CHAPTER 32

STATISTICS

32.1 MEASURES OF CENTRAL TENDENCY

For a given date, a single value of the variable which describes its characteristics is identified. This single value is known as the average. An average value, generally lies in the central part of the distribution and, therefore, such values are called the measures of central tendency. The commonly used measures of central tendency are:

(a) Arithmetic Mean
(b) Geometric Mean
(c) Harmonic Mean
(d) Median
(e) Mode

32.2 TYPES OF DISTRIBUTION

(i) **Individual/Discrete Distribution: (Ungrouped Data):** Here, we are given x_1, x_2, x_3 ,……, x_n different values.

(ii) **Discrete Series with Frequency Distribution (Ungrouped Data with Frequency Distribution):** Here, we are given:

x_i	x_1	x_2	x_3	….	x_n
f_i	f_1	f_2	f_3	…..	f_n

where f_i is frequency of x_i.

(iii) **Continuous series with frequency distribution (grouped data):**
Here, we are given class intervals with corresponding frequencies.

Class interval	0 –10	10 – 20	20 – 30	……
Frequency	f_1	f_2	f_3	…….

Range: Range = Largest observation – smallest observation.

Class size/length of clan-interval: (a – b) is defined as (b – a), e.g., class size of (40 –50) is (50 – 40) = 10

Class–mark of class interval: Mid–point of class interval, e.g., class mark of class interval (40 – 50) is $40+\frac{(50-40)}{2}=45$. In general, class–mark of class interval (a –b) is $a+\frac{(b-a)}{2}=\frac{a+b}{2}$.

32.2.1 Arithmetic Mean

(i) **For discrete series:**

(a) **Direct method:** $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$

(b) **Short-cut method:** $\overline{x} = a + \overline{u} = a + \left(\frac{\sum_{i=1}^{n} u_i}{n} \right); u_i = (x_i - a)$

Here a is a suitable number which makes the greater values of x_i's to smaller values. For example, if given data is 202, 219, 238, 258, 279, 299. It will be convenient to take, a = 250. This method helps to find means by reducing calculations when given values of x_i are larger.

(c) **Step deviation method:** $\overline{x} = a + h\overline{u} = a + h\sum_{i=1}^{n} \frac{u_i}{n}; u_i = \frac{(x_i - a)}{h}$; where a and h are suitable real numbers, e.g., in data, 210, 220, 230, 260, 280, 290, take a = 250 and h = 10.

(ii) **For discrete series with frequency distribution:**

(a) **Direct method:** $\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum f_i}$

(b) **Short-cut method:** $\overline{x} = a + \overline{u} = a + \frac{\sum f_i x_i}{\sum f_i}; u_i = (x_i - a)$; a = suitability chosen real number.

(c) **Step deviation method:** $\overline{x} = a + h\overline{u} = a + h\frac{\sum f_i u_i}{\sum f_i}$; $u_i = \frac{x_i - a}{h}$; where a and h are suitably chosen real number.

(iii) **For continuous series (grouped data):**

(a) **Direct method:** $\overline{x} = \frac{\sum f_i x_i}{\sum f_i}$; where x_i's are class-makes of intervals.

(b) **Short-cut method:** $\overline{x} = a + \overline{u} = a + \frac{\sum f_i u_i}{\sum f_i}; u_1 = (x_1 - a)$; a is suitably chosen real numbers.

(c) **Step deviation method:** $\overline{x} = a + h\overline{u} = a + h\frac{\sum f_i u_i}{\sum f_i}$; $u_i = \frac{x_i - a}{h}$; a and h are suitably choosen real numbers. Generally, h = width of class–intervals. Here, $u_i = \frac{x_i - a}{h}$; defines mean deviation of variate x_i form assumed mean 'a'.

32.2.2 Weighted Arithmetic Mean

If w_1, w_2, w_3,..., w_n are the weights assigned to the values x_1, x_2, x_3,...,x_n respectively, then the weighted average is defined as: Weighted A.M. $= \frac{w_1x_1 + w_2x_2 + w_3x_3 + ... + w_nx_n}{w_1 + w_2 + w_3 + ... + w_n}$

32.3 COMBINED MEAN

If we are given the A.M. of two data sets and their sizes, then the combined A.M. of two data sets can be obtained by the formula: $\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$; where $\bar{x}_{12}$ = combined mean of the two date sets 1 and 2.

$\bar{x}_1$ = Mean of the first data $\bar{x}_2$ = mean of the second data

n_1 = Size of the first data n_2 = Size of the second data

32.3.1 Properties of Arithmetic Mean

(i) In a statistical data, the sum of the deviations of individual values from A.M. is always zero, That is, $\sum_{i=1}^{n} f_i(x_i - \bar{x}) = 0$; where f_i is the frequency of x_i $(1 \le i \le n)$.

(ii) In a statistical data, the sum of square of the deviations of individual values from real number 'a' is the least, when a is mean $(\bar{x})$. That is, $\sum f_i(x_i - a)^2 \ge \sum f_i(x_i - \bar{x})^2$.

(iii) If each observation, x_i is increased (decreased) by 'd', then A.M. also increases (decreases) by 'd'.

$\because$ $\frac{\sum f_i x_i}{\sum f_i} = A$, then $A' = \frac{\sum f_i(x_i \pm d)}{\sum f_i} = \frac{\sum f_i x_i}{\sum f_i} \pm d\frac{\sum f_i}{\sum f_i} = A \pm d.$

(iv) If each observation x_i is multiplied (or divided) by d ($\neq$ 0 for division), then the new A.M. is $d\left(\text{or}\frac{1}{d}\right)$ times of original A.M.

$\because$ $\frac{\sum f_i x_i}{\sum f_i} = A$, then $A' = \frac{\sum f_i(x_i d)}{\sum f_i} = \frac{d\sum f_i x_i}{\sum f_i} = dA.$

32.4 GEOMETRIC MEAN

(a) **For ungrouped data:** G.M. of $x_1, x_2, x_3, \ldots\ldots, x_n$; $x \neq 0$ is given by G.M. = $(x_1, x_2, x_3, \ldots\ldots, x_n)^{1/n}$

(i) If $(x_1, x_2, x_3, \ldots\ldots, x_n) < 0$ and n is even, then G.M. is not defined.

(ii) If $(x_1, x_2, x_3, \ldots\ldots, x_n) < 0$ and n is odd, then G.M. is defined, given by G.M. = $-(|x_1|.|x_2|.|x_3|.\ldots\ldots|x_n|)^{1/n}$.

$\Rightarrow$ G.M. = $-\text{Antilog}\left[\frac{\log|x_1| + \log|x_2| + \ldots + \log|x_n|}{n}\right]$.

(iii) If each $x_i \ge 0$, then G.M. = $\text{Antilog}\left[\frac{\log x_1 + \log x_2 + \ldots + \log x_n}{n}\right]$.

(iv) If each x_i is non-zero and $x_1, x_2, x_3, \ldots, x_n > 0$; then G.M. = $\text{Antilog}\left(\frac{\log|x_1| + \log|x_2| + \ldots + \log|x_n|}{n}\right)$.

(b) **For ungrouped data with frequency distribution or grouped data (continuous series):** It is given by $\text{G.M.}((x_1)^{f_1}.(x_2)^{f_2}\ldots.(x_n)^{f_n})^{1/N}; N = \sum f_i$, when defined. In case of continuous series x_i = class-mark (mid-value of interval).

$$\Rightarrow \quad \text{G.M.} = \text{Antilog}\left(\frac{\sum f_i \log|x_i|}{N}\right) \text{ for } (x_1)^{f_1}.(x_2)^{f_2}....(x_n)^{f_n} > 0,$$

$$\text{and G.M.} = -\text{Antilog}\left(\frac{\sum f_i \log|x_i|}{N}\right) \text{ for } (x_1)^{f_1}.(x_2)^{f_2}....(x_n)^{f_n} < 0;\ N = \text{odd}.$$

32.5 HARMONIC MEAN

The harmonic mean of n observation $x_1, x_2, ..., x_n$ is defined as: H.M: $\frac{n}{1/x_1 + 1/x_2 + ... + 1/x_n}$.

If $x_1, x_2, ..., x_n$ are n observations, which occur with frequencies $f_1, f_2, ..., f_n$ respectively, their H.M. is given by $\text{H.M.} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} (f_i / x_i)}$.

32.6 ORDER OF A.M., G.M. AND H.M.

The arithmetic mean (A.M.) geometric mean (G.M.), and harmonic mean (H.M.) for a given set of observations are related as under: A.M. ≥ G.M. ≥ H.M.

Equality sign holds only when all the observations are equal.

Relation between G.M., H.M, of two numbers a and b G.M. of two numbers a and b is also the G.M. of A.M. and H.M. of a and b.

$$\because \quad (\sqrt{ab})^2 = \left(\frac{a+b}{2}\right).\left(\frac{2ab}{a+b}\right), \text{ i.e., } (\text{G.M.})^2 = (\text{A.M.}).(\text{H.M.}).$$

$$\Rightarrow \quad \text{G.M.} = \sqrt{\text{A.M.} \times \text{H.M.}}$$

32.7 MEDIAN

Median is the middle most or the central value of the variate in a set of observations, when the observations are arranged either in ascending or in descending order of their magnitudes. It divides the arranged series in two equal parts.

(a) **For individual/discrete series:**

Step I: Arrange the variables in ascending or descending order

Step II: $$\text{Median} = \begin{cases} \left(\frac{n+1}{2}\right)^{th} \text{term}; & \text{for } n = \text{odd} \\ \frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2}; & \text{for } n = \text{even} \end{cases}$$

(b) **For discrete series with frequency distribution:**

Step I: Arrange the variables x_i's in ascending or descending order keeping frequencies along with them.

Step II: Prepare a cumulative frequency table and find $\Sigma f_i = N$.

Step: III: Median $= \begin{cases} \left(\frac{N}{2}\right)^{th} \text{ observation if N odd} \\ \left(\frac{N^{th}}{2} + \left(\frac{N}{2}+\right)^{th}\right)^{th} \text{ term if N even} \end{cases}$.

For $\frac{N^{th}}{2}$ terms, see the value of x_i corresponding to $\frac{N^{th}}{2}$ cumulative frequency, similar for the $\left(\frac{N}{2}+1\right)^{th}$ term.

(c) **For continuous series (Grouped data):**

Step I: Prepare the cumulative frequency table.

Step II: Find median class, i.e., class corresponding to $\frac{N^{th}}{2}$ observation.

Step: III: Median $= \ell + \left(\frac{N}{2} - C\right) \times \frac{h}{f}$; where ℓ = lower limit of median class

$N = \sum f_i$; h = width of class-intervals

f = frequency of median class

C = cumulative frequency of class preceding the median class

Remarks:

1. *Median is also known as 2nd quartile (Q_2), i.e., median* $= \ell + \left(2.\frac{N}{4} - C\right) \times \frac{h}{f}$.
2. *1st quartile* $= \ell + \left(1.\frac{N}{4} - C\right) \times \frac{h}{f}$
3. *3rd quartile* $= \ell + \left(3.\frac{N}{4} - C\right) \times \frac{h}{f}$
4. *Similarly, we have deciles D_1, D_2, D_3,.....,D_9; where* $D_i = \ell + \left(i.\frac{N}{10} - C\right) \times \frac{h}{f}$. $\Rightarrow D_5$ *= 5th decile = median*
5. *In the same way, we have percentile P_1, P_2, P_3,.....,P_{99}; where* $P_i = \ell + \left(i\frac{N}{100} - C\right) \times \frac{h}{f}$;

   $\Rightarrow$ P_{50} *= 50th percentile = median. Thus, median* $Q_2 = D_5 = P_{50}$

32.8 MODE

Mode is that value in a series which occurs most frequently. In a frequency distribution, mode is that variate which has the maximum frequency.

32.8.1 Computation of Mode

(a) **Mode of Individual Series:** In the case of individual series, the value which is repeated maximum number of times is the mode of the series.

(b) **Mode of Discrete Series:** In the case of discrete frequency distribution, mode is the value of the variate corresponding to the maximum frequency.

Case (i): If a group has two or more scores with the same frequency and that frequency is the maximum positive distribution is bimodal or multimodal, that is to say, it has several modes, e.g., 1, 1, 1, 4, 4, 5, 5, 5, 7, 8, 9, 9, 9 has modes 1, 5 and 9.

Case (ii): When the scores of a group all have the same frequency, there is no mode, e.g., 2, 2, 3, 3, 6, 6, 9, 9 has no mode.

Case (iii): If two adjacent values are the maximum frequency, the average of two adjacent scores is the mode 0, 1, 3, 3, 5, 5, 7, 8 mode = $\frac{3+5}{2}=4$.

(c) **Mode of Continuous Series:**

Case 1: When classes have the same width:

Step 1: Find the modal class, i.e., the class which has maximum frequency. The modal class can be determined either by inspection or with the help of grouping table.

Step 1: The mode is given by the formula:

Mode = $l+\frac{f_m-f_{m-1}}{2f_m-f_{m-1}-f_{m+1}}\times h$; where l = the lower limit of the modal class

h = the width of the modal class

f_{m-1} = the frequency of the class preceding modal class

f_m = the frequency of the modal class

f_{m+1} = the frequency of the class succeeding modal class

In case, the modal value lies in a class, other than the one, containing maximum frequency, we take the help of the following formula; Mode = $l+\frac{f_{m+1}}{f_{m-1}+f_{m+1}}\times h$; where symbols have usual meaning.

Case (ii): When classes have different width: Let a_i be the width of i^{th} interval

Step I: First, find the heights; $h_i=\frac{f_i}{a_i}$.

The nodal class is the one with the greatest height and mode = $\ell+\frac{h_m-h_{m-1}}{(h_m-h_{m-1})+(h_m-h_{m+1})}.a_i$.

32.9 MEASURES OF DISPERSION

The degree to which numerical values in the set of values tend to spread about an average value is called the dispersion of variation. The commonly used measures of dispersion are:

(a) Range (b) Quartile Deviation or Semi-inter-quartile Range

(c) Mean Deviation (d) Standard Deviation

Range: It is the difference between the greatest and the smallest observations of the distribution.

If L is the largest, and s is the smallest observation in a distribution, then its Range = L – S. Also,

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

Quartile Deviation: Quartile Deviation or semi-inter-quartile range is given by $\text{Q.D.} = \frac{1}{2}(Q_3 - Q_1)$

coefficient of $\text{Q.D.} = \frac{(Q_3 - Q_1)}{(Q_3 + Q_1)}$.

Mean Deviation: For a frequency distribution, the mean deviation from an average (median, or arithmetic mean) is given by

(i) For individual series:

$$\text{M.P. from mean} = \frac{\sum_{i=1}^{n}|x_i - \text{mean}|}{n} \qquad \text{M.D. from median} = \frac{\sum_{i=1}^{n}|x_i - \text{median}|}{n}$$

(ii) For discrete series with frequency distribution and continuous series:

$$\text{M.D. from mean} = \frac{\sum_{i=1}^{n} f_i\,|x_i - \text{median}|}{\sum_{i=1}^{n} f_i} \qquad \text{M.D. from median} = \frac{\sum f_i|x_i - \text{median}|}{\sum f_i}$$

(iii) For continuous series xi = class–mark:

$$\text{Coefficient of M.D. from mean} = \frac{\text{M.D.(Mean)}}{\text{mean}}$$

$$\text{Coefficient of M.D. from median:} = \frac{\text{M.D.(medain)}}{\text{mean}}$$

32.10 STANDARD DEVIATION

The standard deviation of a statistical data is defined as the positive square root of the squared deviations of observations from the A.M. of the series under consideration.

(a) **For ungrouped data/individual/discrete series:**

(i) Direct Method $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum(x_i^2 + \bar{x}^2 - 2\bar{x}x_i)}{n}}$

$$= \sqrt{\frac{\sum x_i^2}{n} + \frac{nx^{-2}}{n} - \frac{2\bar{x}}{n}\cdot\sum x_i} = \sqrt{\frac{\sum x_i^2}{n} + (\bar{x})^2 - 2(\bar{x})^2} = \sqrt{\frac{\sum x_i^2}{n} - (\bar{x})^2} = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$$

Thus, $\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum(x_i)^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$

(ii) Short-cut Method: If observations are larger, select a = any suitable number and take $u_i = (x_i - a)$, then: $\sigma = \sqrt{\frac{\sum\left[(a + u_i) - \left(\frac{\sum u_i}{n} + a\right)\right]}{n}}$

$$\Rightarrow \quad \sigma = \sqrt{\frac{\sum(u_i - \bar{u})}{n}} = \sqrt{\frac{\sum u_i^2}{n} - \left(\frac{\sum u_i}{n}\right)^2};\ u_i = (x_i = a)$$

(iii) Step Deviation Method: Take $u_i = \frac{x_i - a}{h}$; a and h are suitably choosen real numbers, then:

$$\sigma = h\sqrt{\frac{\sum_{i=1}^{n}(u_i - \bar{u})^2}{n}} = h\sqrt{\frac{\sum u_i^2}{n} - \left(\frac{\sum u_i}{n}\right)^2}$$

(b) For discrete series with frequency distribution or continuous series:

(i) **Direct Method**: $\sigma = \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{\sum f_i}} = \sqrt{\frac{\sum f_i x_i^2}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^2}$

(ii) **Short-cut Method:** Take $u_i = (x_i - a)$; $\sigma = \sqrt{\frac{\sum f_i u_i^2}{\sum f_i} - \left(\frac{\sum f_i u_i}{\sum f_i}\right)^2}$

(iii) **Step Deviation Method:** Take $u_i = \frac{x_i - a}{h}$; $\sigma = h\sqrt{\frac{\sum f_i u_i^2}{\sum f_i} - \left(\frac{\sum f_i u_i}{\sum f_i}\right)^2}$

In case of continuous series, x_i = class-mark of i^{th} class-interval.

Remark:

S.D. of first n-natural numbers = $\sqrt{\frac{n^2 - 1}{12}}$

32.11 VARIANCE

That is, variance of a statistical data is square of standard deviation, i.e., variance = $(\text{S.D.})^2 = (\sigma)^2$ or $\sigma = \sqrt{\text{variance}}$. Coefficient of variance (C.V.): $\frac{\text{S.D.}}{\text{Mean}} \times 100 = \frac{\sigma}{x} \times 100$

Note:

C.V. is expressed as per centage.

32.12 COMBINED STANDARD DEVIATION

Let A_1 and A_2 be two series having n_1 and n_2 observations, respectively. Let their A.M be $\bar{x}_1$ and $\bar{x}_2$ and standard deviations be σ_1 and σ_2. Then the combined standard deviation σ or σ_{12} of A_1 and A_2 is given by

$$\sigma \text{ or } \sigma_{12} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2\sigma_2^2}{n_1 + n_2}} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}};$$

where $d_1 = \bar{x}_1 - \bar{x}_{12}, d_2 = \bar{x}_2 - \bar{x}_{12}$ and $\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ is the combined mean.

Remarks:

(i) Coefficient of variation and consistency are reciprocal of each other. Higher is the C.V., lower will be the consistency (stability); again, lower is the C.V, higher will be the stability.

(ii) If we are given scores of two players and the number of matches, in which the given scores were attained, and we are asked to find better run getter, the player with best average (mean). Also we are asked to find most stable player or most consistent player, the player with lower C.V. (Coefficient of variation).