-: Statistics :-

Definition:-

COPSGD

(ascending or descending order)

It is a science of collecting, organizing, Presenting, summarizing, generalizing & draw infrences of a data sets in which we are interested.

Observation:-

Any thing that can be measured or observed is called an observation.

Data:-

Collection of observation is called data

OR

Numbers / ~~observation~~ measurements that are collected as a result of observation.

Population:-

A pop^n is the set of all units of interest in a particular study.

Sample:-

A sample is a True representative part of pop^n. OR

It is a subset of data selected rom the pop^n

• Popⁿ results are based on sample results.

Parameters:-

A parameter is a numerical measurement describing some characteristics of a popⁿ.

e.g μ = popⁿ average, σ^2 = popⁿ variance.

Statistic:-

A statistic is a numerical measurement describing some characteristics of a sample.

e.g $\bar{X}$ = sample average, S^2 = sample variance.

Variable:-

Any characteristic of a person, group or environment that can vary.

⇒ Types of Data:-

(i) Data by nature

(ii) Data by source.

(i) Data by Nature:-

(a) Qualitative Data

(b) Quantitative Data

(c) Discrete Data

(d) Continous Data

(ii) Data by Source:-

(a) Primary Data (b) Secondary Data

By Source:-

(i) Primary Data:- (Ungrouped data)

Data that have been originally collected (raw data) & have not undergone any sort of statistical treatment are called primary data.

- Collected directly from respondent
- Primary data collected through questionare/interview

(b) Secondary Data:-

Data that have undergone any sort of statistical treatment atleast once.

i.e The data that have been collected, classified tabulated or presented in some form for certain purpose.

→ Data collected from NADRA to determine growth rate of Pakistan.

⇒ Data by Nature:-

(a) Qualitative Data:

Qualitative data are observation that are non-numerical. It provides labels or name for a characteristic of an element.

e.g student performance, hair colour

(b) Quantitative Data:-

Quantitative data are observation measured on numerical scale eg income, number of students.

(c) Discrete Data:-

Data whose possible values are Countable is called discrete data. (In whole figure) e.g Number of members in a family, numbers of students in a class.

d) Continous Data:-

Data whose possible values are uncountable & which may assume any value in interval is called continous data

e.g 2-4 infinite numbers.

eg Height both discrete & continous.

Measure of Central Tendency:-

Trend of obs. in a data set, that is ~~also~~ known as measure of location or position. The measures of central Tendency or location are generally known as averages.

Different Type of average:-

(i) Arithmetic Mean (ii) Geometric Mean

(iii) Harmonic Mean (iv) Median

(v) Mode (vi) Quartile

(vii) Decile (viii) Percentile

Mean ↓ A.M G.M H.M

Median ↓ Quartile Decile Percentile

Mode ↓ Uni-mode Bi mode Tri mode

(i) Arithmetic Mean:- (A.M)

The arithmetic mean of "n" obs. are defined as the Total number (sum) of all obs. divided by the number of obs.

- The A.M of sample data is denoted by "$\bar{X}$" & the symbol "μ" is used for mean of popn data

For ungrouped data:-

$$\bar{X} = \frac{x_1 + x_2 + x_3 + --- + x_n}{n}$$

grouped data:- $\bar{X} = \Sigma f_i x_i / n$

$$\bar{X} = \sum_{i=1}^{n} x_i \Big/ n \qquad \bar{X} = \frac{\Sigma X}{n} \text{ For Sample}$$

$$\mu = \frac{\Sigma X}{N} \text{ For Population}$$

Example :-

Find the A.M of the set of numbers. 84, 91, 72, 68, 87, & 78.

So,

$$n = 6$$

$$\bar{X} = \frac{84+91+72+68+87+78}{6} = \frac{480}{6} = 80$$

It means that most values of data lies around 80.

Geometric Mean:-

$$G.M = \text{antilog}\left(\frac{\sum_{i=1}^{n} \log x_i}{n}\right)$$

Harmonic Mean:- Reciprocal of the A.M of the reciprocal of the values

$$H.M = \frac{\sum_{i=1}^{n} 1/x_i}{n} = \frac{n}{\Sigma 1/x_i}$$

$$H.M = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}{n} = \frac{\sum_{i=1}^{n} 1/x_i}{n}$$

edian:-

"The median of a data is the middle item when the item (the obs.) are arranged in ascending or decending order."

Median symbol.. $\tilde{X}$ (X tilda).

When n is ~~even~~ odd:- the value of the $\left(\frac{n+1}{2}\right)^{th}$ obs.

When n is even :- $\left(\frac{n}{2}\right)^{th}$ obs. OR $\left(\frac{n+2}{2}\right)^{th}$.

Example:-

The wages of 5 workers in rupees are 1800, 1900, 1700, 2000, & 2200. Find the median.

: odd no. usually divided data into 2 equal parts

Ascending order:-

↓Median

1700, 1800, $\underline{1900}$, 2000, 2200

$$\tilde{X} = \frac{5+1}{2} = \frac{6}{2} = 3^{rd} \text{ obs.} = 1900 \text{ median}$$

1700, 1800, $\underline{1900}$ & $\underline{2000}$, 2200, 2500.

$$\tilde{X} = \left(\frac{n}{2}\right)^{th} \text{ obs.} -$$

$$\frac{1900+2000}{2} = 1950$$

$$= \left(\frac{6}{2}\right)^{th} = 3^{rd} \text{ obs.} = 1900$$

Example:-(2):- The minimum Temp. in Murree for the 1st 10 days of March was.

-1°C, -2, 1, 0, 3, 3, 4, 3, 2, 6. Find median.

-2, -1, 0, 1, 2, 3, 3, 3, 4, 6

$$\tilde{X} = \left(\frac{n}{2}\right)^{th} \text{ obs.} = \frac{10}{2} = 5^{th} \text{ obs.} = 2$$

For even add central 2 values & divided by 2.

$$\tilde{X} = \frac{2+3}{2} = \frac{5}{2} = 2.5 \Rightarrow \text{Median}$$

Quartiles:- Quantiles:-

Quartiles are the values of variates that divide a set of data into 4 equal parts after arranging the obs. ~~into~~ in ascending order of magnitude.

OR

The 3 values which divide the data set into 4 equal parts.

Q_1: lower quartile (it cover 25% area of data set)

Q_2: Median (it cover 50% area of data set).

Q_3: Upper quartile (it cover (75%) area of data set).

Formulas for ungrouped data:- General $Q_j = \left[\frac{j(n+1)}{4}\right]^{th}$ obs

Q_1 = The value of $\left(\frac{n+1}{4}\right)^{th}$ obs.

Q_2 = The value of $\left[\frac{2(n+1)}{4}\right]^{th}$ obs.

Q_3 = The value of $\left[\frac{3(n+1)}{4}\right]^{th}$ obs.

Example:- Find the Q_1, Q_2, Q_3 from the following data.

26, 22, 14, 30, 18, 11, 35, 41, 12 & 32.

11, 12, 14, 18, 22, 26, 30, 32, 35, 41.

$Q_1 = \left(\frac{n+1}{4}\right)^{th} = \frac{10+1}{4}$

Q_1 = The value of $(2.75)^{th}$ obs.

$= 2^{nd}$ obs. $+ 0.75(3rd \text{ obs} - 2^{nd} \text{ obs})$

$= 12 + 0.75(14-12)$

$= 12 + 0.75(2)$

$Q_1 = 13.5$

$Q_2 = \left[\frac{2(n+1)}{4}\right]^{th}$

$= \left[\frac{2(10+1)}{4}\right]^{th}$

= The value of 5.5^{th} obs.

$= 5^{th}$ obs $+ 0.5(6^{th} \text{ obs} - 5^{th} \text{ obs})$

$= 22 + 0.5(26-22)$

$Q_2 = 24$

$Q_3 = \left[\frac{3(n+1)}{4}\right]^{th}$

$= \frac{3(10+1)}{4}$

= The value of 8.25^{th} obs.

$= 8^{th}$ obs. $+ 0.25(9^{th} \text{ obs} - 8^{th} \text{ obs})$

$= 32 + 0.25(35-32)$

$= 32.75$

Percentile :-

The set of values which divide the group of data into 100 equal parts. $1/100^{th}$ of data set.

P_x = the value of $\frac{x(n+1)^{th}}{100}$ obs

Example:- Percentile Question.

53, 74, 82, 42, 29, 81, 68, 58, 28, 67, 54, 93, 70, 30, 55, 36, 37, 29, 51

Ascending order:-

20, 28, 29, 30, 36, 37, 39, 42, 53, 54, 55, 61, 67, 68, 70, 74, 81, 82, 93.

P_{15} = The value of $15\left(\frac{n+1}{100}\right)$ obs

= 3.15 obs.

3.15 obs

= 3^{rd} obs + 0.15 (4^{th} obs - 3^{rd} obs)

= 29 + 0.15 (30 - 29)

P_{15} = 29.15

Decile:-

The value of variable that divide an ordered data set into Ten equal parts. So, that each part represents $1/10^{th}$ of sample | pop^n.

OR

Decile are the 9th value that divide data set into Ten equal parts.

Formula:-

$$\frac{j(n+1)^{th}}{10} \text{ obs.}$$

Example:-

No. of defective items are:

45, 30, 36, 26, 16, 21, 33, 40, 32, 14, 10, 29, 23, 39, 17, 11, 15, 34, 19, 24, 21, 35, 42, 37.

10, 11, 14, 16, 17, 18, 19, 21, 21, 23, 24, 26, 29, 30, 32, 33, 34, 35, 36, 37, 39, 40, 42, 45.

Find 5th decile.

D_5 = The value of $\frac{5(24+1)}{10}$

$= 12.5^{th}$ obs.

$= 12^{n}$ obs $+ 0.5(13^{th}$ obs $- 12^{th}$ obs$)$

$= 26 + 0.5(29 - 26)$

$= 27.5$

Similarly with D_7.

Show that:-

$Q_2 = D_5 = P_{50} = \tilde{X}$

② :Measures of Dispersion:-

Dispersion = spreadness, difference b/w obs. of data set.

Aim = less dispersion b/w data's obs.

Ideal data set = where dispersion b/w obs. is less.

→ Measure of central tendency does not Tell neither the closeness b/w obs. nor dispersion b/w them but only & only central point.

Types:-

There are two types of measure of dispersion.

(i) Absolute measure of dispersion.

(ii) Relative measure of dispersion.

Absolute measure of dispersion:-

An Absolute measure of dispersion is one that measures the dispersion in terms of the same units or in the square of units, as the units of data.

e.g:- If the units of data are rupees, metres, kg, the unit of measure of dispersion will also be rupees, metres & kg.

③ Relative Measure of Dispersion:-

A relative measure of dispersion one that is expressed in the form of a ratio, co-efficient or percentage and it is independent of the units of measurement. They are also called unit free or unitless measures.

→ It is useful for comparison of data of different nature.

→ Types of Absolute measure of dispersion-

(i) Range (ii) Mean deviation

(iii) Variance (iv) Standard deviation

(v) The Semi-Interquartile range.

→ Types of Relative measure:-

(i) Co-efficient of range (ii) Co-efficient of mean deviation

(iii) Co-efficient of variation (C.V)

(i) Range:-

It is the difference b/w the maximum & minimum obs. of the given data set

Range $= x_m - x_0$

$= 100km - 20km = 80\ km$

unit hence absolute.

Co-efficient of Range:-

$$\text{Co-efficient of Range} = \frac{x_m - x_0}{x_m + x_0}$$

$$= \frac{100\,km - 20\,km}{100\,km + 20\,km}$$

$$= \frac{80\,km}{120\,km} = 0.67$$

This is a pure number & is used for the purposes of comparison.

(2) The Semi-Interquartile Range or the Quartile Deviation:-

The difference b/w the third & first quartile & half of this range is known as semi-interquartile range or quartile deviation.

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

$$\text{Co-efficient of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Advantages:-

The Q.D is superior to range as it is not affected by extremely large or small obs.

It is simple to understand & easy to

calculate.

Disadvantages:-

It gives no information about the position of obs lying outside the two quartile.

(ii) Mean Deviation:-

The mean deviation of a set of data is defined as the arithematic mean of the deviations measured from the mean, all deviations being counted as positive.

$$M.D = \frac{\Sigma|x_i - \bar{x}|}{n}, \text{ for sample data.}$$

$$M.D = \frac{\Sigma|x_i - \mu|}{N}, \text{ for pop}^n \text{ data.}$$

Example:-

x_i	$(x_i - \bar{x})$	$\|x_i - \bar{x}\|$
84	4	4
91	11	11
72	-8	8
68	-12	12
87	7	7
78	-2	2
$\bar{x} = 80$	$\Sigma(x_i - \bar{x}) = 0$	44

$$M.D = \frac{\Sigma|x_i - \bar{x}|}{n}$$

$$= \frac{44}{6}$$

$$= 7.33$$

Co-efficient of M.D:-

Co-efficient of M.D $= \dfrac{M.D}{Mean}$ OR $\dfrac{M.D}{Median}$

Variance:-

The mean of square of deviations of all the obs. from their mean is known as variance.

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}, \text{ for pop}^n \text{ data} \quad = \frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N}\right)^2$$

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n}, \text{ for sample data.}$$

Standard Deviation:-

The positive square root of the variance is called standard deviation.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}, \text{ for pop}^n \text{ data}$$

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}, \text{ for sample data.}$$

Co-efficient of Standard deviation:-

Co-efficient of Standard deviation $= \frac{S.D}{Mean}$

Co-efficient of Variation:- (C.V)

It is used to compare the variation in two or more data sets or distⁿ.

→ A large value of C.V indicates that the variability is great & a small value of C.V indicates less variability.

$$C.V = \frac{s}{\bar{x}} \times 100, \text{ for sample data}$$

$$= \frac{\sigma}{\mu} \times 100, \text{ for pop}^n \text{ data.}$$

Note:-

The C.V is also used to compare the performance of two candidates or of two players given their score in various papers or games, the smaller the co-efficient of variation (C.V) the more consistent is the performance of candidate or player.

Properties of Variance & Standard Deviation:-

1:- The variance of constant is equal to zero.

$$Var(a) = \frac{1}{N} \Sigma (a-a)^2 = 0$$

($\because$ mean of constant is constant itself).

$$S.D(a) = 0$$

2:- The variance & S.D is independent of the origion. i.e it remains unchanged when a constant is added or substracted from each obs. of variable X.

$$Var(X \pm a) = Var(X)$$

3:- The variance is multipied or divided by the square of the constant, when each obs. of the variable X is either multiplied or divided by a constant.

$$Var(aX) = a^2 Var(X).$$

4:- The variance of the sum or difference of the two independent variable variables is equal to the sum or difference of their respective variances.

$$Var(X \pm Y) = Var(X) \pm Var(Y)$$