurtic, if it has a higher peak. A frequency curve is said to be mesokurtic, if it is neither peaked nor flatted. A frequency curve is said to be platykurtic, if it has a lower peak.

The moment coefficient of kurtosis is denoted by β_2 and is defined as

$$\beta_2 = \frac{m_4}{{m_2}^2}$$

The gamma coefficient of kurtosis is defined as $\gamma_2 = \beta_2 - 3$

If $\gamma_2 < 0$ *then the frequency distribution is leptokurtic.*

If $\gamma_2 = 0$ *then the frequency distribution is mesokurtic.*

If $\gamma_2 > 0$ *then the frequency distribution is platykurtic.*

Result: The moment coefficient of kurtosis is independent of change of origin and scale.

Question Bank
Semester – III
US03CSTA01

<table>
<tr><td>1</td><td>Define moments. Establish the relationship between the moments about mean in terms of moments about any arbitrary point. The first four moments about the value 2 of the variable are 1, 16, - 40 and 10. Find the mean and variance. Also find moments about mean, $\beta 1$ and $\beta 2$.</td></tr>
<tr><td>2</td><td>Show that the geometric mean of two positive values X_1 and X_2 is always less than or equal to the arithmetic mean when are the two means equal?</td></tr>
<tr><td>3</td><td>Two workers on the same job show the following results over a long period of time.
<table>
<tr><td></td><td>Worker - A</td><td>Worker - B</td></tr>
<tr><td>Mean time of completing job(in minutes)</td><td>30</td><td>25</td></tr>
<tr><td>S.D (in minutes)</td><td>6</td><td>4</td></tr>
</table>
(i) Which worker appeared to be more consistent in the time he required to complete the job?
(ii) Which worker appears to be faster in completing the job? Explain.</td></tr>
<tr><td>4</td><td>In usual notation, Prove that
$$S^2 = \frac{\sum_{i=1}^{k} ni(Si^2 + di^2)}{\sum_{i=1}^{k} ni}$$</td></tr>
</table>

Where

$$di = \overline{X\imath} - \bar{X}, i = 1,2 \ldots k, \bar{X} = \frac{\sum_{i=1}^{k} ni\overline{X\imath}}{\sum_{i=1}^{k} ni}$$

5 From the following data:

3, 7, 11, 15… 1999

Find (i) n (no. of terms) (ii) Arithmetic mean. State and prove the result which you have used to solve (ii).

6 Calculate coefficient of skewness from the following data. Comment on your finding.

Weights(lbs)	Under 109	110 - 129	130 - 149	150 - 169	170 - 189	≥ 190
No. of persons	15	188	266	96	17	4

7 The following table shows the distribution of the life-time (in hours) of 200 bulbs.

Life-time	100 - 150	150 - 200	200 - 250	250 - 300	300 - 350	350 - 400	400 - 450
No. of bulbs	6	18	73	65	12	22	4

Obtain (a) % of bulbs that have life-time (i) Less than 300 hours (ii) Between 300 and 375 hours

(iii) More than 155 hours. (b) Q_2, D_7, P_{32} and O_4. Comment on your findings.

8 The mean salary of male and female employees in a firm is Rs. 5200 and Rs. 4200 respectively. The mean salary of all employees is Rs. 5000. Find the percentage of male and female employees.

9 The first quartile of the following data is 21.5. Find the missing frequencies and hence find the value of mode.

Class	10 - 15	15 - 20	20 - 25	25 - 30	30 - 35	35 - 40	40 - 45	45 - 50	Total
frequency	24	?	90	122	?	56	20	30	460

10 The following table shows the distribution of income of 200 workers.

Income (In Rs.)	1000 - 1500	1500 - 2000	2000 - 2500	2500 - 3000	3000 - 3500	3500 - 4000	4000 - 4500
No. of workers	6	18	73	65	12	22	4

Obtain or determine graphically (i) the no. of workers that have income between Rs. 3000 to 3500 (ii) % of workers with income more than Rs. 1550 (iii) the minimum income of the richest 50 workers (iv) the minimum and maximum incomes of the middle 60% of workers (v) Karl Pearson’s coefficient of skewness and comment.

11 Which measure of dispersion do you consider the best and why?

12 The following table gives the distribution of daily income of 500 workers in a factory. Calculate an appropriate measure of skewness and comment about the shape of the distribution.

Daily income(Rs.)	50 - 100	100 - 150	150 - 200	200 - 250	250 - 300	≥ 300
No. of workers	10	25	145	220	70	30

13 The following table gives the frequency distribution of the marks of 800 candidates in an examination.

Marks	0 - 20	20 - 40	40 - 60	60 - 80	80 – 100
No. of students	50	220	300	170	60

Obtain or determine graphically (i) Q_1, Q_3, D_5 & comment on it. (ii) Quartile deviation (Q.D) (iii) if the passing standard is 40%, find the result (iv) if it is desired to have 75% result, what grace marks a student be given?

14 What are the desirable properties which an average should possess? Which of the average to your mind possess most of these properties and why?

15 Find the geometric mean of the following data:
2^1, 2^3, $2^5 \ldots 2^{27}$. State and prove the result which you have applied to find the geometric mean.

16 The mean exam score for 31 students in a Geometry class was 79. The median exam score for the same set of students was 75. Two additional students took the exam at a later time and scored 65 and 93. Find the mean and median scores of all 33 students.

17 Prove that the Geometric Mean (G.M.) of n observations in Geometric progression (G.P.) is equal to the Geometric Mean of first and last term.

18 A man travels by a car for 4 days. He traveled for 10 hours each day. He drove on the first day at the rate of 45 kmph, second day at the rate of 40 kmph, third day at the rate of 38 kmph and fourth day at the rate of 37 kmph. Which averages, Arithmetic mean, Geometric mean, Harmonic mean will give us his average speed? Why?

19 Income of employees in an industry given below. The total income of the 10 employees in the class over Rs. 2500 is Rs. 30000. Compute the mean income. Every employee belonging to the top 25% of the earners is required to pay 5% of his income to workers relief fund. What should be the total contribution to this fund?

Monthly Income	0-500	500-1000	1000-1500	1500-2000	2000-2500	2500 & over
No. of workers	90	150	100	80	70	10

20 Weights of the students (in kgs) are recorded by a machine as under.

49	57	50	55	61	54	59	64	58	56

If the weighing machine shows weight more by 3 kg, find the correct values of range, standard deviation and coefficient of variation without calculating the correct weights. State clearly, the results which you have applied.

21 A man having to drive 90 kms wishes to achieve an average speed of 30 kmph. For the first half of the journey his average speed is only 20 kmph. What must be his average for the second half of the second half of the journey if his overall average speed is 30 kmph?

22 What is Skewness? Why there is a need to study Skewness? Differentiate Positive and Negative Skewness by giving figures.

23 Calculate an appropriate measure of central tendency from the following data.

Weights(lbs)	Under 19	110 - 129	130-149	150-169	170-189	190 & above
No. of persons	15	188	266	96	17	4

24 A man climbs up a slope at a speed of 5 kmph and descends it at a speed of 3 kmph. If the distance covered each way is 10 km, find the average speed for the entire journey.

25 From the following table, showing the wage distribution of worker in a factory.

Daily wages (In Rs.)	20 - 40	40 - 60	60 - 80	80 - 100	100 - 120	120 - 140	140 - 160	160 - 180	180 - 200
No. of workers	8	12	20	30	40	35	18	7	5

Determine (i) median wage (ii) the limits for the middle 50% of the wage earners (iii) % of workers who earned less than Rs. 75 (iv) the minimum wages of the 25 higher wage workers.

26 Prove that Geometric mean of series in G.P. (Geometric Progression) is equal to the geometric mean of its first and last term.

27 Prove that Arithmetic mean of series in A.P. (Arithmetic Progression) is equal to the arithmetic mean of its first and last term.

28 What do you mean by measures of central tendency? Write down the characteristics of ideal measures of central tendency. According to you, which is the most ideal measure of central tendency?

29 A cyclist covers his first three kms at a speed of 8 kmph, another 2 kms at 9 kmph and the last 2 kms at 4 kmph. Find the average speed for the entire journey.

30 What is Skewness? Why there is a need to study skewness? Differentiate between positive and negative skewness. State the different methods of studying skewness. Explain any one of them.

31 A study was conducted comparing female adolescents who suffer from bulimia to healthy females with similar body compositions and levels of physical activity.

Listed below are measures of daily caloric intake, recorded in kilocalories per kilogram of samples of adolescents from each group.

Daily caloric intake Kcal/kg)				
Bulimic			Healthy	
15.9	18.9	25.1	20.7	30.6
16.0	19.6	25.2	22.4	33.2
16.5	21.5	25.6	23.1	33.7
17.0	21.6	28.0	23.8	36.6
17.6	22.9	28.7	24.5	37.1
18.1	23.6	29.2	25.3	38.4
18.4	24.1	30.9	25.7	40.8
18.9	24.5		30.6	

(i) Find the median daily caloric intake for both the bulimic adolescents and the healthy ones.

(ii) which group has greater amount of variability in the measurement

(iii) Draw Box – and – Whisker plot for both the groups and comment on it.

32 Define raw moments and central moments. Express raw moments in terms of central moments.

33 The following table gives the frequency distribution of the marks of 800 candidates in an examination.

Marks	0 - 20	20 - 40	40 - 60	60 - 80	80 - 100
No. of students	50	220	300	170	60

Obtain or determine graphically (i) median (ii) the no. of students having marks (a) less than 40 (b) between 45 to 60 (c) more than 75 (ii) Q_1, Q_3, D_5, O3 and P_{25} & comment on it. (iii) If passing standard is 40%, find % of result (iv) if it is desired to have 75% result, what grace marks a student be given? (v) marks obtained by the middle 70 % of students or the range which include the marks of middle 70% students (vi) if first 50 students are to be given direct admission to MCA, what minimum marks a student secure to get admission in MCA (vii) the minimum marks obtained by the upper (higher) 100 students (viii) the maximum marks obtained by the upper 25% of students (ix) the minimum marks obtained by the upper 70% of students. (x) Draw histogram and hence obtain mode.

34 From the following data:

2, 5, 8 …1499

Find (i) n (no. of terms) (ii) Median.

35 Explain the meaning of skewness. State the various methods to determine skewness and its coefficient. Explain any one of them

36 Explain the concept of (i) positive skewness (ii) negative skewness by sketching suitable diagrams locating measures of central tendency.

37 Define moments. Establish the relationship between the moments about mean in terms of moments about any arbitrary point.

SEMESTER - III

USC03STA01

MULTIPLE CHOICE QUESTIONS

1 For comparing the health conditions of two towns, we have to calculate

(a) Crude death rate (b) Crude birth rate

(c) Infant mortality rate (d) Age specific fertility rate

2 If we want to know more about deaths occurring in a different section of the population, we have to calculate

(a) CDR **(b) SDR** (c) STDR (d) None of the above

3 The geometric mean of 2, 4,16 and 32 is

(a) 13.50 **(b) 8** (c) 4.76 (d) None of the above

4 The geometric mean of a set of values lies between arithmetic mean and --------------------------

(a) Arithmetic mean (b) Mode **(c) Harmonic mean** (d) None of the above

5 The relationship between AM, GM, and HM is

(a) AM ≤ GM ≤ HM (b) AM ≥ HM ≥ GM **(c) AM ≥GM ≥ HM** (d) None of the above

6 Median = ----------------- quartile

(a) First **(b) Second** (c) Third (d) Fourth

7 For a symmetrical distribution

(a) $\mu_2 = 0$ (b) $\mu_2 > 0$ **(c)** $\boldsymbol{\mu_3 = 0}$ (d) $\mu_3 > 0$

8 A distribution with two modes is called

(a) unimodal **(b) bimodal** (c) multimodal (d) None of the above

9 Which of the following is not affected by extreme observations

(a) Mean **(b) Median** (c) Mode (d) All of these

10 The arithmetic mean of $1, 2, \ldots\ldots, n$ is

(a) $\frac{n(n+1)(2n+1)}{6}$ (b) $\left(\frac{n(n+1)}{2}\right)^2$ (c) $\frac{n(n+1)}{n}$ (d) **None of the above**

11 Mean – Mode =? (Mean – Median)

(a) 1 (b) 2 **(c) 3** (d) 4

12 If each of a set of observations of a variable is multiplied by a constant (non-zero) value, the variance of the resultant variable is

(a) unaltered (b) increase (c) decrease **(d) both (b) and (c)**

13 If each of a set of observations of a variable is multiplied by a positive constant (non-zero) value, the variance of the resultant variable is

(a) unaltered **(b) increase** (c) decrease (d) none of the above

14 The sum of squares of deviations is least (minimum) when measured from

(a) Mean (b) Median (c) Mode (d) All of the above

15 For a symmetrical distribution, all the odd order central moments are

(a) Equal to zero (b) Greater than zero (c) Less than zero (d) None of the above

16 A.M, G.M and H.M of any series are equal when

(a) the distribution is symmetrical **(b) all the values are same**

(c) the distribution is positively skewed (d) the distribution is unimodal

17 For a symmetrical distribution, $\mu_1 = \mu_3 = \mu_5 =$ ………………. are
(a) Equal to zero (b) Greater than zero (c) Less than zero (d) None of the above

18 The limits for quartile coefficient of skewness (i.e. Bowley's coeff. Of skewness)
(a) ±3 (b) 0 and 3 **(c) ±1** (d) ±∞

19 The statement that the variance is equal to second central moments is
(a) Always true (b) Sometimes true (c) Never true (d) Unambiguous

20 In a frequency curve of scores, the mode was found to be higher than the mean. This shows that the distribution is
(a) Symmetric **(b) Negatively skewed** (c) Positively skewed (d) None of these

21 For any frequency distribution, the coefficient of kurtosis is
(a) Greater than 3 (b) Less than 3 (c) Equal to 3 **(d) All of the above**

22 If 25% of items are less than 10 and 25% are more than 40, the quartile deviation is
(a) 10 (b) 40 **(c) 15** (d) 30

23 In a symmetric distribution, the upper and lower quartiles are equidistant from
(a) Mean (b) Median (c) Mode **(d) All of the above**

24 In a symmetric distribution, the mean and mode are
(a) Same (b) Different (c) Neither (a) nor (b) (d) (a) or (b)

25 For a symmetrical distribution
(a) $Q_2 = \frac{Q_3 - Q_1}{2}$ **(b)** $Q_2 = \frac{Q_3 + Q_1}{2}$ (c) $Q_2 = 0$ (d) None of the above

26 If the mean and mode of a given distribution are equal, then coefficient of skewness is
(a) Greater than zero (b) Less than zero **(c) Equal to zero** (d) All of these

27 The crude death rate usually lies between
(a) 8 and 30 per thousand (b) 5 and 35 per thousand
(c) 2 and 32 per thousand **(d) All of the above**

28 The crude birth rate usually lies between
(a) 8 and 30 per thousand (b) 5 and 35 per thousand
(c) 2 and 32 per thousand **(d) All of the above**

29 Index numbers are also known as
(a) Economic barometers (b) Signs and guide parts
(c) Both (a) and (b) (d) Neither (a) nor (b)

30 Index numbers reveal the state of
(a) inflation (b) deflation **(c) Both (a) and (b)** (d) Neither (a) nor (b)

31 Index numbers are expressed in
(a) Percentage (b) ratio (c) terms of absolute values (d) All of the above

32 Laspeyre's index formula uses the weights of
(a) Base year (b) Current year (c) None of the above (d) (a) and (b)

33 Paasche’s index formula uses the weights of

(a) Base year **(b) Current year** (c) None of the above (d) (a) and (b)

34 The first and foremost step in the construction of index numbers is

(a) Choice of base year

(b) Choice of weights

(c) To delineate the purpose of index numbers

(d) All of the above

35 If Laspeyre’s price index is 324 and Paasche’s price index is 144, then Fisher’s index is

(a) 234 (b) 180 **(c) 216** (d) None of the above

36 Index number of the base year is

(a) 100 (b) 1000 (c) 1 (d) None of the above

37 Fisher’s index number is ------------------------------------ of Laspeyre’s and Paasche’s index numbers

(a) Arithmetic mean **(b) Geometric mean** (c) Harmonic mean (d) Weighted mean