Index

Unit – 4 ⇝ Basic Statistics

Method 1 ⇝ Measure of Central Tendency

Introduction

→ Statistics is the branch of science where we plan, gather and analyze information about a particular collection of objects under investigation.

→ Statistics techniques are used in every other field of science, engineering and humanity, ranging from computer science to industrial engineering to sociology and psychology.

→ For any statistical problem the initial information collection from the sample may look messy, and hence confusing. This initial information needs to be organized first before we make any sense out of it.

Central Tendency

→ The central tendency of a distribution is an estimate of the **center** of a distribution of values.

→ There are three measures to estimate central tendency which is

(1) Mean($\overline{x}$)

(2) Median(M)

(3) Mode(Z)

(4) Quartiles

(5) Percentiles

1.1 Mean

→ The mean means **average**.

→ Mean is denoted by " $\bar{x}$ " and read as x bar.

→ Table of different formulae of mean.

Method	Ungrouped Data	Discrete Grouped Data	Continuous Grouped Data
Direct Method	$\dfrac{\sum x_i}{n}$	$\dfrac{\sum f_i x_i}{\sum f_i}$	
Assumed Mean Method	$A + \dfrac{\sum d_i}{n}$	$A + \dfrac{\sum f_i d_i}{\sum f_i}$	
Step Deviation Method	-----	-----	$A + \dfrac{\sum f_i u_i}{\sum f_i} \times c$

→ $n =$ total number of observations

→ In case of continuous frequency distribution,

$x_i =$ mid value of the respective class.

→ In case of **assumed mean method**, A can be any value from x_i.

→ Use below formula to calculate d_i & u_i

$$d_i = x_i - A \quad ; \quad u_i = \frac{x_i - A}{c}$$

Example of Method-1.1: Examples of Mean

<table>
<tr><td>C</td><td>1</td><td>Find the mean of data 10.2, 9.5, 8.3, 9.7, 9.5, 11.1, 7.8, 8.8, 9.5, 10.

Answer: $\mathbf{9.44}$</td></tr>
<tr><td>C</td><td>2</td><td>Find the mean for following data:

<table>
<tr><td>Marks obtained</td><td>20</td><td>9</td><td>25</td><td>50</td><td>40</td><td>80</td></tr>
<tr><td>Number of students</td><td>6</td><td>4</td><td>16</td><td>7</td><td>8</td><td>2</td></tr>
</table>

Answer: $\mathbf{32.23}$</td></tr>
<tr><td>C</td><td>3</td><td>Find the mean using direct method, assumed mean method and step deviation method:

<table>
<tr><td>Marks</td><td>0 – 10</td><td>10 – 20</td><td>20 – 30</td><td>30 – 40</td><td>40 – 50</td></tr>
<tr><td>No. of students</td><td>5</td><td>10</td><td>40</td><td>20</td><td>25</td></tr>
</table>

Answer: $\mathbf{30}$</td></tr>
<tr><td>C</td><td>4</td><td>Find the missing frequency f_1 and f_2 in the table given below, it is being given that the mean of the given frequency distribution is 50.

<table>
<tr><td>Class</td><td>0 – 20</td><td>20 – 40</td><td>40 – 60</td><td>60 – 80</td><td>80 – 100</td><td>Total</td></tr>
<tr><td>f</td><td>17</td><td>f_1</td><td>32</td><td>f_2</td><td>19</td><td>100</td></tr>
</table>

Answer: $\mathbf{f_1 = 18}$, $\mathbf{f_2 = 14}$</td></tr>
<tr><td>C</td><td>5</td><td>A co-operative bank has two branches employing 50 and 70 workers respectively. The average salaries paid by two respective branches are 360 and 390 rupees per month. Calculate the mean of the salaries of all the employees.

Answer: $\mathbf{377.5}$</td></tr>
</table>

1.2 Median

→ The median is the value found at the **exact middle** of the set of values.

→ Median is denoted by capital letter "**M**".

→ To compute the median, list all observations in ascending order and then locate the value in the center of the sample.

→ Table of formula of median.

<table>
<tr><th>Data</th><th>Formula</th></tr>
<tr><td>Ungrouped Data</td><td rowspan="2">If n is odd, then
$M = \left(\frac{n+1}{2}\right)^{th}$ observation

If n is even, then
$M = \frac{\left(\frac{n}{2}\right)^{th} \text{observation} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$</td></tr>
<tr><td>Discrete Grouped Data</td></tr>
<tr><td>Continuous Grouped Data</td><td>$M = L + \left(\frac{\frac{n}{2} - F}{f}\right) \times c$</td></tr>
</table>

Where,

Median class $=$ Class whose cumulative frequency with property min $\left\{ cf \mid cf \geq \frac{n}{2} \right\}$

L = Lower boundary point of the median class

n = Total number of observation (sum of the frequencies)

F = Cumulative frequency of the class preceding the median class

f = The frequency of the median class

Example of Method-1.2: Median

<table>
<tr><td>C</td><td>1</td><td>Find the median of following data:
20, 25, 30, 15, 17, 35, 26, 18, 40, 45, 50.

Answer: 26</td></tr>
<tr><td>C</td><td>2</td><td>The given observations have been arranged in ascending order. If the median of the data is 63, find the value of x for the following data:
$29,\ 32,\ 48,\ 50,\ x,\ x+2,\ 72,\ 78,\ 84,\ 95.$

Answer: $x = 62$</td></tr>
<tr><td>C</td><td>3</td><td>Calculate the median for the following data:
<table><tr><td>Marks</td><td>20</td><td>9</td><td>25</td><td>50</td><td>40</td><td>80</td></tr><tr><td>No. of students</td><td>6</td><td>4</td><td>16</td><td>7</td><td>8</td><td>2</td></tr></table>Answer: 25</td></tr>
<tr><td>C</td><td>4</td><td>The following table gives marks obtained by 50 students in statistics. Find the median.
<table><tr><td>Marks</td><td>0 – 10</td><td>10 – 20</td><td>20 – 30</td><td>30 – 40</td><td>40 – 50</td></tr><tr><td>No. of students</td><td>16</td><td>12</td><td>18</td><td>3</td><td>1</td></tr></table>Answer: 17.5</td></tr>
<tr><td>C</td><td>5</td><td>The median of 60 observations (following data) is 28.5. Find x and y.
<table><tr><td>Marks</td><td>0 – 10</td><td>10 – 20</td><td>20 – 30</td><td>30 – 40</td><td>40 – 50</td><td>50 – 60</td></tr><tr><td>No. of students</td><td>5</td><td>x</td><td>20</td><td>15</td><td>y</td><td>5</td></tr></table>Answer: $x = 8$, $y = 7$</td></tr>
</table>

1.3 Mode

→ The mode is the **most frequently** occurring value in the set.

→ Mode is denoted by capital letter **"Z"**.

→ The mode is not necessarily unique, like mean and median. we can have data with two modes (bi-modal) or more than two modes (multi-modal).

→ Table of formula of mode.

Data	**Formula**
Ungrouped Data	Most repeated observation among given data
Discrete Grouped Data	Highest frequency among given data
Continuous Grouped Data	$Z = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times c$

Where,

Modal class = A class with highest frequency

L = Lower boundary of modal class

c = Class length

f_1 = Frequency of the modal class

f_0 = Frequency of the class before the modal class

f_2 = Frequency of the class after the modal class

Relation Between Mean, Median and Mode

→ $Z = 3M - 2\bar{x}$; where $\bar{x}$ = Mean, M = Median, Z = Mode

Example of Method-1.3: Mode

C 1 If mean is 16 and median is 20. Calculate the mode.

Answer: 28

C 2 Find the mode of following data:

(a) 2, 4, 2, 5, 7, 2, 8, 9.

(b) 2, 8, 4, 6, 10, 12, 4, 8, 14, 16.

Answer: (a) 2, (b) 4 & 8

C 3 Find the mode of following data:

x	11	22	33	44
f	15	20	19	10

Answer: 22

C 4 Find the mode of following data:

Class	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
f	3	5	7	10	12
	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100
	15	12	6	2	8

Answer: 55

C 5 Find the mode of following data:

Class	200 – 220	220 – 240	240 – 260	260 – 280
f	7	15	21	19
	280 – 300	300 – 320	320 – 340	
	6	4	2	

Answer: 255

C 6 Obtain the mean, mode and median for the following information:

x	< 10	< 20	< 30	< 40	< 50	< 60
f	12	30	57	77	94	100

Answer: $\bar{x} = 28$, $M = 27.407$, $Z = 25.625$

1.4 Quartiles

→ Quartiles are measures which divide a series into **four** equal parts using three quartiles namely Q_1, Q_2 and Q_3.

The quartile Q_1 is known as first quartile or lower quartile,

The quartile Q_2 is known as second quartile or **median** quartile,

The quartile Q_3 is known as third quartile or upper quartile.

→ To compute the quartile, list all observations in ascending order.

→ Table of formula of quartile.

<table>
<tr><th>Data</th><th>Formula</th></tr>
<tr><td>Ungrouped Data</td><td rowspan="2">$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ observation
$Q_2 = \text{Median}$
$Q_3 = 3\left(\frac{n+1}{4}\right)^{th}$ observation</td></tr>
<tr><td>Discrete Grouped Data</td></tr>
<tr><td>Continuous Grouped Data</td><td>$Q_1 = L + \left(\frac{\frac{n}{4} - F}{f}\right) \times c,\qquad Q_2 = L + \left(\frac{\frac{n}{2} - F}{f}\right) \times c$
$Q_3 = L + \left(\frac{\frac{3n}{4} - F}{f}\right) \times c$</td></tr>
</table>

Where,

Quartile class Q_k

$= \text{Class whose cumulative frequency with property } \min\left\{ cf \mid cf \geq \frac{kn}{4} \right\}$

L = Lower boundary point of the quartile class

n = Total number of observation (sum of the frequencies)

F = Cumulative frequency of the class preceding the quartile class

f = The frequency of the quartile class

Example of Method-1.4: Quartiles

C	1	Find the quartiles of the data: 4, 6, 7, 8, 10, 23, 34. **Answer: 6, 8, 23**
C	2	Find the quartile Q_1, and Q_3. x: 2, 4, 6, 8, 10, 12 f: 4, 1, 2, 3, 4, 5 **Answer: 4, 12**
C	3	Compute Q_1, and Q_3 for the data relating to age in years of 543 members in a village. x: 20, 30, 40, 50, 60, 70, 80 f: 3, 61, 132, 153, 140, 51, 3 **Answer: 40, 60**
C	4	Calculate the quartiles Q_1, and Q_3 for the following data. Class: $30-32$, $32-34$, $34-36$, $36-38$, $38-40$, $40-42$, $42-44$ f: 12, 18, 16, 14, 12, 8, 6 **Answer: 33.06, 38.75**
C	5	Calculate the quartiles Q_1, and Q_3 for the following data. Marks: <20, $20-30$, $30-40$, $40<$ Number of Students: 14, 20, 28, 18 **Answer: 23, 39.28**

Table for C 2:

x	2	4	6	8	10	12
f	4	1	2	3	4	5

Table for C 3:

x	20	30	40	50	60	70	80
f	3	61	132	153	140	51	3

Table for C 4:

Class	$30-32$	$32-34$	$34-36$	$36-38$
f	12	18	16	14
	$38-40$	$40-42$	$42-44$	
	12	8	6	

Table for C 5:

Marks	<20	$20-30$	$30-40$	$40<$
Number of Students	14	20	28	18

1.5 Percentiles

→ Percentiles are measures which divide a series into **hundred** equal parts.

→ There are ninety-nine percentiles which is

The percentile P_1 is known as first percentile,

The percentile P_2 is known as second percentile and so on.

→ To compute the percentile, list all observations in ascending order.

→ Table of formula of percentile.

<table>
<tr><th>Data</th><th>Formula</th></tr>
<tr><td>Ungrouped Data</td><td rowspan="2">$P_1 = \left(\frac{n+1}{100}\right)^{th}$ observation
$P_2 = 2\left(\frac{n+1}{100}\right)^{th}$ observation
$\vdots$
$P_{99} = 99\left(\frac{n+1}{100}\right)^{th}$ observation</td></tr>
<tr><td>Discrete Grouped Data</td></tr>
<tr><td>Continuous Grouped Data</td><td>$P_1 = L + \left(\frac{\frac{n}{100} - F}{f}\right) \times c$
$P_2 = L + 2\left(\frac{\frac{n}{100} - F}{f}\right) \times c$
$\vdots$
$P_{99} = L + 99\left(\frac{\frac{n}{100} - F}{f}\right) \times c$</td></tr>
</table>

Where,

Percentile class P_k

$= $ Class whose cumulative frequency with property min $\left\{ cf \mid cf \geq \dfrac{kn}{100} \right\}$

L = Lower boundary point of the percentile class

n = Total number of observation (sum of the frequencies)

F = Cumulative frequency of the class preceding the percentile class

f = The frequency of the percentile class

Example of Method-1.5: Percentiles

<table>
<tr><td>C</td><td>1</td><td>Find 20th Percentile of the data: 4, 6, 7, 8, 10, 23, 34, 55, 60.

Answer: 6</td></tr>
<tr><td>C</td><td>2</td><td>Find the 30th Percentile of the following data.
<table><tr><td>x</td><td>2</td><td>4</td><td>6</td><td>8</td><td>10</td><td>12</td></tr><tr><td>f</td><td>4</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr></table>
Answer: 6</td></tr>
<tr><td>C</td><td>3</td><td>Find the 51th Percentile of the following data:
<table><tr><td>Marks</td><td>15 – 20</td><td>20 – 25</td><td>25 – 30</td><td>30 – 35</td><td>35 – 40</td></tr><tr><td>No. of students</td><td>2</td><td>12</td><td>15</td><td>20</td><td>25</td></tr></table>
Answer: 32.185</td></tr>
</table>

Method 2 ⇝ Measure of Dispersion

Dispersion

→ Dispersion refers to the **spread** of the values around the central tendency.

→ For Example:

$-5, 0, 5$ and $-50, 0, 50$ both have the same mean 0 but clearly the data given in the second case much more widely dispersed than those in the first case.

→ So, measures of central tendency are not sufficient for having some idea about dispersion.

→ Measures of dispersion gives the idea about the degree to which numerical data tend to spread about an average life.

→ There are certain measures of dispersion which is,

(1) Range

(2) Interquartile Range

(3) Standard Deviation

(4) Mean Deviation

Range

→ Range is simply the highest value **minus** the lowest value of a set of data values.

→ For Example:

Range of $-5, 0, 5$ is 10

Reason: Range = Highest value – lowest value

$= 5 - (-5)$

$= 10$

Interquartile Range

→ The difference between the upper and lower quartile is known as the interquartile range.

i.e., Interquartile range = upper Quartile – lower Quartile = $Q_3 - Q_1$.

Standard Deviation

→ Standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

→ It is denoted by "**σ**" and read as "sigma".

→ Table of different formulae of standard deviation.

Method	Ungrouped Data	Discrete Grouped Data	Continuous Grouped Data
Direct Method	$\sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$	$\sqrt{\frac{\sum f_i x_i^2}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^2}$	
Assumed Mean Method	$\sqrt{\frac{\sum d_i^2}{n} - \left(\frac{\sum d_i}{n}\right)^2}$	$\sqrt{\frac{\sum f_i d_i^2}{\sum f_i} - \left(\frac{\sum f_i d_i}{\sum f_i}\right)^2}$	
Step Deviation Method	-----	-----	$\sqrt{\frac{\sum f_i u_i^2}{\sum f_i} - \left(\frac{\sum f_i u_i}{\sum f_i}\right)^2} \times c$

Variance

→ Variance is **expectation** of the squared deviation.

→ It informally measures how far a set of (random) numbers are spread out from their mean.

→ It is denoted by capital letter "**V**" and defined as $V = \sigma^2$.

Coefficient of Variation

→ The Coefficient of Variation is the **ratio** of the standard deviation to the mean and shows the extent of variability in relation to the mean of the population.

→ Coefficient of Variance is defined as

$$C.V. = \frac{\sigma}{\overline{x}} \times 100$$

→ If C.V. is high, then it is less consistent. Similarly, if C.V. is less, then it is more consistent.

→ The higher the Coefficient of Variation, the greater the dispersion.

Mean Deviation

→ The mean deviation is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set.

→ In a simple word, the mean deviation is used to calculate how far the values fall from the middle of the data set.

→ Table of different formulae of mean deviation:

Method	Ungrouped Data	Grouped Data
M.D. about Mean	$\dfrac{\sum \lvert x_i - \bar{x} \rvert}{n}$	$\dfrac{\sum f_i \lvert x_i - \bar{x} \rvert}{\sum f_i}$
M.D. about Median	$\dfrac{\sum \lvert x_i - M \rvert}{n}$	$\dfrac{\sum f_i \lvert x_i - M \rvert}{\sum f_i}$
M.D. about Mode	$\dfrac{\sum \lvert x_i - Z \rvert}{n}$	$\dfrac{\sum f_i \lvert x_i - Z \rvert}{\sum f_i}$

Example of Method-3: Measure of Dispersion

C	1	Determine the interquartile range value for 2, 3, 5, 7, 11, 13, 17, 19, 23, 29. **Answer: 11**
C	2	Find the interquartile range for the following distribution: x: 1, 2, 3, 4, 5 f: 5, 10, 12, 5, 3 **Answer: 10**
C	3	Find the standard deviation for the following data: 6, 7, 10, 12, 13, 4, 8, 12. **Answer: 3.0414**
C	4	Find the standard deviation and variance for the following distribution: x: 0 – 10, 10 – 20, 20 – 30, 30 – 40, 40 – 50, 50 – 60, 60 – 70 f: 6, 14, 10, 8, 1, 3, 8 **Answer: $\sigma = 19.6214$, $V = 384.9993$**

Table for C 2:

x	1	2	3	4	5
f	5	10	12	5	3

Table for C 4:

x	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
f	6	14	10	8	1	3	8

C	5	The arithmetic means of runs scored by three batsmen A, B and C, in the same series of 10 innings, are 50, 48 and 12 respectively. The standard deviations of their runs are 15, 12 and 2 respectively. Who is the most consistent of the three? **Answer: Batsman C is more consistent.**
C	6	Two machines A, B are used to fill a mixture of cement concrete in a beam. Find the standard deviation of each machine & comment on the performances of two machines. A: 32, 28, 47, 63, 71, 39, 10, 60, 96, 14 B: 19, 31, 48, 53, 67, 90, 10, 62, 40, 80 **Answer:** $\sigma_A = 25.4950$, $\sigma_B = 24.4290$ **There is less variability in the performance of the machine B.**
C	7	Find mean deviation about the mean, median and mode for the following data: x: 5, 10, 15, 20, 25 f: 7, 4, 6, 3, 5 **Answer:** $MD(\bar{x}) = 6.32$, $MD(M) = 6.2$, $MD(Z) = 9$
C	8	Find mean deviation about the mean, median and mode for the following data: Class: 5 – 25, 25 – 45, 45 – 65, 65 – 85, 85 – 105 f: 12, 8, 14, 20, 6 **Answer:** $MD(\bar{x}) = 21.33$, $MD(M) = 21.904$, $MD(Z) = 23.466$

Data for C 6:

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Data for C 7:

x	5	10	15	20	25
f	7	4	6	3	5

Data for C 8:

Class	5 – 25	25 – 45	45 – 65	65 – 85	85 – 105
f	12	8	14	20	6

Measure of Association

→ In statistics, various factors or coefficients used to quantify a relationship between two or more variables.

→ Some measures of association are Pearson's correlation coefficient, the Spearman rank correlation coefficient.

→ Measures of association are used in various fields of research.

→ A measure of association determined by correlation analysis and regression analysis.

- Correlation and regression are the most commonly used techniques for investigating the relationship between two quantitative variables.
- Correlation refers to the relationship of two or more variables. Regression establishes a functional relationship between the variables.
- The coefficient of correlation is a relative measure whereas the regression coefficient is an absolute figure.

Correlation

→ Two variables are known as **correlated** if a change in one variable affects a change in the other variable. Such a data connecting two variables is called bivariate data.

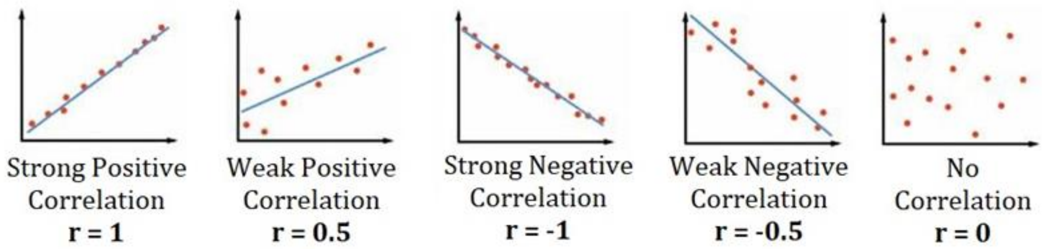
→ For Example:

- Relationship between heights and weights.
- Relationship between price and demand of commodity.
- Relationship between age of husband and age of wife.

→ When two variables are correlated with each other, it is important to know the amount or extent of correlation between them.

→ The numerical measure of correlation of degree of relationship existing between two variables is known as the coefficient of correlation.

→ It is denoted by **r** and it is always lying between -1 and 1.



→ The value of r is ± 0.9 or ± 0.8 etc. shows high degree of relationship between the variables while ± 0.2 or ± 0.1 etc. shows low degree of correlation.

Types of Correlation

→ Correlation is classified into four types.

- Positive Correlation
- Negative Correlation
- Linear Correlation
- Nonlinear Correlation

Positive Correlation

→ If both the variables vary in same direction, then such correlation is known as **positive correlation**.

→ In other words, if the value of one variable increases, the value of other variables also increases, or, if the value of one variable decreases, the value of other variables also decreases.

→ For Example:

- The correlation between heights and weights of group of persons is a positive correlation.

Height(cm)	150	152	155	160	162	165
Weight(kg)	60	62	64	65	67	69

Negative Correlation

→ If both the variables vary in opposite direction, then such correlation is known as **negative correlation**.

→ In other words, if the value of one variable increases, the value of other variables decreases, or, if the value of one variable decreases, the value of other variables increases.

→ For Example:

- The correlation between the price and demand of a commodity is a negative correlation.

Price (₹ per unit)	10	8	6	5	4	1
Demand(units)	100	200	300	400	500	600

Linear Correlation

→ If the ratio of change between two variables is constant, then such correlation is known as **linear correlation**.

→ If such variables are plotted on a graph paper, a straight line is obtained.

→ For Example:

Milk (l)	5	10	15	20	25	30
Curd (kg)	2	4	6	8	10	12

Nonlinear Correlation

→ If the ratio of change between two variables is not constant, then such correlation known as **nonlinear correlation**.

→ If such variables are plotted on a graph paper, a curve is obtained.

→ For Example:

Advertising expenses (₹ in lacs)	3	6	9	12	15
Curd (kg)	10	12	15	15	16

Methods of Studying Correlation

→ There are two different methods of studying correlation:

- Graphical Methods
  (1) Scatter Diagram
  (2) Simple Graph
- Mathematical Methods
  (1) Karl Pearson's coefficient of correlation
  (2) Spearman's rank coefficient of correlation

Method 3 ⇝ Covariance

Covariance

→ Covariance is a measure of the relationship between two random variables X and Y.

→ It is defined as

$$\operatorname{cov}(X, Y) = \frac{1}{n} \cdot \sum (x - \bar{x})(y - \bar{y})$$

Where, $\bar{x}$ and $\bar{y}$ are the mean of the X and Y respectively.

→ Also, we cand find covariance using following formula:

$$\operatorname{cov}(X, Y) = \frac{1}{n} \cdot \sum xy - \frac{1}{n^2} \cdot \sum x \sum y$$

Example of Method-3: Covariance

C	1	Find the covariance of the following data: (15,44), (20,43), (25,45), (30,37), (40,34), (50,37). **Answer:** -39.17
C	2	Determine $\sum xy$ if $n = 5$, $\operatorname{cov}(X, Y) = 7.2$, $\sum x = 25$, $\sum y = 30$ **Answer:** 186
C	3	Compute the covariance between x and y using the following data: <table><tr><td>x</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr><tr><td>y</td><td>3</td><td>2</td><td>5</td><td>4</td><td>6</td></tr></table>**Answer:** 1.6

Method 4 ⇝ Correlation Coefficient

Karl Pearson's Coefficient of Correlation

→ The coefficient of correlation is the measure of correlation between two random variables X and Y, and is denoted by r. It is defined as below:

$$\mathbf{r = \frac{cov(X, Y)}{\sigma_x \, \sigma_y}} \qquad \dots \dots (1)$$

Where,

cov(X, Y) is the covariance of variables X and Y.

σ_x & σ_y are standard deviation of X and Y respectively.

→ This expression is known as Karl Pearson's coefficient of correlation.

→ We have,

$$cov(X, Y) = \frac{1}{n} \cdot \sum (x - \overline{x})(y - \overline{y})$$

$$\sigma_x = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum (y - \overline{y})^2}{n}}$$

→ Substitute the values of cov(X, Y), σ_x and σ_y in equation (1), We get

$$\mathbf{r = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x})^2 \cdot \sum (y - \overline{y})^2}}} \qquad \dots \dots (2)$$

→ Equation (2) can be further reduced to below equation.

$$\mathbf{r = \frac{n \cdot \sum xy - \sum x \sum y}{\sqrt{n \cdot \sum x^2 - (\sum x)^2} \sqrt{n \cdot \sum y^2 - (\sum y)^2}}} \qquad \dots \dots (3)$$

Example of Method-4: Correlation Coefficient

<table>
<tr><td>C</td><td>1</td><td>Compute the coefficient of correlation between x and y using the following data:

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Answer: $\mathbf{r = -0.9203}$</td></tr>
<tr><td>C</td><td>2</td><td>Calculate Karl Pearson's correlation coefficient between age and playing habits:

Age	20	21	22	23	24	25
No. of students	500	400	300	240	200	160
Regular players	400	300	180	96	60	24

Answer: $\mathbf{r = -0.9823}$</td></tr>
<tr><td>C</td><td>3</td><td>Determine the coefficient of correlation if $n = 10, \bar{x} = 5.5, \bar{y} = 4,$
$\sum x^2 = 385, \sum y^2 = 192, \quad \sum (x+y)^2 = 947.$
Answer: $\mathbf{r = -0.6812}$</td></tr>
<tr><td>C</td><td>4</td><td>Given that $n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460$
and $\sum xy = 508$. Later on, it was found that two of the points $(8, 12)$
and $(6, 8)$ were wrongly entered as $(6, 14)$ and $(8, 6)$. Prove that $r = \frac{2}{3}$.</td></tr>
</table>

Method 5 ⇝ Rank Correlation Coefficient

Rank Correlation

→ Let a group of n individuals be arranged in order of merit with respect to some characteristics. The same group would give a different order(rank) for different characteristics.

→ Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is known as **rank correlation** in the characteristics A and B for that group of individuals.

→ It is denoted by $\boldsymbol{\rho}$.

Spearman's Rank Correlation Coefficient

→ Edward Spearman's formula for rank correlation coefficient (ρ) is given by

$$\boldsymbol{\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}}$$

Where, $d =$ Difference of Ranks

Tied Rank

→ If there is a tie between the ranks, then it is known as tied rank.

→ Formula for rank correlation coefficient (ρ) is,

$$\boldsymbol{\rho = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}({m_1}^3 - m_1) + \frac{1}{12}({m_2}^3 - m_2) + \cdots\right]}{n\,(n^2-1)}}$$

Where, $d =$ Difference of Ranks

$m_i =$ number of times a data repeats $= 2, 3, \ldots$

→ Note:

- If any data x_i is repeated 2 times, then $m_1 = 2$.
- If any data x_j is repeated 3 times, then $m_2 = 3$.

→ In case of tie between individuals' ranks, **the rank is divided among equal individuals**.

→ For Example:

- If there is tie with two items at 4th rank, then give average rank 4.5 as rank to both items.

$$\text{Average} = \frac{4+5}{2} = 4.5$$

Example of Method-5: Rank Correlation Coefficient

C 1

In a college, IT department has arranged one competition for IT students to develop an efficient program to solve a problem. Ten students took part in the competition and ranked by two judges given in the following table. Find the degree of agreement between the two judges using rank correlation coefficient.

1st judge	3	5	8	4	7	10	2	1	6	9
2nd judge	6	4	9	8	1	2	3	10	5	7

Answer: $\boldsymbol{\rho = -0.2970}$

C 2

The competitions in a beauty contest are ranked by three judges:

1^{st} judge	1	5	4	8	9	6	10	7	3	2
2^{nd} judge	4	8	7	6	5	9	10	3	2	1
3^{rd} judge	6	7	8	1	5	10	9	2	3	4

Use rank correlation to discuss which pair of judges has nearest approach to beauty.

Answer: **2^{nd} and 3^{rd} judges has nearest approach**

$[\,\boldsymbol{\rho_{12} = 0.5515, \rho_{23} = 0.7333, \rho_{13} = 0.0545}\,]$

C 3

Find the rank correlation coefficient and comment on its value:

Roll no.	1	2	3	4	5	6	7	8	9
Marks in Math.	78	36	98	25	75	82	90	62	65
Marks in Chem.	84	51	91	60	68	62	86	58	53

Answer: $\boldsymbol{\rho = 0.8333}$

C 4

Calculate coefficient of correlation by Spearman's method from following.

Sales	45	56	39	54	45	40	56	60	30	36
Cost	40	36	30	44	36	32	45	42	20	36

Answer: $\boldsymbol{\rho = 0.7636}$

C	5	The coefficient of rank correlation of marks obtained by 10 students in English and Economics was found to be 0.6. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 7 instead of 1. Find the correct coefficient of rank correlation. **Answer: $\mathbf{0.8909}$**

Regression

→ Regression is defined as a method of estimating the value of one variable when the other is known and both are correlated.

→ We use the general form regression line for these algebraic expressions. The algebraic expressions of the regression lines are known as **Regression Equations**.

→ It is highly used in statistical estimation of demand curve, supply curve, production function, cost function, consumption function etc.

Types of Regression

→ Regression is classified into four types:

- Simple Regression
- Multiple Regression
- Linear Regression
- Nonlinear Regression

Simple Regression

→ The regression analysis for studying only two variables at a time is known as **simple regression**.

Multiple Regression

→ The regression analysis for studying more than two variables at a time is known as **multiple regression**.

Linear Regression

→ If the regression curve is a straight line, the regression is known as **linear regression**.

Nonlinear Regression

→ If the regression curve is not a straight line, the regression is known as **nonlinear regression**.

Method of Studying Regression

→ There are two methods of studying regression:

- Method of scatter diagram
- Method of least square

→ We will use method of least square only to find out regression.

Method 6⇝ Linear Regression

Line of Regression (Linear Regression)

→ If the variables, which are highly correlated, are plotted on a graph then the points are around a straight line, the line is known as the **line of regression**.

→ There are two types of line of regression.

- Line of regression of y on x
- Line of regression of x on y

Line of Regression of y on x

→ It is the line which gives the **best estimate for the values of y** for given values of x.

→ The regression equation of y on x is given by

$$\mathbf{y - \overline{y} = b_{yx}\,(x - \overline{x})}$$

Where, b_{yx} = Regression Coefficient $= r\,\dfrac{\sigma_y}{\sigma_x} = \dfrac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$

$$= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2}$$

r = Correlation coefficient between x and y

$\overline{x}, \sigma_x$ = Mean & Standard Deviation of all x_i

$\overline{y}, \sigma_y$ = Mean & Standard Deviation of all y_i

Line of Regression of x on y

→ It is the line which gives the **best estimate for the values of x** for given values of y.

→ The regression equation of x on y is given by

$$\mathbf{x - \overline{x} = b_{xy}\,(y - \overline{y})}$$

Where, b_{xy} = Regression Coefficient $= r\,\dfrac{\sigma_x}{\sigma_y} = \dfrac{n\sum xy - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2}$

$$= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(y-\overline{y})^2}$$

r = Correlation coefficient between x and y

$\overline{x}, \sigma_x$ = Mean & Standard Deviation of all x_i

$\overline{y}, \sigma_y$ = Mean & Standard Deviation of all y_i

Properties of Regression Coefficients

→ The coefficient of correlation is the geometric mean of the coefficients of regression.

i.e., $\mathbf{r} = \sqrt{\mathbf{b_{yx} \cdot b_{xy}}}$

→ The product of both b_{xy} and b_{yx} cannot be more than 1.

→ Both the regression coefficients will have the same sign. They are either both positive and both negative. It means,

If $r < 0$, then $b_{yx} < 0$ & $b_{xy} < 0$.

If $r > 0$, then $b_{yx} > 0$ & $b_{xy} > 0$.

→ The arithmetic mean of the regression coefficients is greater than or equal to the correlation coefficient.

i.e. $\dfrac{b_{xy} + b_{yx}}{2} \geq r$

Example of Method-6: Linear Regression

<table>
<tr><td>C</td><td>1</td><td>Obtain the two lines of regression for the following data:
<table>
<tr><td>Sales (No. of tablets)</td><td>190</td><td>240</td><td>250</td><td>300</td><td>310</td><td>335</td><td>300</td></tr>
<tr><td>Advertising expense (Rs.)</td><td>5</td><td>10</td><td>12</td><td>20</td><td>20</td><td>30</td><td>30</td></tr>
</table>
Answer: $\mathbf{y = 0.1766x - 30.4221}$; $\mathbf{x = 4.7357y + 189.0807}$</td></tr>
<tr><td>C</td><td>2</td><td>A study of amount of rainfall and quantity of air pollution removed is:
<table>
<tr><td>Daily rainfall (0.01 cm)</td><td>4.3</td><td>4.5</td><td>5.9</td><td>5.6</td><td>6.1</td><td>5.2</td><td>3.8</td><td>2.1</td><td>7.5</td></tr>
<tr><td>Particulate removed unit</td><td>126</td><td>121</td><td>116</td><td>118</td><td>114</td><td>118</td><td>132</td><td>141</td><td>108</td></tr>
</table>
a. Find the equation of the regression line to predict the particulate removed from the amount of daily rainfall.
b. Find the amount of particulate removed when daily rainfall is 4.8 units.
Answer: a. $\mathbf{y = -6.3240x + 153.1755}$; b. $\mathbf{122.8203}$</td></tr>
</table>

C	3	The following data regarding the height(y) and weight(x) of 100 students are given: $\sum x = 15000, \sum y = 6800, \sum x^2 = 2272500, \sum y^2 = 463025, \sum xy = 1022250$. Find the equation of regression line of height on weight. **Answer**: $\mathbf{y = 0.1x + 53}$
C	4	The data for advertising and sale given below: <table><tr><td></td><td>Adv. Exp.(x) (Rs. lakh)</td><td>Sales(y) (Rs. lakh)</td></tr><tr><td>Mean</td><td>10</td><td>90</td></tr><tr><td>Standard deviation</td><td>3</td><td>12</td></tr></table>a. Correlation coefficient between prices is 0.8. b. Calculate the two regression lines. c. Find the likely sales when advertising expenditure is 15 lakhs. d. What should be the advertising expenditure if the company wants to attain a sales target of 120 lakhs? **Answer**: **b.** $\mathbf{x = 0.2y - 8 \;;\; y = 3.2x + 58 \;;}$ **c.** $\mathbf{106 \;;}$ **d.** $\mathbf{16}$
C	5	A study of prices of a certain commodity at Raipur and Kanpur yields the below data: <table><tr><td></td><td>Raipur (Rs)</td><td>Kanpur (Rs)</td></tr><tr><td>Average price/kg</td><td>2.463</td><td>2.797</td></tr><tr><td>Standard deviation</td><td>0.326</td><td>0.207</td></tr></table>Correlation coefficient between prices at Raipur and Kanpur is 0.774. Estimate the most likely price at Raipur corresponding to the price of 3.052 per kilo at Kanpur. **Answer**: $\mathbf{2.774}$

Method 7 ⇝ Curve Fitting

Introduction

→ We come across many situations where we often require to find a relationship between two or more variables. For example, weight and height of a person, demand and supply, expenditure depends on income, etc.

→ This relation may be expresses by polynomial or exponential or logarithmic relationship. In order to determine such relationship, first it is requiring to collect the data showing corresponding values of the variables under consideration.

→ Suppose $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be the data showing corresponding values of the variables x and y under consideration. If we plot the above data points on a coordinate system, then the set of points so plotted form a scatter diagram.

→ From this diagram, it is sometimes possible to visualize a smooth curve approximating the data. Such a curve is known as an **approximating curve**.

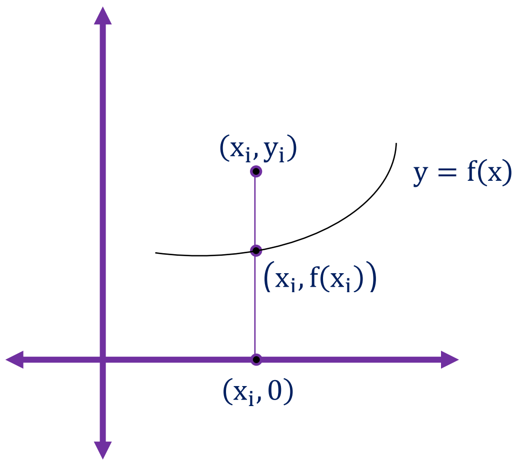
→ In particular, if the data approximate well to a straight line, we say that a linear relationship exists between the variables. It is quite possible that the relationship of the form $y = f(x)$ between two variables x and y, giving the approximating curve and which fit the given data of x and y, is known as **curve fitting**.

Least Square Method

→ Suppose that the data points are $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where x is independent and y is dependent variable.

→ Let the fitting curve f(x) has the following deviations (or errors or residuals) from each data points. i.e., $d_1 = y_1 - f(x_1)$.

→ These $d_i = y_i - f(x_I)$ are known as deviation, error or residual. Its value may be positive, negative or zero.



→ To give equal weightage to each error, we square each of these and form their sum.

$$\text{i.e.}, D = {d_1}^2 + {d_2}^2 + \cdots + {d_n}^2 = \sum_{i=1}^{n} d_i^2$$

$$= \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

7.1 Curve Fitting by Straight Line

→ Let, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be the set of n values and let the relation between x and y be $\mathbf{y = a + bx}$.

→ We have,

$$D = \sum_{i=1}^{n} [y_i - f(x_i)] = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

→ If $D = 0$, then all the n points will lie on $y = f(x)$.

→ If $D \neq 0$, $f(x)$ is chosen such that D is minimum.

→ This will be minimum at,

$$\frac{\partial D}{\partial a} = 0 \Longrightarrow -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\Longrightarrow \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} 1 - b\sum_{i=1}^{n} x_i = 0$$

$$\Longrightarrow \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

$$\Longrightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

Similarly, by $\frac{\partial D}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$

→ We obtain following **Normal Equations** for the best fitting straight line $y = a + bx$.

$$\mathbf{\sum y = na + b\sum x}$$

$$\mathbf{\sum xy = a\sum x + b\sum x^2}$$

→ These equations can be solved simultaneously to give the best value of a and b such that straight line is the best fit to the data.

Example of Method-7.1: Curve Fitting by Straight Line

<table>
<tr><td>C</td><td>1</td><td>Fit a straight line for the given pairs of (x, y) which are $(1,5), (2,7), (3,9), (4,10), (5,11)$.

Answer: $\mathbf{y = 3.9 + 1.5x}$</td></tr>
<tr><td>C</td><td>2</td><td>Fit a straight-line $y = ax + b$ to the following data:
<table><tr><td>x</td><td>-2</td><td>-1</td><td>0</td><td>1</td><td>2</td></tr><tr><td>y</td><td>1</td><td>2</td><td>3</td><td>3</td><td>4</td></tr></table>
Answer: $\mathbf{y = 0.7x + 2.6}$</td></tr>
<tr><td>C</td><td>3</td><td>By method of least squares, fit a linear relation of the form $P = a + bW$ to the following data, P is the pull required to lift a weight W. Also estimate P, when W is 150.
<table><tr><td>P</td><td>50</td><td>70</td><td>100</td><td>120</td></tr><tr><td>W</td><td>12</td><td>15</td><td>21</td><td>25</td></tr></table>
Answer: $\mathbf{P = -11.8005 + 5.3041W}$; $\mathbf{P(150) = 783.8145}$</td></tr>
</table>

7.2 Curve Fitting by Parabola

→ Let, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be the set of n values and let the relation between x and y be $\mathbf{y = a + bx + cx^2}$.

→ We have,

$$D = \sum_{i=1}^{n} [y_i - f(x_i)] = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

→ If $D = 0$, then all the n points will lie on $y = f(x)$. If $D \neq 0$, $f(x)$ is chosen such that D is minimum.

→ Differentiating S with respect to a, b, c and equating with zero (as done while fitting a linear curve). We obtain following **Normal Equations** for the best fitting $\mathbf{y = a + bx + cx^2}$ curve (parabola) of second degree.

$$\sum \mathbf{y = na + b\sum x + c\sum x^2}$$

$$\sum \mathbf{x\,y = a\sum x + b\sum x^2 + c\sum x^3}$$

$$\sum \mathbf{x^2\,y = a\sum x^2 + b\sum x^3 + c\sum x^4}$$

Example of Method-7.2: Curve Fitting by Parabola

C 1 Fit a polynomial of degree two using least square method for the following experimental data. Also, estimate $y(2.4)$.

x	1	2	3	4	5
y	5	12	26	60	97

Answer: $\mathbf{y = 10.4 - 11.0857x + 5.7143x^2}$; $\mathbf{y(2.4) = 16.7087}$

C 2 Fit a second – degree parabola $y = ax^2 + bx + c$ to the following data:

x	-3	-2	-1	0	1	2	3
y	12	4	1	2	7	15	30

Answer: $\mathbf{y = 2.1190x^2 + 2.9286x + 1.6667}$

C	3	Fit a relation of the form $R = a + bV + cV^2$ to the following data, where V is the velocity in km/hr. and R is the resistance in km/quintal. Estimate R when $V = 90$.

V	20	40	60	80	100	120
R	5.5	9.1	14.9	22.8	33.3	46.0

Answer: $\mathbf{R = 4.35 + 0.0024V + 0.0029V^2}$; $\mathbf{R(90) = 28.0560}$

Method 8 ⇝ Fitting of Trend by Moving Average Method

Introduction

→ Moving average method is a simple device of reducing fluctuations and obtaining trend values with a fair degree of accuracy.

→ In this method the average value of number of years (or months, weeks or days) is taken as the trend value for the middle point of the period of moving average.

→ The process of averaging smoothest the curve and reduces the fluctuation.

→ The first thing to be decided in the method is the period of the moving average. What it means is to take a decision about the number of consecutive items whose average would be calculated each time.

Moving Averages of Odd Number of Years

→ Steps to find moving averages of k years (where k is an odd number)

(1) Find the sum of **first k** observations and also find their average (by dividing **k**) and place it against the $\left(\frac{k+1}{2}\right)^{th}$ observation.

(2) **Leave** the first observation, find average of next **k** observations and place it against the $\left(\frac{k+3}{2}\right)^{th}$ observation.

(3) Repeat above steps till last observation of the data is used.

→ i.e.,

Year	Observations	3 yearly moving average
1980	x_1	
1981	x_2	$\frac{1}{3}(x_1+x_2+x_3)$
1982	x_3	$\frac{1}{3}(x_2+x_3+x_4)$
1983	x_4	$\frac{1}{3}(x_3+x_4+x_5)$
1983	x_5	

Moving Averages of Even Number of Years

→ Steps to find moving averages of k years (where k is an even number)

(1) Find the sum of first **k** observations and also find their average (by dividing **k**) and place it against between $\left(\frac{k}{2}\right)^{th}$ & $\left(\frac{k}{2}+1\right)^{th}$ observation.

(2) **Leave** the first observation, find the average of **next k** observations and place it against between $\left(\frac{k}{2}+1\right)^{th}$ & $\left(\frac{k}{2}+2\right)^{th}$ observation.

(3) Repeat above steps till last observation of the data is used and we get **k** yearly moving average.

(4) Find **k** yearly centered moving average(by dividing **2**), which is the trend value of moving average.

→ i.e.,

Year	Observations	4 yearly moving average	4 yearly centered moving average
1980	x_1		
1981	x_2		
		$\frac{1}{4}(x_1+x_2+x_3+x_4)=A_1$	
1982	x_3		$\frac{1}{2}(A_1+A_2)$
		$\frac{1}{4}(x_2+x_3+x_4+x_5)=A_2$	
1983	x_4		$\frac{1}{2}(A_2+A_3)$
		$\frac{1}{4}(x_3+x_4+x_5+x_6)=A_3$	
1984	x_5		

Example of Method-8: Fitting of Trend by Moving Average Method

C 1 Calculate 3 yearly moving averages of the following data.

Years	1971	1972	1973	1974	1975
Value	11	8	12	3	4

Answer: 10.333, 7.666, 6.333

C 2 Calculate 4 yearly moving averages of number of students studying in a higher secondary school in a particular village from the following data.

Years	1991	1992	1993	1994	1995	1996	1997	1998
No. of Students	36	43	43	34	44	54	34	24

Answer: 40, 42.37, 42.62, 40.25

C 3 Calculate 5 yearly moving averages of the following data.

Years	1971	1972	1973	1974	1975	1976	1977	1978	1979
Value	25	49	55	12	52	40	88	12	44

Answer: 38.6, 41.6, 49.4, 40.8, 47.2

C 4 Calculate 6 yearly moving averages of the following data.

Years	2001	2002	2003	2004	2005	2006	2007	2008	2009
Value	124	120	135	140	145	158	162	170	175

Answer: 140.166, 147.5, 155

*** * * * * End of the Unit * * * * ***