DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
Second Year B.Tech. (SEM - III)
COMPUTATIONAL MATHEMATICS (UCSE0301)

Unit No. 4: Statistical Techniques

Correlation: Correlation studies the relationship between two variables in which change in the value of one variable causes change in the other variable.

Types of correlation:

1) Positive and Negative correlation.

Positive correlation: When both variables move in the same direction. If one variable increases other also increases and vice-versa.

Negative correlation: When two variables move in the opposite direction, they are negatively correlated.

2) Linear and non – linear correlation.

Linear Correlation: When two variables change in a constant proportion.

Non- linear correlation: When two variables do not change in the same proportion.

3) Simple and multiple correlations.

Simple correlation: Relationship between two variables is studied.

Multiple Corrections: Relationship between three or more than three variables is studied.

Karl Pearson's Correlation coefficients:

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables (X and Y) for n number of observations. There are several types of correlation coefficient, but the most popular is **Karl Pearson's correlation coefficient** commonly used in **linear regression.**

It is denoted by letter 'r' and defined as,

$$r = \frac{Cov(X, Y)}{S.D(X)\ S.D(Y)}$$

Where,

$$Cov(X,Y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n} \qquad S.D(X) = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} \qquad S.D(Y) = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}$$

Hence,
$$r = \frac{\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}}$$

$$r = \frac{\frac{\Sigma xy}{n} - \left(\frac{\Sigma x}{n}\right)\left(\frac{\Sigma y}{n}\right)}{\sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}\sqrt{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2}}$$

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1.

Degrees of Correlation:

1. Perfect Correlation: When values of both variables changes at a constant rate.

Perfect positive correlation: when values of both variables changes at a constant ratio in the same direction correlation coefficient value (r) are + 1

Perfect negative correlation: When values of both the variables change at a constant ratio in opposite direction. Value of coefficient of correlation is -1

2. Absence of correlation: When there is no relation between the variables r = 0

3. Limited degree correlation: The value of r varies between more than 0 and less than 1

a) High: r his between ± 0.7 & 0.999

b) Moderate: r lies between ± 0.5 and + 0.699

c) Low: r < ± 0.5

Examples

Example 1: Calculate the correlation coefficient for the following series relating price and supply of commodity.

Price (Rs)	22	24	26	28	30	32	34	36	38	40
Supply (Tons)	60	58	58	50	48	48	48	42	36	30

Solution: Let X = Price (Rs) and Y = Supply (Tons)

Here, $n=10 \quad \Sigma x=310 \quad \Sigma y=478 \quad \Sigma x^2=9940 \quad \Sigma y^2=23700 \quad \Sigma xy=14308$

The Karl Pearson's correlation coefficient is,

$$r=\frac{\frac{\Sigma xy}{n}-\left(\frac{\Sigma x}{n}\right)\left(\frac{\Sigma y}{n}\right)}{\sqrt{\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2}\sqrt{\frac{\Sigma y^2}{n}-\left(\frac{\Sigma y}{n}\right)^2}}$$

$$r=\frac{\frac{14308}{10}-\left(\frac{310}{10}\right)\left(\frac{478}{10}\right)}{\sqrt{\frac{9940}{10}-\left(\frac{310}{10}\right)^2}\sqrt{\frac{23700}{10}-\left(\frac{478}{10}\right)^2}}$$

$$r=\frac{-51}{(5.7446)(9.2282)} \qquad r=-0.9620$$

Example 2: Calculate the correlation coefficient between *x* and *y* for the following data:

x	1	3	4	5	7	8	10
y	2	6	8	10	14	16	20

Solution: Here, $n=7 \quad \Sigma x=38 \quad \Sigma y=76 \quad \Sigma x^2=264 \quad \Sigma y^2=1056 \quad \Sigma xy=528$

The Karl Pearson's correlation coefficient is,

$$r=\frac{\frac{\Sigma xy}{n}-\left(\frac{\Sigma x}{n}\right)\left(\frac{\Sigma y}{n}\right)}{\sqrt{\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2}\sqrt{\frac{\Sigma y^2}{n}-\left(\frac{\Sigma y}{n}\right)^2}} \qquad r=\frac{\frac{528}{7}-\left(\frac{38}{7}\right)\left(\frac{76}{7}\right)}{\sqrt{\frac{264}{7}-\left(\frac{38}{7}\right)^2}\sqrt{\frac{1056}{7}-\left(\frac{76}{7}\right)^2}}$$

$$r=\frac{16.4897}{(2.8714)(5.7428)} \qquad r=1.00$$

Examples for Practice

Example 1: Define correlation coefficient and state its properties.

Example 2: From the following data obtain the coefficient of correlation:

X	25	27	30	35	33	28	36
Y	19	22	27	28	30	33	28

Example 3: Following are the result of the percentage growth of plants, which were fully or partially exposed to certain gas.

Fully exposed growth: 12.7 12.6 13.1 13.0 12.5 13.0 13.0 12.8

Partially exposed growth: 10.3 9.3 10.5 10.4 10.0 10.2 10.2 10.0

Find coefficient of correlation for the above data and comment on it.

Example 4: Find the correlation coefficient and the equations of regression lines for the following values of x and y.

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

Example 5: A simply supported beam carries a concentrated load X (lb) at its mid-point. Corresponding to various values of X, the maximum deflection Y (in) is measured.

The data is given below:

X	100	120	140	160	180	200
Y	0.45	0.55	0.60	0.70	0.80	0.85

Find the correlation coefficient between X & Y.

Example 6: Record of test of intelligence ratio (I. R.) and engineering skills (E. S.) of 10 students are given in the following table. Calculate coefficient of correlation.

Student	A	B	C	D	E	F	G	H	I	J
I. R. (x)	105	104	102	101	100	99	98	96	93	92
E. S. (y)	101	103	100	98	95	96	104	92	97	94

Example 7: Calculate the coefficient of correlation between the ages of husband and wife given by the table.

Age of husband (x)	23	27	28	29	30	31	33	35	36	39
Age of wife (y)	18	22	23	24	25	26	28	29	30	32

Regression: A statistical techniques uses a single independent variable to estimate single dependent variable when they are correlated.

Lines of Regression:

It frequently happens that the dots of the scatter diagram generally, tend to cluster along a well defined direction which suggests a linear relationship between the variables *x* and *y*. Such a line of best-fit for the given distribution of dots is called the line of regression.

The Line of Regression of *y* on *x*:

The line giving the best possible mean values of *y* for each specified value of *x* is known as the line of regression of *y* on *x*.

Equation of the line of regression of *y* on *x* is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Where, b_{yx} = regression coefficient of y on x.

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \qquad b_{yx} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}\frac{\sigma_y}{\sigma_x} \qquad b_{yx} = \frac{Cov(X,Y)}{\sigma^2{}_x} \qquad b_{yx} = \frac{\frac{\Sigma xy}{n} - \frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

The Line of Regression of *x* on *y*:

Similarly, the line giving the best possible mean values of *x* for given values of *y* is known as the line of regression of *x* on *y*.

Equation of the line of regression of *x* on *y* is given by

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

Where, b_{xy}= regression coefficient of x on y.

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \qquad b_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}\frac{\sigma_x}{\sigma_y} \qquad b_{xy} = \frac{Cov(X,Y)}{\sigma^2{}_y} \qquad b_{xy} = \frac{\frac{\Sigma xy}{n} - \frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2}$$

Note:

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

The correlation coefficient *r* is positive if both b_{yx} & b_{xy} are positive and *r* is negative if both b_{yx} & b_{xy} are negative.

2) If one of the regression coefficients is greater than unity, the other is less than unity.

3) Two lines of regressions are passing through the point $(\bar{x}, \bar{y})$

Examples

Example 1: Following table gives the data on rainfall and discharge in a certain river. Obtain the lines of regression and coefficient of correlation.

Rainfall x (inches) :	1.53	1.78	2.60	2.95	3.42
Discharge y (1000 cc) :	33.5	36.3	40.0	45.8	53.5

Solution:

Here, $n=5,\ \Sigma x=12.28,\ \Sigma y=209.1,\ \Sigma x^2=32.6682,\ \Sigma y^2=8999.83,\ \Sigma xy=537.949$

$$\therefore \bar{x}=\frac{\Sigma x}{n}=\frac{12.28}{5}=2.456 \qquad \therefore \bar{y}=\frac{\Sigma y}{n}=\frac{209.1}{5}=41.82$$

The regression coefficient of y on x is

$$b_{yx}=\frac{\frac{\Sigma xy}{n}-\frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2}=\frac{\frac{537.949}{5}-(2.456)(41.82)}{\frac{32.6682}{5}-(2.456)^2}=9.7266$$

Equation of the line of regression of y on x is given by

$$y-\bar{y}=b_{yx}(x-\bar{x})$$

$$y-41.82=9.7266(x-2.456)$$

i.e. $y=17.9315+9.7266x$

The regression coefficient of x on y is

$$b_{xy}=\frac{\frac{\Sigma xy}{n}-\frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma y^2}{n}-\left(\frac{\Sigma y}{n}\right)^2}=\frac{\frac{537.949}{5}-(2.456)(41.82)}{\frac{8999.83}{5}-(41.82)^2}=0.0956$$

Equation of the line of regression of x on y is given by

$$x-\bar{x}=b_{xy}(y-\bar{y})$$

$$x-2.456=0.0956(y-41.82)$$

i.e. $x=-1.542+0.0956y$

The correlation coefficient is

$$r=\pm\sqrt{b_{yx}b_{xy}}=r=\sqrt{(9.7266)(0.0956)}=0.9643$$

Example 2: Use the following data, to obtain the regression equations and correlation coefficient. Also find Y for X = 25 and find X for Y = 40

x	14	19	24	21	26	22	15	20	19
y	31	36	48	37	50	45	33	41	39

Solution:

Here, $n=9,\ \Sigma x=180,\ \Sigma y=360,\ \Sigma x^2=3720,\ \Sigma y^2=14746,\ \Sigma xy=7393$

$$\therefore \bar{x}=\frac{\Sigma x}{n}=\frac{180}{9}=20 \qquad \therefore \bar{y}=\frac{\Sigma y}{n}=\frac{360}{9}=40$$

The regression coefficient of y on x is,

$$b_{yx}=\frac{\frac{\Sigma xy}{n}-\frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2}=\frac{\frac{7393}{9}-(20)(40)}{\frac{3720}{9}-(20)^2}=1.6083$$

Equation of the line of regression of y on x is given by

$$y-\bar{y}=b_{yx}(x-\bar{x})$$

$$y-40=1.6083(x-20)$$

i.e. $y=7.834+1.6083x$

when $x=25$

i.e. $y=7.834+1.6083\,(25)=48.0415$

The regression coefficient of x on y is

$$b_{xy}=\frac{\frac{\Sigma xy}{n}-\frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma y^2}{n}-\left(\frac{\Sigma y}{n}\right)^2}=\frac{\frac{7393}{9}-(20)(40)}{\frac{14746}{9}-(40)^2}=0.5578$$

Equation of the line of regression of x on y is given by,

$$x-\bar{x}=b_{xy}(y-\bar{y})$$

$$x-20=0.5578(y-40)$$

i.e. $x=-2.312+0.5578y$

when $y=40$

i.e. $x=-2.312+0.5578(40)=20$

The correlation coefficient is $r=\pm\sqrt{b_{yx}b_{xy}}=\sqrt{(1.6083)(0.5578)}=0.9471$

Example 3: Use the following data, to obtain the regression equations and correlation coefficient.

x	72	98	76	81	56	76	92	88	49
y	124	131	117	132	96	120	136	97	85

Solution:

Here, $n=9, \quad \sum x = 688 \quad \sum x^2 = 54656, \quad \sum y = 1038, \quad \sum y^2 = 122396, \quad \sum xy = 81059$

$$\therefore \bar{x} = \frac{\sum x}{n} = \frac{688}{9} = 76.4444 \qquad \therefore \bar{y} = \frac{\sum y}{n} = \frac{1038}{9} = 115.3333$$

The regression coefficient of y on x is,

$$b_{yx} = \frac{\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \frac{\frac{81059}{9} - (76.4444)(115.3333)}{\frac{54646}{9} - (76.4444)^2} = 0.8331$$

Equation of the line of regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 115.3333 = 0.8331(x - 76.4444)$$

i.e. $\quad y = 51.6474 + 0.8331x$

The regression coefficient of x on y is

$$b_{xy} = \frac{\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2} = \frac{\frac{81059}{9} - (76.4444)(115.3333)}{\frac{122396}{9} - (115.3333)^2} = 0.6379$$

Equation of the line of regression of x on y is given by

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 76.4444 = 0.6379(y - 115.3333)$$

i.e. $\quad x = 2.8732 + 0.6379y$

The correlation coefficient is

$$r = \pm\sqrt{b_{yx}b_{xy}} = \sqrt{(0.8331)(0.6379)} = 0.7289$$

Example 4: Two regression equations of the variables x and y are

$x = 19.13 - 0.87y$ & $y = 11.64 - 0.50x$.

Find (i) mean of x's, (ii) mean of y's & (iii) the correlation coefficient between x & y.

Solution: As the mean of x's & the mean of y's lie on regression lines, we have

$$\bar{x} = 19.13 - 0.87\bar{y} \qquad (1)$$
$$\bar{y} = 11.64 - 0.50\bar{x} \qquad (2)$$

Solving, equations (1) & (2) simultaneously, we get

$$\bar{x} = 15.79$$
$$\bar{y} = 3.74$$

Now, here regression coefficient of y on x is $b_{yx} = -0.5$ &

regression coefficient of x on y is $b_{xy} = -0.87$

So, the correlation coefficient between x & y is

$$r = \sqrt{b_{yx} b_{xy}} = \sqrt{(-0.5)(-0.87)} = -0.66$$

Example 5: The two lines of regression given by $5x - 6y + 90 = 0$ & $15x - 8y - 130$. Use the equation to find the mean of x & the mean of y and the correlation coefficient between x & y. If the variance of x is 16 calculate the variance of y.

Solution: As the mean of x's & the mean of y's lie on regression lines, we have

$$5\bar{x} - 6\bar{y} = -90 \qquad (1)$$
$$15\bar{x} - 8\bar{y} = 130 \qquad (2)$$

Solving, equations (1) & (2) simultaneously, we get

$$\bar{x} = 30$$
$$\bar{y} = 40$$

Now, treating the first equation as the line of regression of y on x, we write it as

$6y = 5x + 90$ i.e. $y = (5/6)\,x + 15$

Here regression coefficient of y on x is $b_{yx} = 5/6$ &

Treating the second equation as the line of regression of x on y, we write it as

$15x = 8y + 130$ i.e. $x = (8/15)\,y + (130/15)$

Here regression coefficient of x on y is $b_{xy} = 8/15$

So, the correlation coefficient between x & y is

$$r = \sqrt{b_{yx} b_{xy}} = \sqrt{\frac{5}{6} \times \frac{8}{15}} = \frac{2}{3} = 0.6666$$

If the variance of x is 16, $\sigma^2{}_x = 16 \;\; \sigma_x = 4$

We know that, $b_{yx} = r\frac{\sigma_y}{\sigma_x} \;\; \therefore \frac{5}{6} = \frac{2}{3}\frac{\sigma_y}{4} \;\;\; \therefore \sigma_y = 5 \;\;\;\; \therefore \sigma^2{}_y = 25$

Example 6: Given the following information about marks of 60 students,

	Mathematics	English
Mean	80	50
S.D	15	10

Coefficient of correlation between marks of Mathematics and English is 0.4. Estimate the marks of the students in Mathematics who scored 60 marks in English.

Solution: Let X= Mathematics and Y = English

Then given that, $\bar{x} = 80 \;\; \bar{y} = 50 \;\; \sigma_x = 15 \;\; \sigma_y = 10 \;\; r = 0.4$

Hence, regression coefficients y on x is, $b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.4\frac{15}{10} = 0.6$

Equation of the line of regression of x on y is given by,

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 80 = 0.6(y - 50)$$

i.e. $x = 50 + 0.6y$

$when \; y = 60$

i.e. $x = 50 + 0.6(60) = 86$

Examples for Practice

Example 1: Following table gives the data on rainfall and discharge in a certain river. Obtain the line of regression of y on x.

Rainfall x (inches) :	1.53	1.78	2.60	2.95	3.42
Discharge y (1000 cc) :	33.5	36.3	40.0	45.8	53.5

Example 2: Associated with a job are two random variables, CPU time required (y) and the number of disk I/O operations (x). Given the following data:

Time(Sec.)y:	40	38	42	50	60	30	20	25	40	39
Number x:	398	390	410	502	590	309	210	252	398	392

Compute the line of regression y on x .

Example 3: Obtain least square regression line of y on x for the following data.

X :	1	2	3	4	5	6	7	8
Y :	9	8	10	12	11	13	14	16

Example 4: Given the following data compute the line of regression y on x.

X :	2	4	6	8	10	12	14	16
Y :	1	2	4	7	8	10	6	8

Example 5: Following data is related to height and weight of sugarcane

Height (in cms) :	105	118	134	110	150	112	140
Weight (in kgs) :	3.2	4.0	5.0	3.8	4.8	4.2	5.2

Find two equations of regression using above data. Also find height of sugarcane having weight 4.5 kg and weight of the sugarcane having height 145 cms.

Example 6: Following are the results of daily irrigation time (in min.) and production of flowers (in kg) for seven different farms of size.

Irrigation time	120	175	155	140	160	140	165
Production	112	136	125	122	128	120	132

Find two equations of regression using above data. Also find production of flowers if irrigation time is 150 min.

Example 7: Out of the two lines of regression given by x+2y-5=0 & 2x+3y-8=0, which one is the regression line of x on y? Use the equation to find the mean of x & the mean of y. If the variance of x is 12 calculate the variance of y.

Example 8: In a partially destroyed laboratory record, only the lines of regression of y on x and x on y are available as $4x-5y+33=0$ an $20x-9y=107$ respectively. Calculate $\bar{x}, \bar{y}$ and the coefficient of correlation between x and y.

Example 9: Two random variables have the regression lines with equations $3x+2y=26$ and $6x+y=31$. Find the mean values and the correlation coefficient between x and y.

Example 10: Given,

	X	Y
Mean	18	100
S.D	14	20

Coefficient of correlation between X and Y is 0.8. Find the most probable value of Y when x = 17 and most probable value of X when Y= 90

Example 11: From the following data, find the two regression equations and estimate the likely value of y when x = 100.

x	72	98	76	81	56	76	92	88	49
y	124	131	117	132	96	120	136	97	85

Curve Fitting:

We may encounter random experiments in which we observe or measure two quantities (random variables) simultaneously, so that we get samples of *pairs* of values (x_1, y_1), (x_2, y_2),…… (x_n, y_n). Curve fitting is the technique used to find the "best fit" line or curve for the given (observed or experimental) set of data points. Most of the time, the curve fit will produce an equation that can be used to find points anywhere along the curve.

Least Squares Principle:

The straight line (or curve) should be fitted through the given points so that the sum of the squares of the distances of those points from the straight line (or curve) is ***minimum***, where the distance is measured in the vertical direction (the *y*-direction).

1.Fitting of Straight Line:

Let (x_1, y_1), (x_2, y_2),…… (x_n, y_n).be the given set of *n* data points. To fit the straight line

$$Y = a + b\,X \qquad (1)$$

to this data, we use normal equations given by,

$$\sum y = na + b\sum x \qquad (2)$$

$$\sum xy = a\sum x + b\sum x^2 \qquad (3)$$

Solving (2) & (3) simultaneously, we get the values of a and b.

Example 1: Use least squares method to fit a straight line to the following data:

X:	0	1	2	3	4
Y:	1	2.9	4.8	6.7	8.9

Solution: Let the straight line to fit the given data be

$$Y = a + b\,X \qquad (1)$$

The normal equations are,

$$\sum y = na + b\sum x \qquad (2)$$

$$\sum xy = a\sum x + b\sum x^2 \qquad (3)$$

Here, $n = 5$, $\sum x = 10$, $\sum x^2 = 30$, $\sum y = 24.3$, $\sum xy = 68.2$

On putting these values in equations (2) & (3), we get

$$24.3 = 5a+10b \qquad (4)$$

$$68.2 = 10a+30b \quad (5)$$

Solving (4) & (5) simultaneously, we get

$$a = 0.94 \text{ and } b = 1.96$$

So, the straight line to best fit is

$$y = 0.94+1.96\,x$$

Example 2: If P is the pull required to lift a load W by means of a pulley block, find a linear law of the form $P = a + bW$ connecting P and W, using the following data:

W:	50	70	100	120
P:	12	15	21	25

Where P and W are taken in kg-wt. Compute P when $W = 150$.

Solution: Let the straight line to fit the given data be

$$P = a + b\,W \quad (1)$$

The normal equations are

$$\sum P = na + b\sum W \quad (2)$$

$$\sum WP = a\sum W + b\sum W^2 \quad (3)$$

Here, $n = 4$, $\sum W = 340$, $\sum W^2 = 31800$, $\sum P = 73$, $\sum WP = 6750$

On putting these values in equations (2) & (3), we get

$$73 = 4a + 340b \quad (4)$$

$$6750 = 340a + 31800b \quad (5)$$

Solving (4) & (5) simultaneously,

$$a = 2.2785 \text{ and } b = 0.1879$$

So, required linear equation to best fit is

$$P = 2.2785 + 0.1879\,W$$

When $W = 150$ kg, $P = 2.2785 + 0.1879\,(150) = 30.4635$

Example 3: Fit a straight line to the following data,

X:	0	5	10	15	20	25
Y:	12	15	17	22	24	30

Solution: Let the straight line to fit the given data be

$$Y = a + b\,X \quad (1)$$

Consider, $u = \left(\frac{x-12.5}{5}\right)$ then u: $-2.5, -1.5, -0.5, 0.5, 1.5, 2.5$

Hence,

u:	-2.5	-1.5	-0.5	0.5	1.5	2.5
Y:	12	15	17	22	24	30

Here, n = 6, $\sum u = 0$, $\sum u^2 = 17.5$, $\sum y = 120$, $\sum uy = 61$

The transformed equation is

$$Y = a + b\,u \qquad (2)$$

The normal equations are,

$$\sum y = na + b\sum u \qquad (3)$$

$$\sum uy = a\sum u + b\sum u^2 \qquad (4)$$

On putting these values in equations (3) & (4), we get

$120 = 6a + 0 \quad \rightarrow a = 20$

$61 = 0 + 17.5\,b \quad \rightarrow b = 3.4857$

On putting these values in equation (2)

$$Y = 20 + 3.4857(u)$$

Now, putting u = (x - 12.5) / 5

$$y = 20 + 3.4857\left(\frac{x-12.5}{5}\right)$$

$$\Rightarrow 5y = 100 + 3.4857x - 43.5712$$

So, the straight line to the best fit is,

$$y = 11.2857 + 0.6971(x)$$

Example 4: Fit a straight line to the following data,

Year(X) :	1951	1961	1971	1981	1991
Production ('000 tons):	10	12	8	10	13

Also estimate the production in 1987.

Solution: Let the straight line to fit the given data be

$$Y = a + b\,X \qquad (1)$$

Consider, $u = \left(\frac{x-1971}{10}\right)$ then u: $-2, -1, 0, 1, 2$

Hence,

u: -2 -1 0 1 2

Y: 10 12 8 10 13

Here, $n = 5$, $\sum u = 0$, $\sum u^2 = 10$, $\sum y = 53$, $\sum uy = 4$

The transformed equation is

$$Y = a + b\,u \qquad (2)$$

The normal equations are,

$$\sum y = na + b\sum u \qquad (3)$$

$$\sum uy = a\sum u + b\sum u^2 \qquad (4)$$

On putting these values in equations (3) & (4), we get

$$53 = 5a + 0 \rightarrow a = 10.6$$

$$4 = 0 + 10\,b \rightarrow b = 0.4$$

On putting these values in equation (2)

$$Y = 10.6 + 0.4(u)$$

Now, putting $u = \left(\frac{x - 1971}{10}\right)$

$$y = 10.6 + 0.4\left(\frac{x - 1971}{10}\right)$$

$$\Rightarrow 10y = 106 + 0.4x - 788.4$$

So, the straight line to the best fit is,

$$y = -68.24 + 0.04(x)$$

When year x =1987 then $y = -68.24 + 0.04(1987) = 11.24$

Examples for Practice

Example 1: The kilometer per liter (km/l) figure for a new engine are recorded for fixed speeds between 56 and 112 km/hr.

Speed (km/hr)	56	104	64	88	112	96	84	68	80	100	60	72
Mileage (km/l)	14.7	13.2	14.5	13.2	12.8	13.4	13.3	14.5	13.8	13	14.6	14.3

Determine the best fit straight line for the mileage results and predict the mileage for a speed of 90 km/hr and 110 km/hr.

Example 2: The following data pertain to the number of jobs per day and the CPU time required.

No. of jobs :X	1	2	3	4	5
CPU Time: Y	2	5	4	9	10

Obtain a least square fit of a line to the observations on CPU time. Also estimate the mean CPU time when X = 3.5.

Example 3: For the following data fit a straight line.

x:	71	68	73	69	67	65	66	67
y:	69	72	70	70	68	67	68	64

$$y = 39.5455 + 0.4242x$$

Example 4: If y is the pull required to lift a load x by means of a pulley block, find a linear law of the form, $y = a + bx$ Connecting x and y, using the following data:

x	50	70	100	120
y	12	15	21	25

Where x and y are taken in kg-wt. Compute y when x = 150.

Example 5: Fit a straight line to the following data:

x	1	2	3	4	6	8
y:	2.4	3	3.4	4	5	6

2.Fitting of Exponential Curves:

a) Exponential curve $y = a\,b^x$

Let (x_1, y_1), (x_2, y_2),…… (x_n, y_n).be the given set of *n* data points. To fit the exponential

Curve, $y = a\,b^x$ (1)

By taking log of both sides, we get

$$\log y = \log a + x \log b$$

i.e. $\log y = A + BX$ (2)

where $A = \log a$ & $B = \log b$

To fit the straight line (2), the normal equations are

$$\sum \log y = nA + B\sum x \quad (3)$$

$$\sum x \log y = A\sum x + B\sum x^2 \quad (4)$$

Solving (3) & (4) simultaneously, we get the values of A and B.

here $a = e^A$ & $b = e^B$

Example 1: For the following data fit a curve of the form $y = ab^x$

x	2	3	4	5	6
y:	144	172.8	207.4	248.8	298.5

Solution: To fit the exponential Curve, $y = a\,b^x$ (1)

By taking log of both sides, we get

$$\log y = \log a + x \log b$$

i.e. $\log y = A + BX$ (2)

where $A = \log a$ & $B = \log b$

To fit the straight line (2), the normal equations are

$$\sum \log y = nA + B\sum x \quad (3)$$

$$\sum x \log y = A\sum x + B\sum x^2 \quad (4)$$

Here, $n = 5$ $\sum x = 20$ $\sum x^2 = 90$ $\sum \log y = 26.6720$ $\sum x \log y = 108.5104$

On putting these values in equations (3) & (4), we get

$$26.6720 = 5A + 20B \quad (5)$$

$$108.5104 = 20A + 90B \quad (6)$$

Solving (5) & (6) simultaneously, we get the values of A and B.

A = 4.6054 and B = 0.1822

here $a = e^{4.6054} = 100.0262$ & $b = e^{0.1822} = 1.9999$

Then from equation (1) the exponential Curve, $y = (100.0262)\ (1.9999)^x$

Example 2: If the growth of a certain kind of bacteria follows the law $y = ab^x$. Then find the best fitting values of *a* & *b* using following data.

x	1	2	3	4	5
y	233.2	253.4	282.3	302.4	332.8

Solution: To fit the exponential Curve, $y = a\ b^x$ (1)

By taking log of both sides, we get

$$\log y = \log a + x \log b$$

$$i.e.\ \log y = A + BX \quad (2)$$

where $A = \log a$ & $B = \log b$

To fit the straight line (2), the normal equations are

$$\Sigma \log y = nA + B\Sigma x \quad (3)$$

$$\Sigma x \log y = A\Sigma x + B\Sigma x^2 \quad (4)$$

Here, $n = 5$ $\Sigma x = 15$ $\Sigma x^2 = 55$ $\Sigma \log y = 28.1491$ $\Sigma x \log y = 85.3354$

On putting these values in equations (3) & (4), we get

$$28.1491 = 5A + 15B \quad (5)$$

$$85.3354 = 15A + 55\ B \quad (6)$$

Solving (5) & (6) simultaneously, we get the values of A and B.

A = 5.3634 and B = 0.0888

here $a = e^{5.3634} = 213.4503$ & $b = e^{0.0888} = 1.0928$

Then from equation (1) the exponential Curve, $y = (213.4503)\ (0.0888)^x$

b) Exponential curve $y = ae^{bx}$

Let (x_1, y_1), (x_2, y_2),…… (x_n, y_n).be the given set of *n* data points. To fit the exponential

Curve, $y = ae^{bx}$ (1)

By taking $\log_e$ of both sides, we get

$$\log y = \log a + bx$$

$$i.e.\ \log y = A + bx \quad (2)$$

where $A = \log a$

To fit the straight line (2), the normal equations are

$$\Sigma \log y = nA + b\Sigma x \quad (3)$$

$$\Sigma x \log y = A\Sigma x + b\Sigma x^2 \quad (4)$$

Solving (3) & (4) simultaneously, we get the values of A and b.

$$here\ a = e^A$$

Example 3: Use least-squares method to fit a curve of the form $y = ae^{bx}$ to the data.

x	1	2	3	4	5	6
y	7.209	5.265	3.846	2.809	2.052	1.499

Solution: To fit the exponential Curve, $y = ae^{bx}$ (1)

By taking log on both sides, we get

$$\log y = \log a + bx$$

$$i.e.\ \log y = A + bx \quad (2)$$

where $A = \log a$

To fit the straight line (2), the normal equations are

$$\Sigma \log y = nA + b\Sigma x \quad (3)$$

$$\Sigma x \log y = A\Sigma x + b\Sigma x^2 \quad (4)$$

Here, $n = 6$ $\Sigma x = 21$ $\Sigma x^2 = 91$ $\Sigma \log y = 7.1399$ $\Sigma x \log y = 19.4928$

On putting these values in equations (3) & (4), we get

$$6A + 21b = 7.1398 \quad (5)$$

$$21A + 91b = 19.4928 \quad (6)$$

Solving (5) & (6) simultaneously, we get the values of A and B.

A = 2.2893 and b = - 0.3141

here $a = e^{2.2893} = 9.8680$ & $b = -0.3141$

Then from equation (1) the exponential Curve, $y = (9.8680)\, e^{-0.3141x}$

Example 4: The values of x and y obtained in an experiment are as follows

X	2.3	3.1	4	4.92	5.91
Y	33	39.1	50.3	67.2	85.6

Fit the probable law $Y = ae^{bx}$

Solution: To fit the exponential Curve, $y = ae^{bx}$ (1)

By taking $\log_e$ of both sides, we get

$\log y = \log a + bx$

i.e. $\log y = A + bx$ (2)

where $A = \log a$

To fit the straight line (2), the normal equations are

$\Sigma \log y = nA + b\Sigma x$ (3)

$\Sigma x \log y = A\Sigma x + b\Sigma x^2$ (4)

Here, $n = 5$ $\Sigma x = 20.23$ $\Sigma x^2 = 90.0345$ $\Sigma \log y = 19.7379$ $\Sigma x \log y = 82.0783$

On putting these values in equations (3) & (4), we get

$5A + 20.23b = 19.7379$ (5)

$20.23A + 90.0345b = 82.0783$ (6)

Solving (5) & (6) simultaneously, we get the values of A and B.

A = 2.8508 and b = 0.2710

here $a = e^{2.8508} = 17.3022$ & $b = 0.2710$

Then from equation (1) the exponential Curve, $y = (17.3022)\, e^{0.2710x}$

Example 5: The values of x and y obtained in an experiment are as follows

X	1	2	3	4	5
Y	1.65	2.7	4.5	7.35	17.52

Fit the probable law $Y = ae^{bx}$

Solution: To fit the exponential Curve, $y = ae^{bx}$ (1)

By taking $\log_e$ of both sides, we get

$$\log y = \log a + bx$$

$$i.e.\ l\log y = A + bx \qquad (2)$$

where $A = \log a$

To fit the straight line (2), the normal equations are

$$\Sigma \log y = nA + b\Sigma x \qquad (3)$$

$$\Sigma x \log y = A\Sigma x + b\Sigma x^2 \qquad (4)$$

Here, $n = 5$ $\Sigma x = 15$ $\Sigma x^2 = 55$ $\Sigma \log y = 7.8561$ $\Sigma x \log y = 29.2950$

On putting these values in equations (3) & (4), we get

$$5A + 15b = 7.8561 \qquad (5)$$

$$15A + 55b = 29.2950 \qquad (6)$$

Solving (5) & (6) simultaneously, we get the values of A and B.

A = -0.1467 and b = 0.5726

here $a = e^{-0.1467} = 0.8635$ & $b = 0.5726$

Then from equation (1) the exponential Curve, $y = (0.8635)\ e^{0.5726x}$

c) Exponential curve $y = a\ x^b$

Let (x_1, y_1), (x_2, y_2),…… (x_n, y_n).be the given set of *n* data points. To fit the exponential

Curve, $y = a\ x^b$ (1)

By taking log of both sides, we get

$$\log y = \log a + b\log x$$

$$i.e.\ l\text{ogy} = A + b\log x \qquad (2)$$

where $A = \log a$

To fit the straight line (2), the normal equations are

$$\Sigma \log y = nA + b\Sigma \log x \qquad (3)$$

$$\Sigma \log x \log y = A\Sigma \log x + b\Sigma (\log x)^2 \qquad (4)$$

Solving (3) & (4) simultaneously, we get the values of A and b.

$$here\ a = e^A$$

Example 6: Fit a least square geometric curve $y = a\,x^b$ to the following data:

x	1	2	3	4	5
y	0.5	2	4.5	8	12.5

Solution: To fit the exponential Curve, $y = a\,x^b$ (1)

By taking log of both sides, we get

$$\log y = \log a + b\log x$$

$$i.e.\ \log y = A + b\log x \quad (2)$$

where $A = \log a$

To fit the straight line (2), the normal equations are

$$\Sigma \log y = nA + b\Sigma \log x \quad (3)$$

$$\Sigma \log x \log y = A\Sigma \log x + b\Sigma (\log x)^2 \quad (4)$$

Here, $n = 5$ $\Sigma \log x = 4.7875$ $\Sigma (\log x)^2 = 6.1995$ $\Sigma \log y = 6.1092$ $\Sigma \log x \log y = 9.0806$

On putting these values in equations (3) & (4), we get

$$5A + 4.7875b = 6.1092 \quad (5)$$

$$4.7875A + 6.1995b = 9.0806 \quad (6)$$

Solving (5) & (6) simultaneously, we get the values of A and b.

A = -0.6931 and b = 2

here $a = e^{-0.6931} = 0.5$ & $b = 2$

Then from equation (1) the exponential Curve, $y = (0.5)\,x^2$

Example 7: Fit a least square geometric curve $y = a\,x^b$ to the following data:

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.1	6.8	7.5

Solution: To fit the exponential Curve, $y = a\,x^b$ (1)

By taking log of both sides, we get

$$\log y = \log a + b\log x$$

$$i.e.\ \log y = A + b\log x \quad (2)$$

where $A = \log a$

To fit the straight line (2), the normal equations are

$$\Sigma \log y = nA + b\Sigma \log x \qquad (3)$$

$$\Sigma \log x \log y = A\Sigma \log x + b\Sigma (\log x)^2 \qquad (4)$$

Here, $n = 6$ $\Sigma \log x = 6.5793$ $\Sigma (\log x)^2 = 9.4099$ $\Sigma \log y = 9.9319$ $\Sigma \log x \log y = 12.0201$

On putting these values in equations (3) & (4), we get

$$6A + 6.5793b = 9.9319 \qquad (5)$$

$$6.5793A + 9.4099b = 12.0201 \qquad (6)$$

Solving (5) & (6) simultaneously, we get the values of A and B.

A = 1.0912 and b = 0.5144

here $a = e^{1.0912} = 2.978$ & $b = 0.5144$

Then from equation (1) the exponential Curve, $y = (2.978)\, x^{0.5144}$

Example 8: Fit a least square geometric curve $y = a\, x^b$ to the following data:

x	1	2	3	4
y	2.5	8	19	50

Solution: To fit the exponential Curve, $y = a\, x^b$ (1)

By taking log of both sides, we get

$$\log y = \log a + b \log x$$

i.e. $\log y = A + b \log x$ (2)

where $A = \log a$

To fit the straight line (2), the normal equations are

$$\Sigma \log y = nA + b\Sigma \log x \qquad (3)$$

$$\Sigma \log x \log y = A\Sigma \log x + b\Sigma (\log x)^2 \qquad (4)$$

Here, $n = 4$ $\Sigma \log x = 3.1781$ $\Sigma (\log x)^2 = 3.6092$ $\Sigma \log y = 9.8522$ $\Sigma \log x \log y = 10.0944$

On putting these values in equations (3) & (4), we get

$$4A + 3.1781b = 9.8522 \qquad (5)$$

$$3.1781A + 3.6092b = 10.0944 \qquad (6)$$

Solving (5) & (6) simultaneously, we get the values of A and B.

A = 0.7983 and b = 2.0952 here $a = e^{0.7983} = 2.2219$ & $b = 2.0952$

Then from equation (1) the exponential Curve, $y = (2.2219)\, x^{2.0952}$

Examples for Practice

Example 1: Fit a curve of the form $y = ab^x$ to the following data,

x	1	2	3	4	5
y	7.1	27.8	62.1	110	161

Example 2: The following data pertains to the demand for a product (in thousands of units) and its price (in cents) charged in 5 different market areas,

Price	X	20	16	10	11	14
Demand	Y	22	41	120	89	56

Fit a power function $y = \propto x^{\beta}$ and use it to estimate the demand when the price of the product is 12 cent.

Example 3: If the growth of a bacteria of a certain kind of bacteria follows the law $y=ab^x$ Then the find the best fitting values of a & b using following data.

x	1	2	3	4	5	6
y:	151	100	61	50	20	8

Example 4: If the growth of a bacteria of a certain kind of bacteria follows the law $N=ab^t$. Then the find the best fitting values of a & b using following data.

Also estimate N when t = 7.

t	0	1	2	3	4	5	6
N	32	47	65	92	132	190	275

Example 5: Fit a least square geometric curve $y = a\,x^b$ to the following data:

x	1	2	3	4	5
y	0.5	2	4.5	8	12.5

Example 6: Fit the curve $y = ax^b$ to the following data:

x:	1	2	3	4	5	6
y:	2.98	4.26	5.21	6.1	6.8	7.5

3. Fitting of parabola:

Let $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ be the given set of *n* data points. To fit the parabola

$$y = a + bx + cx^2 \qquad (1)$$

to this data, we use normal equations given by

$$\sum y = na + b\sum x + c\sum x^2 \qquad (2)$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \qquad (3)$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \qquad (4)$$

Solving (2), (3) & (4) simultaneously, we get the values of a, b and c.

Example 1: Find the best fitting regression equation of type $y = a + bx + cx^2$ to the following data:

x:	3	2	1	0	-1	-2	-3
y:	10	8	3	1	2	6	8

Solution: Let the parabola to fit the given data be

$$y = a + bx + cx^2 \qquad (1)$$

The normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \qquad (2)$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \qquad (3)$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \qquad (4)$$

Here, n = 7, $\sum x = 0$, $\sum x^2 = 28$, , $\sum x^3 = 0$, $\sum x^4 = 196$, $\sum y = 38$, $\sum xy = 11$, $\sum x^2 y = 223$.

On putting these values in equations (2), (3) & (4), we get

$$38 = 7a + 28c \qquad (5)$$

$$11 = 28b \qquad \Rightarrow b = \frac{11}{28} = 0.3929$$

$$223 = 28a + 196c \qquad (6)$$

Solving (5) & (6) simultaneously, we get

$$a = 2.0478, c = 0.8452$$

So, the parabola to best fit is $y = 2.0478 + 0.3929x + 0.8452x^2$

Example 2: Fit a second degree parabola to the following data:

x	0	1	2	3	4
y	7.1	2.4	2.6	2.7	3.4

Solution: Let the parabola to fit the given data be

$$y = a + bx + cx^2 \qquad (1)$$

The normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \qquad (2)$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \qquad (3)$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \qquad (4)$$

Here, n = 5, $\sum x = 10$, $\sum x^2 = 30$, $\sum x^3 = 100$, $\sum x^4 = 354$, $\sum y = 18.2$, $\sum xy = 29.3$, $\sum x^2 y = 91.5$.

On putting these values in equations (2), (3) & (4), we get

$$18.2 = 5a + 10b + 30c \qquad (5)$$

$$29.3 = 10a + 30b + 100c \qquad (6)$$

$$91.5 = 30a + 100b + 354c \qquad (7)$$

By $(6) - 2\times(5)$, we get $10b + 40c = -7.1$ (8)

By $(7) - 6\times(5)$, we get $40b + 174c = -17.7$ (9)

By $(9) - 4\times(8)$, we get $14c = 10.7 \quad \Rightarrow \; 14c = \dfrac{10.7}{14} = 0.7643$

Putting this value in (8), we get $10b = -7.1 - 40(0.7643) \;\Rightarrow\; b = -3.7672$

From equation (5), we get $5a = 18.2 - 10(-3.7672) - 30(0.7643) \Rightarrow \; a = 6.5886$

So, the parabola to best fit is $y = 6.5886 - 3.7672x + 0.7643x^2$

Example .3: Find least squares polynomial approximation of degree two to the following data:

x:	10	15	20	25	30	35	40
y:	11	13	16	20	27	34	41

Solution: Let the parabola to fit the given data be

$$y = a + bx + cx^2 \qquad (1)$$

Consider, $u = \dfrac{x - \bar{x}}{h} = \dfrac{x - 25}{5}$ Then u: $-3, -2, -1, 0, 1, 2, 3$

Hence,

u:	−3	−2	−1	0	1	2	3
Y:	11	13	16	20	27	34	41

So, the transformed equation of parabola is

$$Y = a + bu + cu^2 \qquad (2)$$

The normal equations are

$$\sum y = na + b\sum u + c\sum u^2 \qquad (3)$$

$$\sum uy = a\sum u + b\sum u^2 + c\sum u^3 \qquad (4)$$

$$\sum u^2 y = a\sum u^2 + b\sum u^3 + c\sum u^4 \qquad (5)$$

Here, n = 7, $\sum u = 0$, $\sum u^2 = 28$, , $\sum u^3 = 0$, $\sum u^4 = 196$, $\sum y = 162$, $\sum uy = 143$, $\sum u^2 y = 699$.

On putting these values in equations (3), (4) & (5), we get

$$162 = 7a + 28c \qquad (6)$$

$$143 = 28b \qquad \Rightarrow b = \tfrac{143}{28} = 5.1071$$

$$699 = 28a + 196c \qquad (7)$$

Solving (5) & (6) simultaneously, we get

$$a = 20.7144 \;\&\; C = 0.6071$$

Putting these values in equation (2), we get

$$y = 20.7144 + 5.1071u + 0.6071u^2$$

By substituting, $u = \dfrac{x - \bar{x}}{h} = \dfrac{x - 25}{5}$

$$y = 20.7144 + 5.1071\left(\frac{x-25}{5}\right) + 0.6071\left(\frac{x-25}{5}\right)^2$$

$$y = 20.7144 + 1.02142\,x - 25.5355 + 0.0243x^2 - 1.2142\,x + 15.1775$$

So, the parabola to best fit is $y = 10.3565 - 0.1928x + 0.0243x^2$

Examples for Practice

Example.1: The following tables give the results of measurements of train resistance:

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

Here, V is the velocity in Km / hour and R is the resistance in Kg / quintal. Fit a relation of the form $R = a + bV + cV^2$ to the data. Estimate R when $V = 120$.

Example.2: Find the best fitting regression equation of type $y = a + bx + cx^2$ to the following data:

x:	0	1	2	3	4	5	6
y:	2.4	2.1	3.2	5.6	9.3	14.6	21.9

Example.3: Fit a second degree parabola $y = a + bx + cx^2$ to the following data:

x	1	1.5	2	2.5	3	3.5	4
y	1.1	1.3	1.6	2	2.7	3.4	4.1

Example.4: Find the best fitting regression equation of type $y = a + bx + cx^2$ to the following data:

x:	0	1	2	3	4	5	6
y:	2.4	2.1	3.2	5.6	9.3	14.6	21.9

Example.5: Fit a second degree parabola $y = a + bx + cx^2$ to the following data:

x	1	1.5	2	2.5	3	3.5	4
y	1.1	1.3	1.6	2	2.7	3.4	4.1

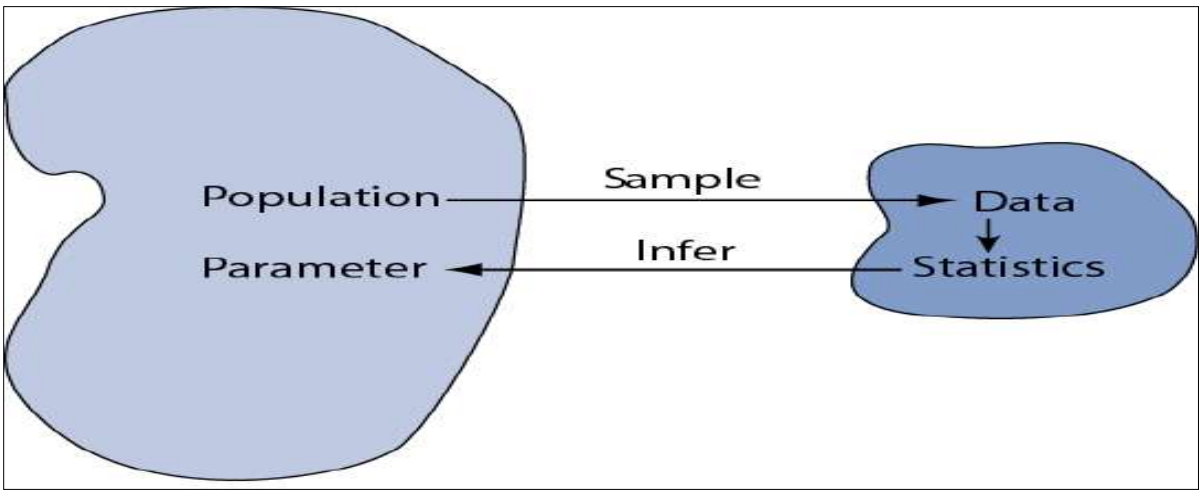
Example.6: Fit a second degree parabola $y = a + bx + cx^2$ to the following data:

x	-3	-2	-1	0	1	2	3
y	4.63	2.11	0.67	0.09	0.63	2.15	4.58

Test of Significance

- **Population:** The group of individuals under study is called the **population.** For example, if we want to have an idea of the average per capita (monthly) income of the people in Maharashtra, we will have to enumerate all the earning individuals in the Maharashtra.
- **Parameter:** A parameter is any numerical quantity that characterizes a given population or some aspects of it. This means the parameter tells us something about the whole population.
- **Samples:** A finite subset of statistical individuals in a population is called a **sample** and the number of individuals in a sample in a sample is called the **sample size (n)**.
  For example, we select 100 people in Kolhapur district for average per capita income in Maharashtra state.
- **Statistic:** A statistic is an estimate of a population parameter based on sample.

- **Sampling:** The process of drawing random samples from population.
  For example, in a shop we assess the quality of wheat by taking a handful of it from the bag and then decide to purchase it or not. A Meal maker normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.
- **Statistical Inference:** It refers to the process of selecting and using a sample to draw inference about population from which sample is drawn is called **statistical inference**.
  For example, we draw 10 diameters of screws from a large lot of screws. Sampling is done in order to see whether a model of the population is accurate enough for practical purposes.



- **Tests of Significance:** The methods of statistical inference used to accept or reject claims based on sample data are known as tests of significance.
- **Hypothesis:** A statement or a claim about a property of a population is called a Hypothesis. There are two types of Hypothesis **Null Hypothesis** and **Alternative Hypothesis**.

- **Null Hypothesis:** A definite Statement about the value of a population parameter. Such a hypothesis, which is usually a hypothesis of no difference, is called null hypothesis. Generally it is represented by H_0. For e.g.H_0:$\mu = 3$

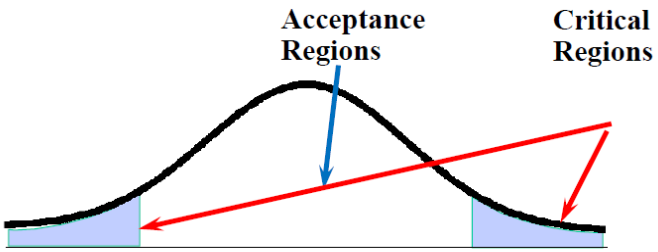
- **Alternative Hypothesis:** Any hypothesis which is complementary to the null hypothesis is called alternative hypothesis. Statement about the value of a population parameter that must be true if the null hypothesis is false. Generally it is represented by H_1. For e.g. H_1: $\mu > 3$ or $\mu < 3$ or $\mu \neq 3$

- **Hypothesis Testing:** Hypothesis Testing is to test the claim or statement of population.
  For example, a statement is made that "the average starting salary for Engineering Graduate student from KITCoEK, is Rs.35000 per month".

- **Errors in Sampling:** The main objective is to draw valid inferences about the population parameters on the basis of the sample results. As such we are liable to commit the following two types of errors.

  - **Type I error:** Reject the null hypothesis when it is true. Probability of type I error is denoted by α.
  - **Type II error:** Accept the null hypothesis when it is wrong. Probability of type II error is denoted by β.
- **Level of significance:** The Probability of type I error (α) is called the level of significance. The level of significance is always fixed in advance before collecting the sample information. Generally it is consider 1 % and 5%.

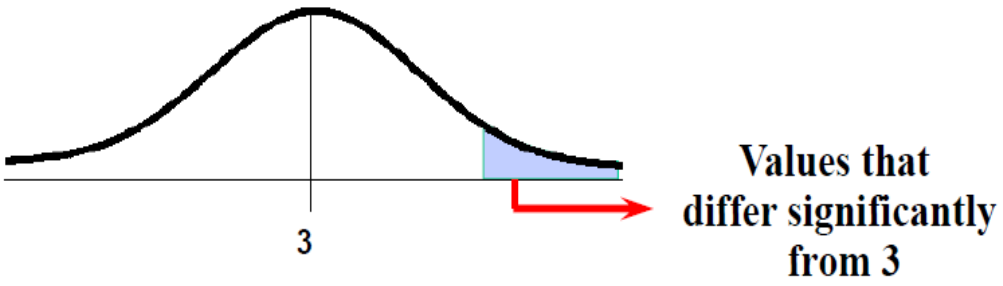
- **Critical Region:** The region (corresponding to a statistic) of rejection of null hypothesis is called **critical region** or **rejection region.** Test statistic falls in some interval which we reject the null hypothesis. This interval is called **rejection region.** The test statistic falls in some interval which we accept the null hypothesis. This interval is called **acceptance region.**



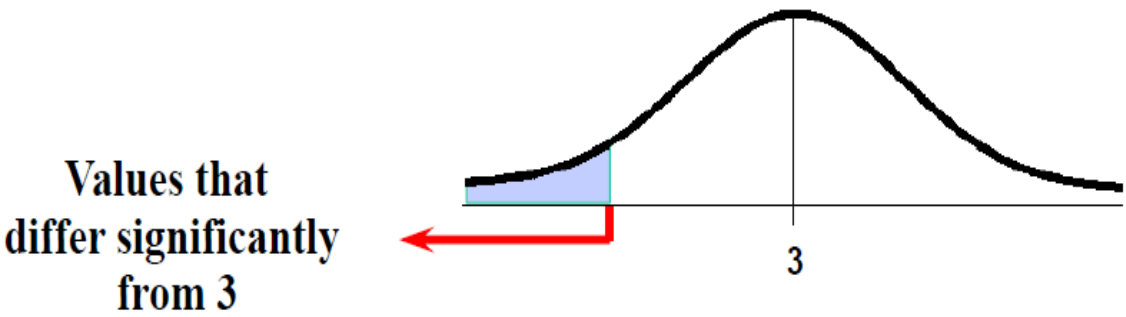
One tailed and two tailed tests:

- **One tailed: A one tailed test is a statistical test in which the critical area of a distribution is one sided so that it is either greater than or less than a certain parameter value.**

For e.g.H_0: $\mu = 3$ vs. H_1: $\mu > 3$ (Right tailed test)

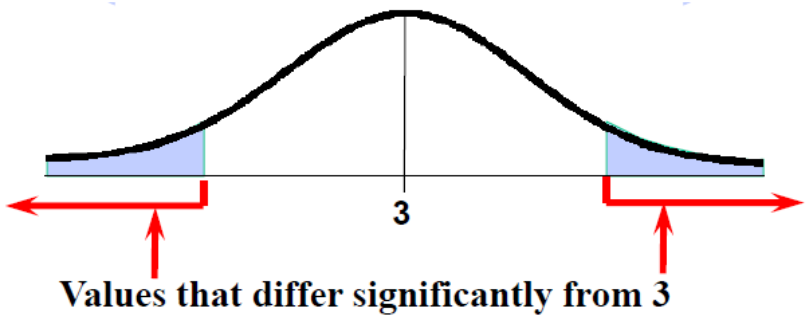


For e.g.H_0: $\mu = 3$ vs. H_1: $\mu < 3$ (Left tailed test)



- **Two tailed: A two tailed test is a statistical test in which the critical area of a distribution lying in both tails of the probability curve of test statistic.**

  For e.g.H_0: $\mu = 3$ vs. H_1: $\mu \neq 3$ (Two tailed test)



- **Critical Values:** The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value. It depends up on,

  (1)The level of significance α.

  (2)The Alternative hypothesis, whether it is one tailed or two tailed.

- **Procedure (Steps) for Testing of Hypothesis:**

Describe in words the population characteristic about which hypothesis are to be tested.

1. Null hypothesis: Set up the null hypothesis H_0.

2. Alternative hypothesis: Set up the alternative hypothesis H_1, this will enable us to decide whether we have to use right, left, or two-tailed test.

3. Level of Significance: Choose the appropriate α in advance.

4. Critical Value: Determine the critical value associated with the Alternative hypothesis and level of significance

5. Test Statistic: Compute the test statistic, (i.e. Z- test or t- test or Chi-square test)

6. Test Criteria: Decide whether to reject the null hypothesis, H_0, or fail to reject the null hypothesis. It depends on the critical value and the test statistic of the test.

7. Conclusion: State your result in the context of the specific problem.

Large Sample Test

Introduction:

The sample size (n) is greater than 30 ($n \geq 30$) it is known as large sample. For large samples the sampling distributions of statistic are normal (Z-test). A test of statistic for large sample is known as large sample test.

Test 1: Test of Significance for Single Mean in Large sample test:

To test the null hypothesis H_0: $\mu = \mu_0$ that the sample has been drawn from a population with mean μ_0 and variance σ^2 i.e. there is no significant difference between the sample mean $\bar{x}$ and population mean μ_0, the test statistic (for large samples), is:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Note:

1. For large samples, if the population standard deviation (S.D) σ is unknown, we use its estimate provided by the sample standard deviation's'. $\sigma^2 = s^2$ so, the test statistic is:

$$Z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

2. Confidence Limits for μ:

95% confidence limits for μ are $\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$.

99% confidence limits for μ are $\bar{x} \pm 2.58\left(\frac{\sigma}{\sqrt{n}}\right)$.

3. Table of Critical Values for Large Samples:

Critical Value Z_α	Level of Significance (α)	
	1% (0.01)	5% (0.05)
Two-tailed test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 1.96$
Right-tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$
Left-tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$

4. If test statistics Z is less than Critical value then accept the null hypothesis otherwise do no accept (Reject the null hypothesis)

Solved Examples

Example 1: A sample of 625 members has a mean of 3.5 cm. Can it be reasonably regarded as a truly random sample from a large population with mean 3.2 and variance 2.25 at 5% level of significance?

Solution: Here $n = 625,\ \bar{x} = 3.5, \mu = 3.2, \sigma^2 = 2.25$

(i) Null Hypothesis: H_0: $\mu = 3.2$ *i.e.* sample is selected from a large population with mean = 3.2.

(ii) Alternative Hypothesis: H_1: $\mu \neq 3.2$ (Two-tailed)

(iii) Level of Significance: $\alpha = 0.05$

(iv)Critical value: For two-tailed test, the value of Z_α at 5% level of significance from the table = 1.96 i.e. $|Z_\alpha| = 1.96$

(v) Test Statistic: Under H_0**,** the test statistic for given sample is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad Z = \frac{3.5 - 3.2}{1.5/\sqrt{625}} = 5 \quad |Z| = 5$$

(vi) Test Criteria: Since test statistic |Z| > critical value $|Z_\alpha|$, the null hypothesis is rejected.

(vii) Conclusion: It cannot be regarded as a random sample from a large population with mean = 3.2.

Example 2: A sample of 40 sugarcane selected at random from farm was checked for the length and it was found that the mean and the variance of the length are 58.5 and 4.2 inches respectively, is it reasonable to say that the average length of the sugarcane is 60 inches. Use 1% level of significance.

Solution: Here $n = 40,\ \bar{x} = 58.5,$ sample variance $s^2 = 4.2$ & $\mu = 60.$

(i) Null Hypothesis: H_0: $\mu = 60$ i.e. the average length of the sugarcane is 60 inches.

(ii) Alternative Hypothesis: H_1: $\mu \neq 60$ (Two-tailed)

(iii) Level of Significance: $\alpha = 0.01$

(iv)Critical value: For two-tailed test, the value of Z_α at 1% level of significance from the table = 2.58 i.e. $|Z_\alpha| = 2.58$

(v) Test Statistic: Since the population S.D. is unknown but sample S.D. *s* is known and the sample is large

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad z = \frac{58.5 - 60}{2.0494/\sqrt{40}} = -4.6291 \qquad |z| = 4.6291$$

(vi)Test Criteria: Since test statistic |Z| > critical value | Z_α |, the null hypothesis is rejected.

(vii) Conclusion: It is not reasonable to say that the average length of the sugarcane is 60 inches.

Example 3: 64 observations are selected at random from a normal distribution whose variance is 25.Their mean is calculated and found to be 11.1. Test the hypothesis that the true value of the population mean is 10. Use 5% level of significance.

Solution: Here $n = 64,\ \bar{x} = 11.1,\ \sigma^2 = 25 \Rightarrow \sigma = 5\ \&\ \mu = 10.$

(i) Null Hypothesis: H_0: μ = 10 (i.e. the true value of the population mean is 10.)

(ii) Alternative Hypothesis: H_1: μ ≠ 10 (Two-tailed)

(iii) Level of Significance: α = 0.05

(iv) Critical value: For two-tailed test, the value of Z_α at 5% level of significance from the table = 1.96 i.e. | Z_α | =1.96

(v) Test Statistic: Under H_0**,** the test statistic for given sample is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad z = \frac{11.1 - 10}{5/\sqrt{64}} \qquad |z| = 1.76$$

(vi) Test Criteria: Since test statistic |Z| < critical value | Z_α |, the null hypothesis is accepted.

(vii) Conclusion: The true value of the population mean is 10.

Example 4: A tire company claims that lives of tires have mean 42000 km with S.D of 4000 km. A change in production process is believed to give better product. A test sample of 81 new tires has mean 42500 km. Test at 5% l. o. s that new product is better than current one.

Solution: Here $n = 81,\ \bar{x} = 42500,\ \mu = 42000\ \&\ \sigma = 4000.$

(i) Null Hypothesis: H_0: μ = 42000 (i.e. the sample mean and the population mean do not differ significantly.)

(ii) Alternative Hypothesis: H_1: μ > 42000 (Right-tailed alternative)

(iii) Level of Significance: α = 0.05

(iv)Critical value: For two-tailed test, the value of Z_α at 5% level of significance from the table = 1.645 i.e. $|Z_\alpha| = 1.645$

(v) Test Statistic: $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ $\quad z = \dfrac{42500 - 42000}{4000/\sqrt{81}} = 1.125 \quad |z| = 1.125$

(vi)Test Criteria: Since test statistic |Z| < critical value $|Z_\alpha|$, the null hypothesis is accepted.

(vii) Conclusion: The new product is not better than current one.

Example 5: A sample of 900 members has a mean 3.4 cm. and S.D = 2.61. Is the sample from a large population of mean 3.25 and S.D.2.61 cm? (Use α = 0.05).If the population is normal and its mean is unknown, find the 95% confidence limits for true mean μ.

Solution: Here $n = 900,\ \bar{x} = 3.4,\ \mu = 3.25, \sigma = s = 2.61$

(i) Null Hypothesis: H_0: μ = 3.25(*i.e.* sample is selected from a large population with mean = 3.25.)

(ii) Alternative Hypothesis: H_1: μ ≠ 3.25 (Two-tailed)

(iii) Level of Significance: α = 0.05

(iv)Critical value: For two-tailed test, the value of Z_α at 5% level of significance from the table = 1.96 i.e. $|Z_\alpha| = 1.96$

(v) Test Statistic: The test statistic for given sample is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad Z = \frac{3.4 - 3.25}{2.61/\sqrt{900}} = 1.73 \quad |Z| = 1.73$$

(vi) Test Criteria: Since test statistic |Z| < critical value $|Z_\alpha|$, the null hypothesis is accepted.

(vii) Conclusion: It can be regarded as a random sample from a large population with mean = 3.25.

If μ is unknown, then 95% confidence limits for μ are as follows:

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} \quad = 3.4 \pm 1.96\frac{2.61}{\sqrt{900}} \quad = 3.4 \pm 0.1705 \quad = (3.2295 \ \& \ 3.5705)$$

$$\therefore \ 3.2295 < \mu < 3.5705$$

Example 6: The mean muscular endurance score of a random sample of 60 subjects was found to 145 with a variance 1600. Construct a 95% confidence interval for the mean. Assume the sample size to be large enough for normal approximation.

Solution : Here $n = 60,\ \bar{x} = 145,\ s^2 = 1600,\ s = \sigma = 40$.

95% confidence limits for μ are

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 145 \pm 1.96\frac{40}{\sqrt{60}} = 145 \pm 10.12 = 134.88 \ \& \ 155.12$$

95% confidence interval for μ is $(134.88, 155.12)$.

Examples for Practice

Example 1: Describe test of significance for single mean with suitable examples. Define its confidence interval.

Example 2: The average marks in mathematics of a sample of 100 students were 51 with a S.D of 6 marks. Could this have been a random sample from a population with average marks 50?

Example 3: The following results are obtained from a sample of 100 boxes of biscuits. Mean weight content = 490 gm S.D. of the weight = 9 gm could the sample come from a population having a mean of 500 gm.

Example 4: The mean and the standard deviation of the eye sight of 50 different persons were found to be 22 feet and 1.5 feet respectively. From these sample results can we say that the hypothesis."Average eye sight of the person in 20 feet" is true?

Example 5: From the survey of 105 people it was found that the mean and the standard deviation of night sleep time are 6.75hrs and 0.35 hrs respectively. Using Z-test is it reasonable to say that "Avg. Sleep time is more than 6 hrs."?

Example 6: A machine is set to produce metal plates of thickness 1.5cm with SD 0.2 cm. A sample of 100 plates produced by the machine gave an average thickness of 1.52cm.is the machine fulfilling the purpose?

Example 7: A sample of 50 items gives the mean 6.2 and S.D 10.24.Can it is regarded as drawn from a normal population with mean 5.4 at 5%level of significance.

Example 8: Can it be concluded that the average life span of an Indian is more than 70 years, if a random sample of 100 Indians has an average life span of 71.8 years with standard deviation of 7.8 years.

Example 9: A sample of 50 pieces of certain type of string was tested. The mean breaking strength turned out to be 14.5 ponds. Test whether the sample is from batch of a string having a mean breaking strength of 15.6pounds and standard deviation of 2.2 pounds.

Example 10: A random sample of size 36 has 53 as mean and sum of squares of deviation from mean is 150.Can this sample be regarded as drawn from the population having 54 as mean.

Example 11: The mean height of random sample of 100 individuals from a population is 160c.m.The S.D of the sample is 10c.m.Would it be reasonable to suppose that the mean height of the population is 165c.m?

Example 12: The mean weight obtained from a random sample of size 100 is 64 gm. The S.D of the weight distribution of the population is 3 gm. Test the statement that the mean weight of the population is 67 gm, at 5% l.o.s.

Small Sample Test

Introduction:

The entire large sample theory was based on the application of normal test .However; if the sample size (n) is small (n < 30), the distribution of the various statistics are far from normality and such normal test cannot be applied it small is sample.

Student's t - Distribution:

Consider a small sample of size n drawn from a normal population with mean μ and variance σ^2.Then student's t is defined as,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \sim t_{\text{n-1 d.f.}}$$

If we calculate t for each sample, we obtain the sampling distribution for t. This distribution is known as student's t - distribution.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Test 2: Test of Significance for Single Mean in Small sample test:

If a random sample x_i (i =1, 2,…, n) of size n has been drawn from normal population. To test the null hypothesis H_0: $\mu = \mu_0$ that the sample has been drawn from a population with mean μ_0 and variance σ^2 i.e. there is no significant difference between the sample mean $\bar{x}$ and population mean μ_0, the test statistic (for small samples), is:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \quad \sim t_{\text{n-1 d.f.}}$$

Where, $sample\,mean = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum x_i^{\,2}}{n} - (\bar{x})^2$

$$hence,\ s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum x_i^{\,2}}{n} - (\bar{x})^2}$$

follows Student's t – distribution with n – 1 degrees of freedom (d.f).

We now compare the calculated value of t with the tabulated value at certain level of significance. If calculated | t | < tabulated t, null hypothesis is accepted otherwise rejected.

Note: Confidence Limits for μ:

95% confidence limits for μ are $\bar{x} \pm t_{0.05}\left(\frac{s}{\sqrt{n-1}}\right)$.

99% confidence limits for μ are $\bar{x} \pm t_{0.01}\left(\frac{s}{\sqrt{n-1}}\right)$.

Assumption for Student's t – test:

1. The parent population from which the sample drawn is normal.
2. The sample observations are independent.
3. The population standard deviation (σ) is unknown.

Example 1: A machinist is making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a standard deviation of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specifications.

Solution: Here $\mu = 0.7, \bar{x} = 0.742, \quad s = 0.04 \ \& \ n = 10.$

(i) Null Hypothesis, H_0: $\mu = 0.7$(i.e. the product is confirming the specifications.)

(ii)Alternative Hypothesis, H_1: $\mu \neq 0.7$ (Two-tailed)

(iii) Level of Significance: $\alpha = 0.05$

(iv)Critical value: For two-tailed test, at 5% level of significance the value of $t_{\alpha,\ n\text{-}1d.f}$ i.e. $t_{0.05,\ 9\ d.f}$ from the table = 2.262

(v) Test Statistic: Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{(n-1)}$$

$$t = \frac{(0.742 - 0.7)}{0.04/\sqrt{9}} \qquad \therefore t = 3.15$$

(vi) Test Criteria: Since calculated | t | > tabulated t, hence the null hypothesis is rejected.

(vii) Conclusion: we conclude that the product is not meeting the specifications.

Example 2: The mean weekly sales of soap bars in departmental stores were 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a s.d. of 17.2. Was the advertising campaign successful?

Solution: Given $n = 22, \ \bar{x} = 153.7, s = 17.2.$

(i) Null Hypothesis, H_0: $\mu = 146.3$(i.e. the advertising campaign is not successful.)

(ii)Alternative Hypothesis, H_1: $\mu > 146.3$ (Right-tailed)

(iii) Level of Significance: $\alpha = 0.05$

(iv)Critical value: For right-tailed test, at 5% level of significance the value of $t_{\alpha,\ n\text{-}1d.f}$ i.e. $t_{0.05,\ 21\ d.f}$ from the table = 1.721

(v) Test Statistic: Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{(n-1)} \quad t = \frac{153.7 - 146.3}{17.2/\sqrt{21}} \quad t = 1.9715$$

(vi)Test Criteria: Since calculated | t | > tabulated t, hence the null hypothesis is rejected.

(vii) Conclusion: We conclude that the advertising campaign is definitely successful in promoting sales.

Example 3: A random sample of 10 boys had the following I.Q.'s: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.

Solution: Let X = I.Q of the boy

From given dada, n =10, μ = 100 $\sum x = 972$ $\sum x^2 = 96312$

$$\bar{x} = \frac{972}{10} = 97.2 \quad s = \sqrt{\frac{\sum x_i^{\,2}}{n} - (\bar{x})^2} = \sqrt{\frac{96312}{10} - (97.2)^2} = \sqrt{183.36} = 13.5410$$

(i) Null Hypothesis, H_0: μ = 100(i.e. the data are consistent with the assumption of a population mean I.Q. of 100)

(ii) Alternative Hypothesis, H_1: μ ≠ 100 (Two-tailed)

(iii) Level of Significance: α = 0.05

(iv) Critical value: For two-tailed test, at 5% level of significance the value of $t_{\alpha\,,\,n\text{-}1d.f}$ i.e. $t_{0.05\,,\,9\,d.f}$ from the table = 2.262

(v) Test Statistic: Under H_0, the test statistic is,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{(n-1)} \quad t = \frac{97.2 - 100}{13.5410/\sqrt{9}} \quad t = -0.6203 \qquad |t| = 0.6203$$

(vi)Test Criteria: Since calculated | t | < tabulated t_α, hence the null hypothesis is accepted.

(vii) Conclusion: We conclude that the data are consistent with the assumption of mean I.Q. of 100 in the population.

<u>Confidence interval for μ:</u>

The 95% confidence limits within which the mean I.Q. values of samples of 10 boys will lie are given by:

$$\bar{x} \pm t_{0.05}\left(\frac{s}{\sqrt{n-1}}\right). \quad = 97.2 \pm 2.262\left(\frac{13.5410}{\sqrt{9}}\right) \quad = 86.9901 \;\&\; 107.4099$$

Hence the required 95% confidence interval is [86.9901< μ < 107.4099].

Example 4: A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43.5 inches for the population is not reasonable. Obtain 95% and 99% confidential limits for the same.

Solution: Given $n=16,\ \bar{x}=41.5\ inches\ \&\ \sum(x-\bar{x})^2=135\ sq.\ inches$

$$hence, s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}=\sqrt{\frac{135}{16}}=2.9047$$

(i) Null Hypothesis, H_0: $\mu = 43.5$(i.e. the data are consistent with the assumption that mean of the population is 43.5 inches.)

(ii) Alternative Hypothesis, H_1: $\mu \neq 43.5$ (Two-tailed)

(iii) Level of Significance: $\alpha = 0.05$

(iv) Critical value: For two-tailed test, at 5% level of significance the value of $t_{\alpha,\ n-1d.f}$ i.e. $t_{0.05,\ 15\ d.f}$ from the table = 2.131

(v) Test Statistic: Under H_0, the test statistic is

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n-1}}\sim t_{(n-1)} \quad t=\frac{41.5-43.5}{2.9047/\sqrt{15}} \quad t=-2.6667 \qquad |t|=2.6667$$

(vi)Test Criteria: Since calculated | t | > tabulated t_α, hence the null hypothesis is rejected.

(vii) Conclusion: We conclude that the assumption of a mean of 43.5 inches for the population is not reasonable.

<u>Confidence interval for μ:</u>

The 95% confidence limit for μ is given by:

$$\bar{x}\pm t_{0.05}\left(\frac{s}{\sqrt{n-1}}\right). \quad =41.5\pm 2.131\left(\frac{2.9047}{\sqrt{15}}\right) \quad =41.5\pm 1.5982$$

Hence the required 95% confidence interval is [39.9018 < μ < 43.0982].

The 99% confidence limit for μ is given by:

$$\bar{x}\pm t_{0.01}\left(\frac{s}{\sqrt{n-1}}\right). \quad =41.5\pm 2.947\left(\frac{2.9047}{\sqrt{15}}\right) \quad =41.5\pm 2.2102$$

Hence the required 95% confidence interval is [39.2898 < μ < 43.7102].

Examples for Practice

Example 1: Define t distribution. State assumption used for testing a specified mean.

Example 2: A random sample of 10observations of animal growth under standard condition give 96.7, 84.8, 101.8, 78.3, 110.6, 93.4, 87.8, 91.3, 98.2, 88.7cm test the hypothesis that they came from distribution with mean100.

Example 3: Tests made on the breaking strength of 10 pieces of a metal were gave the results 578, 572, 570, 568, 572, 570, 570, 572, 596, and 584 kg. Test if the mean breaking strength of the wire can be assumed as 577 kg .

Example 4: On an average the gain in the weight due to certain drug is 2.5 kg, if the gain in the weights of five different subjects after applying this drug are 3, 2.8, 3.6, 3.2,3 using t-test can we say that gain in weight is more than expected?

Example 5: 10 subjects tested for blood clotting time have shown the mean and the standard deviation as 95 and 5 sec respectively. Using t-test, test the hypothesis that "The average blood clotting time is 100 sec"

Example 6: The Company claims that a plant grown by its seed will give average 2kg of production. The sample of such 10 plant has mean and standard deviation of the production as 17 kg and 0.25kg respectively. From the sample results can we say that the claim of the company is not true?

Example 7: A fertilizer mixing machine set to give 12 kg of nitrate for quintal bag of fertilizer .Ten 100 kg bags are examined .The percentage of nitrate per bag are as follows: 11,14,13,12,13,12,13,14,11,12. Is there any reason to believe that the machine is defective?

Example 8: The nine items of a sample having the following values: 45, 47, 50, 52, 48, 47, 49, 53, and 51. Does the mean of these differ significantly from the assumed mean of 47.5?

Example 9: The following values give the lengths of 12 samples of Egyptian cotton taken from a consignment: 48, 46, 49, 46, 52, 45, 43, 47, 47, 46, 45, and 50. Test if the mean length of the consignment can be can be taken as 46.

Example 10: A company supplies tooth paste in a packing of 100 gm. A sample of 10 packing gave the following weights in gms.100.5, 100.3, 100.1, 99.8, 99.7, 100.3, 100.4, 99.2, 99.3, and 99.7.Does the sample support the claim of the company that the packing weight 100 gms.

Example 11: A sample of 20 items has mean 42 units and S.D 5 units. Test the hypothesis that it is a random sample from a normal population with a mean 45.

Example 12: The lifetime of electric bulbs for a random sample of 10from a large consignment gave the following data:

Item	1	2	3	4	5	6	7	8	9	10
Life in‘000hrs’	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can we accept the hypothesis that the average life of bulb is 4000hrs?

Chi-square test for independences of Attributes

Introduction:

The square of a standard normal variate is known as Chi- square variate with 1 degrees of freedom (d. f).

Thus, if $X \sim N(\mu,\sigma^2)$, then $Z = \frac{x-\mu}{\sigma} \sim N(0,1)$ and $Z^2 = \left(\frac{x-\mu}{\sigma}\right)^2$ is a chi-square variate with 1 d. f.

In general if X_i (i = 1, 2, 3…., n) are n independent normal variates with mean μ_i and variances σ_i^2, (i = 1, 2, 3…., n), then

$\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$, is a chi-square variate with n d. f.

Test of Independence of Attributes – Contingency Tables:

Let us consider two attributes A and B, A divided into r classes $A_1, A_2 \ldots A_r$ and B divided into c classes $B_1, B_2 \ldots B_c$. Such a classification in which attributes are divided into more than two classes is known as manifold classification. The various cell frequencies can be expressed in the following table known as r x c contingency table where (A_i) is the number of persons possessing the attribute A_i (i = 1, 2, 3…., r), (B_j) is the number of persons possessing the attribute B_j (j = 1, 2, 3…., c), and $(A_i\ B_j)$ is the number of persons possessing both the attribute A_i and B_j. also $\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{c}(B_j) = N$, Where N is the total frequency.

r x c Contingency Tables

B / A	B_1,	B_2	…….	B_j	…….	B_c	Total
A_1	$(A_1\ B_1)$	$(A_1\ B_2)$	……	$(A_1\ B_j)$	……	$(A_1\ B_c)$	(A_1)
A_2	$(A_2\ B_1)$	$(A_2\ B_2)$	……	$(A_2\ B_j)$	……	$(A_2\ B_c)$	(A_2)
.	.	.	.	.			
A_i	$(A_i\ B_1)$	$(A_i\ B_2)$	……	$(A_i\ B_j)$	……	$(A_i\ B_c)$	(A_i)
.	.	.	.	.			
A_r	$(A_r\ B_1)$	$(A_r\ B_2)$	……	$(A_r\ B_j)$	……	$(A_r\ B_c)$	(A_r)
Total	(B_1)	(B_2)		(B_j)		(B_c)	N

The problem is to test if the two attributes A and B under consideration are independent or not.

Under the **null hypothesis that the attributes are independent, i.e. there is no association between two attributes A and B.**

The expected number of persons possessing both the attributes (A_i) and (B_j) are calculated by using the formula,

$$E(A_i\ B_j) = \frac{(A_i) \times (B_j)}{N}$$

Then **test statistic** under null hypothesis is,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where,

O_{ij} = Observed frequency for contingency table category in row i and column j.

E_{ij} = Expected frequency for contingency table category in row i and column j.

Which is distributed as Chi- square variate with (r -1) (c -1) degrees of freedom (d. f) at α level of significance.

Test criteria:

1. Critical value of chi-square with (r -1) (c -1) degrees of freedom (d. f.) at α level of significance written as, $\chi^2_{\alpha,(r-1)(c-1)}$

2. If the calculated chi-square is **less than** the critical value then **accept the null hypothesis** and we conclude that **there is no association between two attributes A and B. i.e. attributes A and B are independent.**

3. If the calculated chi-square is **greater than** the critical value then **reject the null hypothesis** and we conclude that **there is association between two attributes A and B. i.e. attributes A and B are dependent**.

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Example 1: Following results were collected by a doctor from his 500 patients.

		Preference		
		Liquid	Tablet	Injection
Sex	Males	60	90	100
	Females	110	75	65

Using chi-square test, can we say that the preference of the medicine depends on the sex of the person?

Solution:

(i) Null Hypothesis: Preference of the medicine **independent** on the sex of the person. (i. e. There is no association between preferences of the medicine and the sex of the person.)

(ii) Alternative Hypothesis: Preference of the medicine **dependent** on the sex of the person.

(iii) Level of Significance: $\alpha = 0.05$

(iv) Critical value: For (2-1) (3-1) =2 degrees of freedom the critical value of χ^2 at 5% level of significance from the table = 5.991 $i.e. \chi^2_{0.05,(2-1)(3-1)} = 5.991$

(v) Test Statistic: On the basis of null hypothesis the expected frequencies are as follows:

		Preference			Total
		Liquid	Tablet	Injection	
Sex	Males	60	90	100	250
	Females	110	75	65	250
Total		170	165	165	**500**

$$E(60) = \frac{170 \times 250}{500} = 85 \qquad E(90) = \frac{165 \times 250}{500} = 82.5 \qquad E(100) = \frac{165 \times 250}{500} = 82.5$$

$$E(110) = \frac{170 \times 250}{500} = 85 \qquad E(75) = \frac{165 \times 250}{500} = 82.5 \qquad E(65) = \frac{165 \times 250}{500} = 82.5$$

Calculation of test statistic

O	E	$\frac{(O-E)^2}{E}$
60	85	7.3529
90	82.5	0.6818
100	82.5	3.7121
110	85	7.3529
75	82.5	0.6818
65	82.5	3.7121
Total		23.4936

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 23.4936$$

(vi) Test Criteria: Since calculated chi-square is **greater than** the critical value hence, reject the null hypothesis.

(vii) Conclusion: We conclude that the preference of the medicine is **dependent** on the sex of the person.

Example 2: Two samples of polls of votes for two candidates A and B for a public office are taken, one from the residents of rural areas and one from urban areas. The results are given in the following table. Examine whether the nature of the area is related to voting preference in this election.

		Votes	
		A	B
Area	Rural	620	380
	Urban	550	450

Solution:

(i) Null Hypothesis: The nature of the area is **independent** of the voting preference in this election.

(ii) Alternative Hypothesis: The nature of the area is **dependent** of the voting preference in this election.

(iii) Level of Significance: $\alpha = 0.05$

(iv) Critical value: For (2-1) (2-1) =1 degrees of freedom the critical value of χ^2 at 5% level of significance from the table = 3.841 $i.e. \chi^2_{0.05,(2-1)(2-1)} = 3.841$

(v) Test Statistic: On the basis of null hypothesis the expected frequencies are as follows:

		Votes		
		A	B	Total
Area	Rural	620	380	1000
	Urban	550	450	1000
	Total	1170	830	**2000**

$$E(620) = \frac{1170 \times 1000}{2000} = 585 \qquad E(380) = \frac{830 \times 1000}{2000} = 415$$

$$E(550) = \frac{1170 \times 1000}{2000} = 585 \qquad E(450) = \frac{830 \times 1000}{2000} = 415$$

Calculation of test statistic

O	E	$\frac{(O-E)^2}{E}$
620	585	2.0940
380	415	2.9518
550	585	2.0940
450	415	2.9518
	Total	10.0916

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad 10.0916$$

(vi) Test Criteria: Since calculated chi-square is **greater than** the critical value hence, reject the null hypothesis.

(vii) Conclusion: We conclude that the nature of the area is **dependent** of the voting preference in this election.

Example 3: The following table gives number of good and bad parts:

	Good parts	Bad parts	Total
Day shift	960	40	1000
Evening shift	940	50	990
Night shift	950	45	995
Total	2850	135	2985

Test whether production of bad parts is independent of shift?

Solution:

(i) Null Hypothesis: The nature of the product and shift is **independent.**

(ii) Alternative Hypothesis: The nature of the product and shift is **dependent.**

(iii) Level of Significance: $\alpha = 0.05$

(iv) Critical value: For (3-1) (2-1) =2 degrees of freedom the critical value of χ^2 at 5% level of significance from the table = 5.991 $i.e. \chi^2_{0.05,(2-1)(3-1)} = 5.991$

(v) Test Statistic: On the basis of null hypothesis the expected frequencies are as follows:

$$E(960) = \frac{2850 \times 1000}{2985} = 954.7738 \qquad E(40) = \frac{135 \times 1000}{2985} = 45.2261$$

$$E(940) = \frac{2850 \times 990}{2985} = 945.2261 \qquad E(50) = \frac{135 \times 990}{2985} = 44.7738$$

$$E(950) = \frac{2850 \times 995}{2985} = 950 \qquad E(45) = \frac{135 \times 995}{2985} = 45$$

Calculation of test statistic

O	E	$\frac{(O-E)^2}{E}$
960	954.7738	0.0286
40	45.2261	0.6039
940	945.2261	0.0288
50	44.7738	0.6100
950	950	0
45	45	0
	Total	1.2713

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad 1.2713$$

(vi) Test Criteria: Since calculated chi-square is **less than** the critical value hence, accept the null hypothesis.

(vii) Conclusion: We conclude that the nature of the product and shift is **independent i.e.** production of bad parts is independent of shift.

Example 4: Out of 8000 graduates in a town 800 are females; out of 1600 graduate employees 120 are females. Use χ^2 to determine if any distinction is made in appointment on the basis of gender.

Solution: Observed Frequencies are,

		Status		
		Employed	Not Employed	Total
Sex	Male	1480	5720	7200
	Female	120	680	800
Total		1600	6400	**8000**

(i) Null Hypothesis: No distinction is made in appointment on the basis of gender.

(ii) Alternative Hypothesis: The distinction is made in appointment on the basis of gender.

(iii) Level of Significance: $\alpha = 0.05$

(iv) Critical value: For (2-1) (2-1) =1 degrees of freedom the critical value of χ^2 at 5% level of significance from the table = 3.841 $i.e. \chi^2_{0.05,(2-1)(2-1)} = 3.841$

(v) Test Statistic: On the basis of null hypothesis the expected frequencies are as follows:

		Status		
		Employed	Not Employed	Total
Sex	Male	1480	5720	7200
	Female	120	680	800
Total		1600	6400	**8000**

$$E(1480) = \frac{1600 \times 7200}{8000} = 1440 \qquad E(5720) = \frac{6400 \times 7200}{8000} = 5760$$

$$E(120) = \frac{1600 \times 800}{8000} = 160 \qquad E(680) = \frac{6400 \times 800}{8000} = 640$$

Calculation of test statistic

O	E	$\frac{(O-E)^2}{E}$
1480	1440	1.1111
5720	5760	0.2777
120	160	10
680	640	2.5
	Total	13.8888

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 13.8888$$

(vi) Test Criteria: Since calculated chi-square is **greater than** the critical value hence, reject the null hypothesis.

(vii) Conclusion: We conclude that the distinction is made in appointment on the basis of gender.

Examples for Practice

Example 1: Describe Chi-square test for testing independence of attributes.

Example 2: Test the hypothesis at 5% level of significance that the presence or absence of hypertension is independent on smoking habits from the following experimental data on 180 persons.

	Non Smokers	Moderate Smokers	Heavy Smokers
Hypertension	21	36	30
No Hypertension	48	26	19

Example 3: Test the independence of attributes using following data and write conclusion.

		Drugs		
		Liquid	Pills	Injection
Sex	Males	62	84	70
	Females	70	75	50

Example 4: In a survey of 200 boys of which 75 were intelligent, 40 had educated fathers, while 85 of the unintelligent boys had uneducated fathers. Do these figures support the hypothesis that educated fathers have intelligent boys?

Example 5: Medical Council Collected following data regarding preference of medicine.

	Male	Female	Total
Allopathic	55	65	120
Homeopathic	65	75	140
Ayurvedic	102	58	160
Total	222	198	420

Use Chi-square test and decide whether preference of medicine depends on sex of person.

Example 6: Following is the distribution of 500 people according to drug preference.

	Tablet	Liquid	Injection
Below 10	90	50	25
10-20	45	70	27
20 and above	23	60	110

Using Chi-square test check the independence of above attributes and write conclusion.

Example 7: Using following data examine whether eye colour of son and eye colour of father are independent of each other

		Eye colour of father	
		Black	Blue
Eye colour of son	Black	620	380
	Blue	550	450

Example 8: A drug manufacturing company collected following data regarding preference of drug. Use chi-square test and decide whether preference of drug depends on sex of the person.

	Male	Female
Tablet	55	48
Injection	62	55

Example 9: For the following data apply chi-square test and test of independence of age and type of food.

		Food		
		Chinese	Panjabi	South Indian
Age	Below 15	50	54	60
	15 - 35	70	65	45
	35 and above	35	65	55

Example 10: A certain drug was administered to 500 people out of a total of 800 included in the sample to test its efficiency against typhoid. The results are given below:

	Typhoid	No Typhoid
Drug	200	300
No Drug	280	20

On the basis of these data, can it be concluded that the drug is effective in preventing typhoid.

Example 11: Certain medical council collected following data regarding preference of medicine.

	Male	Female
Allopathic	55	65
Homeopathic	65	75
Ayurvedic	102	58

Use chi-square test and decide whether preference of medicine depends on sex of the person.

Example 12: Test independence of attributes from the following data of height father and height of son and comment on it.

		Height of father	
		Tall	Short
Height of son	Tall	145	75
	Short	60	120

Example 13: In an investigation to health and nutrition of two groups of children of different status, the following results were registered

		Health		
		Below Normal	Normal	Above Normal
Social	Poor	130	102	24
	Rich	20	108	96

Discuss the relation between the health and their social status.

****