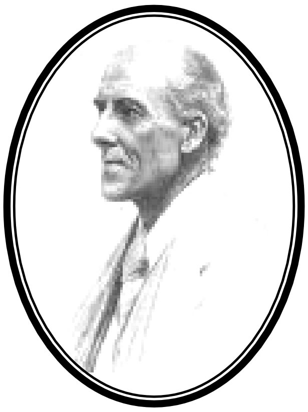
Chapter 2

Statistics

CONTENTS



Karl Pearson

Statistics deals with data collected for specific purposes. Usually the data collected are in raw form, which on processing (organization and classification in the form of ungrouped or grouped data) reveal certain salient features or characteristics of the group. We represent data by bar-charts, pie-charts, histograms, frequency polygons and ogives because such representations are eye-catching and depict glaring features/differences in the data at a glance.

Karl Pearson gave an important formula for coefficient of correlation. Spearman gave the phenomenon of Rank correlation.

2.1 Measures of Central Tendency

2.1.1 Introduction

An average or a central value of a statistical series in the value of the variable which describes the characteristics of the entire distribution.

The following are the five measures of central tendency.

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

2.1.2 Arithmetic Mean

Arithmetic mean is the most important among the mathematical mean.

According to Horace Secrist,

"The arithmetic mean is the amount secured by dividing the sum of values of the items in a series by their number."

(1) **Simple arithmetic mean in individual series (Ungrouped data)**

(i) **Direct method :** If the series in this case be $x_1, x_2, x_3, \ldots\ldots, x_n$ then the arithmetic mean $\bar{x}$ is given by

$$\bar{x} = \frac{\text{Sum of the series}}{\text{Number of terms}}, \quad i.e., \ \bar{x} = \frac{x_1 + x_2 + x_3 + \ldots. + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(ii) **Short cut method**

Arithmetic mean $(\bar{x}) = A + \frac{\Sigma d}{n}$,

where, A = assumed mean, d = deviation from assumed mean = $x - A$, where x is the individual item,

Σd = sum of deviations and n = number of items.

(2) **Simple arithmetic mean in continuous series (Grouped data)**

(i) **Direct method :** If the terms of the given series be $x_1, x_2, \ldots., x_n$ and the corresponding frequencies be $f_1, f_2, \ldots. f_n$, then the arithmetic mean $\bar{x}$ is given by,

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots. + f_n x_n}{f_1 + f_2 + \ldots. + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}.$$

(ii) **Short cut method :** Arithmetic mean $(\bar{x}) = A + \frac{\Sigma f(x - A)}{\Sigma f}$

Where A = assumed mean, f = frequency and $x - A$ = deviation of each item from the assumed mean.

(3) **Properties of arithmetic mean**

(i) Algebraic sum of the deviations of a set of values from their arithmetic mean is zero. If x_i / f_i, i = 1, 2, ..., n is the frequency distribution, then

$\sum_{i=1}^{n} f_i(x_i - \bar{x}) = 0$, $\bar{x}$ being the mean of the distribution.

(ii) The sum of the squares of the deviations of a set of values is minimum when taken about mean.

(iii) **Mean of the composite series :** If $\bar{x}_i, (i = 1, 2, ..., k)$ are the means of k-component series of sizes $n_i, (i = 1, 2, ..., k)$ respectively, then the mean $\bar{x}$ of the composite series obtained on combining the component series is given by the formula $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2 + ... + n_k\bar{x}_k}{n_1 + n_2 + ... + n_k} = \sum_{i=1}^{n} n_i\bar{x}_i \Big/ \sum_{i=1}^{n} n_i$.

2.1.3 Geometric Mean

If $x_1, x_2, x_3, ..., x_n$ are n values of a variate x, none of them being zero, then geometric mean (G.M.) is given by $\text{G.M.} = (x_1.x_2.x_3......x_n)^{1/n} \Rightarrow \log(\text{G.M.}) = \dfrac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n)$.

In case of frequency distribution, G.M. of n values $x_1, x_2, ...x_n$ of a variate x occurring with frequency $f_1, f_2, ..., f_n$ is given by $\text{G.M.} = (x_1^{f_1}.x_2^{f_2}.....x_n^{f_n})^{1/N}$, where $N = f_1 + f_2 + ... + f_n$.

2.1.4 Harmonic Mean

The harmonic mean of n items $x_1, x_2, ..., x_n$ is defined as $\text{H.M.} = \dfrac{n}{\dfrac{1}{x_1} + \dfrac{1}{x_2} + ... + \dfrac{1}{x_n}}$.

If the frequency distribution is $f_1, f_2, f_3, ..., f_n$ respectively, then $\text{H.M.} = \dfrac{f_1 + f_2 + f_3 + ... + f_n}{\left(\dfrac{f_1}{x_1} + \dfrac{f_2}{x_2} + ... + \dfrac{f_n}{x_n}\right)}$

Note : ❑ A.M. gives more weightage to larger values whereas G.M. and H.M. give more weightage to smaller values.

Example: 1 If the mean of the distribution is 2.6, then the value of y is **[Kurukshetra CEE 2001]**

Variate x	1	2	3	4	5
Frequency f of x	4	5	y	1	2

(a) 24 (b) 13 (c) 8 (d) 3

Solution: (c) We know that, Mean $= \dfrac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$

i.e. $2.6 = \dfrac{1\times4 + 2\times5 + 3\times y + 4\times1 + 5\times2}{4 + 5 + y + 1 + 2}$ or $31.2 + 2.6y = 28 + 3y$ or $0.4y = 3.2 \Rightarrow y = 8$

Example: 2 In a class of 100 students there are 70 boys whose average marks in a subject are 75. If the average marks of the complete class are 72, then what are the average marks of the girls **[AIEEE 2002]**

(a) 73 (b) 65 (c) 68 (d) 74

Solution: (b) Let the average marks of the girls students be x, then

$72 = \frac{70 \times 75 + 30 \times x}{100}$ (Number of girls = 100 – 70 = 30)

i.e., $\frac{7200 - 5250}{30} = x$, $\therefore$ $x = 65$.

Example: 3 If the mean of the set of numbers $x_1, x_2, x_3, ..., x_n$ is $\bar{x}$, then the mean of the numbers $x_i + 2i$, $1 \le i \le n$ is **[Pb. CET 1988]**

(a) $\bar{x} + 2n$ (b) $\bar{x} + n + 1$ (c) $\bar{x} + 2$ (d) $\bar{x} + n$

Solution: (b) We know that $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ *i.e.*, $\sum_{i=1}^{n} x_i = n\bar{x}$

$$\therefore \frac{\sum_{i=1}^{n}(x_i + 2i)}{n} = \frac{\sum_{i=1}^{n} x_i + 2\sum_{i=1}^{n} i}{n} = \frac{n\bar{x} + 2(1 + 2 + ...n)}{n} = \frac{n\bar{x} + 2\frac{n(n+1)}{2}}{n} = \bar{x} + (n+1)$$

Example: 4 The harmonic mean of 4, 8, 16 is **[AMU 1995]**

(a) 6.4 (b) 6.7 (c) 6.85 (d) 7.8

Solution: (c) H.M. of 4, 8, 16 $= \frac{3}{\frac{1}{4} + \frac{1}{8} + \frac{1}{16}} = \frac{48}{7} = 6.85$

Example: 5 The average of n numbers $x_1, x_2, x_3, ..., x_n$ is M. If x_n is replaced by x', then new average is **[DCE 2000]**

(a) $M - x_n + x'$ (b) $\frac{nM - x_n + x'}{n}$ (c) $\frac{(n-1)M + x'}{n}$ (d) $\frac{M - x_n + x'}{n}$

Solution: (b) $M = \frac{x_1 + x_2 + x_3 ...... x_n}{n}$ *i.e.*

$$nM = x_1 + x_2 + x_3 + ... x_{n-1} + x_n$$
$$nM - x_n = x_1 + x_2 + x_3 + ... x_{n-1}$$
$$\frac{nM - x_n + x'}{n} = \frac{x_1 + x_2 + x_3 + ... x_{n-1} + x'}{n}$$

$\therefore$ New average $= \frac{nM - x_n + x'}{n}$

Example: 6 Mean of 100 items is 49. It was discovered that three items which should have been 60, 70, 80 were wrongly read as 40, 20, 50 respectively. The correct mean is **[Kurukshetra CEE 1994]**

(a) 48 (b) $82\frac{1}{2}$ (c) 50 (d) 80

Solution: (c) Sum of 100 items $= 49 \times 100 = 4900$

Sum of items added $= 60 + 70 + 80 = 210$

Sum of items replaced $= 40 + 20 + 50 = 110$

New sum $= 4900 + 210 - 110 = 5000$

$\therefore$ Correct mean $= \frac{5000}{100} = 50$

2.1.5 Median

Median is defined as the value of an item or observation above or below which lies on an equal number of observations *i.e.,* the median is the central value of the set of observations provided all the observations are arranged in the ascending or descending orders.

(1) **Calculation of median**

(i) **Individual series :** If the data is raw, arrange in ascending or descending order. Let n be the number of observations.

If n is odd, Median = value of $\left(\frac{n+1}{2}\right)^{th}$ item.

If n is even, Median = $\frac{1}{2}\left[\text{value of}\left(\frac{n}{2}\right)^{\text{th}}\text{ item + value of}\left(\frac{n}{2}+1\right)^{\text{th}}\text{ item}\right]$

(ii) **Discrete series :** In this case, we first find the cumulative frequencies of the variables arranged in ascending or descending order and the median is given by

Median = $\left(\frac{n+1}{2}\right)^{th}$ observation, where n is the cumulative frequency.

(iii) **For grouped or continuous distributions :** In this case, following formula can be used

(a) For series in ascending order, Median = $l+\frac{\left(\frac{N}{2}-C\right)}{f}\times i$

Where l = Lower limit of the median class

f = Frequency of the median class

N = The sum of all frequencies

i = The width of the median class

C = The cumulative frequency of the class preceding to median class.

(b) For series in descending order

Median = $u-\left(\frac{\frac{N}{2}-C}{f}\right)\times i$, where u = upper limit of the median class

$$N=\sum_{i=1}^{n} f_i$$

As median divides a distribution into two equal parts, similarly the quartiles, quantiles, deciles and percentiles divide the distribution respectively into 4, 5, 10 and 100 equal parts. The j^{th} quartile is given by $Q_j=l+\left(\frac{j\frac{N}{4}-C}{f}\right)i; j=1,2,3$. Q_1 is the lower quartile, Q_2 is the median and Q_3 is called the upper quartile.

(2) **Lower quartile**

(i) **Discrete series :** $Q_1 = \text{size of}\left(\frac{n+1}{4}\right)^{\text{th}}\text{ item}$

(ii) **Continuous series :** $Q_1 = l+\frac{\left(\frac{N}{4}-C\right)}{f}\times i$

(3) **Upper quartile**

(i) **Discrete series :** Q_3 = size of $\left[\frac{3(n+1)}{4}\right]^{th}$ item

(ii) **Continuous series :** $Q_3 = l + \frac{\left(\frac{3N}{4} - C\right)}{f} \times i$

(4) **Decile :** Decile divide total frequencies N into ten equal parts.

$$D_j = l + \frac{\frac{N \times j}{10} - C}{f} \times i \quad [j = 1, 2, 3, 4, 5, 6, 7, 8, 9]$$

If $j = 5$, then $D_5 = l + \frac{\frac{N}{2} - C}{f} \times i$. Hence D_5 is also known as median.

(5) **Percentile :** Percentile divide total frequencies N into hundred equal parts

$$P_k = l + \frac{\frac{N \times k}{100} - C}{f} \times i$$

where k = 1, 2, 3, 4, 5,.......,99.

Example: 7 The following data gives the distribution of height of students

Height (in *cm*)	160	150	152	161	156	154	155
Number of students	12	8	4	4	3	3	7

The median of the distribution is

(a) 154 (b) 155 (c) 160 (d) 161

Solution: (b) Arranging the data in ascending order of magnitude, we obtain

Height (in *cm*)	150	152	154	155	156	160	161
Number of students	8	4	3	7	3	12	4
Cumulative frequency	8	12	15	22	25	37	41

Here, total number of items is 41, *i.e.* an odd number. Hence, the median is $\frac{41+1}{2}$ th *i.e.* 21^{st} item.

From cumulative frequency table, we find that median *i.e.* 21^{st} item is 155.

(All items from 16 to 22^{nd} are equal, each = 155)

Example: 8 The median of a set of 9 distinct observation is 20.5. If each of the largest 4 observation of the set is increased by 2, then the median of the new set **[AIEEE 2003]**

(a) Is increased by 2 (b) Is decreased by 2

(c) Is two times the original median (d) Remains the same as that of the original set

Solution: (d) n = 9, then median term $= \left(\frac{9+1}{2}\right)^{th} = 5^{th}$ term . Since last four observation are increased by 2.

∵ The median is 5^{th} observation which is remaining unchanged.

∴ There will be no change in median.

Example: 9 Compute the median from the following table

Marks obtained	**No. of students**

0-10	2
10-20	18
20-30	30
30-40	45
40-50	35
50-60	20
60-70	6
70-80	3

(a) 36.55 (b) 35.55 (c) 40.05 (d) None of these

Solution: (a)

Marks obtained	**No. of students**	**Cumulative frequency**
0-10	2	2
10-20	18	20
20-30	30	50
30-40	45	95
40-50	35	130
50-60	20	150
60-70	6	156
70-80	3	159

$n = \sum f = 159$

Here n = 159, which is odd.

Median number $= \frac{1}{2}(n+1) = \frac{1}{2}(159+1) = 80$, which is in the class 30-40 (see the row of cumulative frequency 95, which contains 80).

Hence median class is 30-40.

$\therefore$ We have l = Lower limit of median class = 30

f = Frequency of median class = 45

C = Total of all frequencies preceding median class = 50

i = Width of class interval of median class = 10

$\therefore$ Required median $= l + \frac{\frac{N}{2} - C}{f} \times i = 30 + \frac{\frac{159}{2} - 50}{45} \times 10 = 30 + \frac{295}{45} = 36.55$.

2.1.6 Mode

Mode : The mode or model value of a distribution is that value of the variable for which the frequency is maximum. For continuous series, mode is calculated as, Mode $= l_1 + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] \times i$

Where, l_1 = The lower limit of the model class

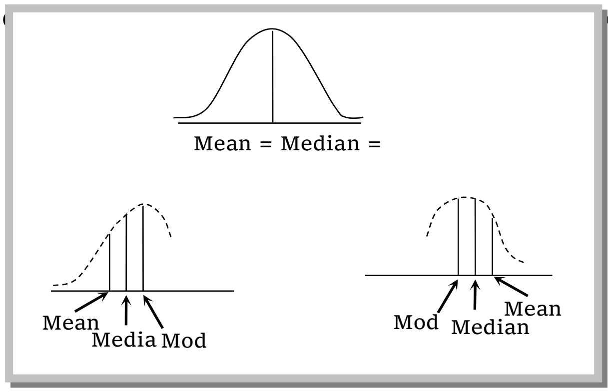
f_1 = The frequency of the model class

f_0 = The frequency of the class preceding the model class

f_2 = The frequency of the class succeeding the model class

i = The size of the model class.

Symmetric distribution : A symmetric is a symmetric distribution if the values of mean, mode and median coincide. In a symmetric distribution frequencies are symmetrically distributed on both sides e.



A distribution which is not symmetric is called a skewed-distribution. In a moderately asymmetric the interval between the mean and the median is approximately one-third of the interval between the mean and the mode *i.e.* we have the following empirical relation between them

Mean – Mode = 3(Mean – Median) $\Rightarrow$ Mode = 3 Median – 2 Mean. It is known as Empirical relation.

Example: 10 The mode of the distribution **[AMU 1988]**

Marks	4	5	6	7	8
No. of students	6	7	10	8	3

(a) 5 (b) 6 (c) 8 (d) 10

Solution: (b) Since frequency is maximum for 6

$\therefore$ Mode = 6

Example: 11 Consider the following statements **[AIEEE 2004]**

(1) Mode can be computed from histogram

(2) Median is not independent of change of scale

(3) Variance is independent of change of origin and scale

Which of these is/are correct

(a) (1), (2) and (3) (b) Only (2) (c) Only (1) and (2) (d) Only (1)

Solution: (d) It is obvious.

Important Tips

☞ ***Some points about arithmetic mean***

- *Of all types of averages the arithmetic mean is most commonly used average.*
- *It is based upon all observations.*
- *If the number of observations is very large, it is more accurate and more reliable basis for comparison.*

☞ ***Some points about geometric mean***

- *It is based on all items of the series.*
- *It is most suitable for constructing index number, average ratios, percentages etc.*
- *G.M. cannot be calculated if the size of any of the items is zero or negative.*

☞ ***Some points about H.M.***

- *It is based on all item of the series.*
- *This is useful in problems related with rates, ratios, time etc.*
- $A.M. \geq G.M. \geq H.M.$ and also $(G.M.)^2 = (A.M.)(H.M.)$

☞ ***Some points about median***

- *It is an appropriate average in dealing with qualitative data, like intelligence, wealth etc.*
- *The sum of the deviations of the items from median, ignoring algebraic signs, is less than the sum from any other point.*

☞ ***Some points about mode***

- *It is not based on all items of the series.*
- *As compared to other averages mode is affected to a large extent by fluctuations of sampling,.*
- *It is not suitable in a case where the relative importance of items have to be considered.*

2.1.7 Pie Chart (Pie Diagram)

Here a circle is divided into a number of segments equal to the number of components in the corresponding table. Here the entire diagram looks like a pie and the components appear like slices cut from the pie. In this diagram each item has a sector whose area has the same percentage of the total area of the circle as this item has of the total of such items. For example if N be the total and n_1 is one of the components of the figure corresponding to a particular item, then the angle of the sector for this item $= \left(\frac{n_1}{N}\right) \times 360°$, as the total number of degree in the angle subtended by the whole circular arc at its centre is 360°.

Example: 12 If for a slightly assymetric distribution, mean and median are 5 and 6 respectively. What is its mode**[DCE 199**

(a) 5 (b) 6 (c) 7 (d) 8

Solution: (d) We know that

Mode = 3Median – 2Mean

= 3(6) – 2(5) = 8

Example: 13 A pie chart is to be drawn for representing the following data

Items of expenditure	Number of families
Education	150
Food and clothing	400
House rent	40
Electricity	250
Miscellaneous	160

The value of the central angle for food and clothing would be **[NDA 1998]**

(a) 90° (b) 2.8° (c) 150° (d) 144°

Solution: (d) Required angle for food and clothing $= \frac{400}{1000} \times 360° = 144°$

2.1.8 Measure of Dispersion

The degree to which numerical data tend to spread about an average value is called the dispersion of the data. The four measure of dispersion are

(1) Range (2) Mean deviation (3) Standard deviation (4) Square deviation

(1) **Range :** It is the difference between the values of extreme items in a series. Range = $X_{max} - X_{min}$

The coefficient of range (scatter) $= \frac{x_{max} - x_{min}}{x_{max} + x_{min}}$.

Range is not the measure of central tendency. Range is widely used in statistical series relating to quality control in production.

(i) **Inter-quartile range :** We know that quartiles are the magnitudes of the items which divide the distribution into four equal parts. The inter-quartile range is found by taking the difference between third and first quartiles and is given by the formula

Inter-quartile range $= Q_3 - Q_1$

Where Q_1 = First quartile or lower quartile and Q_3 = Third quartile or upper quartile.

(ii) **Percentile range :** This is measured by the following formula

Percentile range $= P_{90} - P_{10}$

Where P_{90} = 90th percentile and P_{10} = 10th percentile.

Percentile range is considered better than range as well as inter-quartile range.

(iii) **Quartile deviation or semi inter-quartile range :** It is one-half of the difference between the third quartile and first quartile *i.e.*, $\text{Q.D.} = \frac{Q_3 - Q_1}{2}$ and coefficient of quartile deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$.

Where, Q_3 is the third or upper quartile and Q_1 is the lowest or first quartile.

(2) **Mean deviation :** The arithmetic average of the deviations (all taking positive) from the mean, median or mode is known as mean deviation.

(i) **Mean deviation from ungrouped data (or individual series)**

Mean deviation $= \frac{\sum |x - M|}{n}$

Where $|x - M|$ means the modulus of the deviation of the variate from the mean (mean, median or mode). M and n is the number of terms.

(ii) **Mean deviation from continuous series :** Here first of all we find the mean from which deviation is to be taken. Then we find the deviation $dM = |x - M|$ of each variate from the mean M so obtained.

Next we multiply these deviations by the corresponding frequency and find the product $f.dM$ and then the sum $\sum f\,dM$ of these products.

Lastly we use the formula, mean deviation $= \frac{\sum f |x - M|}{n} = \frac{\sum f\,dM}{n}$, where $n = \Sigma f$.

Important Tips

☞ *Mean coefficient of dispersion* $= \frac{\text{Mean deviation from the mean}}{\text{Mean}}$

- *Median coefficient of dispersion* $= \dfrac{\text{Mean deviation from the median}}{\text{Median}}$
- *Mode coefficient of dispersion* $= \dfrac{\text{Mean deviation from the mode}}{\text{Mode}}$
- *In general, mean deviation (M.D.) always stands for mean deviation about median.*

(3) **Standard deviation :** Standard deviation (or S.D.) is the square root of the arithmetic mean of the square of deviations of various values from their arithmetic mean and is generally denoted by σ read as sigma.

(i) **Coefficient of standard deviation :** To compare the dispersion of two frequency distributions the relative measure of standard deviation is computed which is known as coefficient of standard deviation and is given by

Coefficient of S.D. $= \dfrac{\sigma}{\bar{x}}$, where $\bar{x}$ is the A.M.

(ii) **Standard deviation from individual series**

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$$

where, $\bar{x}$ = The arithmetic mean of series

N = The total frequency.

(iii) **Standard deviation from continuous series**

$$\sigma = \sqrt{\frac{\Sigma f_i(x_i-\bar{x})^2}{N}}$$

where, $\bar{x}$ = Arithmetic mean of series

x_i = Mid value of the class

f_i = Frequency of the corresponding x_i

$N = \Sigma f$ = The total frequency

Short cut method

(i) $\sigma = \sqrt{\dfrac{\Sigma fd^2}{N} - \left(\dfrac{\Sigma fd}{N}\right)^2}$ (ii) $\sigma = \sqrt{\dfrac{\Sigma d^2}{N} - \left(\dfrac{\Sigma d}{N}\right)^2}$

where, $d = x - A$ = Deviation from the assumed mean A

f = Frequency of the item

$N = \Sigma f$ = Sum of frequencies

(4) **Square deviation**

(i) **Root mean square deviation**

$$S = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^2}$$

where A is any arbitrary number and S is called mean square deviation.

(ii) **Relation between S.D. and root mean square deviation :** If σ be the standard deviation and S be the root mean square deviation.

Then $S^2 = \sigma^2 + d^2$.

Obviously, S^2 will be least when $d = 0$ *i.e.* $\bar{x} = A$

Hence, mean square deviation and consequently root mean square deviation is least, if the deviations are taken from the mean.

2.1.9 Variance

The square of standard deviation is called the variance.

Coefficient of standard deviation and variance : The coefficient of standard deviation is the ratio of the S.D. to A.M. *i.e.*, $\frac{\sigma}{x}$. Coefficient of variance = coefficient of S.D.$\times 100 = \frac{\sigma}{\overline{x}} \times 100$.

Variance of the combined series : If $n_1; n_2$ are the sizes, $\overline{x}_1; \overline{x}_2$ the means and $\sigma_1; \sigma_2$ the standard deviation of two series, then $\sigma^2 = \frac{1}{n_1 + n_2}[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)]$

Where, $d_1 = \overline{x}_1 - \overline{x}$, $d_2 = \overline{x}_2 - \overline{x}$ and $\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}$.

Important Tips

- *Range is widely used in statistical series relating to quality control in production.*
- *Standard deviation ≤ Range i.e., variance ≤ (Range)².*
- *Empirical relations between measures of dispersion*
  - *Mean deviation* $= \frac{4}{5}$ *(standard deviation)*
  - *Semi interquartile range* $= \frac{2}{3}$ *(standard deviation)*
- *Semi interquartile range* $= \frac{5}{6}$ *(mean deviation)*
- *For a symmetrical distribution, the following area relationship holds good*

  $\overline{X} \pm \sigma$ *covers 68.27% items*

  $\overline{X} \pm 2\sigma$ *covers 95.45% items*

  $\overline{X} \pm 3\sigma$ *covers 99.74% items*
- *S.D. of first n natural numbers is* $\sqrt{\frac{n^2 - 1}{12}}$.
- *Range is not the measure of central tendency.*

2.1.10 Skewness

"Skewness" measures the lack of symmetry. It is measured by $\gamma_1 = \frac{\Sigma(x_i - \mu)^3}{\{\Sigma(x_i - \mu^2)\}^{3/2}}$ and is denoted by γ_1.

The distribution is skewed if,

(i) Mean ≠ Median ≠ Mode

(ii) Quartiles are not equidistant from the median and

(iii) The frequency curve is stretched more to one side than to the other.

(1) **Distribution :** There are three types of distributions

(i) **Normal distribution :** When $\gamma_1 = 0$, the distribution is said to be normal. In this case

Mean = Median = Mode

(ii) **Positively skewed distribution :** When $\gamma_1 > 0$, the distribution is said to be positively skewed. In this case

Mean > Median > Mode

(iii) **Negative skewed distribution :** When $\gamma_1 < 0$, the distribution is said to be negatively skewed. In this case

Mean < Median < Mode

(2) **Measures of skewness**

(i) **Absolute measures of skewness :** Various measures of skewness are

(a) $S_K = M - M_d$ (b) $S_K = M - M_o$ (c) $S_k = Q_3 + Q_1 - 2M_d$

where, M_d = median, M_o = mode, M = mean

Absolute measures of skewness are not useful to compare two series, therefore relative measure of dispersion are used, as they are pure numbers.

(3) **Relative measures of skewness**

(i) **Karl Pearson's coefficient of skewness :** $S_k = \frac{M - M_o}{\sigma} = 3\frac{(M - M_d)}{\sigma}$, $-3 \le S_k \le 3$, where σ is standard deviation.

(ii) **Bowley's coefficient of skewness :** $S_k = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$

Bowley's coefficient of skewness lies between –1 and 1.

(iii) **Kelly's coefficient of skewness :** $S_K = \frac{P_{10} + P_{90} - 2M_d}{P_{90} - P_{10}} = \frac{D_1 + D_9 - 2M_d}{D_9 - D_1}$

Example: 14 A batsman scores runs in 10 innings 38, 70, 48, 34, 42, 55, 63, 46, 54, 44, then the mean deviation is **[Kerala En**

(a) 8.6 (b) 6.4 (c) 10.6 (d) 9.6

Solution: (a) Arranging the given data in ascending order, we have

34, 38, 42, 44, 46, 48, 54, 55, 63, 70,

Here median M = $\frac{46+48}{2} = 47$ ($\because n = 10$, median is the mean of 5th and 6th items)

$\therefore$ Mean deviation $= \frac{\Sigma|x_i - M|}{n} = \frac{\Sigma|x_i - 47|}{10} = \frac{13+9+5+3+1+1+7+8+16+23}{10} = 8.6$

Example: 15 S.D. of data is 6 when each observation is increased by 1, then the S.D. of new data is **[Pb. CET 1994]**

(a) 5 (b) 7 (c) 6 (d) 8

Solution: (c) S.D. and variance of data is not changed, when each observation is increased (OR decreased) by the same constant.

Example: 16 In a series of 2*n* observations, half of them equal *a* and remaining half equal –*a*. If the standard deviation of the observations is 2, then |*a*| equals **[AIEEE 2004]**

(a) $\frac{\sqrt{2}}{n}$ (b) $\sqrt{2}$ (c) 2 (d) $\frac{1}{n}$

Solution: (c) Let *a*, *a*, ...*n* times – *a*, – *a*, – *a*, – *a*, ----- *n* time *i.e.* mean = 0 and S.D. $= \sqrt{\frac{n(a-0)^2 + n(-a-0)^2}{2n}}$

$2 = \sqrt{\frac{na^2 + na^2}{2n}} = \sqrt{a^2} = \pm a$. Hence $|a| = 2$

Example: 17 If μ is the mean of distribution (y_i, f_i), then $\Sigma f_i(y_i - \mu) =$ **[Kerala PET 2001]**

(a) M.D. (b) S.D. (c) 0 (d) Relative frequency

Solution: (c) We have, $\Sigma f_i(y_i - \mu) = \Sigma f_i y_i - \mu \Sigma f_i = \mu \Sigma f_i - \mu \Sigma f_i = 0$ $\left[\because \mu = \frac{\Sigma f_i y_i}{\Sigma f_i}\right]$

Example: 18 What is the standard deviation of the following series **[DCE 1996]**

Measurements	0-10	10-20	20-30	30-40
Frequency	1	3	4	2

(a) 81 (b) 7.6 (c) 9 (d) 2.26

Solution: (c)

Class	Frequency	y_i	$u_i = \frac{y_i - A}{10}$, $A = 25$	f_iu_i	$f_iu_i^2$
0-10	1	5	– 2	– 2	4
10-20	3	15	– 1	– 3	3
20-30	4	25	0	0	0
30-40	2	35	1	2	2
	10			– 3	9

$$\sigma^2 = c^2\left[\frac{\sum f_iu_i^2}{\sum f_i} - \left(\frac{\sum f_iu_i^2}{\sum f_i}\right)^2\right] = 10^2\left[\frac{9}{10} - \left(\frac{-3}{10}\right)^2\right] = 90 - 9 = 81 \Rightarrow \sigma = 9$$

Example: 19 In an experiment with 15 observations on x, the following results were available $\sum x^2 = 2830$, $\sum x = 170$. On observation that was 20 was found to be wrong and was replaced by the correct value 30. Then the corrected variance is **[AIEEE 2003]**

(a) 78.00 (b) 188.66 (c) 177.33 (d) 8.33

Solution: (a) $\sum x = 170$, $\sum x^2 = 2830$

Increase in $\sum x = 10$, then $\sum x' = 170 + 10 = 180$

Increase in $\sum x^2 = 900 - 400 = 500$, then $\sum x' = 2830 + 500 = 3330$

Variance $= \frac{1}{n}\sum x'^2 - \left(\frac{\sum x'}{n}\right)^2 = \frac{3330}{15} - \left(\frac{180}{15}\right)^2 = 222 - 144 = 78$

Example: 20 The quartile deviation of daily wages (in Rs.) of 7 persons given below 12, 7, 15, 10, 17, 19, 25 is **[Pb. CET 1991, 96; Kurukshetra CEE 1997]**

(a) 14.5 (b) 5 (c) 9 (d) 4.5

Solution: (d) The given data in ascending order of magnitude is 7, 10, 12, 15, 17, 19, 25

Here $Q_1 = \text{size of} \left(\frac{n+1}{4}\right)^{th}$ item = size of 2^{nd} item = 10

$Q_3 = \text{size of} \left(\frac{3(n+1)}{4}\right)^{th}$ item = size of 6^{th} item = 19

Then Q.D. $= \frac{Q_3 - Q_1}{2} = \frac{19 - 10}{2} = 4.5$

Example: 21 Karl-Pearson's coefficient of skewness of a distribution is 0.32. Its S.D. is 6.5 and mean 39.6. Then the median of the distribution is given by **[Kurukshetra CEE 1991]**

(a) 28.61 (b) 38.81 (c) 29.13 (d) 28.31

Solution: (b) We know that $S_k = \frac{M - M_o}{\sigma}$, Where M = Mean, M_o = Mode, σ = S.D.

i.e. $0.32 = \frac{39.6 - M_o}{6.5} \Rightarrow M_o = 37.52$ and also know that, $M_o =$ 3median – 2mean

$37.52 =$ 3(Median) – 2(39.6)

Median = 38.81 (approx.)

Example: 22 The S.D. of a variate x is σ. The S.D. of the variate $\frac{ax+b}{c}$ where a, b, c are constant, is **[Pb. CET 1996]**

(a) $\left(\frac{a}{c}\right)\sigma$ (b) $\left|\frac{a}{c}\right|\sigma$ (c) $\left(\frac{a^2}{c^2}\right)\sigma$ (d) None of these

Solution: (b) Let $y = \frac{ax+b}{c}$ *i.e.*, $y = \frac{a}{c}x + \frac{b}{c}$ *i.e.* $y = Ax + B$, where $A = \frac{a}{c}$, $B = \frac{b}{c}$

$\therefore\ \bar{y} = A\bar{x} + B$

$\therefore\ y - \bar{y} = A(x - \bar{x}) \Rightarrow (y - \bar{y})^2 = A^2(x - \bar{x})^2 \Rightarrow \Sigma(y - \bar{y})^2 = A^2\,\Sigma(x - \bar{x})^2 \Rightarrow n.\sigma_y^2 = A^2.n\sigma_x^2 \Rightarrow \sigma_y^2 = A^2\sigma_x^2$

$\Rightarrow \sigma_y = |A|\,\sigma_x \Rightarrow \sigma_y = \left|\frac{a}{c}\right|\sigma_x$

Thus, new S.D. $= \left|\frac{a}{c}\right|\sigma$.

2.2 Correlation & Regression

2.2.1 Introduction

"If it is proved true that in a large number of instances two variables tend always to fluctuate in the same or in opposite directions, we consider that the fact is established and that a relationship exists. This relationship is called correlation."

(1) **Univariate distribution :** These are the distributions in which there is only one variable such as the heights of the students of a class.

(2) **Bivariate distribution :** Distribution involving two discrete variable is called a bivariate distribution. For example, the heights and the weights of the students of a class in a school.

(3) **Bivariate frequency distribution :** Let x and y be two variables. Suppose x takes the values $x_1, x_2, ..., x_n$ and y takes the values $y_1, y_2, ..., y_n$, then we record our observations in the form of ordered pairs (x_1, y_1), where $1 \le i \le n, 1 \le j \le n$. If a certain pair occurs f_{ij} times, we say that its frequency is f_{ij}.

The function which assigns the frequencies f_{ij}'s to the pairs (x_i, y_j) is known as a bivariate frequency distribution.

Example: 1 The following table shows the frequency distribution of age (x) and weight (y) of a group of 60 individuals

x (yrs) / y (yrs.)	40 – 45	45 – 50	50 – 55	55 – 60	60 – 65
45 – 50	2	5	8	3	0
50 – 55	1	3	6	10	2
55 – 60	0	2	5	12	1

Then find the marginal frequency distribution for x and y.

Solution: Marginal frequency distribution for x

x	40 – 45	45 – 50	50 – 55	55 – 60	60 – 65
f	3	10	19	25	3

Marginal frequency distribution for y

y	45 – 50	50 – 55	55 – 60
f	18	22	20

2.2.2 Covariance

Let $(x_1, x_i); i = 1, 2, ..., n$ be a bivariate distribution, where $x_1, x_2, ..., x_n$ are the values of variable x and $y_1, y_2, ..., y_n$ those of y. Then the covariance $Cov\ (x, y)$ between x and y is given by

$$Cov\,(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad \text{or} \quad Cov\,(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\,\bar{y}) \quad \text{where,} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{are}$$

means of variables x and y respectively.

Covariance is not affected by the change of origin, but it is affected by the change of scale.

Example: 2 Covariance (x, y) between x and y, if $\sum x = 15$, $\sum y = 40$, $\sum x.y = 110$, $n = 5$ is **[DCE 2000]**

(a) 22 (b) 2 (c) – 2 (d) None of these

Solution: (c) Given, $\sum x = 15, \sum y = 40$

$\sum x.y = 110,\ n = 15$

We know that, $Cov\,(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i.y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum x.y - \left(\frac{1}{n}\sum x\right)\left(\frac{1}{n}\sum y\right)$

$= \frac{1}{5}(110) - \left(\frac{15}{5}\right)\left(\frac{40}{5}\right) = 22 - 3\times 8 = -2$.

2.2.3 Correlation

The relationship between two variables such that a change in one variable results in a positive or negative change in the other variable is known as correlation.

(1) **Types of correlation**

(i) **Perfect correlation :** If the two variables vary in such a manner that their ratio is always constant, then the correlation is said to be perfect.

(ii) **Positive or direct correlation :** If an increase or decrease in one variable corresponds to an increase or decrease in the other, the correlation is said to be positive.

(iii) **Negative or indirect correlation :** If an increase or decrease in one variable corresponds to a decrease or increase in the other, the correlation is said to be negative.

(2) **Karl Pearson's coefficient of correlation :** The correlation coefficient $r(x,y)$, between two variable x and y is given by, $r(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$ or $\frac{Cov(x,y)}{\sigma_x\sigma_y}$, $r(x,y) = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{\Sigma dxdy}{\sqrt{\Sigma dx^2}\sqrt{\Sigma dy^2}}.$$

(3) **Modified formula :** $r = \frac{\sum dxdy - \frac{\sum dx.\sum dy}{n}}{\sqrt{\left\{\sum dx^2 - \frac{\left(\sum dx\right)^2}{n}\right\}\left\{\sum dy^2 - \frac{\left(\sum dy\right)^2}{n}\right\}}}$, where $dx = x - \bar{x}; dy = y - \bar{y}$

Also $r_{xy} = \frac{Cov(x,y)}{\sigma_x\sigma_y} = \frac{Cov(x,y)}{\sqrt{\mathrm{var}(x).\,\mathrm{var}(y)}}$.

Example: 3 For the data

x:	4	7	8	3	4
y:	5	8	6	3	5

The Karl Pearson's coefficient is **[Kerala (Engg.) 2002]**

(a) $\frac{63}{\sqrt{94\times 66}}$ (b) 63 (c) $\frac{63}{\sqrt{94}}$ (d) $\frac{63}{\sqrt{66}}$

Solution: (a) Take $A = 5,\ B = 5$

x_i	y_i	$u_i = x_i - 5$	$v_i = y_i - 5$	u_i^2	v_i^2	$u_i v_i$
4	5	– 1	0	1	0	0
7	8	2	3	9	9	6
8	6	3	1	1	1	3
3	3	– 2	– 2	4	4	4
4	5	– 1	0	0	0	0
Total		$\sum u_i = 1$	$\sum v_i = 2$	$\sum u_i^2 = 19$	$\sum v_i^2 = 14$	$\sum u_i v_i = 13$

$$\because \quad r(x,y) = \frac{\sum u_i v_i - \frac{1}{n}\sum u_i \sum v_i}{\sqrt{\sum u_i^2 - \frac{1}{n}\left(\sum u_i\right)^2}\sqrt{\sum v_i - \frac{1}{n}\left(\sum v_i\right)^2}} = \frac{13 - \frac{1 \times 2}{5}}{\sqrt{19 - \frac{1^2}{5}}\sqrt{14 - \frac{2^2}{5}}} = \frac{63}{\sqrt{94}\sqrt{66}}.$$

Example: 4 Coefficient of correlation between observations (1, 6),(2, 5),(3, 4), (4, 3), (5, 2), (6, 1) is **[Pb. CET 1997; Him. CET 2001; DCE 2002]**

(a) 1 (b) – 1 (c) 0 (d) None of these

Solution: (b) Since there is a linear relationship between x and y, *i.e.* $x + y = 7$

$\therefore$ Coefficient of correlation = – 1.

Example: 5 The value of co-variance of two variables x and y is $-\frac{148}{3}$ and the variance of x is $\frac{272}{3}$ and the variance of y is $\frac{131}{3}$. The coefficient of correlation is

(a) 0.48 (b) 0.78 (c) 0.87 (d) None of these

Solution : (d) We know that coefficient of correlation $= \frac{Cov\,(x,y)}{\sigma_x . \sigma_y}$

Since the covariance is – *ive.*

$\therefore$ Correlation coefficient must be – *ive*. Hence (d) is the correct answer.

Example: 6 The coefficient of correlation between two variables x and y is 0.5, their covariance is 16. If the S.D of x is 4, then the S.D. of y is equal to **[AMU 1988, 89, 90]**

(a) 4 (b) 8 (c) 16 (d) 64

Solution: (b) We have, $r_{xy} = 0.5$, $Cov(x,y) = 16$. S.D of x *i.e.*, $\sigma_x = 4$, $\sigma_y = ?$

We know that, $r(x,y) = \frac{Cov(x,y)}{\sigma_x . \sigma_y}$

$$0.5 = \frac{16}{4 . \sigma_y}; \ \therefore \ \sigma_y = 8 .$$

Example: 7 For a bivariate distribution (x, y) if $\sum x = 50$, $\sum y = 60$, $\sum xy = 350$, $\bar{x} = 5, \bar{y} = 6$ variance of x is 4, variance of y is 9, then $r(x,y)$ is **[AMU 1991; Pb. CET 1998; DCE 1998]**

(a) 5/6 (b) 5/36 (c) 11/3 (d) 11/18

Solution: (a) $\bar{x} = \frac{\sum x}{n} \Rightarrow 5 = \frac{50}{n} \Rightarrow n = 10$.

$$\therefore Cov\ (x,y) = \frac{\sum xy}{n} - \bar{x}.\bar{y} = \frac{350}{10} - (5)(6) = 5.$$

$$\therefore \ r(x,y) = \frac{Cov(x,y)}{\sigma_x . \sigma_y} = \frac{5}{\sqrt{4}.\sqrt{9}} = \frac{5}{6}.$$

Example: 8 *A, B, C, D* are non-zero constants, such that

(i) both *A* and *C* are negative. (ii) *A* and *C* are of opposite sign.

If coefficient of correlation between x and y is r, then that between $AX+B$ and $CY+D$ is

(a) r (b) $-r$ (c) $\frac{A}{C}r$ (d) $-\frac{A}{C}r$

Solution : (a,b) (i) Both A and C are negative.

Now $Cov(AX+B, CY+D) = AC\ Cov.(X,Y)$

$\sigma_{AX+B} = |A|\ \sigma_x$ and $\sigma_{CY+D} = |C|\ \sigma_y$

Hence $\rho(AX+B, CY+D) = \frac{AC.Cov(X,Y)}{(|A|\ \sigma_x)(|C|\ \sigma_y)} = \frac{AC}{|AC|}\rho(X,Y) = \rho(X,Y) = r$, $(\because AC>0)$

(ii) $\rho(AX+B, CY+D) = \frac{AC}{|AC|}\rho(X,Y)$, $(\because AC<0)$

$$= \frac{AC}{-AC}\rho(X,Y) = -\rho(X,Y) = -r.$$

2.2.4 Rank Correlation

Let us suppose that a group of n individuals is arranged in order of merit or proficiency in possession of two characteristics A and B.

These rank in two characteristics will, in general, be different.

For example, if we consider the relation between intelligence and beauty, it is not necessary that a beautiful individual is intelligent also.

Rank Correlation : $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$, which is the Spearman's formulae for rank correlation coefficient.

Where $\sum d^2$ = sum of the squares of the difference of two ranks and n is the number of pairs of observations.

Note : ❑ We always have, $\sum d_i = \sum(x_i - y_i) = \sum x_i - \sum y_i = n(\bar{x}) - n(\bar{y}) = 0$, $(\because \bar{x} = \bar{y})$

If all d's are zero, then $r = 1$, which shows that there is perfect rank correlation between the variable and which is maximum value of r.

❑ If however some values of x_i are equal, then the coefficient of rank correlation is given by

$$r = 1 - \frac{6\left[\sum d^2 + \left(\frac{1}{12}\right)(m^3 - m)\right]}{n(n^2-1)}$$

where m is the number of times a particular x_i is repeated.

Positive and Negative rank correlation coefficients

Let r be the rank correlation coefficient then, if

- $r > 0$, it means that if the rank of one characteristic is high, then that of the other is also high or if the rank of one characteristic is low, then that of the other is also low. *e.g.*, if the two characteristics be height and weight of persons, then $r > 0$ means that the tall persons are also heavy in weight.
- $r = 1$, it means that there is perfect correlation in the two characteristics *i.e.*, every individual is getting the same ranks in the two characteristics. Here the ranks are of the type (1, 1), (2, 2),....., (n, n).
- $r < 1$, it means that if the rank of one characteristics is high, then that of the other is low or if the rank of one characteristics is low, then that of the other is high. *e.g.*, if the two characteristics be richness and slimness in person, then $r < 0$ means that the rich persons are not slim.

- $r = -1$, it means that there is perfect negative correlation in the two characteristics *i.e*, an individual getting highest rank in one characteristic is getting the lowest rank in the second characteristic. Here the rank, in the two characteristics in a group of n individuals are of the type $(1, n), (2, n-1), ..., (n, 1)$.
- $r = 0$, it means that no relation can be established between the two characteristics.

Important Tips

- *If $r = 0$, the variable x and y are said to be uncorrelated or independent.*
- *If $r = -1$, the correlation is said to be negative and perfect.*
- *If $r = +1$, the correlation is said to be positive and perfect.*
- *Correlation is a pure number and hence unitless.*
- *Correlation coefficient is not affected by change of origin and scale.*
- *If two variate are connected by the linear relation $x + y = K$, then x, y are in perfect indirect correlation. Here $r = -1$.*
- *If x, y are two independent variables, then* $\rho(x+y, x-y) = \dfrac{\sigma_x^2 - \sigma_y^2}{\sigma_x^2 + \sigma_y^2}$.
- $r(x,y) = \dfrac{\sum u_i v_i - \frac{1}{n}\sum u_i . \sum v_i}{\sqrt{\sum u_i^2 - \frac{1}{n}\left(\sum u_i\right)^2}\sqrt{\sum v_i^2 - \frac{1}{n}\left(\sum v_i\right)^2}}$, where $u_i = x_i - A, v_i = y_i - B$.

Example: 9 Two numbers within the bracket denote the ranks of 10 students of a class in two subjects (1, 10), (2, 9), (3, 8), (4, 7), (5, 6), (6, 5), (7, 4), (8, 3), (9, 2), (10, 1). The rank of correlation coefficient is **[MP PET 1996]**

(a) 0 (b) – 1 (c) 1 (d) 0.5

Solution: (b) Rank correlation coefficient is $r = 1 - 6.\dfrac{\sum d^2}{n(n^2-1)}$, Where $d = y - x$ for pair (x, y)

$\therefore \ \Sigma d^2 = 9^2 + 7^2 + 5^2 + 3^2 + 1^2 + (-1)^2 + (-3)^2 + (-5)^2 + (-7)^2 + (-9)^2 = 330$

Also $n = 10$; $\therefore \ r = 1 - \dfrac{6 \times 330}{10\,(100-1)} = -1$.

Example : 10 Let $x_1, x_2, x_3, ..., x_n$ be the rank of n individuals according to character A and $y_1, y_2, ..., y_n$ the ranks of same individuals according to other character B such that $x_i + y_i = n + 1$ for $i = 1, 2, 3, ..., n$. Then the coefficient of rank correlation between the characters A and B is

(a) 1 (b) 0 (c) – 1 (d) None of these

Solution: (c) $x_i + y_i = n + 1$ for all $i = 1, 2, 3, ..., n$

Let $x_i - y_i = d_i$. Then, $2x_i = n + 1 + d_i \Rightarrow d_i = 2x_i - (n+1)$

$$\therefore \ \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}[2x_i - (n+1)]^2 = \sum_{i=1}^{n}[4x_i^2 + (n+1)^2 - 4x_i(n+1)]$$

$$\sum_{i=1}^{n} d_i^2 = 4\sum_{i=1}^{n} x_i^2 + (n)(n+1)^2 - 4(n+1)\sum_{i=1}^{n} x_i = 4\,\frac{n(n+1)(2n+1)}{6} + (n)(n+1)^2 - 4(n+1)\frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} d_i^2 = \frac{n(n^2-1)}{3}.$$

$$\therefore \ r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(n)(n^2-1)}{3(n)(n^2-1)} \ \ i.e., \ r = -1.$$

Regression

2.2.5 Linear Regression

If a relation between two variates x and y exists, then the dots of the scatter diagram will more or less be concentrated around a curve which is called the **curve of regression**. If this curve be a straight line, then it is known as line of regression and the regression is called **linear regression**.

Line of regression: The line of regression is the straight line which in the least square sense gives the best fit to the given frequency.

2.2.6 Equations of lines of Regression

(1) **Regression line of y on x :** If value of x is known, then value of y can be found as

$$y-\bar{y}=\frac{Cov(x,y)}{\sigma_x^2}(x-\bar{x}) \text{ or } y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

(2) **Regression line of x on y :** It estimates x for the given value of y as

$$x-\bar{x}=\frac{Cov(x,y)}{\sigma_y^2}(y-\bar{y}) \text{ or } x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

(3) **Regression coefficient :** (i) Regression coefficient of y on x is $b_{yx}=\frac{r\sigma_y}{\sigma_x}=\frac{Cov(x,y)}{\sigma_x^2}$

(ii) Regression coefficient of x on y is $b_{xy}=\frac{r\sigma_x}{\sigma_y}=\frac{Cov(x,y)}{\sigma_y^2}$.

2.2.7 Angle between Two lines of Regression

Equation of the two lines of regression are $y-\bar{y}=b_{yx}(x-\bar{x})$ and $x-\bar{x}=b_{xy}(y-\bar{y})$

We have, m_1 = slope of the line of regression of y on x $=b_{yx}=r.\frac{\sigma_y}{\sigma_x}$

m_2 = Slope of line of regression of x on y = $\frac{1}{b_{xy}}=\frac{\sigma_y}{r.\sigma_x}$

$$\therefore \tan\theta=\pm\frac{m_2-m_1}{1+m_1m_2}=\pm\frac{\frac{\sigma_y}{r\sigma_x}-\frac{r\sigma_y}{\sigma_x}}{1+\frac{r\sigma_y}{\sigma_x}.\frac{\sigma_y}{r\sigma_x}}=\pm\frac{(\sigma_y-r^2\sigma_y)\sigma_x}{r\sigma_x^2+r\sigma_y^2}=\pm\frac{(1-r^2)\sigma_x\sigma_y}{r(\sigma_x^2+\sigma_y^2)}.$$

Here the positive sign gives the acute angle θ, because $r^2\le 1$ and σ_x,σ_y are positive.

$$\therefore \quad \tan\theta=\frac{1-r^2}{r}.\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \qquad .....(i)$$

Note : ❑ If $r=0$, from (i) we conclude $\tan\theta=\infty$ or $\theta=\pi/2$ *i.e.*, two regression lines are at right angels.

❑ If $r=\pm1$, $\tan\theta=0$ *i.e.*, $\theta=0$, since θ is acute *i.e.*, two regression lines coincide.

2.2.8 Important points about Regression coefficients b_{xy} and b_{yx}

(1) $r=\sqrt{b_{yx}.b_{xy}}$ *i.e.* the coefficient of correlation is the geometric mean of the coefficient of regression.

(2) If $b_{yx}>1$, then $b_{xy}<1$ *i.e.* if one of the regression coefficient is greater than unity, the other will be less than unity.

(3) If the correlation between the variable is not perfect, then the regression lines intersect at $(\bar{x}, \bar{y})$.

(4) b_{yx} is called the slope of regression line y on x and $\frac{1}{b_{xy}}$ is called the slope of regression line x on y.

(5) $b_{yx} + b_{xy} > 2\sqrt{b_{yx} b_{xy}}$ or $b_{yx} + b_{xy} > 2r$, *i.e.* the arithmetic mean of the regression coefficient is greater than the correlation coefficient.

(6) Regression coefficients are independent of change of origin but not of scale.

(7) The product of lines of regression's gradients is given by $\frac{\sigma_y^2}{\sigma_x^2}$.

(8) If both the lines of regression coincide, then correlation will be perfect linear.

(9) If both b_{yx} and b_{xy} are positive, the r will be positive and if both b_{yx} and b_{xy} are negative, the r will be negative.

Important Tips

- *If $r = 0$, then $\tan\theta$ is not defined i.e. $\theta = \frac{\pi}{2}$. Thus the regression lines are perpendicular.*
- *If $r = +1$ or -1, then $\tan\theta = 0$ i.e. $\theta = 0$. Thus the regression lines are coincident.*
- *If regression lines are $y = ax + b$ and $x = cy + d$, then $\bar{x} = \frac{bc + d}{1 - ac}$ and $\bar{y} = \frac{ad + b}{1 - ac}$.*
- *If b_{yx}, b_{xy} and $r \geq 0$ then $\frac{1}{2}(b_{xy} + b_{yx}) \geq r$ and if b_{xy}, b_{yx} and $r \leq 0$ then $\frac{1}{2}(b_{xy} + b_{yx}) \leq r$.*
- *Correlation measures the relationship between variables while regression measures only the cause and effect of relationship between the variables.*
- *If line of regression of y on x makes an angle α, with the +ive direction of X-axis, then $\tan\alpha = b_{yx}$.*
- *If line of regression of x on y makes an angle β, with the +ive direction of X-axis, then $\cot\beta = b_{xy}$.*

Example : 11 The two lines of regression are $2x - 7y + 6 = 0$ and $7x - 2y + 1 = 0$. The correlation coefficient between x and y is **[DCE 1999]**

(a) – 2/7 (b) 2/7 (c) 4/49 (d) None of these

Solution: (b) The two lines of regression are $2x - 7y + 6 = 0$ .....(i) and $7x - 2y + 1 = 0$ ......(ii)

If (i) is regression equation of y on x, then (ii) is regression equation of x on y.

We write these as $y = \frac{2}{7}x + \frac{6}{7}$ and $x = \frac{2}{7}y - \frac{1}{7}$

$\therefore b_{yx} = \frac{2}{7}, b_{xy} = \frac{2}{7}$; $\therefore b_{yx}.b_{xy} = \frac{4}{49} < 1$, So our choice is valid.

$\therefore r^2 = \frac{4}{49} \Rightarrow r = \frac{2}{7}$. $[\because b_{yx} > 0, b_{xy} > 0]$

Example: 12 Given that the regression coefficients are – 1.5 and 0.5, the value of the square of correlation coefficient is **[Kurukshetra CEE 2002]**

(a) 0.75 (b) 0.7
(c) – 0.75 (d) – 0.5

Solution: (c) Correlation coefficient is given by $r^2 = b_{yx}.b_{xy} = (-1.5)(0.5) = -0.75$.

Example: 13 In a bivariate data $\sum x = 30, \sum y = 400$, $\sum x^2 = 196, \sum xy = 850$ and $n = 10$. The regression coefficient of y on x is **[Kerala (Engg.) 2002]**

(a) – 3.1 (b) – 3.2 (c) – 3.3 (d) – 3.4

Solution: (c) $Cov(x,y) = \frac{1}{n}\sum xy - \frac{1}{n^2}\sum x.\sum y = \frac{1}{10}(850) - \frac{1}{100}(30)(400) = -35$

$$Var(x) = \sigma_x^2 = \frac{1}{n}\sum x^2 - \left(\frac{\sum x}{n}\right)^2 = \frac{196}{10} - \left(\frac{30}{10}\right)^2 = 10.6$$

$$b_{yx} = \frac{Cov(x,y)}{Var(x)} = \frac{-35}{10.6} = -3.3.$$

Example: 14 If two lines of regression are $8x - 10y + 66 = 0$ and $40x - 18y = 214$, then $(\bar{x}, \bar{y})$ is **[AMU 1994; DCE 1994]**

(a) (17, 13) (b) (13, 17) (c) (– 17, 13) (d) (– 13, – 17)

Solution: (b) Since lines of regression pass through $(\bar{x}, \bar{y})$, hence the equation will be $8\bar{x} - 10\bar{y} + 66 = 0$ and $40\bar{x} - 18\bar{y} = 214$

On solving the above equations, we get the required answer $\bar{x} = 13, \bar{y} = 17$.

Example: 15 The regression coefficient of y on x is $\frac{2}{3}$ and of x on y is $\frac{4}{3}$. If the acute angle between the regression line is θ, then $\tan\theta =$

(a) $\frac{1}{18}$ (b) $\frac{1}{9}$ (c) $\frac{2}{9}$ (d) None of these

Solution: (a) $b_{yx} = \frac{2}{3}, b_{xy} = \frac{4}{3}$. Therefore, $\tan\theta = \left|\frac{b_{xy} - \frac{1}{b_{yx}}}{1 + \frac{b_{xy}}{b_{yx}}}\right| = \left|\frac{\frac{4}{3} - \frac{3}{2}}{1 + \frac{4/3}{2/3}}\right| = \frac{1}{18}$.

Example: 16 If the lines of regression of y on x and x on y make angles 30^o and 60^o respectively with the positive direction of X-axis, then the correlation coefficient between x and y is **[MP PET 2002]**

(a) $\frac{1}{\sqrt{2}}$ (b) $\frac{1}{2}$

(c) $\frac{1}{\sqrt{3}}$ (d) $\frac{1}{3}$

Solution: (c) Slope of regression line of y on x = $b_{yx} = \tan 30^o = \frac{1}{\sqrt{3}}$

Slope of regression line of x on y = $\frac{1}{b_{xy}} = \tan 60^o = \sqrt{3}$

$\Rightarrow b_{xy} = \frac{1}{\sqrt{3}}$. Hence, $r = \sqrt{b_{xy}.b_{yx}} = \sqrt{\left(\frac{1}{\sqrt{3}}\right)\left(\frac{1}{\sqrt{3}}\right)} = \frac{1}{\sqrt{3}}$.

Example: 17 If two random variables x and y, are connected by relationship $2x + y = 3$, then $r_{xy} =$ **[AMU 1991]**

(a) 1 (b) – 1 (c) – 2 (d) 3

Solution: (b) Since $2x + y = 3$

$\therefore \quad 2\bar{x} + \bar{y} = 3$; $\therefore \quad y - \bar{y} = -2(x - \bar{x})$. So, $b_{yx} = -2$

Also $x - \bar{x} = -\frac{1}{2}(y - \bar{y})$, $\therefore \quad b_{xy} = -\frac{1}{2}$

$\therefore r_{xy}^2 = b_{yx}.b_{xy} = (-2)\left(-\frac{1}{2}\right) = 1 \Rightarrow r_{xy} = -1$. ($\because$ both b_{yx}, b_{xy} are –*ive*)

2.2.9 Standard error and Probable error

(1) **Standard error of prediction :** The deviation of the predicted value from the observed value is known as the standard error prediction and is defined as $S_y = \sqrt{\left\{\frac{\sum (y - y_p)^2}{n}\right\}}$

where y is actual value and y_p is predicted value.

In relation to coefficient of correlation, it is given by

(i) Standard error of estimate of x is $S_x = \sigma_x \sqrt{1 - r^2}$ (ii) Standard error of estimate of y is $S_y = \sigma_y \sqrt{1 - r^2}$.

(2) **Relation between probable error and standard error :** If r is the correlation coefficient in a sample of n pairs of observations, then its standard error S.E.$(r) = \frac{1 - r^2}{\sqrt{n}}$ and probable error P.E. (r) = 0.6745 (S.E.)= 0.6745 $\left(\frac{1 - r^2}{\sqrt{n}}\right)$. The probable error or the standard error are used for interpreting the coefficient of correlation.

(i) If $r < P.E.(r)$, there is no evidence of correlation.

(ii) If $r > 6P.E.(r)$, the existence of correlation is certain.

The square of the coefficient of correlation for a bivariate distribution is known as the "Coefficient of determination".

Example: 18 If $Var(x) = \frac{21}{4}$ and $Var(y) = 21$ and $r = 1$, then standard error of y is

(a) 0 (b) $\frac{1}{2}$ (c) $\frac{1}{4}$ (d) 1

Solution: (a) $S_y = \sigma_y \sqrt{1 - r^2} = \sigma_y \sqrt{1 - 1} = 0$.