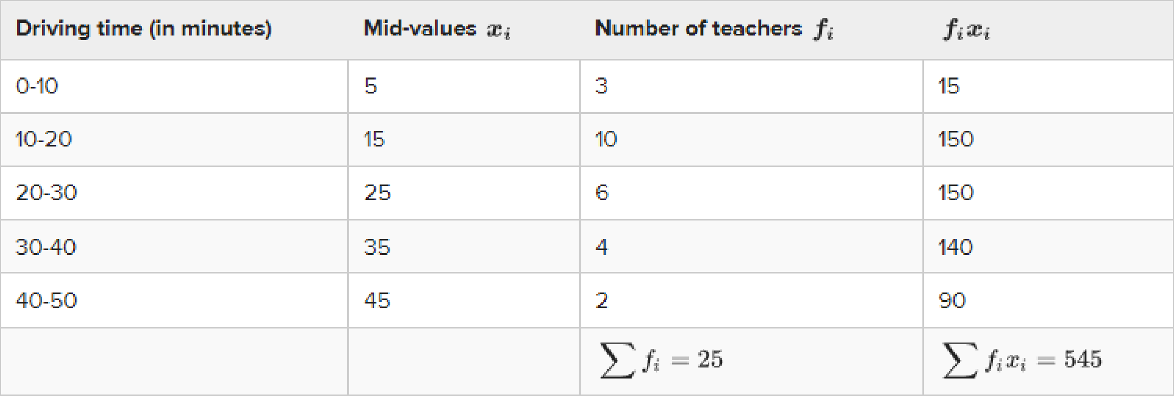
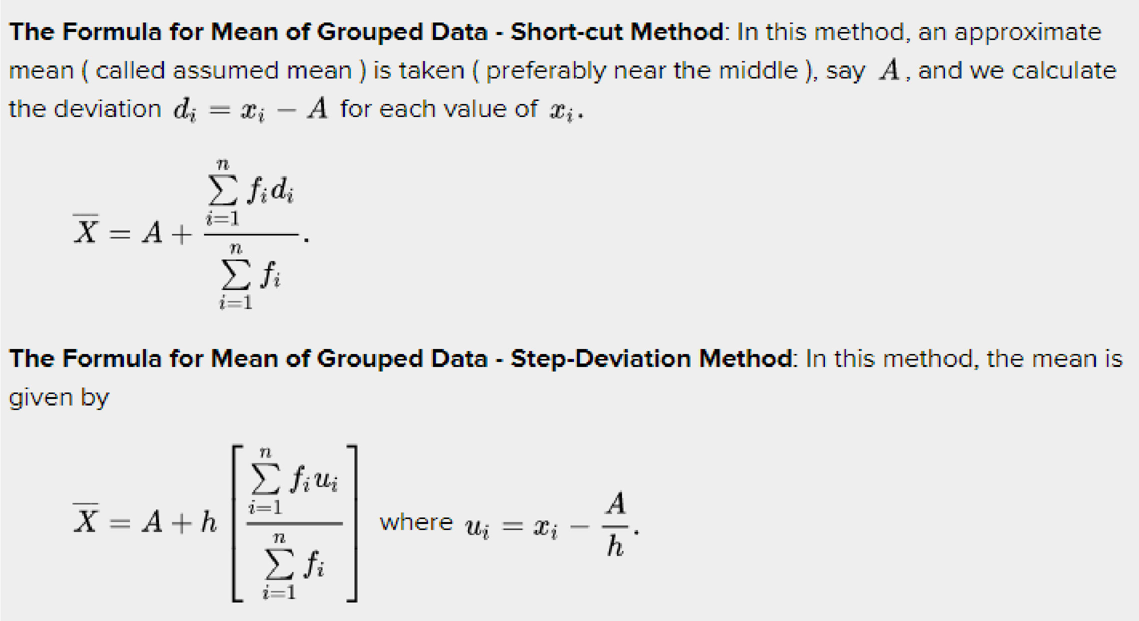
Mean



Driving time (in minutes)	Mid-values x_i	Number of teachers f_i	f_ix_i
0-10	5	3	15
10-20	15	10	150
20-30	25	6	150
30-40	35	4	140
40-50	45	2	90
		$\sum f_i = 25$	$\sum f_ix_i = 545$

$$\overline{X} = \frac{\sum f_ix_i}{\sum f_i} = \frac{545}{25} \quad . \overline{X} = \frac{\sum_{i=1}^{6} x_i f_i}{\sum_{i=1}^{6} f_i}$$

$$\overline{X} = 21.8$$



The Formula for Mean of Grouped Data - Short-cut Method: In this method, an approximate mean (called assumed mean) is taken (preferably near the middle), say A, and we calculate the deviation $d_i = x_i - A$ for each value of x_i.

$$\overline{X} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}.$$

The Formula for Mean of Grouped Data - Step-Deviation Method: In this method, the mean is given by

$$\overline{X} = A + h\left[\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}\right] \quad \text{where } u_i = x_i - \frac{A}{h}.$$

Median:

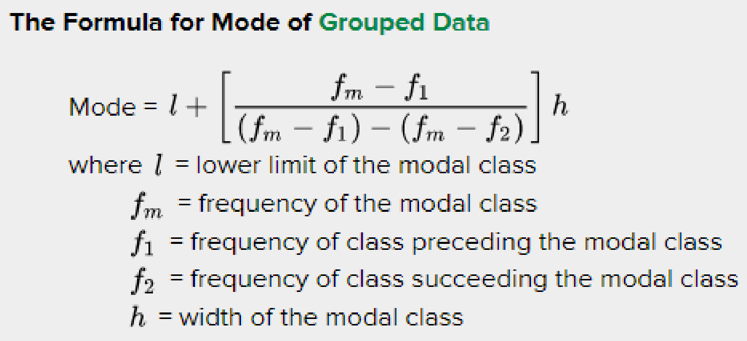
If n is odd then the median = value of $\left(\frac{n+1}{2}\right)^{\text{th}}$ observation.

If n is even then the median = arithmetic mean of the value of $\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$

$$\text{Median} = l + \frac{\frac{N}{2} - F}{f} \times h$$

N/2 is median class; l = lower limit of the median class; f = frequency of the median class

F = cumulative frequency of the class preceding the median class;
N = total number of observations; h = width of the median class



The Formula for Mode of Grouped Data

$$\text{Mode} = l + \left[\frac{f_m - f_1}{(f_m - f_1) - (f_m - f_2)}\right] h$$

where l = lower limit of the modal class

f_m = frequency of the modal class

f_1 = frequency of class preceding the modal class

f_2 = frequency of class succeeding the modal class

h = width of the modal class

negatively skewed: (Mean < Median < Mode)

positively skewed: (Mean > Median > Mode)

Mode = 3 Median - 2 Mean

Variance:

$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{n} \qquad \sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2$$

$$\sigma^2 = \frac{\sum fx^2}{n} - \bar{x}^2$$

$$\text{where} \quad \bar{x} = \sum \frac{fx}{n} \quad \text{(The Mean)}$$

The co-efficient of variation shows the extent of variability of data in a sample in relation to the mean of the population. Rho = standard deviation; mew = mean

$$\mathrm{CV} = \frac{\sigma}{\mu}$$

$$r = \frac{\Sigma xy - \bar{n}\bar{x}\bar{y}}{(n-1)SD(x)SD(y)}$$

;

coefficient of corelation: helps in establishing a relation between predicted and actual values obtained in a statistical experiment. The calculated value of the correlation coefficient explains the exactness between the predicted and actual values.

ρ (X,Y) = cov (X,Y) / σX.σY.

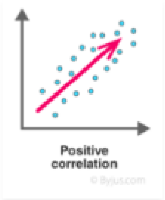
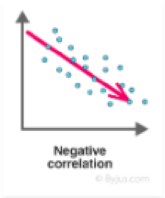
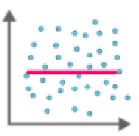
Karl Pearson's Correlation Coefficient:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$$

Assumed Mean Method:

$$r = \frac{N\sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{N\sum d_x^2 - (\sum d_x)^2} - \sqrt{N\sum d_y^2 - (\sum d_y)^2}}$$

dx = X – A; **dy** = Y - A

Positive Correlation	The value of one variable increases linearly with increase in another variable. This indicates a similar relation between both the variables. So its correlation coefficient would be positive or 1 in this case.	
Negative Correlation	When there is a decrease in values of one variable with increase in values of other variable. In that case, correlation coefficient would be negative.	
Zero Correlation or No Correlation	There is one more situation when there is no specific relation between two variables.	

Rank of corelation:

ρ= $$1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$ where d=(Rx-Ry),

Rx, Ry is rank in largest value having 1st rank.

Regression model evaluation metrics: The MSE, MAE, RMSE, and R-Squared metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis.

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.

MSE (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

R-squared (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is. The above metrics can be expressed.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

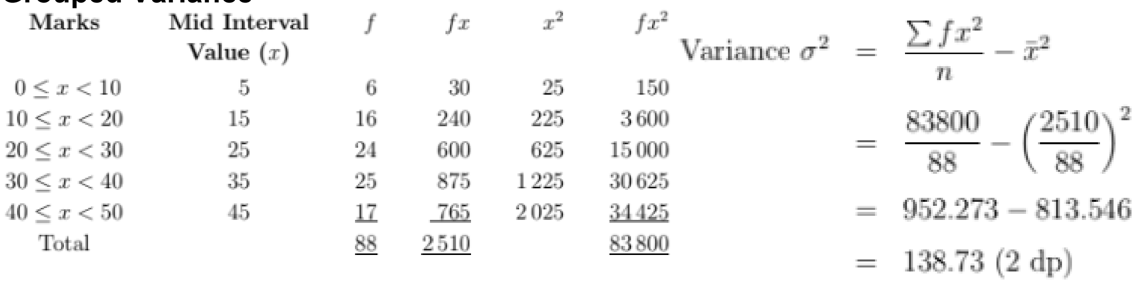
$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \qquad R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

Where,
$\hat{y}$ $-$ *predicted value of y*
$\bar{y}$ $-$ *mean value of y*

Where,
$\hat{y}$ $-$ *predicted value of y*
$\bar{y}$ $-$ *mean value of y*

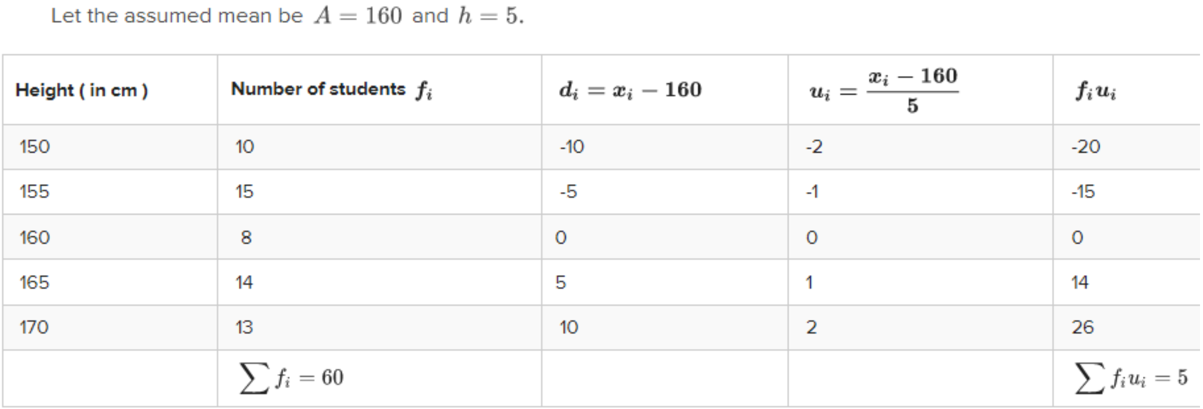
Grouped Variance



Marks	Mid Interval Value (x)	f	fx	x^2	fx^2
$0 \le x < 10$	5	6	30	25	150
$10 \le x < 20$	15	16	240	225	3 600
$20 \le x < 30$	25	24	600	625	15 000
$30 \le x < 40$	35	25	875	1 225	30 625
$40 \le x < 50$	45	17	765	2 025	34 425
Total		88	2 510		83 800

$$\begin{aligned} \text{Variance } \sigma^2 &= \frac{\sum fx^2}{n} - \bar{x}^2 \\ &= \frac{83800}{88} - \left(\frac{2510}{88}\right)^2 \\ &= 952.273 - 813.546 \\ &= 138.73 \text{ (2 dp)} \end{aligned}$$

Mean by Step-Deviation Method



Let the assumed mean be $A = 160$ and $h = 5$.

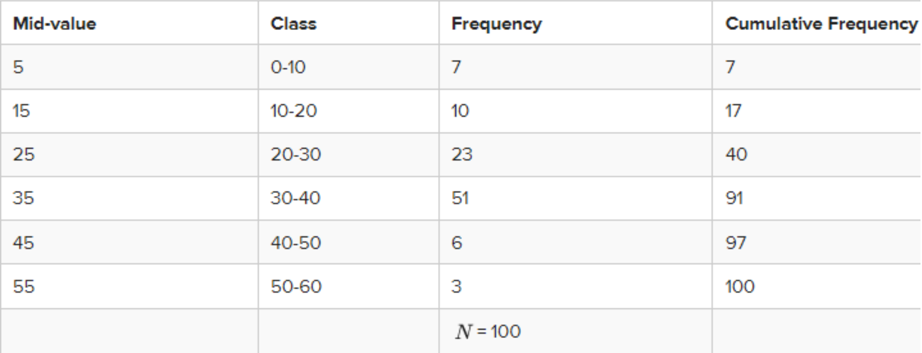
Height (in cm)	Number of students f_i	$d_i = x_i - 160$	$u_i = \frac{x_i - 160}{5}$	$f_i u_i$
150	10	-10	-2	-20
155	15	-5	-1	-15
160	8	0	0	0
165	14	5	1	14
170	13	10	2	26
	$\sum f_i = 60$			$\sum f_i u_i = 5$

$$\begin{aligned} \text{Median} &= l + \frac{\frac{N}{2} - F}{f} \times h \\ &= 30 + \frac{\frac{100}{2} - 40}{51} \times 10 \\ &= 30 + \frac{50 - 40}{51} \times 10 = \\ &= 31.96 \end{aligned}$$

Mean formula:

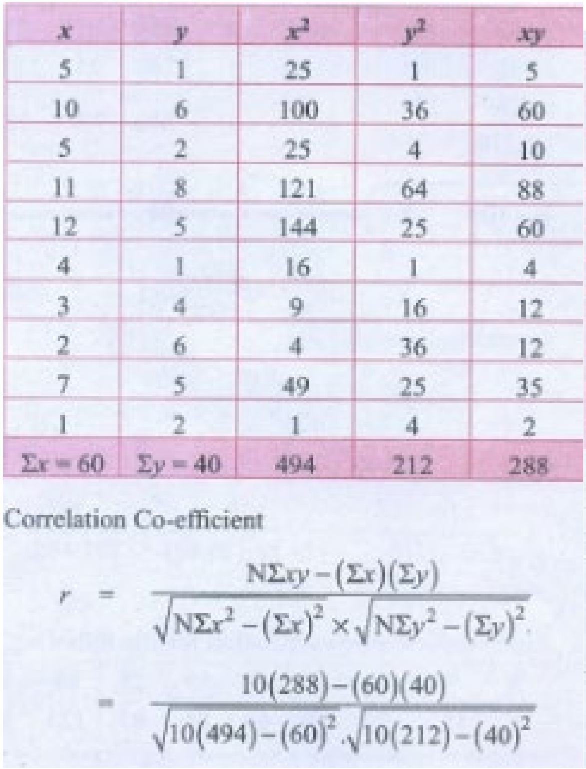
$$\overline{X} = A + h\left[\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}\right]$$

Median:



Mid-value	Class	Frequency	Cumulative Frequency
5	0-10	7	7
15	10-20	10	17
25	20-30	23	40
35	30-40	51	91
45	40-50	6	97
55	50-60	3	100
		$N = 100$	

Corelation coefficient:

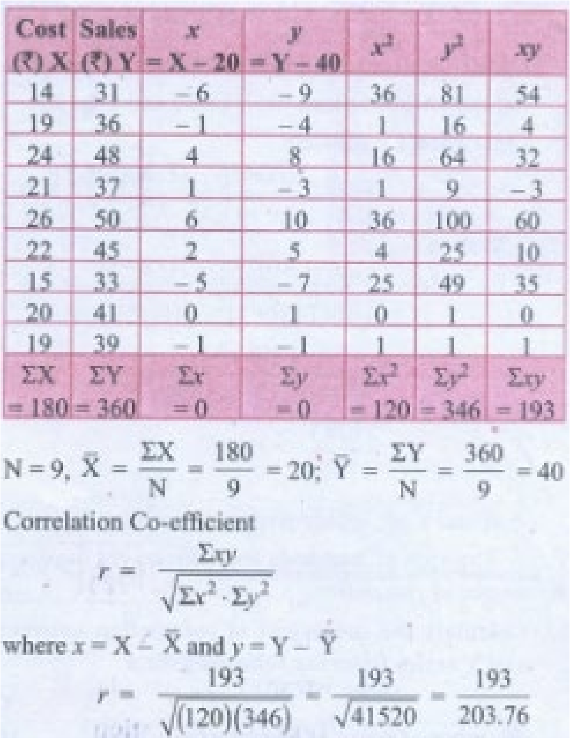


x	y	x^2	y^2	xy
5	1	25	1	5
10	6	100	36	60
5	2	25	4	10
11	8	121	64	88
12	5	144	25	60
4	1	16	1	4
3	4	9	16	12
2	6	4	36	12
7	5	49	25	35
1	2	1	4	2
$\Sigma x = 60$	$\Sigma y = 40$	494	212	288

Correlation Co-efficient

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{N\Sigma x^2 - (\Sigma x)^2} \times \sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{10(288) - (60)(40)}{\sqrt{10(494) - (60)^2} . \sqrt{10(212) - (40)^2}}$$



Cost (₹) X	Sales (₹) Y	$x = X - 20$	$y = Y - 40$	x^2	y^2	xy
14	31	−6	−9	36	81	54
19	36	−1	−4	1	16	4
24	48	4	8	16	64	32
21	37	1	−3	1	9	−3
26	50	6	10	36	100	60
22	45	2	5	4	25	10
15	33	−5	−7	25	49	35
20	41	0	1	0	1	0
19	39	−1	−1	1	1	1
$\Sigma X = 180$	$\Sigma Y = 360$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 120$	$\Sigma y^2 = 346$	$\Sigma xy = 193$

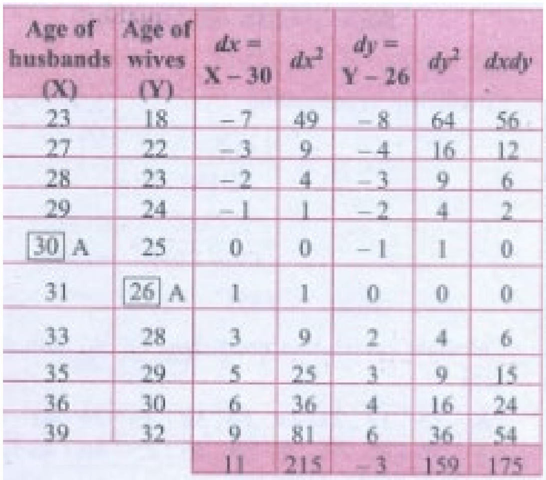
$$N = 9,\ \bar{X} = \frac{\Sigma X}{N} = \frac{180}{9} = 20;\ \bar{Y} = \frac{\Sigma Y}{N} = \frac{360}{9} = 40$$

Correlation Co-efficient

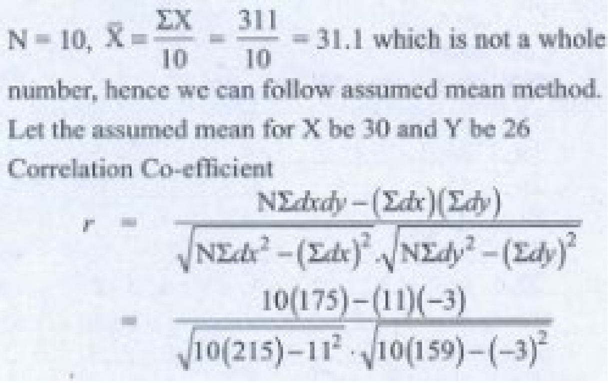
$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$

$$r = \frac{193}{\sqrt{(120)(346)}} = \frac{193}{\sqrt{41520}} = \frac{193}{203.76}$$



Age of husbands (X)	Age of wives (Y)	$dx = X - 30$	dx^2	$dy = Y - 26$	dy^2	$dxdy$
23	18	−7	49	−8	64	56
27	22	−3	9	−4	16	12
28	23	−2	4	−3	9	6
29	24	−1	1	−2	4	2
$\boxed{30}$ A	25	0	0	−1	1	0
31	$\boxed{26}$ A	1	1	0	0	0
33	28	3	9	2	4	6
35	29	5	25	3	9	15
36	30	6	36	4	16	24
39	32	9	81	6	36	54
		11	215	−3	159	175



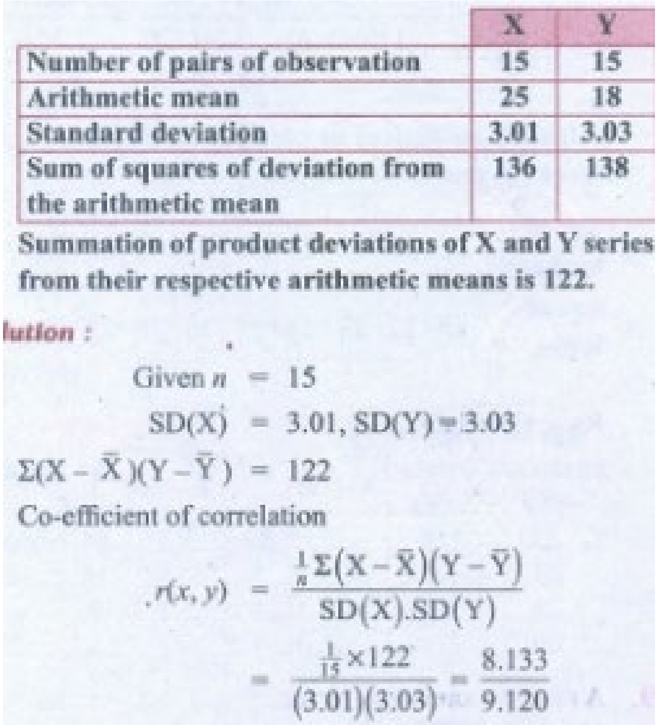
$N = 10,\ \bar{X} = \frac{\Sigma X}{10} = \frac{311}{10} = 31.1$ which is not a whole number, hence we can follow assumed mean method.

Let the assumed mean for X be 30 and Y be 26

Correlation Co-efficient

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

$$= \frac{10(175) - (11)(-3)}{\sqrt{10(215) - 11^2} \cdot \sqrt{10(159) - (-3)^2}}$$



	X	Y
Number of pairs of observation	15	15
Arithmetic mean	25	18
Standard deviation	3.01	3.03
Sum of squares of deviation from the arithmetic mean	136	138

Summation of product deviations of X and Y series from their respective arithmetic means is 122.

lution :

$$\text{Given } n = 15$$

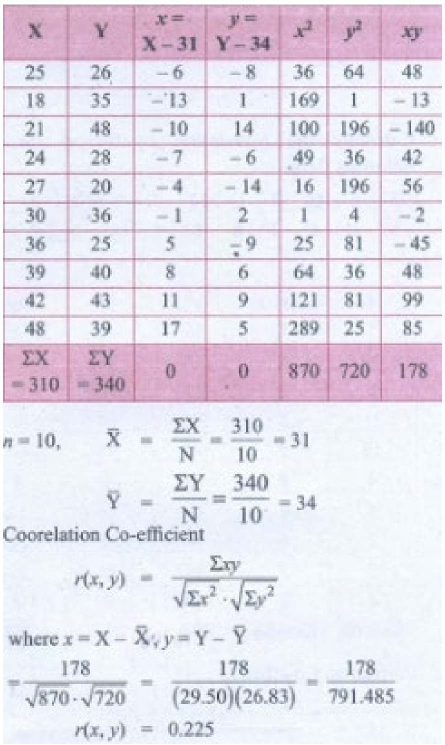
$$SD(X) = 3.01, SD(Y) = 3.03$$

$$\Sigma(X-\bar{X})(Y-\bar{Y}) = 122$$

Co-efficient of correlation

$$r(x,y) = \frac{\frac{1}{n}\Sigma(X-\bar{X})(Y-\bar{Y})}{SD(X).SD(Y)}$$

$$= \frac{\frac{1}{15}\times 122}{(3.01)(3.03)} = \frac{8.133}{9.120}$$



X	Y	$x = X-31$	$y = Y-34$	x^2	y^2	xy
25	26	−6	−8	36	64	48
18	35	−13	1	169	1	−13
21	48	−10	14	100	196	−140
24	28	−7	−6	49	36	42
27	20	−4	−14	16	196	56
30	36	−1	2	1	4	−2
36	25	5	−9	25	81	−45
39	40	8	6	64	36	48
42	43	11	9	121	81	99
48	39	17	5	289	25	85
$\Sigma X = 310$	$\Sigma Y = 340$	0	0	870	720	178

$$n = 10, \quad \bar{X} = \frac{\Sigma X}{N} = \frac{310}{10} = 31$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{340}{10} = 34$$

Coorelation Co-efficient

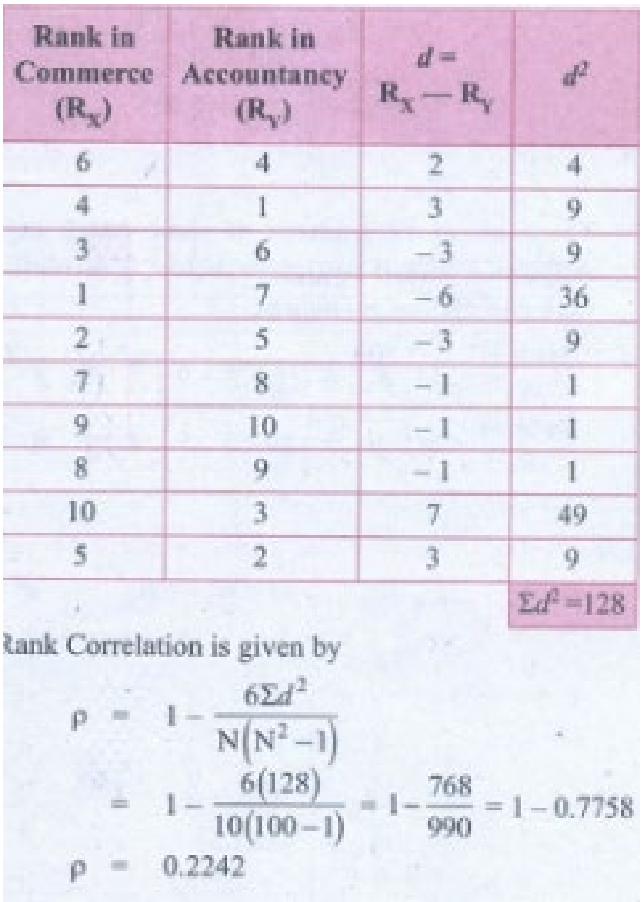
$$r(x,y) = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\cdot\sqrt{\Sigma y^2}}$$

where $x = X - \bar{X}, y = Y - \bar{Y}$

$$= \frac{178}{\sqrt{870}\cdot\sqrt{720}} = \frac{178}{(29.50)(26.83)} = \frac{178}{791.485}$$

$$r(x,y) = 0.225$$

Rank



Rank in Commerce (R_X)	Rank in Accountancy (R_Y)	$d = R_X - R_Y$	d^2
6	4	2	4
4	1	3	9
3	6	−3	9
1	7	−6	36
2	5	−3	9
7	8	−1	1
9	10	−1	1
8	9	−1	1
10	3	7	49
5	2	3	9
			$\Sigma d^2 = 128$

Rank Correlation is given by

$$\rho = 1 - \frac{6\Sigma d^2}{N(N^2-1)}$$

$$= 1 - \frac{6(128)}{10(100-1)} = 1 - \frac{768}{990} = 1 - 0.7758$$

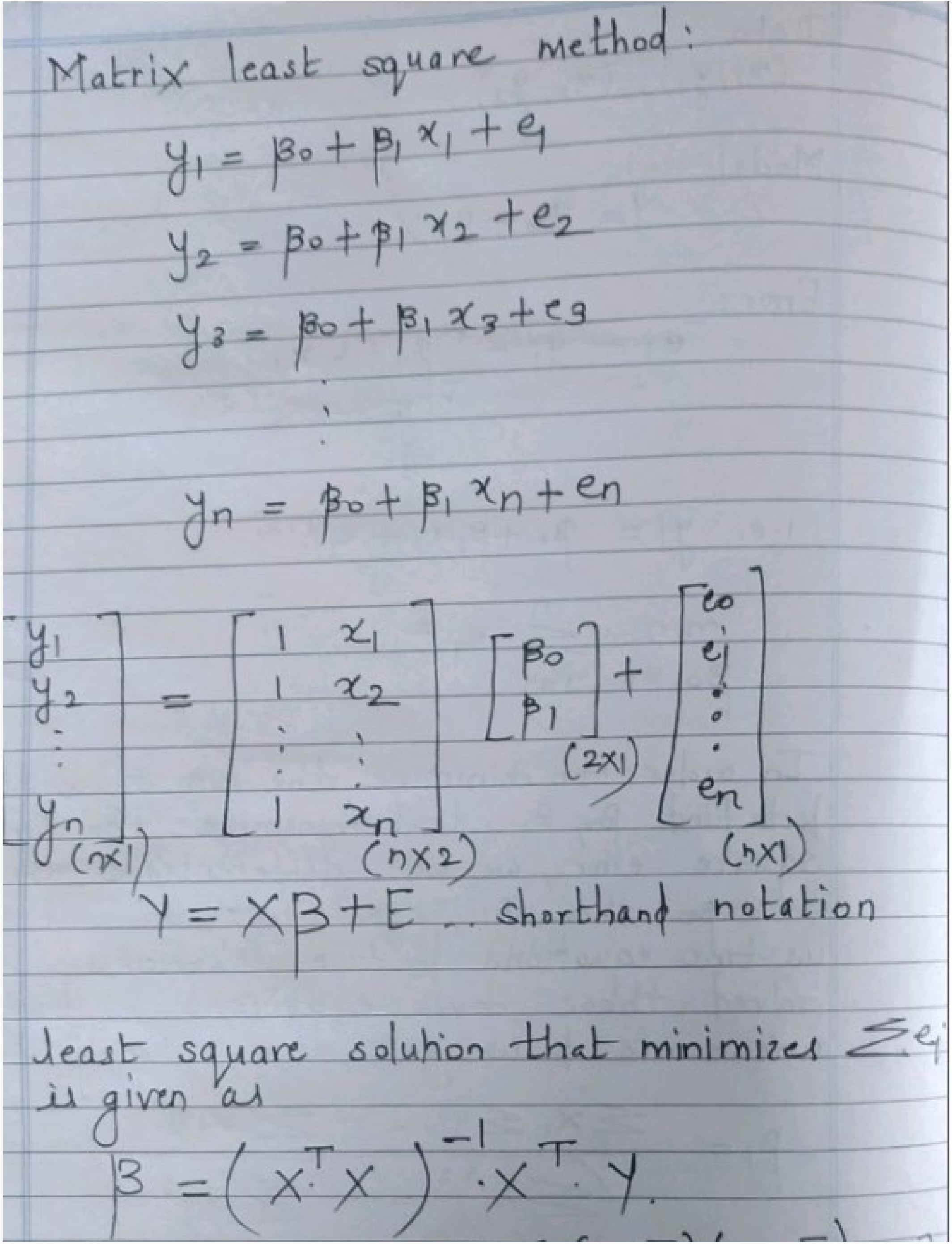
$$\rho = 0.2242$$

Simple Linear Regression: y = a + bx

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b(\sum x)}{n}$$

b = Slope of the line; **a =** Y-intercept of the line.
X = Values of the first data set.; **Y =** Values of the second data set.



Matrix least square method:

$$y_1 = \beta_0 + \beta_1 x_1 + e_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + e_2$$

$$y_3 = \beta_0 + \beta_1 x_3 + e_3$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_n + e_n$$

$$\underset{(n\times1)}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} = \underset{(n\times2)}{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}} \underset{(2\times1)}{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}} + \underset{(n\times1)}{\begin{bmatrix} e_0 \\ e_j \\ \vdots \\ e_n \end{bmatrix}}$$

$Y = X\beta + E$ .. shorthand notation

least square solution that minimizes $\sum e_i$ is given as

$$\beta = (X^T X)^{-1} \cdot X^T \cdot Y.$$

simple linear regression using matrix form.

X = 2 4 6
Y = 3 6 7

Find regression equation for y given x.

Soln

$$\beta = (X^T . X)^{-1} . X^T . Y.$$

$$X^T . X = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 4 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 12 \\ 12 & 56 \end{bmatrix}$$

$$(X^T . X)^{-1} = \frac{1}{\det(X^T . X)} . Adj(X^T . X)$$

$$= \frac{1}{24} \begin{bmatrix} 56 & -12 \\ -12 & 3 \end{bmatrix}$$

$$X^T . Y = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 4 & 6 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \\ 7 \end{bmatrix}$$

$$= \begin{bmatrix} 16 \\ 72 \end{bmatrix}$$

$$\beta = \frac{1}{24} \begin{bmatrix} 56 & -12 \\ -12 & 3 \end{bmatrix} \begin{bmatrix} 16 \\ 72 \end{bmatrix}.$$

$$= \frac{1}{24} \begin{bmatrix} 32 \\ 24 \end{bmatrix} = \begin{bmatrix} 1.33 \\ 1. \end{bmatrix}$$

$$\boxed{Y = 1.33 + X}$$

Multiple Linear Regression.

$$(\underbrace{x_i, u_i, w_i,}_{\text{independent variable}} \; y_i) \quad \nwarrow \text{dependent variable}$$

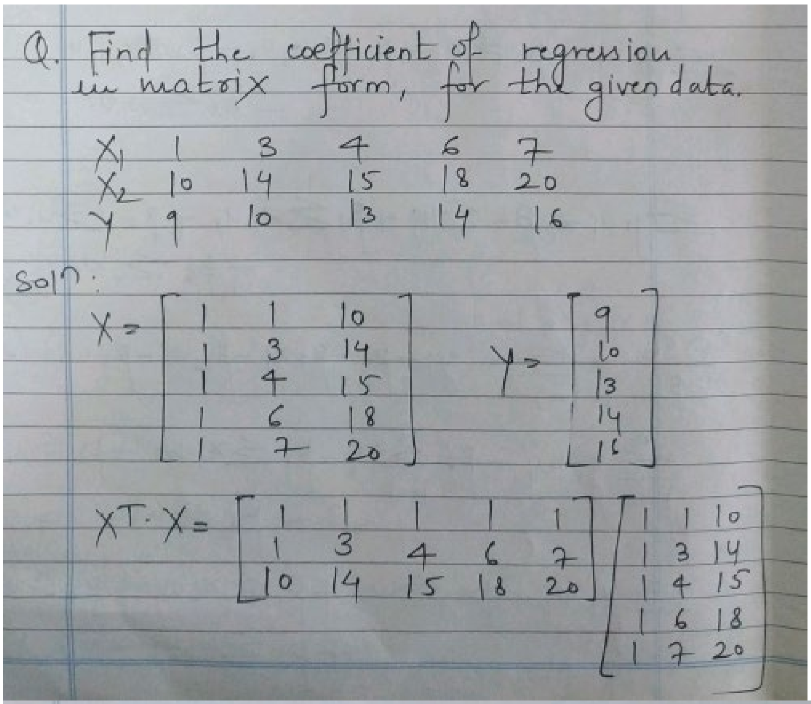
$$\begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i u_i \\ \sum y_i w_i \end{bmatrix} = \begin{bmatrix} n & \sum x_i & \sum u_i & \sum w_i \\ \sum x_i & \sum x_i x_i & \sum u_i x_i & \sum w_i x_i \\ \sum u_i & \sum x_i u_i & \sum u_i u_i & \sum w_i u_i \\ \sum w_i & \sum x_i w_i & \sum u_i w_i & \sum w_i w_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

By Gauss elimination method we can get $\beta_0, \beta_1, \beta_2, \beta_3$

This can get extended for any number of independent variable

$$Y = X\beta + E$$

$$\beta = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$



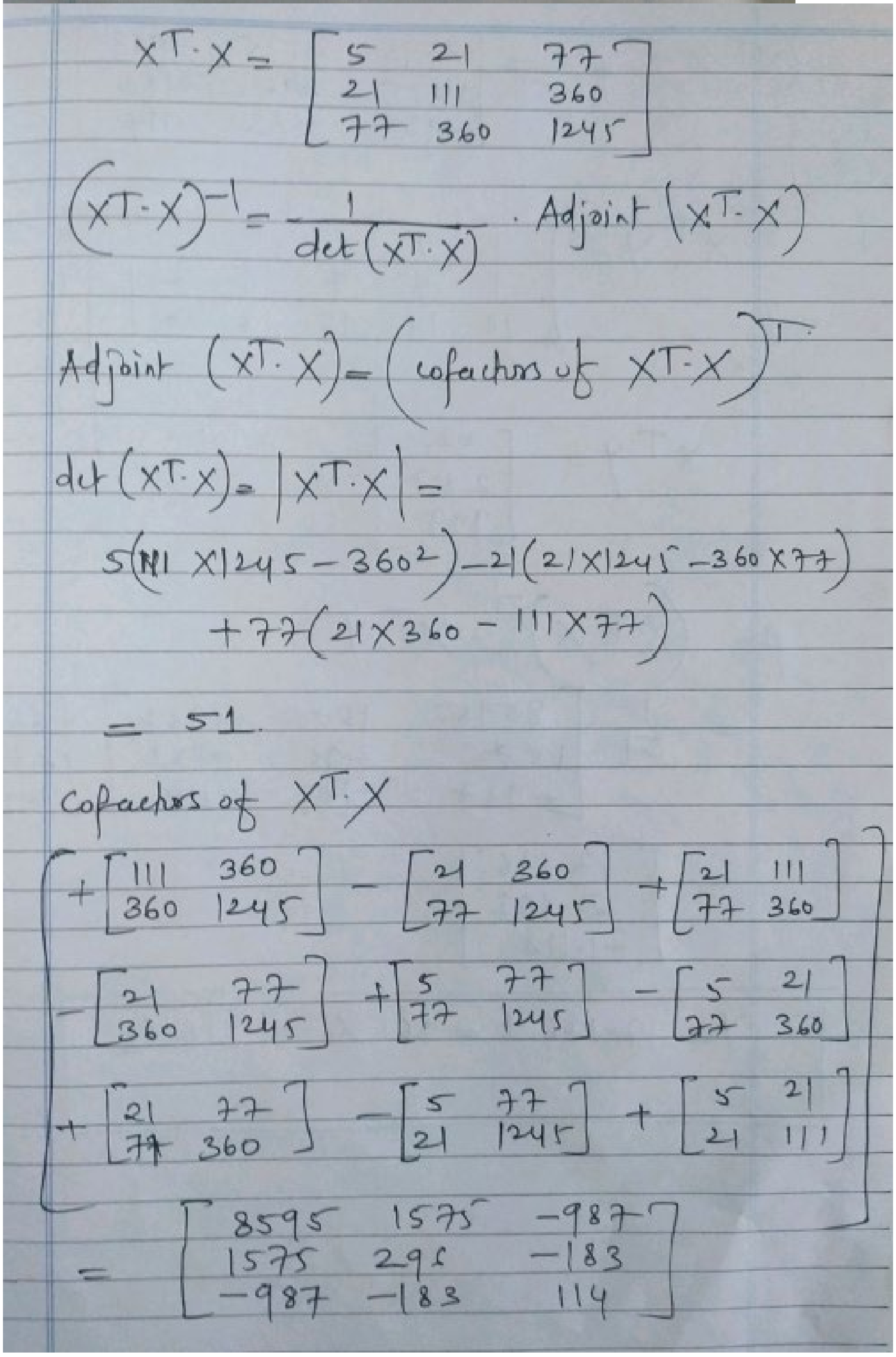
Q. Find the coefficient of regression in matrix form, for the given data.

X_1	1	3	4	6	7
X_2	10	14	15	18	20
Y	9	10	13	14	16

Soln:

$$X = \begin{bmatrix} 1 & 1 & 10 \\ 1 & 3 & 14 \\ 1 & 4 & 15 \\ 1 & 6 & 18 \\ 1 & 7 & 20 \end{bmatrix} \qquad Y = \begin{bmatrix} 9 \\ 10 \\ 13 \\ 14 \\ 16 \end{bmatrix}$$

$$X^T \cdot X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 6 & 7 \\ 10 & 14 & 15 & 18 & 20 \end{bmatrix} \begin{bmatrix} 1 & 1 & 10 \\ 1 & 3 & 14 \\ 1 & 4 & 15 \\ 1 & 6 & 18 \\ 1 & 7 & 20 \end{bmatrix}$$



$$X^T \cdot X = \begin{bmatrix} 5 & 21 & 77 \\ 21 & 111 & 360 \\ 77 & 360 & 1245 \end{bmatrix}$$

$$\left(X^T \cdot X\right)^{-1} = \frac{1}{\det\left(X^T \cdot X\right)} \cdot \text{Adjoint}\left(X^T \cdot X\right)$$

$$\text{Adjoint}\left(X^T \cdot X\right) = \left(\text{cofactors of } X^T \cdot X\right)^T$$

$$\det\left(X^T \cdot X\right) = \left|X^T \cdot X\right| =$$

$$5\left(111 \times 1245 - 360^2\right) - 21\left(21 \times 1245 - 360 \times 77\right) + 77\left(21 \times 360 - 111 \times 77\right)$$

$$= 51.$$

Cofactors of $X^T X$

$$\left(\begin{matrix} +\begin{bmatrix} 111 & 360 \\ 360 & 1245 \end{bmatrix} & -\begin{bmatrix} 21 & 360 \\ 77 & 1245 \end{bmatrix} & +\begin{bmatrix} 21 & 111 \\ 77 & 360 \end{bmatrix} \\ -\begin{bmatrix} 21 & 77 \\ 360 & 1245 \end{bmatrix} & +\begin{bmatrix} 5 & 77 \\ 77 & 1245 \end{bmatrix} & -\begin{bmatrix} 5 & 21 \\ 77 & 360 \end{bmatrix} \\ +\begin{bmatrix} 21 & 77 \\ 77 & 360 \end{bmatrix} & -\begin{bmatrix} 5 & 77 \\ 21 & 1245 \end{bmatrix} & +\begin{bmatrix} 5 & 21 \\ 21 & 111 \end{bmatrix} \end{matrix} \right]$$

$$= \begin{bmatrix} 8595 & 1575 & -987 \\ 1575 & 296 & -183 \\ -987 & -183 & 114 \end{bmatrix}$$

$$\left(X^{T}\cdot X\right)^{-1} = \frac{1}{51}\begin{bmatrix} 8595 & 1575 & -987 \\ 1575 & 296 & -183 \\ -987 & -183 & 114 \end{bmatrix}$$

$$X^{T}\cdot Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 6 & 7 \\ 10 & 14 & 15 & 18 & 20 \end{bmatrix} \begin{bmatrix} 9 \\ 10 \\ 13 \\ 14 \\ 16 \end{bmatrix}$$

$$X^{T}\cdot Y = \begin{bmatrix} 62 \\ 287 \\ 997 \end{bmatrix}$$

$$\beta = \left(X^{T}\cdot X\right)^{-1}\cdot X^{T}\cdot Y$$

$$= \frac{1}{51}\begin{bmatrix} 8595 & 1575 & -987 \\ 1575 & 296 & -183 \\ -987 & -183 & 114 \end{bmatrix}\begin{bmatrix} 62 \\ 287 \\ 997 \end{bmatrix}$$

$$= \begin{bmatrix} 17\cdot176 \\ 2\cdot961 \\ -1\cdot118 \end{bmatrix}$$

$$\boxed{Y = 17\cdot176 + 2\cdot961\,X_1 - 1\cdot118\,X_2}$$

Moments:

In a frequency distribution, to simplify calculation we can use short-cut method.

If $d_i = \frac{(x_i - A)}{h}$ or $(x_i - A) = hd_i$ then, we get the moments about an arbitrary point A are

First order moment $\mu_1' = \frac{\sum_{i=1}^{k} f_i d_i^1}{N} \times h$

Second order moment $\mu_2' = \frac{\sum_{i=1}^{k} f_i d_i^2}{N} \times h^2$

Third order moment $\mu_3' = \frac{\sum_{i=1}^{k} f_i d_i^3}{N} \times h^3$

Fourth order moments $\mu_4' = \frac{\sum_{i=1}^{k} f_i d_i^4}{N} \times h^4$

Similarly, rth order moment about A is given by

$$\mu_r' = \frac{\sum_{i=1}^{k} f_i d_i^r}{N} \times h^r; \quad \text{for } r = 1, 2, \ldots$$

1. β_1 is defined as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

It is used as measure of skewness. For a symmetrical distribution, β_1 shall be zero.

β_1 as a measure of skewness does not tell about the direction of skewness, i.e. positive or negative. Because μ_3 being the sum of cubes of the deviations from mean may be positive or negative but $\mu^2{}_3$ is always positive. Also μ_2 being the variance always positive. Hence β_1 would be always positive. This drawback is removed if we calculate Karl Pearson's coefficient of skewness γ_1 which is the square root of β_1 ,i. e.

$$\gamma_1 = \pm\sqrt{\beta_1} = \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{\mu 3}{\sigma^2}$$

Then the sign of skewness would depend upon the value of μ_3 whether it is positive or negative. It is advisable to use γ_1 as measure of skewness.

2. β_2 measures kurtosis and it is defined by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

and similarly, coefficient of kurtosis γ_2 is defined as

$$\gamma_2 = \beta_2 - 3$$

Statistics: Statistics is a form of mathematical analysis that uses quantified models, representations and synopses for a given set of experimental data or real-life studies. Statistics studies methodologies to gather, review, analyze and draw conclusions from data. Some statistical measures include mean, regression analysis, skewness, kurtosis, variance and analysis of variance. **Function: 1. Summarizing Data:** Statistics helps in organizing and summarizing large amounts of data, making it easier to understand and interpret. The complex mass of figures can be made simple and understandable with the help of statistical methods. Statistical techniques such as averages, dispersion, graph, diagram etc. make huge mass of figures easily understandable. So, the function of statistics is to reduce the complexity of the huge mass of figures to a simpler form.
2. Statistics facilities comparison: The science of statistics does not mean only counting but also comparison. Unless the figures are compared with other figures with the same kind, they are meaningless. Statistical methods such as averages, ratios, percentages, rates, coefficients etc. offer the best way of comparison between two phenomena which will enable to draw valid conclusion. So, statistics helps in the comparison of two phenomena. For example: The statement that "the per capita income of Nepal is $160" is not so clear unless it is compared with the per capita income of any other country. **3. Making Inferences:** While preparing suitable policies and plans, it is necessary to have the knowledge of future tendency. This is mostly in case of industry, commerce and so on. Statistical methods provide helpful means in forecasting the future by studying and analyzing the tendencies based on passed records. For example: Suppose a businessman wants to know the expected sales of T.V. for the next year, the better method for him would be to analyze the sales data of the past years for the estimation of the sales volume for the next year. **4.Testing Hypotheses:** Statistical methods are helpful not only in estimating the present forecasting the future but also helpful in formulating and testing the hypothesis for the development of new theories. Hypothesis like 'whether a particular fertilizer is effective for the production of a particular commodity' 'whether a dice is biased or not' can be tested with the help of statistical tools**. 5. To help in formulation of policies:** Statistics helps in formulating the policies in different fields mainly in economics, business etc. The government policies are also framed on the basis of statistics. In fact, without statistics, suitable policies cannot be framed. Eg: The quantity of food grains to be imported in a particular year depends upon the expected internal production and the expected consumption. That is if the expected wheat production in the particular year be 701 thousands metric tons and that of consumption 710 thousand metric tons so we must import 9 thousand metric tons of food grains.
Limitations: 1) Statistics does not deal with individuals: A part of the definition of statistics is that it must be the aggregates of facts. That is, it deals only with the mass phenomena. A single item or the isolated figure cannot be regarded as statistics. This is a serious limitation of statistics. For example: the mark obtained by a student in English is 75 does not constitute statistics but the average of a group of students in English is 75 forms statistics. **2) Statistics does not study qualitative phenomena:** The science of statistics studies only the quantitative aspect of the problem. Statistics cannot directly be used for the study of qualitative phenomena such as honesty, intelligence, beauty, poverty etc. however, some statistical techniques can be used to study such qualitative phenomena indirectly by expressing them into numbers. For example: the intelligence of the boys can be studied with the help of marks obtained by them in an examination.
3) Statistical laws are not exact: 100% accuracy is rare in statistical work because statistical laws are true only on the average. They are not exact as, are the laws of Physics and Mathematics. For example: the probability of getting a head in a single toss of a coin is ½. This does not imply that 3 heads will be obtained if a coin is tossed 6 times. Only one head, 2 times head or all the times head or no head may be obtained.
4) Statistics is only a means: Statistical methods provide only a method of studying problem. There are other methods also. These methods should be used to supplement the conclusions derived with the help of statistics. **5) Statistics is liable to be misused:** The most important limitation of statistics is that it must be handled by experts. Statistical methods are the most dangerous tools in the hands of inexpert. Since statistics deals with masses of figures, so it can easily be manipulated by inexperienced and skilled persons. Statistical methods if properly be used, may conclude useful results and if misused by inexpert, unskilled persons, it may lead to fallacious conclusion. We have the following example consisting the result concluded by an inexpert and unskilled person.

Importance: (i) Planning: Statistics is indispensable in planning may it be in business, economics or government level. The modern age is termed as the ‘age of planning’ and almost all organizations in the government or business or management are resorting to planning for efficient working and for formulating policy decision. To achieve this end, the statistical data relating to production, consumption, birth, death, investment, income are of paramount importance. Today efficient planning is a must for almost all countries, particularly the developing economies for their economic development.

(ii) Mathematics: Statistics is intimately related to and essentially dependent upon mathematics. The modern theory of Statistics has its foundations on the theory of probability which in turn is a particular branch of more advanced mathematical theory of Measures and Integration. Ever increasing role of mathematics into statistics has led to the development of a new branch of statistics called Mathematical Statistics.

(iii) Economics: Statistics and Economics are so intermixed with each other that it looks foolishness to separate them. Development of modern statistical methods has led to an extensive use of statistics in Economics. All the important branches of Economics—consumption, production, exchange, distribution, public finance—use statistics for the purpose of comparison, presentation, interpretation, etc. Problem of spending of income on and by different sections of the people, production of national wealth, adjustment of demand and supply, effect of economic policies on the economy etc. simply indicate the importance of statistics in the field of economics and in its different branches. Statistics of Public Finance enables us to impose tax, to provide subsidy, to spend on various heads, amount of money to be borrowed or lent etc. So we cannot think of Statistics without Economics or Economics without Statistics.

(iv) Statistics in Social Sciences: Every social phenomenon is affected to a marked extent by a multiplicity of factors which bring out the variation in observations from time to time, place to place and object to object. Statistical tools of Regression and Correlation Analysis can be used to study and isolate the effect of each of these factors on the given observation. Sampling Techniques and Estimation Theory are very powerful and indispensable tools for conducting any social survey, pertaining to any strata of society and then analyzing the results and drawing valid inferences. The most important application of statistics in sociology is in the field of Demography for studying mortality (death rates), fertility (birth rates), marriages, population growth and so on.

(v) Statistics in Trade: As already mentioned, statistics is a body of methods to make wise decisions in the face of uncertainties. Business is full of uncertainties and risks. We have to forecast at every step. Speculation is just gaining or losing by way of forecasting. Can we forecast without taking into view the past? Perhaps, no. The future trend of the market can only be expected if we make use of statistics. Failure in anticipation will mean failure of business. Changes in demand, supply, habits, fashion etc. can be anticipated with the help of statistics. Statistics is of utmost significance in determining prices of the various products, determining the phases of boom and depression etc. Use of statistics helps in smooth running of the business, in reducing the uncertainties and thus contributes towards the success of business.

(vi) Statistics in Research Work: The job of a research worker is to present the result of his research before the community. The effect of a variable on a particular problem, under differing conditions, can be known by the research worker only if he makes use of statistical methods. Statistics are everywhere basic to research activities. To keep alive his research interests and research activities, the researcher is required to lean upon his knowledge and skills in statistical methods.

Classification of Data: It is the process of arranging data into homogeneous (similar) groups according to their common characteristics. Raw data cannot be easily understood, and it is not fit for further analysis and interpretation. Arrangement of data helps users in comparison and analysis.**For example**, the population of a town can be grouped according to sex, age, marital status, etc. The method of arranging data into homogeneous classes according to the common features present in the data is known as classification. A planned data analysis system makes the fundamental data easy to find and recover. This can be of particular interest for legal discovery, risk management, and compliance. Written methods and sets of guidelines for data classification should determine what levels and measures the company will use to organise data and define the roles of employees within the business regarding input stewardship. Once a data - classification scheme has been designed, the security standards that stipulate proper approaching practices for each division and the storage criteria that determines the data's lifecycle demands should be discussed.

Functions/Objectives:
(1) Geographical classification: This is according to place, area or region.
(2) Condensation of data: Classification presents the huge raw data in a condensed form. And it is easily comprehensible to the mind and also highlights the main features contained in the data. **(3) It facilitates comparison:** Classification allows us to make meaningful comparisons depending on the basis of classification. (**4) It helps to study relationships**: The classification of data with respect to two or more comparisons enables us to study the relationship between the two criterion. For example, sex of the students and faculty they join in the university. **(5) It gives statistical treatment of the data:** Classification arranges the huge heterogeneous data into relatively homogeneous groups according to their points of similarities. This way data is made more intelligible and useful.
Types: (i) Content: Content-based classification inspects and interprets files looking for sensitive information**. (ii) Context:** Context-based classification looks at application, location or creator among other variables as indirect indicators of sensitive information. **(iii) User:** User-based classification depends on a manual, end-user selection of each document. User-based classification relies on user knowledge and discretion at creation, edit, review or dissemination to flag sensitive documents. Content-context and user-based approaches can be both right or wrong depending on the business-need and data-type.

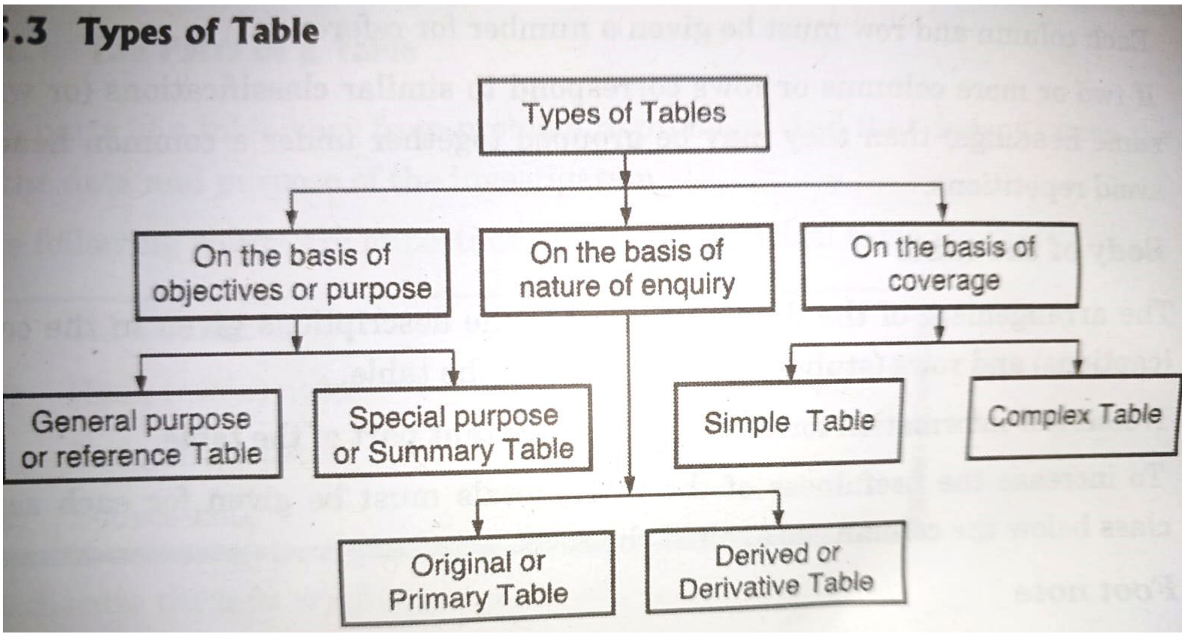
Basis/methods of classification: (1) Geographical classification: When data are classified with reference to geographical locations such as countries, states, cities, districts, etc., it is known as geographical classification. It is also known as 'spatial classification'. **(2) Chronological classification:** A classification where data are grouped according to time is known as a chronological classification. In such a classification, data are classified either in ascending or in descending order with reference to time such as years, quarters, months, weeks, etc. It is also known as temporal classification'. **(3) Qualitative classification:** Under this classification, data are classified on the basis of some attributes or qualities like honesty, beauty, intelligence, literacy, marital status, etc. For example, the population can be divided on the basis of marital status (as married or unmarried) **(4) Quantitative classification:** This type of classification is made on the basis of some measurable characteristics like height, weight, age, income, marks of students, etc.

Variable: The term variable is derived from the word 'vary' that means to differ or change. Hence, variable means the characteristic that varies, differs, or changes from person to person, time to time, place to place, etc. A variable refers to a quantity or attribute whose value varies from one investigation to another. **1) Discrete variables:** Variables that are capable of taking only an exact value and not any fractional value are termed as discrete variables. For example, the number of workers or the number of students in a class is a discrete variable as they cannot be in fraction. Similarly, the number of children in a family can be 1, 2, and so on, but cannot be 1.5, 2.75.
2) Continuous variables: Variables that can take all the possible values (integral as well as fractional) in a given specified range are termed as continuous variables.For example, temperature, height, weight, marks, etc.

Tabulation is a systematic and logical representation of numeric data in rows and columns to facilitate comparison and statistical analysis. It facilitates comparison by bringing related information close to each other and helps in statistical analysis and interpretation. In other words, the method of placing organised data into a tabular form is known as tabulation. It may be complex, double, or simple, depending upon the nature of categorisation.

Objectives Of Tabulation: (1) To simplify complex data: It reduces the bulk of information, i.e., it reduces raw data in a simplified and meaningful form so that it can be easily interpreted by a common man in less time. **(2) To bring out essential features of data:** It brings out the chief/main characteristics of data. It presents facts clearly and precisely without textual explanation**. (3) To facilitate comparison:** The representation of data in rows and columns is helpful in simultaneous detailed comparison on the basis of several parameters. **(4) To facilitate statistical analysis**: Tables serve as the best source of organised data for statistical analysis.The task of computing average, dispersion, correlation, etc., becomes easier if data is presented in the form of a table. **(5) To save space:** A table presents facts in a better way than the textual form. It saves space without sacrificing the quality and quantity of data.

Parts of a Table: 1) Table Number: This is the first part of a table and is given on top of any table to facilitate easy identification and for further reference. **2) Title of the Table:** One of the most important parts of any table is its title. The title is either placed just below the table number or at its right.It is imperative for the title to be brief, crisp and carefully-worded to describe the tables' contents effectively. **3) Headnote:** The headnote of a table is presented in the portion just below the title. It provides information about the unit of data in the table, like "amount in Rupees" or "quantity in kilograms", etc.
4) Column Headings or Captions: The headings of the columns are referred to as the caption. It consists of one or more column heads. A caption should be brief, short, and self-explanatory, Column heading is written in the middle of a column in small letters.
5) Row Headings or Stubs: The title of each horizontal row is called a stub.
6) Body of a Table: This is the portion that contains the numeric information collected from investigated facts. The data in the body is presented in rows which are read horizontally from left to right and in columns, read vertically from top to bottom.
7) Footnote: Given at the bottom of a table above the source note, a footnote is used to state any fact that is not clear from the table's title, headings, caption or stub. For instance, if a table represents the profit earned by a company, a footnote can be used to state if said profit is earned before, or after tax calculations. **8) Source Note:** It refers to the source from where the table's information has been collected.


.3 Types of Table
Types of Tables
On the basis of objectives or purpose
On the basis of nature of enquiry
On the basis of coverage
General purpose or reference Table
Special purpose or Summary Table
Simple Table
Complex Table
Original or Primary Table
Derived or Derivative Table

Diagrammatic and graphic presentation of data means visual representation of the data. It shows a comparison between two or more sets of data and helps in the presentation of highly complex data in its simplest form. Diagrams and graphs are clear and easy to read and understand. In the **diagrammatic** presentation of data, bar charts, rectangles, sub-divided rectangles, pie charts, or circle diagrams are used.

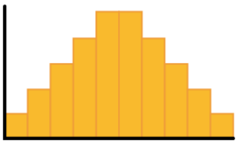
In the **graphic** presentation of data, graphs like histograms, frequency polygon, frequency curves, cumulative frequency polygon, and graphs of time series are used.

Merits: 1. To Simplify the Data; **2.** Appealing presentation; **3.** Helps with Comparison of Data; **4.** Helps in Forecasting; **5.** Saves Time and Labour; **6.** Universally Acceptable; **7.** Helps in Decision Making; **Demerits: 1.** Handle with Care; **2.** Specific Information; **3.** Low precision

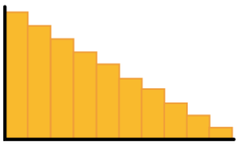
Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

Types of Histogram:

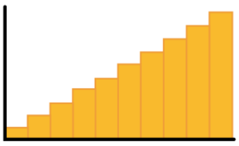
Symmetric (normal) vs skewed and uniform distriutions



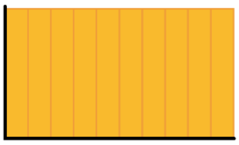
Normal distribution (unimodal, symmetric, the “bell curve”)



Right-skewed distribution (Positively-skewed)

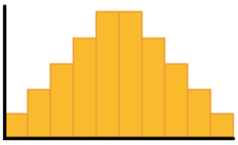


Left-skewed distribution (Negatively-skewed)

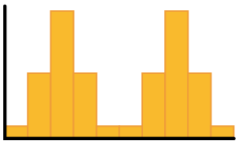


Uniform distribution (equal spread, no peaks)

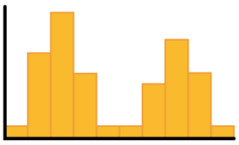
Unimodal vs bimodal distributions



Normal distribution (unimodal, symmetric, the “bell curve”)



Symmetric bimodal distribution (two modes)



Non-symmetric bimodal distribution (two modes)

Feature	Bar Chart/Graph	Histogram
Data Type	Categorical or Discrete Data	Continuous Data
X-Axis	Represents categories or groups	Represents the range of values
Y-Axis	Represents the frequency or count	Represents the frequency or count
Bar Width	Usually has equal width and spacing	Width may vary based on data intervals
Gaps	May have gaps between bars	No gaps between bars
Data Analysis	Used to compare different categories	Used to visualize data distribution
Examples	Comparison of sales by product categories	Distribution of heights of individuals

The sources of data can be classified into two types: statistical and non-statistical. Statistical sources refer to data that is gathered for some official purposes, incorporate censuses, and officially administered surveys. Non-statistical sources refer to the collection of data for other administrative purposes or for the private sector.

Internal data is generated and used within a company or organization. This data is usually produced by the company's operations, such as sales, customer service, or production. You can find that internal data occurs in various formats, such as spreadsheets, databases, or customer relationship management (CRM) systems.

External data is information businesses collect from external sources, including customers, partners, competitors, and industry reports. This data can be purchased from third-party providers or gathered from publicly available sources. Examples of external data include market research reports, social media data, and government data.

Census in statistics refers to the process of collecting data from an entire population rather than a sample. **1. Definition:** A census is a complete enumeration or count of all individuals or elements in a population. It aims to collect data on various characteristics of every member of the population. **2. Purpose:** The main purpose of a census is to obtain accurate and detailed information about the entire population. It provides a comprehensive snapshot of the population at a given point in time. **3. Data Collection:** Census data is typically collected through questionnaires or surveys. Enumerators may visit households, businesses, or other locations to collect the required information. In some cases, online or mail-in questionnaires are used. **4. Population Coverage:** Census data aims to cover the entire population within a defined geographic area or jurisdiction. It includes people of all ages, genders, socioeconomic backgrounds, and other relevant characteristics. **5. Key Variables:** Census data often includes demographic information such as age, sex, race, ethnicity, education level, occupation, income, housing, and other variables specific to the research objectives or government requirements. **6. Government and Policy Planning:** Census data plays a crucial role in government decision-making, policy planning, resource allocation, and understanding population trends. It helps in determining social welfare programs, infrastructure development, healthcare services & electoral representation. **7. Accuracy & Reliability:** Census data aims to be accurate and reliable by capturing information from the entire population. Efforts are made to ensure data quality through training of enumerators, data validation processes, and data cleaning techniques.**8. Privacy and Confidentiality:** Census data collection adheres to strict privacy and confidentiality guidelines. Personal information collected is protected and kept confidential to maintain respondent privacy. **9. Challenges:** Conducting a census can be a complex and resource-intensive process. It requires significant planning, coordination, and financial investment. Ensuring high response rates, reaching remote or hard-to-reach populations, and addressing language barriers are common challenges. **10. Census Frequency:** Census operations are typically conducted at regular intervals, such as once every ten years in many countries. However, the frequency may vary depending on the specific needs and policies of a country.

Primary Data: Data that has been generated by the researcher himself/herself, surveys, interviews, experiments, specially designed for understanding and solving the research problem at hand.
Secondary Data: Using existing data generated by large government Institutions, healthcare facilities etc. as part of organizational record keeping. The data is then extracted from more varied datafiles.

Primary Data	**Secondary Data**
Refers to the data that is collected directly from the source	Refers to the data that has already been collected by someone else
Generally collected through surveys, observations, experiments, etc.	Generally collected through sources such as government publications, books, journals, reports, etc.
Original and unique data	Data that has already been processed and analyzed by someone else
Generally more accurate and reliable as it is collected for a specific purpose	May be less accurate and reliable as it may not be collected for a specific purpose
Time-consuming and expensive to collect	Less time-consuming and less expensive to collect
Provides a deeper insight into the problem at hand	Provides a broader overview of the problem at hand

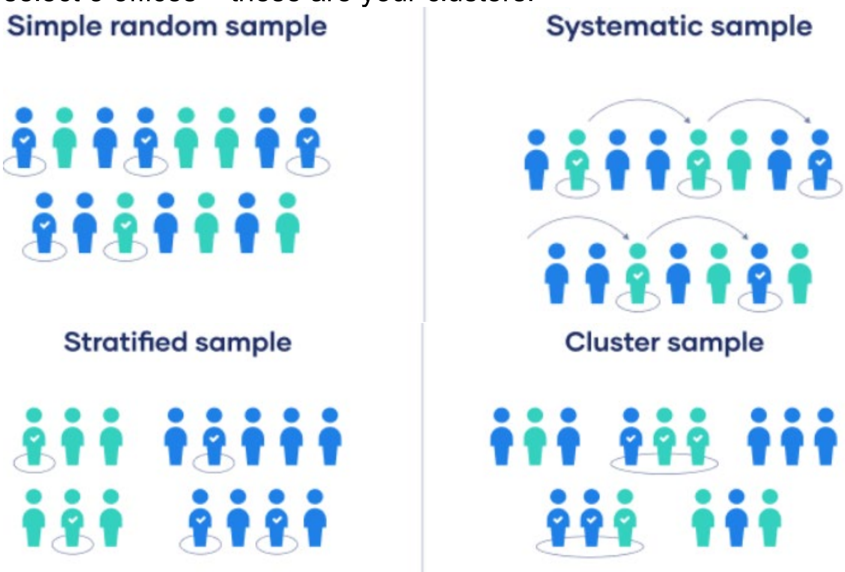
Probability sampling refers to the selection of a sample from a population, when this selection is based on the principle of randomization, that is, random selection or chance. Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice. There are 4 types of probability sample.

1. Simple random sampling: every member of the population has an equal chance of being selected. Your sampling frame should include the whole population. To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance. **Example:** You want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

2. Systematic sampling: Similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals. **Example:** All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

3. Stratified sampling: It involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample. To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role). Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup. **Example:** The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

4. Cluster sampling: Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It’s difficult to guarantee that the sampled clusters are really representative of the whole population.**Example:** The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don’t have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.



Non-probability sampling individuals are selected based on non-random criteria, and not every individual has a chance of being included. This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias.

1. Convenience sampling:A convenience sample simply includes the individuals who happen to be most accessible to the researcher. This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias. **Example:** You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

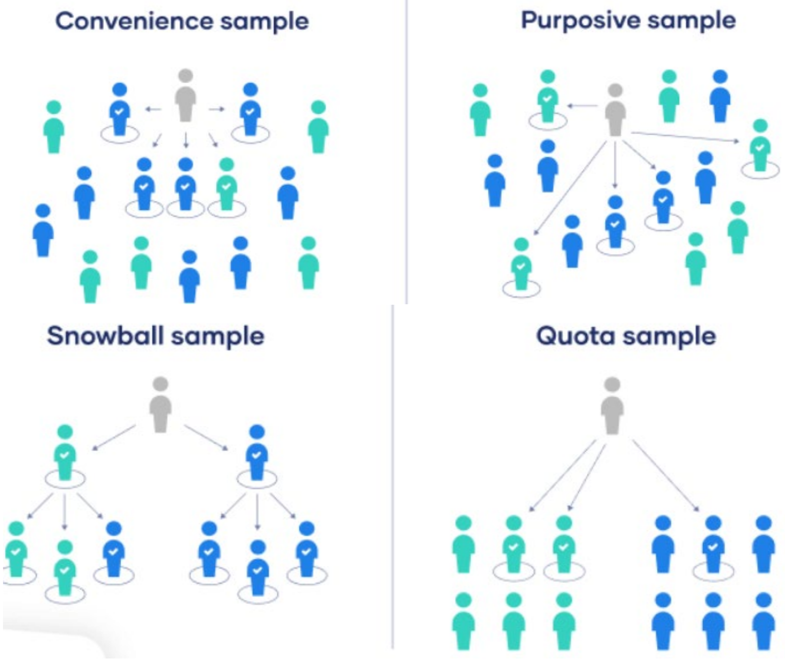
2. Voluntary response sampling: Similar to a convenience sample it is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g.responding to a public online survey). Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others, leading to self-selection bias.

Example: Voluntary response sampling: You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can't be sure that their opinions are representative of all students.

3. Purposive sampling: This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research. It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your inclusion and exclusion criteria and beware of observer bias affecting your arguments. **Example:** You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

4. Snowball sampling: If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to sampling bias.

Example: You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.



Primary data collection methods involve gathering new and original data directly from the source:**1. Surveys:** Conducting surveys through questionnaires, interviews, or online forms to gather information directly from respondents. Surveys can be administered in person, over the phone, via email, or through online survey platforms. **2. Observations:** Making direct observations of people, objects, or events to collect data. This can involve structured observations with predefined criteria or unstructured observations to capture natural behavior. **3. Experiments:** Designing controlled experiments to gather data by manipulating variables and observing the outcomes. Experiments are often used in scientific research to establish cause-and-effect relationships. **4. Interviews:** Conducting one-on-one or group interviews to collect detailed information and insights from participants. Interviews can be structured (using a fixed set of questions) or unstructured (allowing for open-ended discussions). **5. Focus groups:** Organizing small group discussions with selected participants to explore their opinions, attitudes, and experiences regarding a specific topic. Focus groups encourage interaction and generate rich qualitative data. **6. Case studies:** Conducting in-depth investigations of individuals, groups, organizations, or events to gather comprehensive and detailed data. Case studies often involve multiple data collection methods, such as interviews, observations, and document analysis. **7. Document analysis:** Examining existing documents, records, reports, or other written materials to extract relevant data. This can include reviewing official documents, archival records, organizational documents, or personal diaries. **8. Field experiments:** Conducting experiments in real-world settings, such as a natural environment or a specific community, to collect data that reflects real-life conditions and behaviors. **9. Ethnography:** Immersing oneself in a particular culture or community to observe and understand their behaviors, practices, and social interactions. Ethnographic research involves extended periods of observation and participant engagement. **10. Sensor data collection:** Using various sensors and data collection devices, such as GPS trackers, wearable devices, or environmental sensors, to collect real-time data on physical parameters, location, or other variables.

Secondary data collection methods involve gathering information that has been previously collected and published by other sources: **1. Published sources:** Utilizing books, journals, research papers, magazines, newspapers, and other publications to extract relevant data. **2. Government sources:** Accessing data from government agencies, such as census reports, statistical publications, official websites, and public records. **3. Online databases:** Exploring various online databases, repositories, and archives that contain a wide range of data, including academic research, industry reports, and government publications. **4. Research studies:** Reviewing existing research studies, surveys, and experiments conducted by other researchers in the field. **5. Institutional sources:** Gathering data from educational institutions, libraries, research centers, and non-profit organizations that may have collected and shared relevant data. **6. Online platforms and websites:** Extracting data from websites, online forums, social media platforms, and other digital sources that provide information related to the research topic. **7. Data aggregators:** Accessing data aggregators or data providers that compile and organize data from various sources, such as market research firms or specialized data companies. **8. Historical records:** Analyzing historical documents, records, archives, and artifacts that contain valuable information about the research subject. **9. Secondary surveys:** Utilizing data collected by other researchers through surveys, questionnaires, or interviews and made available for further analysis. **10. Meta-analysis:** Conducting a systematic review and analysis of existing research studies and their findings to derive conclusions or identify patterns.

	Probability Sampling	**Non-Probability Sampling**
Selection Method	Random selection of participants from the population	Non-random selection of participants from the population
Representative	Provides a representative sample of the population	May not provide a representative sample of the population
Generalizability	Results can be generalized to the population	Results may have limited generalizability
Bias	Less prone to bias and systematic errors	More prone to bias and systematic errors
Sample Size	Requires a larger sample size for accurate estimation	Smaller sample size may be sufficient
Statistical	Statistical techniques can be applied for inference	Statistical inference may be limited or not possible
Precision	Results tend to have higher precision and accuracy	Results may have lower precision and accuracy
Cost	Can be more time-consuming and expensive to implement	Can be less time-consuming and less expensive

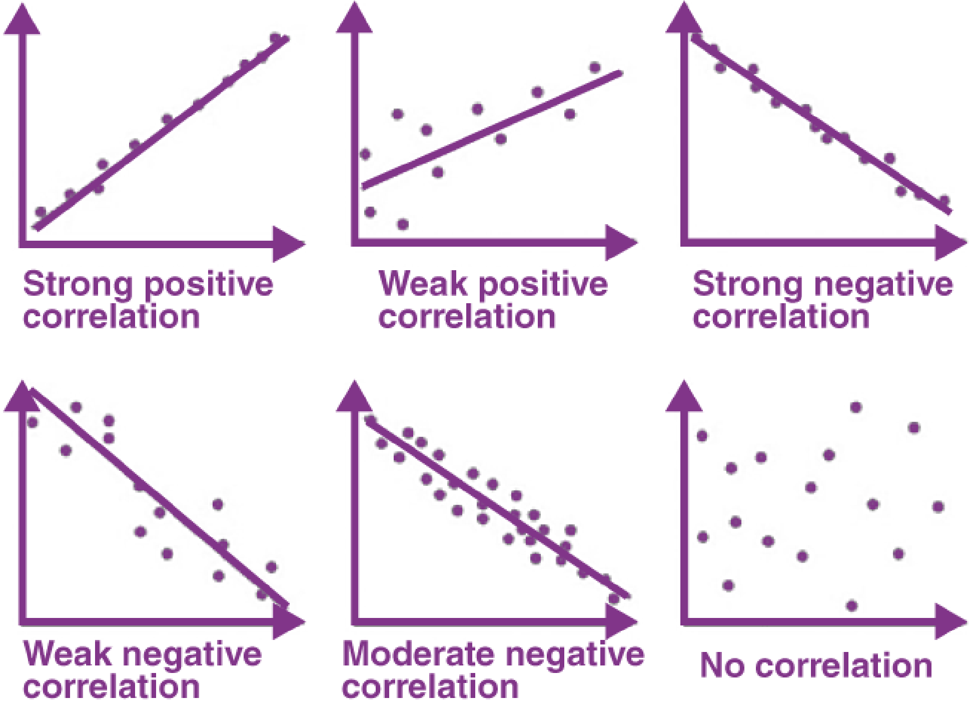
Advantages of Probability Sampling: 1. Representative Sample: Probability sampling methods aim to provide a representative sample of the population, ensuring that each individual has an equal chance of being selected. This enhances the generalizability of the findings. **2. Statistical Inference:** Probability sampling allows for the application of statistical techniques to estimate population parameters and make valid inferences. The results obtained from probability samples can be used to make reliable conclusions about the population. **3. Reduced Bias:** Probability sampling methods are designed to minimize bias and ensure that the sample is unbiased. This helps in obtaining accurate and reliable results. **4. Precision:** Probability sampling tends to result in higher precision and accuracy in estimation, as it allows for the calculation of sampling errors and confidence intervals.

Disadvantages of Probability Sampling: 1. Resource Intensive: Probability sampling methods often require a larger sample size to achieve accurate estimates. This can be time-consuming and costly, especially when dealing with large populations. **2. Infeasible for Certain Populations:** Probability sampling may not be feasible or practical for certain populations, such as rare or hard-to-reach populations. This can limit the applicability of probability sampling methods.

Advantages of Non-Probability Sampling: 1. Convenience and Accessibility: Non-probability sampling methods are often easier to implement and require less time and resources. They can be useful when access to the entire population is difficult or when resources are limited. **2. Quick Data Collection:** Non-probability sampling methods allow for rapid data collection, as they do not require extensive sampling frames or random selection procedures. **3. Flexibility:** Non-probability sampling provides flexibility in selecting participants, allowing researchers to target specific groups or individuals of interest.

Disadvantages of Non-Probability Sampling: 1. Biased Sample: Non-probability sampling methods may result in biased samples, as certain segments of the population may be overrepresented or underrepresented. This can limit the generalizability of the findings. **2. Limited Statistical Inference:** Non-probability sampling does not allow for precise statistical inference about the population. The results obtained from non-probability samples may not be representative of the entire population. **3. Sampling Errors Unknown:** Non-probability sampling does not provide a known sampling error or a way to estimate it. This can affect the accuracy and reliability of the results.

Correlation refers to a process for establishing the relationships between two variables. The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables. The two variables are often given the symbols X and Y. In order to illustrate how the two variables are related, the values of X and Y are pictured by drawing the scatter diagram, graphing combinations of the two variables. The scatter diagram is given first, and then the method of determining Pearson's r is presented. **Types of Correlation:** The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables**: 1) Positive Correlation:** when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.**2)Negative Correlation:** when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable**. 3)No Correlation:** when there is no linear dependence or no relation between the two variables.



Karl Pearson's coefficient of correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linearly related variables. The coefficient of correlation is expressed by "r".

$$\mathrm{r} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}} \qquad \mathrm{Cov(X,Y)} = \frac{\Sigma(\mathrm{X}-\overline{\mathrm{X}})(\mathrm{Y}-\overline{\mathrm{Y}})}{\mathrm{N}} = \frac{\Sigma\,\mathrm{xy}}{\mathrm{N}}$$

Pearson Correlation Coefficient Formula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Sample Correlation Coefficient Formula: $r_{xy} = S_{xy}/S_xS_y$
Where Sx and Sy are the sample standard deviations, and Sxy is the sample covariance.

Population Correlation Coefficient Formula: rxy **= $\sigma_{xy}/\sigma_x\sigma_y$**
The population correlation coefficient uses σx and σy as the population standard deviations and σxy as the population covariance.

The **slope** indicates the steepness of a line and the **intercept** indicates the location where it intersects an axis. The slope and the intercept define the linear relationship between two variables, and can be used to estimate an average rate of change. The greater the magnitude of the slope, the steeper the line and the greater the rate of change. By examining the equation of a line, you quickly can discern its slope and y-intercept (where the line crosses the y-axis). Usually, this relationship can be represented by the equation y = b0 + b1x, where b0 is the y-intercept and b1 is the slope. **For example**, a company determines that job performance for employees in a production department can be predicted using the regression model y = 130 + 4.3x, where x is the hours of in-house training they receive (from 0 to 20) and y is their score on a job skills test. The value of the y-intercept (130) indicates the average job skill score for an employee with no training. The value of the slope (4.3) indicates that for each hour of training, the job skill score increases, on average, by 4.3 points.

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc.
Uses: 1. Relationship Assessment: Regression analysis allows us to assess the strength, direction, and significance of the relationship between a dependent variable and one or more independent variables. It helps us understand how changes in independent variables influence the dependent variable. **2. Prediction:** Regression analysis enables us to make predictions or forecasts based on the identified relationship between variables. By fitting a regression model to the data, we can estimate the value of the dependent variable for given values of the independent variables. **3. Variable Selection:** Regression analysis aids in identifying the most significant variables that impact the dependent variable. It helps in determining which independent variables are statistically significant predictors and should be included in the regression model.
4. Hypothesis Testing: Regression analysis allows us to test hypotheses about the relationship between variables. By examining the coefficients and associated p-values, we can assess whether the relationship is statistically significant and draw conclusions about the underlying population. **5. Model Evaluation:** Regression analysis provides measures to evaluate the quality and fit of the regression model. Techniques such as R-squared, adjusted R-squared & residual analysis help in assessing how well the model captures the variation in the data & whether the assumptions of the model are met.

Linear regression is a statistical regression method which is used for predictive analysis. It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables. It is used for solving the regression problem in machine learning. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression. If there is only one input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variable, then such linear regression is called multiple linear regression. The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of the year of experience. y = a + bx

$$\text{b=} \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$\text{a=} \frac{\sum y - b(\sum x)}{n}$$

b = Slope of the line; **a =** Y-intercept of the line.
X = Values of the first data set.; **Y =** Values of the second data set.

The essential step in any machine learning model is to evaluate the accuracy of the model. **The Mean Squared Error, Mean absolute error, Root Mean Squared Error, and R-Squared or Coefficient of determination** metrics are used to evaluate the performance of the model in regression analysis. **1) The Mean absolute error (MAE)** represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Where,
$\hat{y}$ − *predicted value of y*
$\bar{y}$ − *mean value of y*

2) Mean Squared Error (MSE) represents the average of the squared difference b/w the original & predicted values in the data set. It measures the variance of the residuals.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

3) Root Mean Squared Error (RMSE) is the square root of Mean Squared error. It measures the standard deviation of residuals.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

4) The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

5) Adjusted R squared is a modified version of R square, and it is adjusted for the number of independent variables in the model, and it will always be less than or equal to R^2.In the formula below n is the number of observations in the data and k is the number of the independent variables in the data.

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1}\right]$$

Differences among these evaluation metrics:

1) Mean Squared Error(MSE) and Root Mean Square Error penalizes the large prediction errors vi-a-vis Mean Absolute Error (MAE). However, RMSE is widely used than MSE to evaluate the performance of the regression model with other random models as it has the same units as the dependent variable (Y-axis).

2) MSE is a differentiable function that makes it easy to perform mathematical operations in comparison to a non-differentiable function like MAE. Therefore, in many models, RMSE is used as a default metric for calculating Loss Function despite being harder to interpret than MAE.**3)** The lower value of MAE, MSE, and RMSE implies higher accuracy of a regression model. However, a higher value of R square is considered desirable.

4) R Squared & Adjusted R Squared are used for explaining how well the independent variables in the linear regression model explains the variability in the dependent variable. R Squared value always increases with the addition of the independent variables which might lead to the addition of the redundant variables in our model. However, the adjusted R-squared solves this problem. **5)** Adjusted R squared takes into account the number of predictor variables, and it is used to determine the number of independent variables in our model. The value of Adjusted R squared decreases if the increase in the R square by the additional variable isn't significant enough. **6)** For comparing the accuracy among different linear regression models, RMSE is a better choice than R Squared.

The mean absolute percentage error (MAPE), also called the mean absolute percentage deviation (MAPD): measures accuracy of a forecast system. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

n is the number of fitted points,
A_t is the actual value,
F_t is the forecast value.

Mathematical Equations: 1. Definition: Mathematical equations express a relationship between variables using mathematical symbols and operations. They describe mathematical laws, principles, and formulas**. 2. Types:** Mathematical equations can be algebraic, differential, integral, or a combination of these. They may involve basic arithmetic operations, algebraic manipulations, trigonometric functions, logarithms, exponentials, and more**. 3. Solving:** Mathematical equations are solved by applying mathematical operations to isolate the variable of interest. This is done using algebraic techniques, like simplification, factoring, substitution, or solving systems of equations.

Statistical Equations: 1. Definition: Statistical equations are mathematical equations used in statistical analysis to model and analyze data. They describe the relationship between variables and provide measures of central tendency, dispersion, correlation, regression, and more. **2. Types:** Statistical equations include formulas for calculating measures like mean, median, mode, variance, standard deviation, correlation coefficient, regression coefficients, hypothesis tests, and confidence intervals. **3. Applications:** Statistical equations are used to summarize, interpret, and draw conclusions from data. They help in understanding patterns, relationships, and variability in data, and are crucial in making informed decisions based on data analysis.

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know: 1) How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth). 2) The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).
Assumptions of multiple linear regression: 1) Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable. **2) Independence of observations:** the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among variables. In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model. **3) Normality:** The data follows a normal distribution.
4) Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor. $y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$
y = the predicted value of the dependent variable
B_0 = the y-intercept (value of y when all other parameters are set to 0)
B_1X_1 = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value) … = do the same for however many independent variables you are testing
B_nX_n = the regression coefficient of the last independent variable
\epsilon = model error (a.k.a. how much variation there is in our estimate of y)
outcome: The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model, even though the predictors may not be related to the outcome variable.
Uses: MLR examines how multiple independent variables are related to one dependent variable. Once each of the independent factors has been determined to predict the dependent variable, the information on the multiple variables can be used to create an accurate prediction on the level of effect they have on the outcome variable.

Partial relation coefficient: Partial regression coefficients, also known as standardized regression coefficients or beta coefficients, are important measures in multiple regression analysis. **1. Definition:** Partial regression coefficients represent the unique contribution of each independent variable in a multiple regression model while controlling for the effects of other independent variables. They indicate the change in the dependent variable associated with a one-unit change in the respective independent variable, holding all other independent variables constant. **2. Standardization:** Partial regression coefficients are standardized, which means they are expressed in standard deviation units. This allows for comparing the relative importance and contribution of different independent variables, regardless of their scales or units of measurement.
3. Interpretation: The value of a partial regression coefficient indicates the strength and direction of the relationship between an independent variable and the dependent variable, after accounting for the effects of other independent variables. A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship. **4. Importance:** Partial regression coefficients help in assessing the unique impact of each independent variable on the dependent variable. By considering these coefficients, researchers can identify the most influential predictors and understand their relative contributions to the overall regression model.
5. Usefulness: Partial regression coefficients are useful for comparing the relative importance of independent variables and determining their individual significance in explaining the variation in the dependent variable. They also help in identifying potential multicollinearity issues & understanding the independent variables' specific effects.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

(a) Partial regression coefficient: The partial regression coefficient, also known as the partial slope coefficient or standardized regression coefficient, measures the change in the dependent variable associated with a one-unit change in a specific independent variable while holding all other independent variables constant. It quantifies the unique contribution of the independent variable to the regression model. The range of the partial regression coefficient is from negative infinity to positive infinity.

(b) Partial correlation coefficient: The partial correlation coefficient measures the strength and direction of the linear relationship between two variables while controlling for the effects of other variables. It represents the correlation between two variables after removing the influence of other variables. The range of the partial correlation coefficient is from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.

(c) Multiple correlation coefficient: The multiple correlation coefficient, also known as the coefficient of multiple determination, measures the strength and direction of the linear relationship between a dependent variable and multiple independent variables in a regression model. It represents the proportion of the total variation in the dependent variable that can be explained by the independent variables. The range of the multiple correlation coefficient is from 0 to 1, where 0 indicates no relationship between the variables and 1 indicates a perfect relationship where all the variation in the dependent variable is explained by the independent variables.

Test of Significance for Overall Fit of the Model: 1. The overall fit of a regression model can be assessed using the F-test or the R-squared value. **2.** The F-test evaluates whether the regression model, as a whole, significantly explains the variation in the dependent variable. **3.** The F-test compares the explained variation (due to the regression model) to the unexplained variation (residuals) in the data. **4.** The null hypothesis states that the regression model has no explanatory power, while the alternative hypothesis suggests that the model is a significant predictor of the dependent variable. **5.** If the calculated F-statistic is greater than the critical F-value at a chosen significance level, we reject the null hypothesis and conclude that the regression model has a significant overall fit.

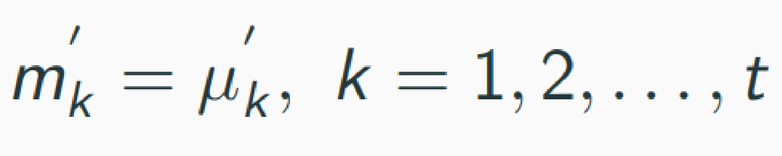
Test of Significance for Individual Coefficients: 1. The significance of individual coefficients in a regression model can be assessed using t-tests or p-values. **2.** Each coefficient represents the effect of a specific independent variable on the dependent variable while holding other variables constant. **3.** The null hypothesis for each coefficient test states that the coefficient is equal to zero, indicating no effect of the corresponding independent variable. **4.** The alternative hypothesis suggests that the coefficient is significantly different from zero, indicating a significant effect of the independent variable. **5.** The t-test calculates the t-value by dividing the estimated coefficient by its standard error. The t-value is compared to the critical t-value at a chosen significance level. If the t-value is greater than the critical t-value, we reject the null hypothesis and conclude that the coefficient is statistically significant.

Software Requirements Specification (SRS) is a document that outlines the functional and non-functional requirements of a software system. It serves as a communication bridge between stakeholders, project managers, and development teams.
Advantages of SRS: 1. Clarity and Understanding: SRS provides a clear and concise understanding of the software requirements, ensuring that all stakeholders have a common understanding of what needs to be developed. **2. Requirement Verification:** SRS helps in verifying the accuracy and completeness of the requirements. It acts as a reference document for validating the software against the specified requirements.
3. Scope Management: SRS defines the boundaries and scope of the software project, helping in managing expectations and preventing scope creep. **4. Requirement Baseline:** SRS serves as a baseline for the project, providing a documented agreement on the requirements that can be referred to throughout the development lifecycle.
5. Risk Mitigation: By clearly defining requirements, SRS helps identify potential risks and challenges early in the project, allowing for effective risk management.
Features of SRS: 1. Functional Requirements: SRS captures the functional requirements of the software, describing what the system should do and the desired behaviors. **2. Non-Functional Requirements:** SRS also includes non-functional requirements such as performance, security, usability, and reliability criteria that the software should meet. **3. User Interface Design:** SRS may include user interface specifications, describing the layout, navigation, and visual elements of the software.
4. System Constraints: SRS specifies any constraints or limitations that need to be considered during the development process, such as hardware or software compatibility requirements. **5. Use Cases and Scenarios:** SRS may include use cases or scenarios that describe how the system will be used, helping to understand the interaction between users and the software.

Parametric point estimation: 1) Unbiasedness: Unbiasedness is a desirable property of a point estimator in parametric point estimation. An estimator is unbiased if, on average, it produces an estimate that is equal to the true parameter value. In other words, the expected value of the estimator is equal to the true parameter value. Mathematically, an estimator $\bar{Y}$ is unbiased for a parameter θ if $E(\bar{Y}) = \theta$.
2) Efficiency: Efficiency measures the precision or the information content of an estimator in parametric point estimation. An efficient estimator achieves the smallest possible variance among all unbiased estimators. In simple terms, it means that the estimator provides estimates that are close to the true parameter value with minimum variability. An estimator is considered efficient if it has the smallest possible variance among all unbiased estimators. **3) Consistency: C**onsistency is another important property of a point estimator. A consistent estimator converges to the true parameter value as the sample size increases. In other words, as more data is collected, the estimates become more accurate and approach the true value. Mathematically, an estimator $\bar{Y}$ is consistent for a parameter θ if, as the sample size n approaches infinity, the probability that $|\bar{Y} - \theta| > \varepsilon$ (where ε is a small positive value) approaches zero.
4) Sufficiency: A sufficient statistic contains all the information needed to estimate a parameter in parametric point estimation. It summarizes the relevant information from the data and discards unnecessary details. A sufficient statistic is a function of the data that does not lose any information about the parameter. It allows for efficient estimation by reducing the dimensionality of the data. **5) Mean-squared error:** Mean-squared error (MSE) is a measure of the average squared difference between an estimator and the true parameter value. It combines the bias and variance of an estimator into a single measure. A good estimator should have a small MSE, indicating both low bias and low variability. The formula for mean-squared error (MSE) is MSE = E((Ȳ - θ)^2), where $\bar{Y}$ is the estimator and θ is the true parameter value. It is the average of the squared difference between the estimator and the true value. **6) Best Asymptotically Normal (BAN) Estimators:** BAN estimators are estimators that achieve the best possible asymptotic properties, particularly in large sample sizes. They are consistent, asymptotically unbiased, and have the smallest variance. BAN estimators play a crucial role in statistical inference when dealing with large datasets. They are estimators that achieve the best possible asymptotic properties in large sample sizes, such as consistency and efficiency. **7) Completeness:** Completeness is a property of a statistic in parametric point estimation. A statistic is complete if it captures all the information about the parameter. In other words, there are no unbiased estimators of the parameter that do not depend on the statistic. Completeness ensures that there are no alternative estimators that provide additional information beyond what is already captured by the statistic. Completeness is determined based on the properties of the statistic and does not have a specific formula. **8) Admissibility:** Admissibility refers to the optimality of an estimator in parametric point estimation. An estimator is admissible if there is no other estimator that dominates it. Dominance means that the dominating estimator performs at least as well as the dominated estimator for all possible parameter values and performs strictly better for some parameter values. Admissibility ensures that an estimator is not outperformed by any other estimator in terms of both bias and variance. Admissibility is not defined by a formula but by the concept of dominance. An estimator is admissible if there is no other estimator that dominates it, meaning it performs at least as well for all parameter values and strictly better for some parameter values.

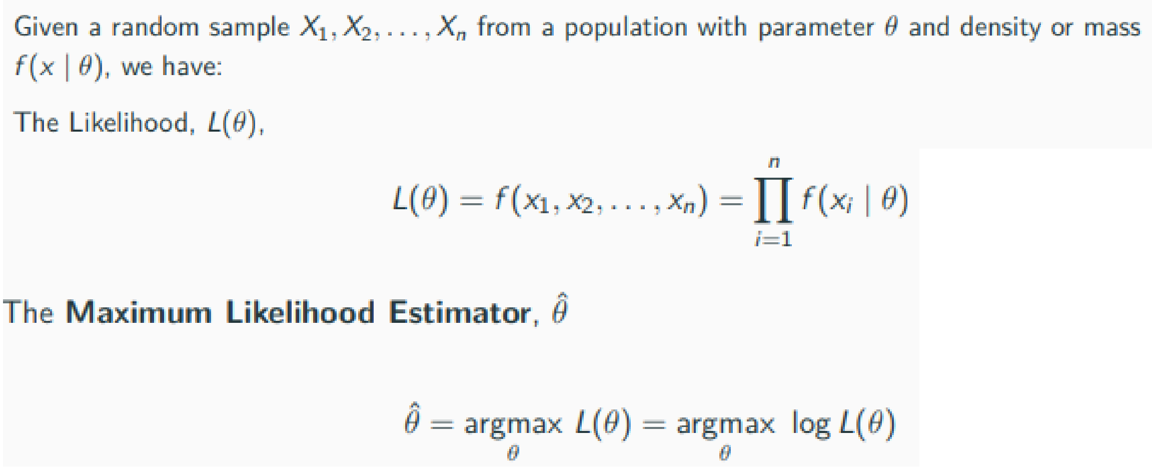
Method of Moments: 1) The Method of Moments is a technique used to estimate the parameters of a statistical model by equating sample moments with population moments. **2)** It assumes that the population moments can be expressed as functions of the unknown parameters. **3)** The method involves setting up equations based on sample moments and solving them to obtain estimators for the parameters. **4)** It aims to find the parameter values that make the sample moments match the corresponding population moments. **5)** The estimators obtained through the method of moments are consistent, meaning they converge to the true parameter values as the sample size increases. **6)** However, the method may not always yield unique or efficient estimators, and it may not be applicable for all statistical models. **Advantages: 1)** Easy to understand and implement. **2)** Provides consistent estimators, meaning they converge to the true parameter values as the sample size increases. **3)** Can be applied to a wide range of statistical models. **4)** Requires fewer assumptions compared to other estimation methods. **Disadvantages: 1)** May not always yield unique or efficient estimators. **2)** Relies on the assumption that the population moments can be expressed as functions of the unknown parameters, which may not always hold. **3)** The method may not be applicable or suitable for complex or high-dimensional models. **4)** Can be sensitive to outliers in the data.

The Method Of Moments (MOM) consists Of equating sample moments and population moments. If a population has t parameters, the MOM consists of solving the system of equations.



$$m_k^{'} = \mu_k^{'}, \;\; k = 1, 2, \ldots, t$$

Method of Maximum Likelihood: 1) The Method of Maximum Likelihood is a statistical technique used to estimate the parameters of a model by maximizing the likelihood function. **2)** It assumes that the data are generated from a specific probability distribution with unknown parameters. **3)** The likelihood function is a measure of how likely the observed data are under the assumed distribution and parameter values. **4)** The method involves finding the parameter values that maximize the likelihood function, or equivalently, the log-likelihood function. **5)** The estimators obtained through the method of maximum likelihood are asymptotically efficient, meaning they achieve the smallest possible variance among consistent estimators as the sample size increases. **6)** The method relies on assumptions about the underlying distribution and may not be appropriate for small sample sizes or when the model assumptions are violated. **Advantages: 1)** Asymptotically efficient estimators, meaning they achieve the smallest possible variance among consistent estimators as the sample size increases. **2)** Utilizes the full likelihood function, making use of all available information in the data. **3)** Can provide interval estimates and hypothesis tests based on likelihood ratio tests. **4)** Well-established theoretical properties and widely used in statistical inference. **Disadvantages: 1)** Requires assumptions about the underlying distribution and model structure, which may not always be realistic or accurate. **2)** May not have closed-form solutions, requiring numerical optimization methods. **3)** Can be sensitive to the choice of starting values for the optimization process. **4)** Can be computationally intensive for complex models or large datasets.



Given a random sample $X_1, X_2, \ldots, X_n$ from a population with parameter θ and density or mass $f(x \mid \theta)$, we have:

The Likelihood, $L(\theta)$,

$$L(\theta) = f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

The **Maximum Likelihood Estimator**, $\hat{\theta}$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\; L(\theta) = \underset{\theta}{\operatorname{argmax}}\; \log L(\theta)$$

Neyman-Fisher Factorization Theorem is a fundamental result in statistical theory that provides a way to factorize the likelihood function into two components: one depending on the parameter of interest and the other depending on the nuisance parameters. This theorem is useful in statistical inference, hypothesis testing, and model selection.
The Neyman-Fisher Factorization Theorem states that the likelihood function can be factorized as follows: $L(\theta; x) = g(T(x); \theta)h(x)$, where:

- $L(\theta; x)$ is the likelihood function, which is the probability density function (pdf) or probability mass function (pmf) of the observed data x, given the parameter θ.
- θ is the parameter of interest; - x is the observed data.
- $T(x)$ is a statistic, which is a function of the observed data.
- $g(T(x); \theta)$ is a function that depends only on parameter of interest θ & the statistic $T(x)$
- $h(x)$ is a function that depends only on the observed data x but does not depend on the parameter of interest θ.

The factorization of the likelihood function allows for the separation of the parameter of interest from the nuisance parameters, which are the parameters that are not of direct interest but still need to be accounted for in the analysis. This separation simplifies the inference process and facilitates the estimation of the parameter of interest.
By utilizing the Neyman-Fisher Factorization Theorem, various statistical procedures can be developed, such as maximum likelihood estimation, likelihood ratio tests, and the construction of confidence intervals.
It is important to note that the Neyman-Fisher Factorization Theorem holds under certain regularity conditions, including the existence of a unique maximum likelihood estimator and the continuity of the likelihood function.

Neyman-Pearson Lemma is a fundamental result in statistical hypothesis testing that provides an optimal criterion for constructing tests of hypotheses. It states that, under certain conditions, the likelihood ratio test is the most powerful test with a given level of significance. The Neyman-Pearson Lemma can be stated as follows:
Consider a hypothesis testing problem with two hypotheses:-

- Null hypothesis (H0): $\theta \in \Theta0$
- Alternative hypothesis (H1): $\theta \in \Theta1$

where θ is the unknown parameter, and Θ0 and Θ1 are two disjoint parameter spaces corresponding to the null and alternative hypotheses, respectively.
Let $L(x; \theta)$ denote the likelihood function, which is the probability density function (pdf) or probability mass function (pmf) of the observed data x given the parameter θ.

The Neyman-Pearson Lemma Proposition:

Suppose we want to test H0 against H1 at a significance level α. The likelihood ratio test is the most powerful test at level α if the critical region is of the form:

$C = \{x: L(x; \theta) \leq k\}$

where k is chosen such that the probability of rejecting H0 at level α is exactly α.
In equation form, the likelihood ratio is given by:

$\lambda(x) = L(x; \theta1) / L(x; \theta0)$

The likelihood ratio test compares the likelihood of the data under the alternative hypothesis (θ1) to the likelihood under the null hypothesis (θ0). If the likelihood ratio is greater than a certain threshold, the null hypothesis is rejected in favor of the alternative hypothesis.
The Neyman-Pearson Lemma provides a way to construct tests that maximize the power of the test for a given level of significance. It is particularly useful in situations where the alternative hypothesis is specified, and one wants to detect a specific effect or difference. However, it should be noted that the Neyman-Pearson Lemma assumes specific conditions and may not be applicable in all testing scenarios.

Likelihood ratio test is a statistical test used to compare the fit of two competing statistical models, often referred to as the null model and the alternative model. It is based on the likelihood function, which measures the likelihood of observing the given data under each model. The likelihood ratio test compares the likelihood of the data under the alternative model to the likelihood under the null model. The test statistic is the ratio of these likelihoods, known as the likelihood ratio: LR = L(θ1) / L(θ0)
where L(θ1) is the likelihood under the alternative model (with parameter θ1) and L(θ0) is the likelihood under the null model (with parameter θ0).
To perform the test, we calculate the likelihood ratio and compare it to a critical value determined based on the desired level of significance (α). If the likelihood ratio exceeds the critical value, we reject the null hypothesis in favor of the alternative hypothesis. Otherwise, we fail to reject the null hypothesis.

The interpretation of the likelihood ratio test is as follows:

- If the likelihood ratio is significantly greater than 1, it suggests that the alternative model provides a significantly better fit to the data compared to the null model.
- If the likelihood ratio is close to 1, it indicates that both models provide similar fits to the data, and there is no strong evidence to favor one over the other.
- If the likelihood ratio is significantly less than 1, it implies that the null model provides a significantly better fit to the data compared to the alternative model.

The critical value for the likelihood ratio test depends on the desired level of significance and the degrees of freedom. It can be obtained from statistical tables or calculated using asymptotic distributions such as the chi-square distribution.

Population	**Sample**
Refers to the entire group or set of individuals, objects, or events of interest.	Refers to a subset of the population selected for study.
Represents the entire target population, including all its characteristics.	Represents a smaller portion of the population, which may or may not accurately reflect all the characteristics of the population.
Parameters, such as mean and variance, are used to describe the population.	Statistics, like sample mean and sample variance, are used to estimate population parameters.
Results obtained from studying the population are typically more accurate and precise.	Results obtained from studying a sample are subject to sampling error and may not perfectly represent the population.
Used for making inferences and generalizations about the entire population.	Used to draw conclusions about the population based on the characteristics observed in the sample.
Examples: All students in a school, all customers of a company.	Examples: Randomly selected students from a school, a survey of customers.

Standard Deviation	**Variance**
Measures the average deviation of data points from the mean.	Measures the average squared deviation of data points from the mean.
Provides a measure of the spread or dispersion of the data.	Provides a measure of the variability or dispersion of the data.
Calculated by taking the square root of the variance.	Calculated by averaging the squared deviations from the mean.
Expressed in the same unit as the data.	Expressed in squared units, which may not be as easily interpretable.
More commonly used in practice as it is in the same unit as the data.	Often used in theoretical calculations and statistical formulas.
Sensitive to outliers or extreme values in the data.	Also sensitive to outliers, but to a greater extent than standard deviation.

The Student's t-test is a statistical hypothesis test used to determine if there is a significant difference between the means of two groups. Specifically, the t-test is appropriate when the sample sizes are small, and the population standard deviations are unknown. The t-test calculates a test statistic called the t-value, which is then compared to the critical values from the t-distribution to make a decision about the null hypothesis. Here's an explanation of the Student's t-test for (n-2) degrees of freedom (df):

1. Null Hypothesis: The null hypothesis assumes that there is no significant difference between the means of the two groups under comparison. It is typically denoted as H0.

2. Alternative Hypothesis: The alternative hypothesis represents the possibility of a significant difference between the means of the two groups. It is denoted as Ha.

3. Test Statistic: The t-value is the test statistic used in the t-test. It is calculated using the following formula: t = (x1 - x2) / sqrt((s1^2 / n1) + (s2^2 / n2))

where x1 and x2 are the sample means of the two groups, s1 and s2 are the sample standard deviations, and n1 and n2 are the sample sizes of the two groups.

4. Degrees of Freedom: The degrees of freedom for the t-test is calculated as (n1 + n2 - 2). It represents the number of independent pieces of information available to estimate the population parameters.

5. Critical Region: The critical region is the range of t-values that would lead to the rejection of the null hypothesis. The critical values depend on the significance level (α) chosen for the test and the degrees of freedom. They can be obtained from the t-distribution table or using statistical software.

6. Decision: The decision about the null hypothesis is based on comparing the calculated t-value to the critical values. If the calculated t-value falls within the critical region, the null hypothesis is rejected, suggesting a significant difference between the means of the two groups. If the calculated t-value falls outside the critical region, the null hypothesis is not rejected, indicating no significant difference between the means.

The F-distribution, also known as Fisher's F-distribution, is a probability distribution that arises in statistical inference. It has several features and applications, including:

Features: 1. Skewed distribution: The F-distribution is a positively skewed distribution, meaning that it has a long tail on the right side. **2. Shape parameter:** The F-distribution has two shape parameters, denoted as df1 and df2. These parameters determine the shape of the distribution and influence the degrees of freedom for the numerator and denominator. **3. Non-negative values: The** values of the F-distribution are always non-negative, ranging from 0 to positive infinity. **4. Variability:** The F-distribution is sensitive to variability. As the variability in the numerator and denominator increases, the F-distribution becomes more spread out.

Applications of the F-distribution: 1. Analysis of variance (ANOVA): The F-distribution is extensively used in ANOVA, which is a statistical technique used to compare means across multiple groups. ANOVA tests the null hypothesis that the means of the groups are equal using the F-test. **2. Regression analysis:** In regression analysis, the F-distribution is employed to assess the overall significance of a regression model. The F-test is used to determine if the regression model as a whole explains a significant amount of variability in the dependent variable. **3. Quality control:** The F-distribution is utilized in quality control processes to compare the variances of two or more populations. It helps determine if there are significant differences in variability between the groups.

4. Experimental design: The F-distribution plays a crucial role in experimental design to evaluate the effect of different factors on a response variable. It helps determine the statistical significance of the factors and their interactions. **5. Hypothesis testing:** The F-distribution is employed for hypothesis testing when comparing variances or testing the equality of regression slopes in multiple regression models.

MP test/ Mann-Whitney U test or Wilcoxon rank-sum test, is a non-parametric statistical test used to compare two independent samples and determine if they come from the same population. **1. Non-parametric test:** The MP test is a non-parametric test, meaning that it does not require any assumptions about the distribution of the data. **2. Ordinal data:** The MP test is suitable for ordinal or continuous data. It compares the ranks or relative positions of the observations in the two samples. **3. Null hypothesis:** The null hypothesis for the MP test states that there is no difference between the distributions of the two samples. **4. Test statistic:** The test statistic for the MP test is U, also known as the Mann-Whitney U statistic. It represents the sum of ranks of one of the samples and can be calculated using the formula: U = R - (n1 * (n1 + 1)) / 2
where R is the sum of ranks of one sample and n1 is the sample size of that group. **5. Decision rule:** The decision to reject or fail to reject the null hypothesis is based on comparing the calculated U value to critical values from the Mann-Whitney U distribution table or by using a statistical software. **6. Interpretation:** If the calculated U value is significantly different from the critical values, it indicates that there is a statistically significant difference between the two samples. **7. Effect size:** The MP test does not provide an effect size estimate directly. However, researchers can calculate the probability of superiority (P) or the probability of a random observation from one sample being greater than a random observation from the other sample. **8. Sample size requirements:** The MP test is suitable for small to moderate sample sizes. It is more robust to outliers and violations of normality assumptions compared to parametric tests like the t-test.

UMP (Uniformly Most Powerful) test/ Neyman-Pearson test, is a hypothesis test that aims to maximize the power of the test for a given significance level. **1. Hypothesis testing:** The UMP test is used to test a specific hypothesis by comparing it to an alternative hypothesis. **2. Null and alternative hypothesis:** The null hypothesis (H0) represents a specific claim or assumption about the population, while the alternative hypothesis (H1) represents the opposite claim or assumption. **3. Power of a test:** The power of a test refers to the ability of the test to correctly reject the null hypothesis when the alternative hypothesis is true. The UMP test aims to maximize the power for a given significance level. **4. Test statistic:** The test statistic used in the UMP test depends on the specific hypothesis being tested. Different test statistics are used for different types of data and hypotheses. **5. Significance level:** The significance level, denoted by α, represents the maximum probability of rejecting the null hypothesis when it is true. The UMP test allows researchers to control the significance level and choose the most powerful test for a given level. **6. Likelihood ratio:** The UMP test often relies on the likelihood ratio, which is the ratio of the likelihood of the data under the alternative hypothesis to the likelihood under the null hypothesis. The likelihood ratio is used to calculate the test statistic. **7. Decision rule:** The decision to reject or fail to reject the null hypothesis is based on comparing the test statistic to critical values obtained from the distribution of the test statistic. The critical values are chosen to control the significance level and maximize the power of the test. **8. Optimal test:** The UMP test is considered optimal because it maximizes the power for a given significance level. It is designed to be the most sensitive to detecting deviations from the null hypothesis among all possible tests with the same significance level.

Selection of a Simple Random Sample: 1. Definition: A simple random sample is a sampling technique where each individual or element in the population has an equal chance of being selected. **2. Randomization:** The selection process involves using a random mechanism to ensure that every member of the population has an equal probability of being chosen. This can be achieved through methods like random number generators or lottery systems. **3. Sampling Frame:** A sampling frame is a list or representation of the population from which the sample will be drawn. It should include all members of the population and be well-defined to ensure the random selection process is implemented correctly. **Method of Drawing a Simple Random Sample: 1. Assigning Numbers:** Each member of the population is assigned a unique number or identifier. **2. Random Selection:** Randomly select the desired number of individuals or elements from the sampling frame. This can be done using various randomization techniques, such as random number tables, computer-generated random numbers, or random sampling software. **Merits of a Simple Random Sample: 1. Unbiased Representation:** A simple random sample provides an unbiased representation of the population as every individual has an equal chance of being selected. This helps to reduce sampling bias and ensures that the sample is representative of the entire population. **2. Simplicity:** The process of selecting a simple random sample is straightforward and easy to understand. It does not require complex sampling techniques or specialized knowledge. **3. Generalizability:** The results obtained from a simple random sample can be generalized to the entire population, making it useful for making inferences about the larger population. **Demerits of a Simple Random Sample: 1. Time and Cost:** Selecting a simple random sample from a large population can be time-consuming and expensive, especially if the population size is significant. **2. Infeasibility**: In some cases, it may be challenging or impractical to create a sampling frame that includes all members of the population. This can make the implementation of a simple random sample difficult. **3. Lack of Precision:** While a simple random sample provides unbiased estimates, it may not always yield the most precise results compared to more sophisticated sampling techniques. Other sampling methods, like stratified or cluster sampling, can provide more efficient estimates with smaller sample sizes.

Multistage Sampling: It is a sampling technique that involves selecting samples in multiple stages, where each stage involves a different sampling method. It is commonly used when it is not feasible or practical to directly sample from the entire population. **1. Stage 1:** In the first stage, clusters or primary sampling units (PSUs) are selected. Clusters are groups of individuals or elements that share common characteristics or are geographically grouped. The selection of clusters can be done using methods such as random sampling or stratified sampling.**2. Stage 2:** In the second stage, a subset of elements is selected from within each selected cluster. The selection within clusters can be done using simple random sampling, systematic sampling, or other sampling techniques. **3. Further Stages:** Depending on the complexity of the sampling design, additional stages can be included. Each stage involves selecting smaller subsets of elements until the final sample is obtained. **Advantages:1)** Cost-effective: Multistage sampling can be more cost-effective compared to other sampling methods, as it allows for sampling from larger populations without the need to access or enumerate every individual. **2) Practicality:** Multistage sampling is often practical in situations where the population is large, geographically dispersed, or lacks a comprehensive sampling frame. **3) Efficiency:** By selecting clusters in the initial stages, the sampling process becomes more efficient, especially when the clusters are internally homogeneous. **Disadvantages:1) Complex Design:** Multistage sampling requires careful planning and coordination, as each stage introduces additional complexities in terms of sample selection, estimation, and analysis.**2) Increased Variability:** The use of clusters in multistage sampling can lead to increased variability within clusters, which may affect the precision of estimates**. 3) Potential Bias:** If the clusters or stages are not appropriately selected, or if there is variability between clusters, multistage sampling can introduce bias in the estimates.

Stratified Random Sampling: 1. Definition: Stratified random sampling is a sampling technique where the population is divided into homogeneous subgroups called strata, and a random sample is then selected from each stratum. **2. Stratum Formation:** Strata are formed based on specific characteristics or attributes of the population, such as age, gender, location, or any other relevant variable. Each stratum should be internally homogeneous and externally heterogeneous. **3.Sample Selection:** Within each stratum, a random sample is selected using simple random sampling or other sampling methods. The sample size from each stratum can be proportional to the size of the stratum or can be fixed. **Allocation of Sample Size: 1. Proportional Allocation:** In proportional allocation, the sample size for each stratum is determined in proportion to the size of the stratum relative to the total population. This ensures that each stratum is adequately represented in the sample. **2. Optimum Allocation:** Optimum allocation involves allocating the sample size to strata in a way that minimizes the total sampling error. This allocation takes into account the variability within each stratum and aims to achieve more precise estimates. **Merits of Stratified Random Sampling: 1. Increased Precision:** Stratified random sampling can improve the precision of estimates compared to simple random sampling. By sampling from each stratum, the variability within each stratum is accounted for, leading to more accurate results. **2. Representation of Subgroups:** Stratified sampling ensures that important subgroups or strata within the population are adequately represented in the sample. This allows for more meaningful and detailed analysis of the data. **3. Efficiency:** Stratified sampling can be more efficient in terms of cost and time compared to other sampling methods like simple random sampling. By focusing resources on targeted strata, researchers can achieve desired precision with a smaller sample size. **Demerits of Stratified Random Sampling: 1. Complex Design:** The design and implementation of stratified sampling can be more complex compared to simple random sampling. It requires prior knowledge about the population characteristics and careful selection of relevant stratification variables. **2. Increased Administrative Effort:** Stratified sampling may require additional administrative effort in terms of data collection, stratification, and sample selection. Proper documentation and coordination are crucial to ensure the integrity of the sampling process. **3. Potential Bias:** If the stratification variables are not well-chosen or if there is incomplete information about the population, stratified sampling can introduce bias. It is important to carefully select stratification variables that are strongly correlated with the study variables of interest.

Type I error is a false positive conclusion, while a Type II error is a false negative conclusion. Making a statistical decision always involves uncertainties, so the risks of making these errors are unavoidable in hypothesis testing. The probability of making a Type I error is the significance level, or alpha (α), while the probability of making a Type II error is beta (β). These risks can be minimized through careful planning in your study design. **Example: Type I vs Type II error:** You decide to get tested for COVID-19 based on mild symptoms. There are two errors that could potentially occur: **Type I error (false positive):** the test result says you have coronavirus, but you actually don't. **Type II error (false negative):** the test result says you don't have coronavirus, but you actually do. **Type I error:** means rejecting the null hypothesis when it's actually true. It means concluding that results are statistically significant when, in reality, they came about purely by chance or because of unrelated factors. The risk of committing this error is the significance level (alpha or α) you choose. That's a value that you set at the beginning of your study to assess the statistical probability of obtaining your results (p value). The significance level is usually set at 0.05 or 5%. This means that your results only have a 5% chance of occurring, or less, if the null hypothesis is actually true. If the p value of your test is lower than the significance level, it means your results are statistically significant and consistent with the alternative hypothesis. If your p value is higher than the significance level, then your results are considered statistically non-significant. **Type II error** means not rejecting the null hypothesis when it's actually false. This is not quite the same as "accepting" the null hypothesis, because hypothesis testing can only tell you whether to reject the null hypothesis. Instead, a Type II error means failing to conclude there was an effect when there actually was. In reality, your study may not have had enough statistical power to detect an effect of a certain size. Power is the extent to which a test can correctly detect a real effect when there is one. A power level of 80% or higher is usually considered acceptable. The risk of a Type II error is inversely related to the statistical power of a study. The higher the statistical power, the lower the probability of making a Type II error.

Null hypothesis is the claim that there’s no effect in the population. If the sample provides enough evidence against the claim that there’s no effect in the population ($p \leq \alpha$), then we can reject the null hypothesis. Otherwise, we fail to reject the null hypothesis. Although “fail to reject” may sound awkward, it’s the only wording that statisticians accept. Be careful not to say you “prove” or “accept” the null hypothesis. Null hypotheses often include phrases such as “no effect,” “no difference,” or “no relationship.” When written in mathematical terms, they always include an equality (usually =, but sometimes ≥ or ≤).

Alternative hypothesis (Ha) is the other answer to your research question. It claims that there’s an effect in the population. Often, your alternative hypothesis is the same as your research hypothesis. In other words, it’s the claim that you expect or hope will be true. The alternative hypothesis is the complement to the null hypothesis. Null and alternative hypotheses are exhaustive, meaning that together they cover every possible outcome. They are also mutually exclusive, meaning that only one can be true at a time. Alternative hypotheses often include phrases such as “an effect,” “a difference,” or “a relationship.” When alternative hypotheses are written in mathematical terms, they always include an inequality (usually ≠, but sometimes < or >). As with null hypotheses, there are many acceptable ways to phrase an alternative hypothesis.

	Null hypotheses (H_0)	**Alternative hypotheses (H_a)**
Definition	A claim that there is **no effect** in the population.	A claim that there is **an effect** in the population.
Also known as	H_0	H_a H_1
Typical phrases used	• No effect • No difference • No relationship • No change • Does not increase Does not decrease	• An effect • A difference • A relationship • A change • Increases Decreases
Symbols used	Equality symbol (=, ≥, or ≤)	Inequality symbol (≠, <, or >)
$p \leq \alpha$	Rejected	Supported
$p > \alpha$	Failed to reject	Not supported

Degree of freedom (df) is a concept in statistics that represents the number of values in a calculation that are free to vary. It is often denoted by "df" and is an important parameter in various statistical tests and distributions. In general, the degree of freedom is the number of independent pieces of information available in a sample or population that are relevant to the statistical analysis. It determines the number of values that are allowed to vary in order to estimate or test certain parameters.