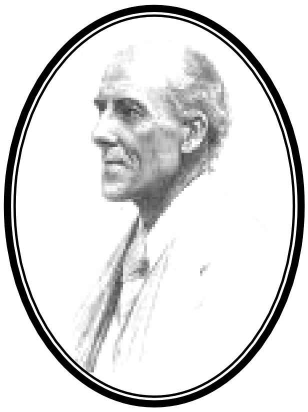
Chapter 2

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

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

(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, $\text{Mean} = \dfrac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$

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

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

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

Example: 3 If the mean of the set of numbers $x_1, x_2, x_3, \ldots\ldots, x_n$ 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.*, $\sum_{i=1}^{n} x_i = n\bar{x}$

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

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 $x_1, x_2, x_3, \ldots\ldots, x_n$ 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) $M = \frac{x_1 + x_2 + x_3 \ldots\ldots x_n}{n}$ *i.e.*

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

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

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

(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 $D_5 = l + \dfrac{\frac{N}{2} - C}{f} \times i$. 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.

$\because$ The median is 5^{th} observation which is remaining unchanged.

$\therefore$ 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

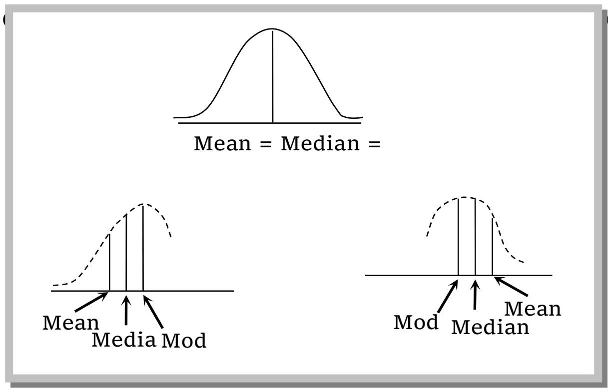
f_1 = The frequency of the model class

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 19**

(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) $= \dfrac{x_{\max} - x_{\min}}{x_{\max} + x_{\min}}$.

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.*, $\text{Q.D.} = \dfrac{Q_3 - Q_1}{2}$ and coefficient of quartile deviation $= \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$.

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 $= \dfrac{\Sigma |x - M|}{n}$

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 $\Sigma f\, dM$ of these products.

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

Important Tips

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

☞ *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.$\times 100 = \frac{\sigma}{\bar{x}} \times 100$.

Variance of the combined series : If $n_1; n_2$ are the sizes, $\bar{x}_1; \bar{x}_2$ 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, $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$ and $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

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*

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

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

  $\bar{X} \pm 3\sigma$ *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 $\sum f_i(y_i - \mu) =$ **[Kerala PET 2001]**

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

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

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 2nd item = 10

$Q_3 = \text{size of}\left(\frac{3(n+1)}{4}\right)^{th}$ item = size of 6th 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 $A = \frac{a}{c}$, $B = \frac{b}{c}$

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

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

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

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 $x_1, x_2, \ldots\ldots, x_n$ and y takes the values $y_1, y_2, \ldots\ldots, y_n$, 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

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

$Cov\,(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ or $Cov\,(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\,\bar{y})$ where, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ 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 $\rho(AX+B, CY+D) = \frac{AC.Cov\,(X,Y)}{(|A|\ \sigma_x)(|C|\ \sigma_y)} = \frac{AC}{|AC|}\rho(X,Y) = \rho(X,Y) = r,\quad (\because AC>0)$

(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

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

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 $(1, n), (2, n-1), \ldots\ldots, (n, 1)$.
- $r = 0$, 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)

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

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

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

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 $r^2\leq 1$ and σ_x,σ_y are positive.

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

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 $r=\pm 1$, $\tan\theta=0$ *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 $2x + y = 3$

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

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

$\therefore\ r_{xy}^2 = b_{yx}.b_{xy} = (-2)\left(-\frac{1}{2}\right) = 1 \Rightarrow r_{xy} = -1$. ($\because$ 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 $\text{S.E.}(r) = \frac{1-r^2}{\sqrt{n}}$ 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) $S_y = \sigma_y \sqrt{1-r^2} = \sigma_y \sqrt{1-1} = 0$.

Assignment

Mean

Basic Level

1. If the mean of 3, 4, x, 7, 10 is 6, then the value of x is
   (a) 4 (b) 5 (c) 6 (d) 7

2. The mean of a set of numbers is $\bar{x}$. If each number is multiplied by λ, then the mean of new set is
   (a) $\bar{x}$ (b) $\lambda + \bar{x}$ (c) $\lambda \bar{x}$ (d) None of these

3. The mean of discrete observations $y_1, y_2, ..., y_n$ is given by **[DCE 1999]**
   (a) $\frac{\sum_{i=1}^{n} y_i}{n}$ (b) $\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} i}$ (c) $\frac{\sum_{i=1}^{n} y_i f_i}{n}$ (d) $\frac{\sum_{i=1}^{n} y_i f_i}{\sum_{i=1}^{n} f_i}$

4. If the mean of numbers 27, 31, 89, 107, 156 is 82, then the mean of 130, 126, 68, 50, 1 is **[Pb. CET 1989; Kurukshetra CEE**
   (a) 75 (b) 157 (c) 82 (d) 80

5. d_i is the deviation of a class mark y_i from 'a' the assumed mean and f_i is the frequency, if $M_g = x + \frac{1}{\Sigma f_i}(\Sigma f_i d_i)$, then x is
   (a) Lower limit (b) Assumed mean (c) Number of observations (d) Class size

6. The mean of a set of observation is $\bar{x}$. If each observation is divided by α, $\alpha \neq 0$ and then is increased by 10, then the n
   (a) $\frac{\bar{x}}{\alpha}$ (b) $\frac{\bar{x}+10}{\alpha}$ (c) $\frac{\bar{x}+10\alpha}{\alpha}$ (d) $\alpha\bar{x}+10$

7. If the mean of the numbers $27+x$, $31+x$, $89+x$, $107+x, 156+x$ is 82, then the mean of $130+x, 126+x, 68+x, 50+x, 1+x$ is **[Kerala PET 2001]**
   (a) 75 (b) 157 (c) 82 (d) 80

8. Consider the frequency distribution of the given numbers

Value :	1	2	3	4
Frequency :	5	4	6	f

   If the mean is known to be 3, then the value of f is **[NDA 2001]**
   (a) 3 (b) 7 (c) 10 (d) 14

9. If the arithmetic mean of the numbers $x_1, x_2, x_3, ..., x_n$ is $\bar{x}$, then the arithmetic mean of numbers $ax_1+b, ax_2+b, ax_3+b, ... ax_n+b$, where a, b are two constants would be **[NDA Sept. 1998]**
   (a) $\bar{x}$ (b) $n\,a\bar{x}+nb$ (c) $a\bar{x}$ (d) $a\bar{x}+b$

10. The mean of n items is $\bar{x}$. If the first term is increased by 1, second by 2 and so on, then new mean is **[DCE 1998]**
   (a) $\bar{x}+n$ (b) $\bar{x}+\frac{n}{2}$ (c) $\bar{x}+\frac{n+1}{2}$ (d) None of these

11. The G.M. of the numbers $3, 3^2, 3^3, ..., 3^n$ is **[Pb. CET 1997]**

(a) $3^{2/n}$ (b) $3^{(n-1)/2}$ (c) $3^{n/2}$ (d) $3^{(n+1)/2}$

12. The reciprocal of the mean of the reciprocals of n observations is their **[AMU 1985]**

(a) A.M. (b) G.M. (c) H.M. (d) None of these

13. The harmonic mean of 3, 7, 8, 10, 14 is

(a) $\frac{3+7+8+10+14}{5}$ (b) $\frac{1}{3}+\frac{1}{7}+\frac{1}{8}+\frac{1}{10}+\frac{1}{14}$ (c) $\frac{\frac{1}{3}+\frac{1}{7}+\frac{1}{8}+\frac{1}{10}+\frac{1}{14}}{4}$ (d) $\frac{5}{\frac{1}{3}+\frac{1}{7}+\frac{1}{8}+\frac{1}{10}+\frac{1}{14}}$

14. If the algebraic sum of deviations of 20 observations from 30 is 20, then the mean of observations is **[NDA (Sept.) 2000]**

(a) 30 (b) 30.1 (c) 29 (d) 31

15. The weighted mean of first n natural numbers whose weights are equal to the squares of corresponding numbers is **[Pb. CET 1989]**

(a) $\frac{n+1}{2}$ (b) $\frac{3n(n+1)}{2(2n+1)}$ (c) $\frac{(n+1)(2n+1)}{6}$ (d) $\frac{n(n+1)}{2}$

16. The mean of the values 0, 1, 2,......,n having corresponding weight ${}^nc_0, {}^nc_1, {}^nc_2, ..., {}^nc_n$ respectively is **[AMU 1990; CET 19**

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

17. If the values $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, ... \frac{1}{n}$ occur at frequencies 1, 2, 3, 4, 5, ... n in a distribution, then the mean is **[NDA 2000]**

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

18. The number of observations in a group is 40. If the average of first 10 is 4.5 and that of the remaining 30 is 3.5, then the average of the whole group is **[AMU 1992; DCE 1996]**

(a) $\frac{1}{5}$ (b) $\frac{15}{4}$ (c) 4 (d) 8

19. A student obtain 75%, 80% and 85% in three subjects. If the marks of another subject are added, then his average cannot be less than

[NDA 2000]

(a) 60% (b) 65% (c) 80% (d) 90%

20. The mean age of a combined group of men and women is 30 years. If the means of the age of men and women are respectively 32 and 27, then the percentage of women in the group is **[NDA Sept. 1998]**

(a) 30 (b) 40 (c) 50 (d) 60

21. The mean monthly salary of the employees in a certain factory is Rs. 500. The mean monthly salary of male and female employees are respectively Rs. 510 and Rs. 460. The percentage of male employees in the factory is **[NDA (Sept.**

(a) 60 (b) 70 (c) 80 (d) 90

22. The A.M. of a 50 set of numbers is 38. If two numbers of the set, namely 55 and 45 are discarded, the A.M. of the remaining set of numbers is **[Kurukshetra CEE 1993]**

(a) 38.5 (b) 37.5 (c) 36.5 (d) 36

23. Mean of 100 observations is 45. It was later found that two observations 19 and 31 were incorrectly recorded as 91 and 13. The correct mean is **[NDA 2001]**

(a) 44.0 (b) 44.46 (c) 45.00 (d) 45.54

24. A car completes the first half of its journey with a velocity v_1 and the rest half with a velocity v_2. Then the average velocity of the car for the whole journey is **[AMU 1989; DCE 1995]**

(a) $\frac{v_1+v_2}{2}$ (b) $\sqrt{v_1 v_2}$ (c) $\frac{2v_1v_2}{v_1+v_2}$ (d) None of these

25. An automobile driver travels from plane to a hill station 120 *km* distant at an average speed of 30 *km* per hour. He then makes the return trip at an average speed of 25 *km* per hour. He covers another 120 *km* distance on plane at an average speed of 50 *km* per hour. His average speed over the entire distance of 300 *km* will be

(a) $\frac{30+25+50}{3}$ *km/hr* (b) $(30, 25, 50)^{\frac{1}{3}}$ (c) $\frac{3}{\frac{1}{30}+\frac{1}{25}+\frac{1}{50}}$ *km/hr* (d) None of these

26. The average weight of students in a class of 35 students is 40 *kg*. If the weight of the teacher be included, the average rises by $\frac{1}{2}$ *kg*; the weight of the teacher is **[Kerala (Engg.) 2002]**

(a) 40.5 *kg* (b) 50 *kg* (c) 41 *kg* (d) 58 *kg*

27. If $\overline{X}_1$ and $\overline{X}_2$ are the means of two distributions such that $\overline{X}_1 < \overline{X}_2$ and $\overline{X}$ is the mean of the combined distribution, then

(a) $\overline{X} < \overline{X}_1$ (b) $\overline{X} > \overline{X}_2$ (c) $\overline{X} = \frac{\overline{X}_1 + \overline{X}_2}{2}$ (d) $\overline{X}_1 < \overline{X} < \overline{X}_2$

28. If a variable takes values 0, 1, 2, ..., *n* with frequencies $q^n, \frac{n}{1}q^{n-1}p, \frac{n(n-1)}{1.2}q^{n-2}p^2, ..., p^n$, where *p* + *q* = 1, then the mean is

(a) *np* (b) *nq* (c) *n*(*p* + *q*) (d) None of these

29. The A.M. of *n* observations is *M*. If the sum of *n* – 4 observations is *a*, then the mean of remaining 4 observations is

(a) $\frac{nM-a}{4}$ (b) $\frac{nM+a}{2}$ (c) $\frac{nM-A}{2}$ (d) *n M* + *a*

Median

Basic Level

30. Which one of the following measures of marks is the most suitable one of central location for computing intelligence of students

[Kurukshetra CEE 1995]

(a) Mode (b) Arithmetic mean (c) Geometric mean (d) Median

31. The central value of the set of observations is called

(a) Mean (b) Median (c) Mode (d) G.M.

32. For a frequency distribution 7th decile is computed by the formula

(a) $D_7 = l + \frac{\left(\frac{N}{7} - C\right)}{f} \times i$ (b) $D_7 = l + \frac{\left(\frac{N}{10} - C\right)}{f} \times i$ (c) $D_7 = l + \frac{\left(\frac{7N}{10} - C\right)}{f} \times i$ (d) $D_7 = l + \frac{\left(\frac{10N}{7} - C\right)}{f} \times i$

33. Which of the following, in case of a discrete data, is not equal to the median

(a) 50th percentile (b) 5th decile (c) 2nd quartile (d) Lower quartile

34. The median of 10, 14, 11, 9, 8, 12, 6 is **[Kurukshetra CEE 1997]**

(a) 10 (b) 12 (c) 14 (d) 11

35. The relation between the median *M*, the second quartile Q_2, the fifth decile D_5 and the 50th percentile P_{50}, of a set of observations is

[AMU 1990]

(a) $M = Q_2 = D_5 = P_{50}$ (b) $M < Q_2 < D_5 < P_{50}$ (c) $M > Q_2 > D_5 > P_{50}$ (d) None of these

36. For a symmetrical distribution $Q_1 = 25$ and $Q_3 = 45$, the median is

(a) 20 (b) 25 (c) 35 (d) None of these

Advance Level

37. If a variable takes the discrete values $\alpha-4, \alpha-\frac{7}{2}, \alpha-\frac{5}{2}, \alpha-3, \alpha-2, \alpha+\frac{1}{2}, \alpha-\frac{1}{2}, \alpha+5\,(\alpha>0)$, then the median is **[DCE 1997; Pb. CET 1988]**

(a) $\alpha-\frac{5}{4}$ (b) $\alpha-\frac{1}{2}$ (c) $\alpha-2$ (d) $\alpha+\frac{5}{4}$

38. The upper quartile for the following distribution

Size of items	1	2	3	4	5	6	7
Frequency	2	4	5	8	7	3	2

is given by the size of

(a) $\left(\frac{31+1}{4}\right)$th item (b) $\left[2\left(\frac{31+1}{4}\right)\right]$th item (c) $\left[3\left(\frac{31+1}{4}\right)\right]$th item (d) $\left[4\left(\frac{31+1}{4}\right)\right]$th item

Mode

Basic Level

39. For a continuous series the mode is computed by the formula

(a) $l+\frac{f_{m-1}}{f_m-f_{m-1}-f_{m+1}}\times C$ or $l+\left(\frac{f_1}{f_m-f_1-f_2}\right)\times i$

(b) $l=\frac{f_m-f_{m-1}}{f_m-f_{m-1}-f_{m+1}}\times C$ or $l+\frac{f_m-f_1}{f_m-f_1-f_2}\times i$

(c) $l+\frac{f_m-f_{m-1}}{2f_m-f_{m-1}-f_{m+1}}\times C$ or $l+\frac{f_m-f_1}{2f_m-f_1-f_2}\times i$

(d) $l+\frac{2f_m-f_{m-1}}{f_m-f_{m-1}-f_{m+1}}\times C$ or $l+\frac{2f_m-f_1}{f_m-f_1-f_2}\times i$

40. A set of numbers consists of three 4's, five 5's, six 6's, eight 8's and seven 10's. The mode of this set of numbers is **[AMU 1989]**

(a) 6 (b) 7 (c) 8 (d) 10

41. The mode of the following items is 0, 1, 6, 7, 2, 3, 7, 6, 6, 2, 6, 0, 5, 6, 0 **[AMU 1995]**

(a) 0 (b) 5 (c) 6 (d) 2

Relation between mean, median and mode

Basic Level

42. If mean = (3 median – mode) k, then the value of k is

(a) 1 (b) 2 (c) $\frac{1}{2}$ (d) $\frac{3}{2}$

43. In a moderately asymmetrical distribution the mode and mean are 7 and 4 respectively. The median is **[NDA Sept. 1998]**

(a) 4 (b) 5 (c) 6 (d) 7

44. If in a moderately asymmetrical distribution mode and mean of the data are 6λ and 9λ respectively, then median is **[Pb. CET 1988]**

(a) 8λ (b) 7λ (c) 6λ (d) 5λ

45. Which of the following is not a measure of central tendency **[Pb. CET 1989]**

(a) Mean (b) Median (c) Mode (d) Range

46. The most stable measure of central tendency is **[AMU 1994]**

(a) Mean (b) Median (c) Mode (d) None of these

47. Which of the following average is most affected of extreme observations **[DCE 1995]**

(a) Mode (b) Median (c) Arithmetic mean (d) Geometric mean

48. The following data was collected from the newspaper : (percentage distribution)

Country	Agriculture	Industry	Services	Others
India	45	19	28	8
U.K.	3	40	44	13
Japan	6	48	43	3
U.S.A.	3	35	61	1

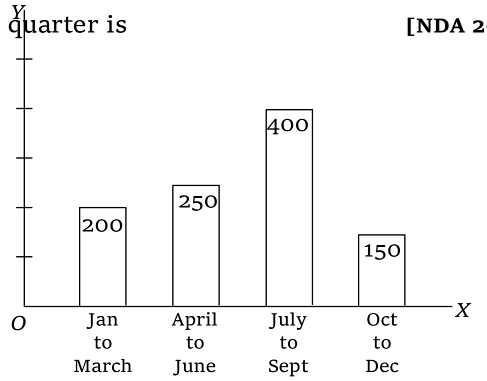
It is an example of **[NDA Sept. 1998]**

(a) Data given in text form (b) Data given in diagrammatic form

(c) Primary data (d) Secondary data

49. The mortality in a town during 4 quarters of a year due to various causes is given below :

Based on this data, the percentage increase in mortality in the third quarter is **[NDA 2000]**



(a) 40
(b) 50
(c) 60
(d) 75

50. A market with 3900 operating firms has the following distribution for firms arranged according to various income groups of workers

Income group	**No. of firms**
150-300	300
300-500	500
500-800	900
800-1200	1000
1200-1800	1200

If a histogram for the above distribution is constructed the highest bar in the histogram would correspond to the class **[NDA Sept. 1998]**

(a) 500-800 (b) 1200-1800 (c) 800-1200 (d) 150-300

51. The total expenditure incurred by an industry under different heads is best presented as a **[NDA 2000]**

(a) Bar diagram (b) Pie diagram (c) Histogram (d) Frequency polygon

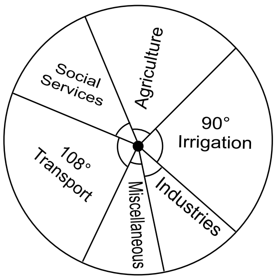
52. The expenditure of a family for a certain month were as follows :

Food – Rs.560, Rent – Rs.420, Clothes – Rs.180, Education – Rs.160, Other items – Rs.120

A pie graph representing this data would show the expenditure for clothes by a sector whose angle equals

(a) 180° (b) 90° (c) 45° (d) 64°

53. Section-wise expenditure of a State Govt. is shown in the given figure. The expenditure incurred on transport is**[NDA (**



(a) 25% (b) 30% (c) 32% (d) 35%

Measures of dispersion

Basic Level

54. The measure of dispersion is **[DCE 1998]**

(a) Mean deviation (b) S.D. (c) Quartile deviation (d) All of these

55. The mean deviation from the median is **[Kurukshetra CEE 1995, 98]**

(a) Greater than that measured from any other value (b) Less than that measured from any other value

(c) Equal to that measured from any other value (d) Maximum if all observations are positive

56. The S.D. of 5 scores 1 2 3 4 5 is **[AMU 1991; DCE 2000]**

(a) $\frac{2}{5}$ (b) $\frac{3}{5}$ (c) $\sqrt{2}$ (d) $\sqrt{3}$

57. The variance of the data 2, 4, 6, 8, 10 is **[AMU 1992]**

(a) 6 (b) 7 (c) 8 (d) None of these

58. The mean deviation of the numbers 3, 4, 5, 6, 7 is **[AMU 1993; DCE 1998]**

(a) 0 (b) 1.2 (c) 5 (d) 25

59. If the standard deviation of 0, 1, 2, 3, ...,9 is K, then the standard deviation of 10, 11, 12, 13 .....19 is

(a) K (b) $K + 10$ (c) $K+\sqrt{10}$ (d) $10K$

60. For a normal distribution if the mean is M, mode is M_0 and median is M_d, then

(a) $M > M_d > M_0$ (b) $M < M_d < M_0$ (c) $M = M_d M_0$ (d) $M = M_d = M_0$

61. For a frequency distribution mean deviation from mean is computed by **[DCE 1994]**

(a) $\text{M.D.} = \frac{\Sigma d}{\Sigma f}$ (b) $\text{M.D.} = \frac{\Sigma fd}{\Sigma f}$ (c) $\text{M.D.} = \frac{\Sigma f|d|}{\Sigma f}$ (d) $\text{M.D.} = \frac{\Sigma f}{\Sigma f|d|}$

62. Let s be the standard deviation of n observations. Each of the n observations is multiplied by a constant c. Then the standard deviation of the resulting numbers is

(a) s (b) cs (c) $s\sqrt{c}$ (d) None of these

63. The S.D. of the first n natural numbers is

(a) $\frac{n+1}{2}$ (b) $\sqrt{\frac{n(n+1)}{2}}$ (c) $\sqrt{\frac{n^2-1}{12}}$ (d) None of these

64. Quartile deviation for a frequency distribution **[DCE 1998]**

(a) $Q = Q_3 - Q_1$ (b) $Q = \frac{1}{2}(Q_3 - Q_1)$ (c) $Q = \frac{1}{3}(Q_3 - Q_1)$ (d) $Q = \frac{1}{4}(Q_2 - Q_1)$

65. The variance of the first n natural numbers is **[AMU 1994; SCRA 2001]**

(a) $\frac{n^2-1}{12}$ (b) $\frac{n^2-1}{6}$ (c) $\frac{n^2+1}{6}$ (d) $\frac{n^2+1}{12}$

66. For a moderately skewed distribution, quartile deviation and the standard deviation are related by **[AMU 1996]**

(a) $\text{S.D.} = \frac{2}{3}\text{Q.D.}$ (b) $\text{S.D.} = \frac{3}{2}\text{Q.D.}$ (c) $\text{S.D.} = \frac{3}{4}\text{Q.D.}$ (d) $\text{S.D.} = \frac{4}{3}\text{Q.D.}$

67. For a frequency distribution standard deviation is computed by applying the formula **[Kurukshetra CEE 1999]**

(a) $\sigma = \sqrt{\left(\frac{\sum fd}{\sum f}\right) - \frac{\sum fd^2}{\sum f}}$ (b) $\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd^2}{\sum f}\right)^2}$ (c) $\sigma = \sqrt{\left(\frac{\sum fd}{\sum f}\right)^2 - \frac{\sum fd^2}{\sum f}}$ (d) $\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$

68. For a frequency distribution standard deviation is computed by

(a) $\sigma = \frac{\sum f(x-\bar{x})}{\sum f}$ (b) $\sigma = \frac{\sqrt{\sum f(x-\bar{x})^2}}{\sum f}$ (c) $\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}}$ (d) $\sigma = \sqrt{\frac{\sum f(x-\bar{x})}{\sum f}}$

69. If Q.D is 16, the most likely value of S.D. will be

(a) 24 (b) 42 (c) 10 (d) None of these

70. If M.D. is 12, the value of S.D. will be

(a) 15 (b) 12 (c) 24 (d) None of these

71. The range of following set of observations 2, 3, 5, 9, 8, 7, 6, 5, 7, 4, 3 is

(a) 11 (b) 7 (c) 5.5 (d) 6

72. If v is the variance and σ is the standard deviation, then **[Kurukshetra CEE 1995]**

(a) $v^2 = \sigma$ (b) $v = \sigma^2$ (c) $v = \frac{1}{\sigma}$ (d) $v = \frac{1}{\sigma^2}$

73. If each observation of a raw data whose variance is σ^2, is increased by λ, then the variance of the new set is

(a) σ^2 (b) $\lambda^2\sigma^2$ (c) $\lambda + \sigma^2$ (d) $\lambda^2 + \sigma^2$

74. If each observation of a raw data whose variance is σ^2, is multiplied by λ, then the variance of the new set is**[Pb. CET**

(a) σ^2 (b) $\lambda^2\sigma^2$ (c) $\lambda + \sigma^2$ (d) $\lambda^2 + \sigma^2$

75. The standard deviation for the set of numbers 1, 4, 5, 7, 8 is 2.45 nearly. If 10 are added to each number, then the new standard deviation will be

(a) 2.45 nearly (b) 24.45 nearly (c) 0.245 nearly (d) 12.45 nearly

76. For a given distribution of marks mean is 35.16 and its standard deviation is 19.76. The co-efficient of variation is

(a) $\frac{35.16}{19.76}$ (b) $\frac{19.76}{35.16}$ (c) $\frac{35.16}{19.76} \times 100$ (d) $\frac{19.76}{35.16} \times 100$

77. If 25% of the item are less than 20 and 25% are more than 40, the quartile deviation is

(a) 20 (b) 30 (c) 40 (d) 10

78. For a normal curve, the greatest ordinate is

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

79. If the variance of observations $x_1, x_2, ... x_n$ is σ^2, then the variance of $ax_1, ax_2 ...., ax_n$, $\alpha \neq 0$ is

(a) σ^2 (b) $a\sigma^2$ (c) $a^2\sigma^2$ (d) $\frac{\sigma^2}{a^2}$

80. The mean deviation from the mean for the set of observations –1, 0, 4 is

(a) $\sqrt{\frac{14}{3}}$ (b) 2 (c) $\frac{2}{3}$ (d) None of these

81. The mean and S.D. of 1, 2, 3, 4, 5, 6 is

(a) $\frac{7}{2}, \sqrt{\frac{35}{12}}$ (b) 3, 3 (c) $\frac{7}{2}, \sqrt{3}$ (d) $3, \frac{35}{12}$

82. The standard deviation of 25 numbers is 40. If each of the numbers is increased by 5, then the new standard deviation will be

[DCE 1995]

(a) 40 (b) 45 (c) $40+\frac{21}{25}$ (d) None of these

83. The S.D of 15 items is 6 and if each item is decreased by 1, then standard deviation will be **[Pb. CET 1998]**

(a) 5 (b) 7 (c) $\frac{91}{15}$ (d) 6

84. The quartile deviation for the data

x :	2	3	4	5	6
f :	3	4	8	4	1

is **[AMU 1988; Kurukshetra CEE 1999]**

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

85. The sum of squares of deviations for 10 observations taken from mean 50 is 250. The co-efficient of variation is**[DCE 1**

(a) 50% (b) 10% (c) 40% (d) None of these

86. One set containing five numbers has mean 8 and variance 18 and the second set containing 3 numbers has mean 8 and variance 24. Then the variance of the combined set of numbers is

(a) 42 (b) 20.25 (c) 18 (d) None of these

87. The means of five observations is 4 and their variance is 5.2. If three of these observations are 1, 2 and 6, then the other two are

[AMU 1994]

(a) 2 and 9 (b) 3 and 8 (c) 4 and 7 (d) 5 and 6

88. The mean of 5 observations is 4.4 and their variance is 8.24. If three observations are 1, 2 and 6, the other two observations are

[AMU 1998]

(a) 4 and 8 (b) 4 and 9 (c) 5 and 7 (d) 5 and 9

89. Consider any set of observations $x_1, x_2, .x_3, ..., x_{101}$; it being given that $x_1 < x_2 < x_3 < .... < x_{100} < x_{101}$; then the mean deviation of this set of observations about a point k is minimum when k equals **[DCE 1997]**

(a) x_1 (b) x_{51} (c) $\frac{x_1+x_2+...+x_{101}}{101}$ (d) x_{50}

90. The mean and S.D of the marks of 200 candidates were found to be 40 and 15 respectively. Later, it was discovered that a score of 40 was wrongly read as 50. The correct mean and S.D respectively are

(a) 14.98, 39.95 (b) 39.95, 14.98 (c) 39.95, 224.5 (d) None of these

91. Let r be the range and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ be the S.D. of a set of observations $x_1, x_2, ...x_n$, then

(a) $S \le r\sqrt{\frac{n}{n-1}}$ (b) $S = r\sqrt{\frac{n}{n-1}}$

(c) $S \ge r\sqrt{\frac{n}{n-1}}$ (d) None of these

92. In any discrete series (when all values are not same) the relationship between M.D. about mean and S.D. is

(a) M.D. = S.D. (b) M.D. ≥ S.D. (c) M.D. < S.D. (d) M.D. ≤ S.D.

93. For (2n +1) observations $x_1, -x_1, x_2, -x_2, ...x_n, -x_n$ and 0 where x's are all distinct. Let S.D. and M.D. denote the standard deviation and median respectively. Then which of the following is always true **[Orissa JEE 2002]**

(a) S.D. < M.D.

(b) S.D. > M.D.

(c) S.D. = M.D.

(d) Nothing can be said in general about the relationship of S.D. and M.D.

94. Suppose values taken by a variable X are such that $a \le x_i \le b$ where x_i denotes the value of X in the i^{th} case for i = 1, 2, ... n. Then

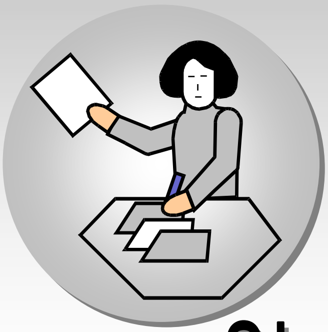
[Kurukshetra CEE 1995, 2000]

(a) $a \le \text{Var}(X) \le b$ (b) $a^2 \le \text{Var}(X) \le b^2$ (c) $\frac{a^2}{4} \le \text{Var}(X)$ (d) $(b-a)^2 \ge \text{Var}(X)$

95. The variance of α, β and γ is 9, then variance of 5α, 5β and 5γ is **[AMU 1998]**

(a) 45 (b) $\frac{9}{5}$ (c) $\frac{5}{9}$ (d) 225

* * *



Answer Sheet

Measures of Central Tendency ***Assignment (Basic and Advance Level)***

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
c	c	a	a	b	c	a	d	d	c	d	c	d	d	b	d	d	b	a	b
21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
c	b	d	c	c	d	d	a	a	d	b	c	d	a	a	c	a	c	c	c
41	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
c	c	b	a	d	a	c	c	c	b	b	c	b	d	b	c	c	b	a	d
61	**62**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**
c	b	c	b	a	b	d	c	a	a	b	b	b	b	a	d	d	d	c	b
81	**82**	**83**	**84**	**85**	**86**	**87**	**88**	**89**	**90**	**91**	**92**	**93**	**94**	**95**					
a	a	d	d	b	b	c	b	b	b	a	b	b	d	d					

Assignment

Basic Level

1. For the bivariate frequency table for x and y

y \ x	0 – 10	10 – 20	20 – 30	30 – 40	Sum
0 – 10	3	2	4	2	11
10 – 20	–	1	3	1	5
20 – 30	3	2	–	–	5
30 – 40	–	6	7	–	13
Sum	6	11	14	3	34

Then the marginal frequency distribution for y is given by

(a)

0 – 10	6
10 – 20	11
20 – 30	14
30 – 40	3

(b)

0 – 10	11
10 – 20	5
20 – 30	5
30 – 40	13

(c)

0 – 10	10
10 – 20	12
20 – 30	11

30 – 40	1

(d) None of these

2. The variables x and y represent height in *cm* and weight in *gm* respectively. The correlation between x and y has the unit **[MP PET 2003]**

(a) *gm* (b) *cm* (c) *gm.cm* (d) None of these

3. The value of $\sum[(x-\bar{x})(y-\bar{y})]$ is

(a) $n.r_{xy}.\sigma_x\sigma_y$ (b) $r_{xy}.\sigma_x^2\sigma_y^2$ (c) $r_{xy}\sqrt{\sigma_x\sigma_y}$ (d) None of these

4. Karl Pearson's coefficient of correlation is dependent **[MP PET 1993]**

(a) Only on the change of origin and not on the change of scale (b) Only on the change of scale and not on the change of origin

(b) On both the change of origin and the change of scale (d) Neither on the change of scale nor on the change of origin

5. If X and Y are independent variable, then correlation coefficient is

(a) 1 (b) – 1 (c) $\frac{1}{2}$ (d) 0

6. The value of the correlation coefficient between two variable lies between **[Kurukshetra CEE 1998]**

(a) 0 and 1 (b) – 1 and 1 (c) 0 and ∞ (d) $-\infty$ and 0

7. The coefficient of correlation between two variables x and y is given by

(a) $r=\frac{\sigma_x^2+\sigma_y^2+\sigma_{x-y}^2}{2\sigma_x\sigma_y}$ (b) $r=\frac{\sigma_x^2+\sigma_y^2-\sigma_{x-y}^2}{2\sigma_x\sigma_y}$ (c) $r=\frac{\sigma_x^2+\sigma_y^2+\sigma_{x-y}^2}{\sigma_x\sigma_y}$ (d) $r=\frac{\sigma_x^2+\sigma_y^2-\sigma_{x-y}^2}{\sigma_x\sigma_y}$

8. If r is the correlation coefficient between two variables, then **[MP PET 1995; Pb. CET 1995]**

(a) $r\ge 1$ (b) $r\le 1$ (c) $|r|\le 1$ (d) $|r|\ge 1$

9. When the correlation between two variables is perfect, then the value of coefficient of correlation r is

(a) -1 (b) $+1$ (c) 0 (d) ± 1

10. If correlation between x and y is r, then between y and x correlation will be

(a) $-r$ (b) $\frac{1}{r}$ (c) r (d) 1- r

11. If r is the coefficient of correlation and $Y=a+bX$, then $|r|=$

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

12. If coefficient of correlation between the variables x and y is zero, then

(a) Variables x and y have no relation (b) y decreases as x increases

(c) y increases as x increases (d) There may be a relation between x and y

13. When the origin is changed, then the coefficient of correlation

(a) Becomes zero (b) Varies (c) Remains fixed (d) None of these

14. If $r=-0.97$, then

(a) Correlation is negative and curved (b) Correlation is linear and negative

(c) Correlation is in third and fourth quadrant (d) None of these

15. In a scatter diagram, if plotted points form a straight line running from the lower left to the upper right corner, then there exists a

(a) High degree of positive correlation (b) Perfect positive correlation

(c) Perfect negative correlation (d) None of these

16. If the two variables x and y of a bivariate distribution have a perfect correlation, they may be connected by **[Kurukshet**

(a) $xy = 1$ (b) $\frac{a}{x} + \frac{b}{y} = 1$ (c) $\frac{x}{a} + \frac{y}{b} = 1$ (d) None of these

17. If x and y are related as $y - 4x = 3$, then the nature of correlation between x and y is

(a) Perfect positive (b) Perfect negative (c) No correlation (d) None of these

18. If $\sum x = 15, \sum y = 36$, $\sum xy = 110$, $n = 5$ then $Cov(x, y)$ equals **[AI CBSE 1991]**

(a) $\frac{1}{5}$ (b) $\frac{-1}{5}$ (c) $\frac{2}{5}$ (d) $-\frac{2}{5}$

19. For a bivariable distribution (x, y), if $\sum xy = 350, \sum x = 50, \sum y = 60, \bar{x} = 5, \bar{y} = 6$, then $Cov(x, y)$ equals **[Pb. CET 1997, AMU 1992]**

(a) 5 (b) 6 (c) 22 (d) 28

20. For covariance the number of variate values in the two given distribution should be **[AMU 1989]**

(a) Unequal (b) Any number in one and any number in the other

(c) Equal (d) None of these

21. If x and y are independent variables, then **[AMU 1994]**

(a) $Cov(x, y) = 1$ (b) $Cov(x, y) = -1$ (c) $Cov(x, y) = 0$ (d) $Cov(x, y) = \pm\frac{1}{2}$

22. If

x :	3	4	8	6	2	1
y :	5	3	9	6	9	2

then the coefficient of correlation will be approximately **[AI CBSE 1990]**

(a) 0.49 (b) 0. 40 (c) – 0. 49 (d) – 0. 40

23. The coefficient of correlation for the following data

x	20	25	30	35	40	45
y	16	10	8	20	5	10

will be **[AI CBSE 1988]**

(a) 0. 32 (b) – 0.32 (c) 0. 35 (d) None of these

24. Coefficient of correlation from the following data

x :	1	2	3	4	5
y :	2	5	7	8	10

will be **[DSSE 1983, AI CBSE 1991]**

(a) 0. 97 (b) – 0.97 (c) 0. 90 (d) None of these

25. Coefficient of correlation between x and y for the following data

x : 15 16 17 17 18 20 10

y : 12 17 15 16 12 15 11

will be approximately **[DSSE 1979, 81; AI CBSE 1990]**

(a) 0. 50 (b) 0. 53 (c) – 0. 50 (d) – 0. 53

26. Karl Pearson's coefficient of correlation between x and y for the following data **[AISSE 1983, 85, 90]**

x : 3 4 8 9 6 2 1

y : 5 3 7 7 6 9 2

(a) 0. 480 (b) – 0. 480 (c) 0. 408 (d) – 0. 408

27. The coefficient of correlation for the following data

x : 1 2 3 4 5 6 7 8 9 10

y : 3 10 5 1 2 9 4 8 7 6

will be **[AISSE 1986, 1990]**

(a) 0. 224 (b) 0. 240 (c) 0. 30 (d) None of these

28. Karl Pearson's coefficient of correlation between the marks in English and Mathematics by ten students

Marks in English	20	13	18	21	11	12	17	14	19	15
Marks in Maths	17	12	23	25	14	8	19	21	22	19

will be **[AISSE 1979, 82]**

(a) 0. 75 (b) – 0. 75 (c) 0. 57 (d) None of these

29. Coefficient of correlation between x and y for the following data

x	–4	–3	–2	–1	0	1	2	3	4
y	16	9	4	1	0	1	4	9	16

will be **[Mathematics Olympiad 1981; DSSE 1980]**

(a) 1 (b) –1 (c) 0 (d) None of these

30. If the variances of two variables x and y are respectively 9 and 16 and their covariance is 8, then their coefficient of correlation is

[MP PET 1998]

(a) $\frac{2}{3}$ (b) $\frac{8}{3\sqrt{2}}$ (c) $\frac{9}{8\sqrt{2}}$ (d) $\frac{2}{9}$

31. If the co-efficient of correlation between x and y is 0. 28, covariance between x and y is 7.6 and the variance of x is 9, then the S.D. of y series is

(a) 9.8 (b) 10. 1 (c) 9.05 (d) 10. 05

32. If $Cov(x, y) = 0$, then $\rho(x,y)$ equals **[AMU 1993]**

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

33. Karl Pearson's coefficient of correlation between the heights (in inches) of teachers and students corresponding to the given data

Height of teachers x :	66	67	68	69	70
Height of students y :	68	66	69	72	70

is **[MP PET 1993]**

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

34. The coefficient of correlation between x and y is 0.6, then covariance is 16. Standard deviation of x is 4, then the standard deviation of y is

(a) 5 (b) 10 (c) 20/3 (d) None of these

35. If $Cov(u,v) = 3,\ \sigma_u^2 = 4.5, \sigma_v^2 = 5.5$, then $\rho(u,v)$ is **[AMU 1988]**

(a) 0.121 (b) 0.603 (c) 0.07 (d) 0.347

36. Given $n = 10, \sum x = 4, \sum y = 3, \sum x^2 = 8, \sum y^2 = 9$ and $\sum xy = 3$, then the coefficient of correlation is **[Pb. CET 1999]**

(a) $\frac{1}{4}$ (b) $\frac{7}{12}$ (c) $\frac{15}{4}$ (d) $\frac{14}{3}$

37. Let r_{xy} be the coefficient of correlation between two variables x and y. If the variable x is multiplied by 3 and the variable y is increased by 2, then the correlation coefficient of the new set of variables is

(a) r_{xy} (b) $3r_{xy}$ (c) $3r_{xy} + 2$ (d) None of these

38. Coefficient of correlation between the two variates X and Y is

X	1	2	3	4	5
Y	5	4	3	2	1

(a) 0 (b) –1 (c) 1 (d) None of these

39. The coefficient of correlation between two variables X and Y is 0.5, their covariance is 15 and $\sigma_x = 6$, then $\sigma_y =$ **[AMU 1998]**

(a) 5 (b) 10 (c) 20 (d) 6

40. Karl Pearson's coefficient of rank correlation between the ranks obtained by ten students in Mathematics and Chemistry in a class test as given below

Rank in Mathematics :	1	2	3	4	5	6	7	8	9	10
Rank in Chemistry :	3	10	5	1	2	9	4	8	7	6

is **[AISSE 1990]**

(a) 0.224 (b) 0.204 (c) 0.240 (d) None of these

41. The sum of squares of differences in ranks of marks obtained in Physics and Chemistry by 10 students in a test is 150, then the co-efficient of rank-correlation is given by

(a) 0.909 (b) 0.091 (c) 0.849 (d) None of these

Advance Level

42. If a, b, h, k are constants, while U and V are $U = \frac{X-a}{h}, V = \frac{Y-b}{k}$, then **[DCE 1999]**

(a) $Cov\ (X, Y) = Cov\ (U, V)$ (b) $Cov\ (X, Y) = hk\ Cov\ (U, V)$

(c) $Cov(X, Y) = ab\, Cov(U, V)$ (d) $Cov(U, V) = hk\, Cov(X, Y)$

43. Let X, Y be two variables with correlation coefficient $\rho(X, Y)$ and variables U, V be related to X, Y by the relation $U = 2X$, $V = 3Y$, then $\rho(U, V)$ is equal to **[AMU 1999]**

(a) $\rho(X, Y)$ (b) $6\rho(X, Y)$ (c) $\sqrt{6}\rho(X, Y)$ (d) $\frac{3}{2}\rho(X, Y)$

44. If X and Y are two uncorrelated variables and if $u = X + Y$, $v = X - Y$, then $r(u, v)$ is equal to **[DCE 1998]**

(a) $\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2 - \sigma_y^2}$ (b) $\frac{\sigma_x^2 - \sigma_y^2}{\sigma_x^2 + \sigma_y^2}$ (c) $\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x \sigma_y}$ (d) None of these

45. If $\bar{x} = \bar{y} = 0, \sum x_i y_i = 12, \sigma_x = 2, \sigma_y = 3$ and $n = 10$, then the coefficient of correlation is **[MP PET 1999]**

(a) 0.4 (b) 0.3 (c) 0.2 (d) 0.1

46. Let X and Y be two variables with the same variance and U and V be two variables such that $U = X + Y$, $V = X - Y$. Then $Cov(U, V)$ is equal to

(a) $Cov(X, Y)$ (b) 0 (c) 1 (d) – 1

Regression

Basic Level

47. If there exists a linear statistical relationship between two variables x and y, then the regression coefficient of y on x is **[MP PET 1998]**

(a) $\frac{cor(x,y)}{\sigma_x . \sigma_y}$

(b) $\frac{cor(x,y)}{\sigma_y^2}$

(c) $\frac{cor(x,y)}{\sigma_x^2}$

(d) $\frac{cor(x,y)}{\sigma_x}$, where σ_x, σ_y are standard deviations of x and y respectively.

48. If $ax + by + c = 0$ is a line of regression of y on x and $a_1 x + b_1 y + c_1 = 0$ that of x on y, then

(a) $a_1 b \le ab_1$ (b) $aa_1 = bb_1$ (c) $ab_1 \le a_1 b$ (d) None of these

49. Least square lines of regression give best possible estimates, when $\rho(X, Y)$ is **[DCE 1996]**

(a) <1 (b) > -1 (c) -1 or 1 (d) None of these

50. Which of the following statement is correct **[Kurukshetra CEE 1995]**

(a) Correlation coefficient is the arithmetic mean of the regression coefficient

(b) Correlation coefficient is the geometric mean of the regression coefficient

(c) Correlation coefficient is the harmonic mean of the regression coefficient

(d) None of these

51. The relationship between the correlation coefficient r and the regression coefficients b_{xy} and b_{yx} is**[MP PET 2003; Pb. C**

(a) $r = \frac{1}{2}(b_{xy} + b_{yx})$ (b) $r = \sqrt{b_{xy} . b_{yx}}$ (c) $r = (b_{xy} b_{yx})^2$ (d) $r = b_{xy} + b_{yx}$

52. If the coefficient of correlation is positive, then the regression coefficients **[Pb. CET 1998; PU CET 2002]**

(a) Both are positive

(b) Both are negative

(c) One is positive and another is negative

(d) None of these

53. If b_{yx} and b_{xy} are both positive (where b_{yx} and b_{xy} are regression coefficients), then **[MP PET 2001]**

(a) $\frac{1}{b_{yx}}+\frac{1}{b_{xy}}<\frac{2}{r}$ (b) $\frac{1}{b_{yx}}+\frac{1}{b_{xy}}>\frac{2}{r}$

(c) $\frac{1}{b_{yx}}+\frac{1}{b_{xy}}<\frac{r}{2}$ (d) None of these

54. If x_1 and x_2 are regression coefficients and r is the coefficient of correlation, then

(a) $x_1-x_2>r$ (b) $x_1+x_2<r$ (c) $x_1+x_2\geq 2r$ (d) None of these

55. If one regression coefficient be unity, then the other will be

(a) Greater than unity (b) Greater than or equal to unity (c) Less than or equal to unity (d)

56. If one regression coefficient be less than unity, then the other will be

(a) Less than unity (b) Equal to unity (c) Greater than unity (d) All of the above

57. If regression coefficient of y on x is 2, then the regression coefficient of x on y is **[AMU 1990]**

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

58. The lines of regression of x on y estimates **[AMU 1993]**

(a) x for a given value of y (b) y for a given value of x (c) x from y and y from x(d)

59. The statistical method which helps us to estimate or predict the unknown value of one variable from the known value of the related variable is called **[Pb. CET 1995]**

(a) Correlation (b) Scatter diagram (c) Regression (d) Dispersion

60. The coefficient of correlation between two variables x and y is 0.8 while regression coefficient of y on x is 0.2. Then the regression coefficient of x on y is **[MP PET 1993]**

(a) –3.2 (b) 3.2 (c) 4 (d) 0.16

61. If the lines of regression coincide, then the value of correlation coefficient is

(a) 0 (b) 1 (c) 0.5 (d) 0.33

62. Two lines of regression are $3x+4y-7=0$ and $4x+y-5=0$. Then correlation coefficient between x and y is**[AI CBSE 199**

(a) $\frac{\sqrt{3}}{4}$ (b) $-\frac{\sqrt{3}}{4}$ (c) $\frac{3}{16}$ (d) $-\frac{3}{16}$

63. If the two lines of regression are $4x+3y+7=0$ and $3x+4y+8=0$, then the means of x and y are **[AI CBSE 1990]**

(a) $-\frac{4}{7},-\frac{11}{7}$ (b) $-\frac{4}{7},\frac{11}{7}$ (c) $\frac{4}{7},-\frac{11}{7}$ (d) 4, 7

64. The two regression lines for a bivariate data are $x+y+50=0$ and $2x+3y+K=0$. If $\bar{x}=0$, then $\bar{y}$ is

[BCA Delhi Entrance Exam. 1999]

(a) 50 (b) $K-100$ (c) – 50 (d) $50+K$

65. The two regression lines are $2x-9y+6=0$ and $x-2y+1=0$. What is the correlation coefficient between x and y **[DCE 1999]**

(a) $-\frac{2}{3}$ (b) $\frac{2}{3}$ (c) $\frac{4}{9}$ (d) None of these

66. If the two regression coefficient between x and y are 0.8 and 0.2, then the coefficient of correlation between them is **[MP PET 2000]**

(a) 0.4 (b) 0.6 (c) 0.3 (d) 0.5

67. The two lines of regression are given by $3x+2y=26$ and $6x+y=31$. The coefficient of correlation between x and y is **[DCE 2000]**

(a) $-\frac{1}{3}$ (b) $\frac{1}{3}$ (c) $-\frac{1}{2}$ (d) $\frac{1}{2}$

68. If the lines of regression be $x-y=0$ and $4x-y-3=0$ and $\sigma_x^2=1$, then the coefficient of correlation is

(a) - 0.5 (b) 0.5 (c) 1.0 (d) - 1.0

69. A student obtained two regression lines as $L_1 \equiv x-5y+7=0$ and $L_2 \equiv 3x+y-8=0$. Then the regression line of y on x is

(a) L_1 (b) L_2 (c) Neither of the two (d) $x-5y=0$

70. If b_{yx} and b_{xy} are regression coefficients of y on x and x on y respectively, then which of the following statement is true

[Pb. CET 1996]

(a) $b_{xy}=1.5, b_{yx}=1.4$ (b) $b_{xy}=1.5, b_{yx}=0.9$ (c) $b_{xy}=1.5, b_{yx}=0.8$ (d) $b_{xy}=1.5, b_{yx}=0.6$

71. Angle between two lines of regression is given by **[Kurukshetra CEE 2000; DCE 1998]**

(a) $\tan^{-1}\left(\dfrac{b_{yx}-\dfrac{1}{b_{xy}}}{1+\dfrac{b_{xy}}{b_{yx}}}\right)$ (b) $\tan^{-1}\left(\dfrac{b_{yx}-b_{xy}-1}{b_{yx}+b_{xy}}\right)$ (c) $\tan^{-1}\left(\dfrac{b_{xy}-\dfrac{1}{b_{yx}}}{1+\dfrac{b_{xy}}{b_{yx}}}\right)$ (d) $\tan^{-1}\left(\dfrac{b_{yx}-b_{xy}}{1+b_{yx}.b_{xy}}\right)$

72. If acute angle between the two regression lines is θ, then

(a) $\sin\theta \ge 1-r^2$ (b) $\tan\theta \ge 1-r^2$ (c) $\sin\theta \le 1-r^2$ (d) $\tan\theta \le 1-r^2$

73. If the angle between the two lines of regression is 90°, then it represents **[DCE 1999]**

(a) Perfect correlation (b) Perfect negative correlation (c) No linear correlation (d)

74. If $2x+y=7$ and $x+2y=7$ are the two regression lines respectively, then the correlation co-efficient between x and y is

[DCE 1983; AMU 1993]

(a) + 1 (b) −1 (c) $+\frac{1}{2}$ (d) $-\frac{1}{2}$

75. For a perfect correlation between the variables x and y, the line of regression is $ax+by+c=0$ where $a,b,c>0$; then $\rho(x,y)=$

[AMU 1999]

(a) 0 (b) −1 (c) 1 (d) None of these

76. If two random variables X and Y of a bivariate distribution are connected by the relationship $3x+2y=4$, then correlation coefficient r_{xy} equals **[AMU 1999]**

(a) 1 (b) −1 (c) 2/3 (d) −2/3

77. Two variables x and y are related by the linear equation $ax+by+c=0$. The coefficient of correlation between the two is +1, if

[DCE 2002]

(a) a is positive (b) b is positive (c) a and b both are positive (d)a and b are of opposite sig

78. If the two lines of regression are $5x+3y=55$ and $7x+y=45$, then the correlation coefficient between x and y is **[AMU 1**

(a) +1 (b) –1 (c) $-\sqrt{\frac{5}{21}}$ (d) $-\sqrt{\frac{21}{5}}$

79. The error of prediction of x from the required line of regression is given by,

(where ρ is the co-efficient of correlation) **[AMU 1992]**

(a) $\sigma_x(1-\rho^2)$ (b) $n\sigma_x^2(1-\rho^2)$ (c) $\sigma_x^2(1-\rho^2)$ (d) $n\sigma_y^2(1-\rho^2)$

80. Probable error of r is

(a) $0.6745\left(\frac{1-r^2}{\sqrt{n}}\right)$ (b) $0.6754\left(\frac{1+r^2}{\sqrt{n}}\right)$ (c) $0.6547\left(\frac{1-r^2}{n}\right)$ (d) $0.6754\left(\frac{1-r^2}{n}\right)$

Advance Level

81. For the following data

	x	y
Mean	65	67
Standard deviation	5.0	2.5
Correlation coefficient	0.8	

Then the equation of line of regression of y on x is

(a) $y-67=\frac{2}{5}(x-65)$ (b) $y-67=\frac{1}{5}(x-65)$ (c) $x-65=\frac{2}{5}(y-67)$ (d) $x-65=\frac{1}{5}(y-67)$

82. If the lines of regression of y on x and that of x on y are $y=kx+4$ and $x=4y+5$ respectively, then

(a) $k\le 0$ (b) $k\ge 0$ (c) $0\le k\le \frac{1}{4}$ (d) $0\le k\le 1$

83. From the following observations $\{(x,y)\}=\{(1,7),(4,5),(7,2),(10,6),(13,5)\}$. The line of regression of y on x is **[AI CBSE 1991]**

(a) $7x+30y-187=0$ (b) $7x-30y-187=0$ (c) $7x-30y+187=0$ (d) None of these

84. If the variance of $x=9$ and regression equations are $4x-5y+33=0$ and $20x-9y-10=0$, then the coefficient of correlation between x and y and the variance of y respectively are **[AMU1997, 2002]**

(a) 0.6; 16 (b) 0.16; 16 (c) 0.3; 4 (d) 0.6; 4

85. If the two lines of regression are $x+4y=3$ and $3x+y=15$, then value of x for $y=3$ is **[DCE 1998]**

(a) 4 (b) –9 (c) – 4 (d) None of these

86. Which of the following two sets of regression lines are the true representative of the information from the bivariate population

I. $x+4y=15$ and $y+3x=12, \bar{x}=3, \bar{y}=3$ II. $3x+4y=9$ and $4x+y=1, \bar{x}=-\frac{5}{10}, \bar{y}=\frac{30}{13}$ **[AMU 2000]**

(a) Both I and II (b) II only (c) I only (d) None of these

87. Out of the two lines of regression given by $x+2y=4$ and $2x+3y-5=0$, the regression line of x on y is **[Kurukshetra CE**

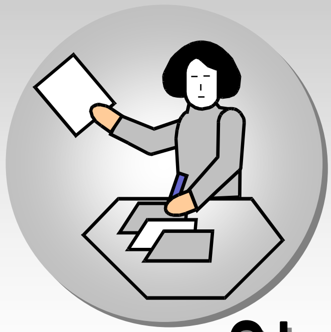
(a) $x+2y=4$ (b) $2x+3y-5=0$

(c) The given lines cannot be the regression lines (d) $x+2y=0$

88. Regression of savings (S) of a family on income Y may be expressed as $S = a + \frac{Y}{m}$, where a and m are constants. In a random sample of 100 families the variance of savings is one-quarter of the variance of incomes and the correlation coefficient is found to be 0.4. The value of m is

(a) 2 (b) 5 (c) 8 (d) None of these

* * *



Answer Sheet

Correction and Regression ***Assignment (Basic and Advance Level)***

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
b	d	a	d	d	b	b	c	d	c	c	a	c	b	b	c	a	c	a	c
21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
c	a	b	a	b	c	a	a	c	a	c	a	a	c	b	b	a	b	a	a
41	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
b	b	a	b	c	b	c	c	c	b	b	a	b	c	c	d	c	a	c	b
61	**62**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**
b	b	a	c	b	a	c	b	c	d	c	c	c	d	b	b	d	c	b	a
81	**82**	**83**	**84**	**85**	**86**	**87**	**88**												
a	c	d	a	a	c	b	b												