CORRELATION

Upto now we have considered data in single variable, better known as **univariate population**. We have found the measures of central tendency, dispersion, skewness and kurtosis of the distribution. Now we shall study data on two variables for all individuals in a region better known as **bivariate population**. Here the question will arise, does there exist 'association' between the two variables? If the value of one variable increases (or decreases) does it affect the other ? If yes, then how much and in what direction? If two or more quantities vary in sympathy so that movement in one tends to be accompanied by the corresponding movements in the other(s) then they are said to be **correlated**.

In a bivariate distribution, if an increase (or decrease) in the values of one variable results in an increase (or decrease) in the other, the **correlation** is said to be *positive*, whereas, if an increase (or decrease) in one variable results in decrease (or increase) of the other, the correlation is said to be *negative*. If there is no relationship observed between the variables they are said to be *independent* or *uncorrelated*.

SCATTER DIAGRAM AND COEFFICIENT OF CORRELATION

If we have data as n pairs of values of two variables x and y, say, we plot these values on the graph taking one of the variables along the X-axis and the other along the Y-axis. Then the resulting diagram, consisting of n points on the graph, is called a **scatter diagram**. If $\bar{x}$ and $\bar{y}$ denote the means of x and y respectively, then the point $(\bar{x}, \bar{y})$ has a strategic position on the scatter diagram. Let as shift the origin to the point $(\bar{x}, \bar{y})$ then the point P (x, y) has new coordinates (X, Y) where $X = x - \bar{x}$, $Y = y - \bar{y}$.

The points (X, Y) will be so distributed over all the four quadrants of the XY-plane that the product XY will be positive in the first and the third quadrants and negative in the second and the fourth quadrants.

Now ΣXY will determine the trend of the dots in the scatter diagram as described below :

(*i*) If ΣXY is positive then the trend of the dots will be through the first and the third quadrants.

(*ii*) If ΣXY is negative then the trend of the dots will be through the second and the fourth quadrants.

(*iii*) If ΣXY is zero there is no trend of the dots.

As such $\frac{1}{n}\Sigma XY$ can be taken as the measure of correlation. However, if we non-dimensionalize X and Y by dividing these by σ_x and σ_y respectively, we finally get the numerical measure of correlation between x and y as

$$\frac{1}{n}\Sigma\frac{X}{\sigma_x}\frac{Y}{\sigma_y} = \frac{\Sigma XY}{n\sigma_x\sigma_y} = \rho.$$

This is called Karl Pearson's **coefficient of correlation, denoted by ρ or by r, given by**

$$\rho \text{ (or } r) = \frac{1}{n}\frac{\Sigma XY}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\Sigma XY}{\sqrt{\frac{1}{n}\Sigma X^2}\sqrt{\frac{1}{n}\Sigma Y^2}} = \frac{\Sigma XY}{\sqrt{(\Sigma X^2)(\Sigma Y^2)}}$$

$$\rho = \frac{1}{n}\frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$$

Substituting for X and Y in terms of x and y and simplifying, we can write

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

This may become very lengthy and uncomfortable if the means of x and y are not integers. In such cases, we use the step deviation process and write

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{\left\{n\Sigma d_x^2 - (\Sigma d_x)^2\right\}\left\{n\Sigma d_y^2 - (\Sigma d_y)^2\right\}}}$$

where $d_x = \frac{x - A}{h}$, $d_y = \frac{y - B}{k}$, and A and B are assumed means of x and y respectively and h and k are the uniform class intervals of x and y respectively.

EXAMPLE 32.14. The table below gives the marks of 8 students in a test in two subjects A and B.

Student	1	2	3	4	5	6	7	8
Marks in subject A(x)	10	6	9	10	12	13	11	9
Marks in subject B(y)	9	4	6	9	11	13	8	4

Calculate the coefficient of correlation between x and y.

SOLUTION: Here the means of x and y are integers and hence we can apply the direct method to compute ρ.

Student	x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
1	10	9	0	1	0	1	0
2	6	4	− 4	− 4	16	16	16
3	9	6	− 1	− 2	1	4	2
4	10	9	0	1	0	1	0
5	12	11	2	3	4	9	6
6	13	13	3	5	9	25	15
7	11	8	1	0	1	0	0
8	9	4	− 1	− 4	1	16	4
N = 8	$\Sigma x = 80$	$\Sigma y = 64$	$\Sigma X = 0$	$\Sigma Y = 0$	$\Sigma X^2 = 32$	$\Sigma Y^2 = 72$	$\Sigma XY = 43$

$$\therefore \quad \bar{x} = \frac{\Sigma x}{N} = 10, \quad \bar{y} = \frac{\Sigma y}{N} = 8, \quad \rho = \frac{\Sigma XY}{\sqrt{\Sigma X^2}\sqrt{Y^2}} = \frac{43}{\sqrt{32 \times 72}} = 0.896. \quad \textbf{Ans.}$$

EXAMPLE 32.15. The ages of husbands and their wives are given in the following table:

x (age of husband)	23	27	28	29	30
y (age of wife)	18	22	23	24	25

Calculate the coefficient of correlation between x and y from the above table.

SOLUTION: Let us adopt the step deviation method and take $d_x = x - 28$, $dy = y - 23$.

x	y	$d_x = x - 28$	$d_y = y - 23$	d_x^2	d_y^2	$d_x d_y$
23	18	−5	−5	25	25	25
27	22	−1	−1	1	1	1
28	23	0	0	0	0	0
29	24	1	1	1	1	1
30	25	2	2	4	4	4
Total		−3	−3	31	31	31

$$\therefore \quad \rho = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{\left[n\Sigma d_x^2 - (\Sigma d_x)^2\right]\left[n\Sigma d_y^2 - (\Sigma d_y)^2\right]}} = \frac{5(31) - (-3)(-3)}{\sqrt{[5\times 31 - 9][5\times 31 - 9]}} = 1. \quad \textbf{Ans.}$$

This means the ages of husbands and wives are perfectly correlated.

EXAMPLE 32.16. Calculate Karl Pearson's coefficient of correlation between per capita national income and per capita consumer expenditure as per the following data

Year	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Per capita N.I.	249	251	248	252	258	269	271	272	280	275
Per capita C.E.	237	238	236	240	245	255	254	252	258	251

SOLUTION: Let x and y denote the per capita national income and per capita consumer expenditure respectively. Consider the tabular data:

Year	x	y	$d_x = x - 260$	$d_y = y - 240$	d_x^2	d_y^2	$d_x d_y$
1974	249	237	− 11	− 3	121	9	33
1975	251	238	− 9	− 2	81	4	18
1976	248	236	− 12	− 4	144	16	48
1977	252	240	− 8	0	64	0	0
1978	258	245	− 2	5	4	25	− 10
1979	269	255	9	15	81	225	135
1980	271	251	11	14	121	196	154
1981	272	252	12	12	144	144	144
1982	280	258	20	18	400	324	360
1983	275	251	15	11	225	121	165
N = 10			25	66	1385	1064	1047

$$\rho = \frac{\Sigma d_x d_y - \frac{1}{N}\Sigma d_x \Sigma d_y}{\sqrt{\left[\left\{\Sigma d_x^2 - \frac{1}{N}(\Sigma d_x)^2\right\}\left\{\Sigma d_y^2 - \frac{1}{N}(\Sigma d_y)^2\right\}\right]}}$$

$$= \frac{1047 - \frac{1}{10}(25)(66)}{\sqrt{\left[\left\{1385 - \frac{1}{10}(25)^2\right\}\left\{1064 - \frac{1}{10}(66)^2\right\}\right]}}$$

$$= \frac{1047 - 165}{\sqrt{(1385 - 62.5)(1064 - 435.6)}}$$

$$= \frac{882}{\sqrt{1322.5 \times 628.4}} = 0.967. \quad \textbf{Ans.}$$

BIVARIATE FREQUENCY DISTRIBUTION

EXAMPLE 32.17. Following table gives a bivariate distribution showing frequency of marks according to age by a group of 100 students in an intelligence test.

Marks	Age				
	18	19	20	21	Total
10-20	4	2	2	—	8
20-30	5	4	6	4	19
30-40	6	8	10	11	35
40-50	4	4	6	8	22
50-60		2	4	4	10
60-70		2	3	1	6
Total	19	22	31	28	100

Find the coefficient of correlation between age and marks.

SOLUTION: We give below the correlation table

$v = \frac{y-35}{10}$	x / y	$u = x - 19$ Marks x	−1 18	0 19	1 20	2 21	Total f	fv	fv^2	fuv
−2	15	10-20	4 [8]	2 [0]	2 [−4]	— [1]	8	−16	32	4
−1	25	20-30	5 [5]	4 [0]	6 [−6]	4 [−8]	19	−19	19	−9
0	35	30-40	6 [0]	8 [0]	10 [0]	11 [0]	35	0	0	0
1	45	40-50	4 [−4]	4 [0]	6 [6]	8 [16]	22	22	22	18
2	55	50-60	—	2 [0]	4 [8]	4 [16]	10	20	40	24
3	65	60-70	—	2 [0]	3 [9]	1 [6]	6	18	54	15
		Total	19	22	31	28	100	25	167	52
		fu	−19	0	31	56	68			
		fu^2	19	0	31	112	162			
		fuv	9	0	13	30	52			

Here $u = x - 19$ and $v = \frac{y - 35}{10}$

$$\therefore \quad \bar{x} = 19 + \frac{1}{N}\Sigma fu = 19 + \frac{68}{100} = 19.68$$

$$\bar{y} = 35 + \frac{1}{N}\Sigma fv = 35 + \frac{25}{100} = 35.25$$

$$\therefore \quad r = \frac{\frac{1}{N}\Sigma fuv - \frac{\Sigma fu \Sigma fv}{N^2}}{\sqrt{\left\{\frac{1}{N}\Sigma fu^2 - \left(\frac{\Sigma fu}{N}\right)^2\right\}\left\{\frac{1}{N}\Sigma fv^2 - \left(\frac{\Sigma fv}{N}\right)^2\right\}}}$$

$$= \frac{\frac{1}{100} \times 52 - \frac{68}{100} \cdot \frac{25}{100}}{\sqrt{\left\{\frac{162}{100} - \left(\frac{68}{100}\right)^2\right\}\left\{\frac{167}{100} - \left(\frac{25}{100}\right)^2\right\}}} = \frac{0.52 - 0.68 \times 0.25}{\sqrt{(1.62 - 0.68^2)(1.67 - 0.25^2)}}$$

$$= \frac{0.35}{\sqrt{1.1576 \times 1.6075}} = 0.25. \quad \textbf{Ans.}$$

The steps followed above, are

(*i*) Take the step deviations of the variable x and denote these by d_x (or u).

(*ii*) Take the step deviations of the variable y and denote these by d_y (or v).

(*iii*) Multiply d_x, d_y (or u, v) and the respective frequency of each cell and write the figure obtained in the right hand upper corner of the cell.

(*iv*) Add together all the cornered values as calculated in step (*iii*) and obtain the total $\Sigma f d_x d_y$ (or $\Sigma f u v$).

(*v*) Multiply the frequencies of the variable x by the deviations of x and obtain the total $\Sigma f d_x$ (or $\Sigma f u$).

(*vi*) Take the square of the deviations of the variable x and multiply them by the respective frequencies and obtain $\Sigma f d_x^2$ (or $\Sigma f u^2$).

(*vii*) Multiply the frequencies of the variable y by the deviations of y and obtain the total $\Sigma f d_y$ (or $\Sigma f v$).

(*viii*) Take the square of the deviations of the variable y and multiply them by the respective frequencies and obtain $\Sigma f d_y^2$ (or $\Sigma f v^2$).

(*ix*) Substitute the values of $\Sigma f d_x d_y$ (= $\Sigma f u v$), $\Sigma f d_x$ (or $\Sigma f u$), $\Sigma f d_y$ (or $\Sigma f v$), $\Sigma f d_x^2$ (or $\Sigma f u^2$) and $\Sigma f d_y^2$ (or $\Sigma f v^2$) in the formula for r and get the value of r.

Properties About the Correlation Coefficient

I. The coefficient of correlation r (or ρ) lies between -1 and 1, that is, $|r| < 1$.

$$\Sigma\left(\frac{x-\bar{x}}{\sigma_x} + \frac{y-\bar{y}}{\sigma_y}\right)^2 = \Sigma\left[\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} + \frac{2(x-\bar{x})(y-\bar{y})}{\sigma_x\,\sigma_y}\right]$$

$$= \frac{1}{\sigma_x^2}\Sigma(x-\bar{x})^2 + \frac{1}{\sigma_x^2}\Sigma(y-\bar{y})^2 + \frac{2}{\sigma_x\,\sigma_y}\Sigma(x-\bar{x})(y-\bar{y})$$

$$= N + N + 2N\rho = 2N(1+\rho).$$

Since $\Sigma\left(\frac{x-\bar{x}}{\sigma_x} + \frac{y-\bar{y}}{\sigma_y}\right)^2$ is positive, being the sum of the squares, we have $2N(1+\rho) \geq 0$ hence $\rho \geq -1$.

Next, on similar lines, we can have

$$\Sigma\left(\frac{x-\bar{x}}{\sigma_x} - \frac{y-\bar{y}}{\sigma_y}\right)^2 = 2N(1-\rho)$$

which is again positive hence $2N(1-\rho) \geq 0$ or $\rho \leq 1$.

Hence $-1 \leq \rho \leq 1$.

II. The coefficient of correlation is independent of the change of scale and change of origin of the variables x and y.

By change of origin we mean subtracting some constant from every given value of x and y and by change of scale we mean multiplying or dividing every value of x and y by some constant. We know that the coefficient of co-relation is given by

$$r_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}} \quad \ldots(1)$$

Let us consider the changes $u = \frac{x-a}{h}$ and $v = \frac{y-b}{k}$

hence $\bar{u} = \frac{\bar{x}-a}{h}$ and $\bar{v} = \frac{\bar{y}-b}{k}$.

It follows that $u - \bar{u} = \frac{x-\bar{x}}{h}$ and $v - \bar{v} = \frac{y-\bar{y}}{k}$ $\ldots(2)$

Therefore, substituting for $(x-\bar{x})$ and $(y-\bar{y})$ from (2) in (1), we get

$$r_{xy} = \frac{hk\,\Sigma(u-\bar{u})(v-\bar{v})}{\sqrt{h^2\,\Sigma(u-\bar{u})^2\,k^2\Sigma(v-\bar{v})^2}} = \frac{\Sigma(u-\bar{u})(v-\bar{v})}{\sqrt{\Sigma(u-\bar{u})^2\,\Sigma(v-\bar{v})^2}} = r$$

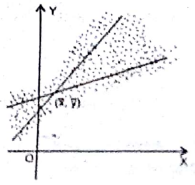
which shows that the coefficient of correlation is independent of change of scale and change of origin.

REGRESSION

If two variables are co-related (closely or otherwise), one may be interested in estimating (or predicting) the value of one variable, given the value of the other. For example, if we have established that expenditure on advertisement and amount of sales are correlated, we can find out the expected amount of expenditure on advertising for attaining a given amount of sales. Thus the regression analysis is a statistical device with the help of which we are in a position to estimate (or predict) the unknown values of one variable from the known values of another variable.

SCATTER DIAGRAM AND LINES OF REGRESSION

The dots of the scatter diagram generally tend to cluster about a well defined direction which implies that there should exist 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**. (See the adjoining figure). Actually, **there will be two such lines,** not one, as shown in the figure. One line will give the most probable values of y for each specified values of x and is called the **line of regression of y on x.**



while the other line gives the most probable values of x for each specified values of y and this second line is called the **line of regression of x on y.**

We first obtain the "line of regression of y on x." Let the straight line, satisfying the general trend of x dots in the scatter diagram, have the equation $y = a + bx$ ...(1)

The unknown constants a and b will be determined by using the principle of least squares which has been discussed earlier while dealing with the curve of best fit. Under that principle the normal equations for the unknowns a and b, are

$$\Sigma y = na + b\Sigma x \quad ...(2)$$

and $$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad ...(3)$$

Dividing (2) by n throughout, we get $\bar{y} = a + b\bar{x}$ ...(4)

where $\bar{x}$ and $\bar{y}$ are means of x and y respectively.

This shows that the point $(\bar{x}, \bar{y})$ lies on (1).

Subtracting (4) from (1), we get $y - \bar{y} = b(x - \bar{x})$. ...(5)

Multiplying (5) throughout by $(x - \bar{x})$ and taking summation for all points, we get

$$\Sigma[(x-\bar{x})(y-\bar{y})] = b\Sigma(x-\bar{x})^2.$$

$$\therefore \text{ slope of line (5), } b = \frac{\Sigma[(x-\bar{x})(y-\bar{y})]}{\Sigma(x-\bar{x})^2} = \frac{\Sigma XY}{\Sigma X^2}, \quad \text{where } X = x - \bar{x},\ Y = y - \bar{y},$$

$$= \frac{\Sigma XY}{n\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} \text{ since } \sigma_x^2 = \frac{1}{n}\Sigma X^2,\ \sigma_y^2 = \frac{1}{n}\Sigma Y^2 \text{ and } r = \frac{1}{n}\frac{\Sigma XY}{\sigma_x\sigma_y}.$$

Thus, (5) becomes $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$ which is the line of regression of y on x.

Similarly, the line of regression of x on y can be derived as $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$.

and the slope of line of regression of x on $y = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$.

Next, if θ is the angle between the two lines of regression, then

$$\tan\theta = \frac{\left|r\frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}\right|}{1 + \frac{\sigma_y^2}{\sigma x^2}} = \frac{\frac{\sigma_y}{\sigma_x}\left|r - \frac{1}{r}\right|\sigma_x^2}{\sigma_x^2 + \sigma_y^2} = \frac{\sigma_x\sigma_y(1-r^2)}{r(\sigma_x^2 + \sigma_y^2)}.$$

or $$\tan\theta = \left(\frac{1-r^2}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}.$$ [GGSIPU IV Sem II Term 2015]

When $r = 1$ or -1, $\tan\theta = 0$ and lines of regression are parallel, rather coincident. and when $r = 0$, $\tan\theta = \infty$ and lines of regression are perpendicular to each other.

Thus, the **line of best fit to predict the values of y for each specified values of x has the equation**

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad ...(6)$$

which is the required **line of regression of y on x.** Its slope is called the ***coefficient of regression of y on x* and is denoted by b_{yx}.**

$$\therefore \quad b_{yx} = r\frac{\sigma_y}{\sigma_x}, \quad ...(7)$$

Simply interchanging x and y in (6), we get the equation of **the line of regression of x on y** as $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$ ...(8)

so that **the coefficient of regression of x on y, denoted by b_{xy}** is given by

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}, \quad ...(9)$$

Further, $b_{yx} \cdot b_{xy} = r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y} = r^2$,

which implies that **r is the geometric mean between b_{yx} and b_{xy}**, that is, b_{yx}, r and b_{xy} are in geometric progression.

Also, obviously b_{yx} and b_{xy} are of same sign and since r^2 is less than one, one regression coefficient has to be less than one and the other greater than one. Further the sign of r has to be the same as that of the regression coefficients.

Next, let us investigate if the regression coefficients are independent of change of origin and change of scale or not.

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \quad \text{and} \quad b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

We have already shown that r, the correlation coefficient is independent of change of origin as well as change of scale.

Since $\sigma_x^2 = \frac{1}{n}\sum(x-\bar{x})^2$ and $\sigma_y^2 = \frac{1}{n}\sum(y-\bar{y})^2$

Changing the origin and scale, we write

$$u = \frac{x-a}{h}, \quad v = \frac{y-b}{k}, \text{ i.e., } x = a + hu \text{ and } y = b + kv$$

$$\Rightarrow \quad \bar{x} = a + h\bar{u} \text{ and } \bar{y} = b + k\bar{v}$$

$$\therefore \quad x - \bar{x} = h(u - \bar{u}), \quad y - \bar{y} = k(v - \bar{v}).$$

and $$\sigma_x^2 = \frac{h^2}{n}\sum(u-\bar{u})^2 \text{ and } \sigma_y^2 = \frac{k^2}{n}\sum(v-\bar{v})^2$$

$$\therefore \quad \frac{\sigma_y}{\sigma_x} = \frac{k}{h}\sqrt{\frac{\sum(v-\bar{v})^2}{\sum(u-\bar{u})^2}}$$

which means that regression coefficients are independent of change of origin but not independent of change of scale.

EXAMPLE 32.18. (*a*) From the following data, obtain the two lines of regression

x	6	2	10	4	8
y	9	11	5	8	7

(*b*) From the data given below, find the equations of lines of regression of x on y and of y on x. Also calculate the correlation coefficient.

x	2	4	6	8	10
y	5	7	9	8	11

(*c*) If the coefficient of correlation between two variables x and y is 0.5 and acute angle between their lines of regression is $\tan^{-1}(3/8)$, show that

$$\sigma_x = \frac{\sigma_y}{2+\sqrt{3}}.$$

SOLUTION: (*a*) $d_x = x - A_x$, $d_y = y - A_y$ where A_x, A_y are assumed means for x and y.

x	y	$d_x = x-5$	$d_y = y-7$	d_x^2	d_y^2	$d_x d_y$
6	9	1	2	1	4	2
2	11	−3	4	9	16	−12
10	5	5	−2	25	4	−10
4	8	−1	1	1	1	−1
8	7	3	0	9	0	0
$\Sigma x = 30$	$\Sigma y = 40$	$\Sigma d_x = 5$	$\Sigma d_y = 5$	$\Sigma d_x^2 = 45$	$\Sigma d_y^2 = 25$	$\Sigma d_x d_y = -21$

The line of regression of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$.

Here $\bar{x} = \frac{1}{n}\Sigma x = \frac{30}{5} = 6$ and $\bar{y} = \frac{1}{n}\Sigma y = \frac{40}{5} = 8$.

$$b_{xy} = \frac{\Sigma d_x d_y - \frac{1}{N}\Sigma d_x \Sigma d_y}{\Sigma d_x^2 - \frac{1}{N}(\Sigma d_x)^2} = \frac{-21 - \frac{1}{5}(5)(5)}{25 - \frac{1}{5}(5)^2} = \frac{-26}{20} = -1.3.$$

$\therefore$ $x - 6 = -1.3(y - 8)$ or $x = -1.3y + 16.4$.

Next, the line of regression of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

or $y - 8 = b_{yx}(x - 6)$.

Here $b_{yx} = \frac{\Sigma d_x d_y - \frac{1}{N}\Sigma d_x d_y}{\Sigma d_x^2 - \frac{1}{N}(\Sigma d_x)^2} = \frac{-21 - \frac{1}{5}(5)(5)}{45 - \frac{1}{5}(5)^2} = \frac{-26}{40} = -0.65$

$\therefore$ regression lines are $x - 6 = -1.3(y - 8)$ and $y - 8 = -0.65(x - 6)$

or $y = -0.65x + 11.9$ and $x = -1.3y + 16.4$. **Ans.**

(*b*) Let $A_x = 5$ and $A_y = 7$

x	y	$d_x = x-5$	$d_y = y-7$	d_x^2	d_y^2	$d_x d_y$
2	5	−3	−2	9	4	6
4	7	−1	0	1	0	0
6	9	1	2	1	4	2
8	8	3	1	9	1	3
10	11	5	4	25	16	20
$\Sigma x = 30$	$\Sigma y = 40$	$\Sigma d_x = 5$	$\Sigma d_y = 5$	$\Sigma d_x^2 = 45$	$\Sigma d_y^2 = 25$	$\Sigma d_x d_y = 31$

Now $b_{xy} = \dfrac{\Sigma d_x d_y - \frac{1}{N}\Sigma d_x \Sigma d_y}{\Sigma d_y^2 - \frac{1}{N}(\Sigma d_y)^2} = \dfrac{31 - \frac{1}{5}\cdot 5\cdot 5}{25 - \frac{1}{5}(5)^2} = 1.3$

Here $\bar{x} = \frac{\Sigma x}{N} = \frac{30}{5} = 6$, $\bar{y} = \frac{\Sigma y}{N} = \frac{40}{5} = 8$.

$\therefore$ The line of regression of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$ or $x = (1.3)(y - 8) + 6$

Next, $$b_{yx} = \frac{\Sigma d_x d_y - \frac{1}{N}\Sigma d_x \Sigma d_y}{\Sigma d_x^2 - \frac{1}{N}(\Sigma d_x)^2} = \frac{31 - \frac{1}{5}\cdot 5.5}{45 - \frac{1}{5}5.5} = \frac{26}{40} = 0.65$$

$\therefore$ The line of regression of y on x is $y - \bar{y} = b_{xy}(x - \bar{x})$ or $y = 0.65(x - 6) + 8$

Now, the correlation coefficient

$$r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{(0.65)(1.3)} = 0.919 \quad \textbf{Ans.}$$

(*c*) Given coefficient of correlation = 0.5 and angle between the lines of regression $= \tan^{-1}\left(\frac{3}{8}\right)$.

We know that $$\tan\theta = \frac{\sigma_x \sigma_y (1 - r^2)}{r(\sigma_x^2 + \sigma_y^2)} \quad \text{or} \quad \frac{3}{8} = \frac{\sigma_x \sigma_y [1 - (0.5)^2]}{0.5[\sigma_x^2 + \sigma_y^2]}$$

or $$\sigma_x^2 + \sigma_y^2 = \frac{8}{3}\left[\frac{1 - (0.5)^2}{0.5}\right]\sigma_x \sigma_y$$

or $$\sigma_x^2 + \sigma_y^2 - 4\sigma_x\sigma_y = 0 \quad \text{or} \quad \left(\frac{\sigma_x}{\sigma_y}\right)^2 - 4\left(\frac{\sigma_x}{\sigma_y}\right) + 1 = 0.$$

$\therefore$ $$\sigma_x = \left(\frac{1}{2 \pm \sqrt{3}}\right)\sigma_y$$ **Hence Proved.**

EXAMPLE 32.19. Given the bivariate data

x	1	5	3	2	1	1	7	3
y	6	1	0	0	1	2	1	5

(*i*) fit a regression line of y on x and predict y when $x = 5$

(*ii*) fit a regression line of x on y and predict x when $y = 2.5$

(*iii*) calculate Karl Pearson's coefficient of correlation.

SOLUTION :

x	y	$d_x = x - 3$	$d_y = y - 2$	d_x^2	d_y^2	$d_x d_y$
1	6	− 2	4	4	16	− 8
5	1	2	− 1	4	1	− 2
3	0	0	− 2	0	4	0
2	0	− 1	− 2	1	4	2
1	1	− 2	− 1	4	1	2
1	2	− 2	0	4	0	0
7	1	4	− 1	16	1	− 4
3	5	0	3	0	9	0
Total		$\Sigma d_x = -1$	$\Sigma d_y = 0$	$\Sigma d_x^2 = 33$	$\Sigma d_y^2 = 36$	$\Sigma d_x d_y = -10$

Here $\bar{x} = 3 + \frac{1}{8}\Sigma d_x = 3 - \frac{1}{8} = 2.875$ and $\bar{y} = 2 + \frac{1}{8}(0) = 2$

(*i*) Next, for the line of regression of y on x

$$b_{yx} = \frac{\Sigma d_x d_y - \frac{1}{N}\Sigma d_x d_y}{\Sigma d_x^2 - \frac{1}{N}(\Sigma d_x)^2} = \frac{-10 - \frac{1}{8}(-1)(0)}{33 - \frac{1}{8}(-1)^2} = -0.304.$$

Equation of regression line of y on x, is $y - \bar{y} = b_{yx}(x - \bar{x})$

or $y - 2 = -0.304\,(x - 2.875)$ or $y = -0.304\,x + 2.874$.

$\therefore$ at $x = 5$ the predicted value of $y = 2.874 - 5 \times 0.304 = 1.354$. **Ans.**

(*ii*) For the line of regression of x on y

$$b_{xy} = \frac{\Sigma d_x d_y - \frac{1}{N}\Sigma d_x \Sigma d_y}{\Sigma d_y^2 - \frac{1}{N}(\Sigma d_y)^2} = \frac{-10 - \frac{1}{8}(-1)(0)}{36 - \frac{1}{8}(0)} = -0.278.$$

$\therefore$ Equation of regression line of x on y, is $x - \bar{x} = b_{xy}(y - \bar{y})$

or $x - 2.875 = -0.278\,(y - 2)$ or $x = -0.278\,y + 3.431$.

$\therefore$ At $y = 25$ the predicted value of $x = 3.431 - 0.278\,(2.5) = 2.736$. **Ans.**

(*iii*) $\rho^2 = b_{xy}\, b_{yx} = (-0.278)(-0.304) = 0.845$ $\therefore$ $\rho = 0.291$ **Ans.**

EXAMPLE 32.20. (a) *From the following data*

x	23	27	28	28	29	30	31	33	35	36
y	18	20	22	27	21	29	27	29	28	29

Estimate y when $x = 32$ by using suitable line of regression.

(b) *In a partially destroyed laboratory record for analysing correlation data, following results only are left. Variance $x = 9$,*

Regression lines are $8x - 10y + 66 = 0$ and $40x - 18y = 214$.

Determine from the above information

(*i*) *mean values of x and y*

(*ii*) *coefficient of correlation between x and y*

(*iii*) *standard deviation of y.*

SOLUTION: (a) Here let $A = 30$ and $B = 27$ so $u = x - 30$ and $v = y - 27$.

x	y	u	v	u^2	v^2	uv
23	18	−7	−9	49	81	63
27	20	−3	−7	9	49	21
28	22	−2	−5	4	25	10
28	27	−2	0	4	0	0
29	21	−1	−6	1	36	6
30	29	0	2	0	4	0
31	27	1	0	1	0	0
33	29	3	2	9	4	6
35	28	5	1	25	1	5
36	29	6	2	36	4	12
		$\Sigma u = 0$	$\Sigma v = -7$	$\Sigma u^2 = 138$	$\Sigma v^2 = 204$	$123 = \Sigma uv$

$\therefore$ $\bar{x} = 30$, $\bar{y} = 27 + \left(\frac{-20}{10}\right) = 25$

Line of regression of y on x, is

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

$$b_{yx} = \frac{\Sigma uv - \frac{1}{n}\Sigma u \Sigma v}{\Sigma u^2 - \frac{1}{n}(\Sigma u)^2} = \frac{123 - \frac{1}{10}(0)(-20)}{138 - \frac{1}{10}(0)} = \frac{123}{138}$$

$\therefore$ Line of regression of y on x becomes

$$y - 25 = \frac{123}{138}(x - 30)$$

At $x = 32$ the estimated value of $y = 25 + 2\left(\frac{123}{138}\right) = 26.78$ **Ans.**

(b) (*i*) Since the lines of regression intersect at $(\bar{x}, \bar{y})$, we solve for x and y, the given equations $8x - 10y + 66 = 0$ ...(1)

and $40x - 18y = 214$ ...(2)

to get $x = 13$, $y = 17$ Thus, $\bar{x} = 13$ and $\bar{y} = 17$.

(*ii*) To find the correlation coefficient we have to find the regression coefficients. As we do not know which line is regression line of y and x and which one is of x on y.

Assume that (1) is line of regression of x on y, then we have

$$x = \frac{-66}{8} + \frac{10}{8}y \quad \therefore \quad b_{xy} = \frac{10}{8} = 1.25.$$

From equation (2), we can write $y = \frac{-214}{18} + \frac{40}{18}x$ hence $b_{yx} = \frac{40}{18} = 2.22.$

Since both the regression coefficients are greater than one, hence the assumption is wrong. Therefore equation (1) is line of regression of y on x. From equation (1), we have

$$y = \frac{8}{10}x + \frac{66}{10} \quad \text{hence} \quad b_{yx} = \frac{8}{10}.$$

In turn, equation (2) represents line of regression of x on y hence $b_{xy} = \frac{18}{40}$.

Next, $r^2 = b_{xy}\, b_{yx} = \frac{8}{10} \times \frac{18}{40} = 0.36 \quad \Rightarrow \quad r = 0.6$

(*iii*) Also we know that $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ hence $0.8 = 0.6\frac{\sigma_y}{\sigma_x}$

$\Rightarrow \quad 4\sigma_x = 3\sigma_y$. We are given that $\sigma_x^2 = 9$ hence $\sigma_x = 3$

$\therefore \quad \sigma_y = \frac{4}{3}\sigma_x = 4.$ **Ans.**

RANK CORRELATION

This measure introduced by Spearman is especially useful when quantitative measures for certain factors, like evaluation of leadership ability or the judgement in beauty contest, cannot be fixed, but the individuals in the group can be arranged in order, thereby obtaining the rank for each in the group. Spearman's correlation coefficient, denoted by r, is defined as

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

where d represents the difference of ranks between paired items in two series.

EXAMPLE 32.21. Two ladies Deepti and Nancy, were asked to rank lipsticks from 7 known companies. The ranks given by them are as follows

Lipsticks companies	A	B	C	D	E	F	G
Rank X by Deepti	2	1	4	3	5	7	6
Rank Y by Nancy	1	3	2	4	5	6	7

Compute the Spearman's rank correlation coefficient.

SOLUTION :

	X R_1	Y R_2	D $= \lvert R_1 - R_2 \rvert$	D^2
A	2	1	1	1
B	1	3	− 2	4
C	4	2	2	4
D	3	4	− 1	1
E	5	5	0	0
F	7	6	1	1
G	6	7	− 1	1
			Total	$\Sigma D^2 = 12$

$$r = \frac{6\Sigma D^2}{N(N^2 - 1)} = \frac{6 \times 12}{7(49 - 1)} = \frac{3}{14}. \quad \textbf{Ans.}$$

statistics. Quite often, the parameter values are not known and the statistics obtained from the samples are used for analysing the population. Note that the statistics obtained from different samples can vary from one sample to another sample.

One vital problem of sampling theory is to find out whether these variations in the statistic of the samples (which may be due to fluctuations in the sampling processes) are significant or insignificant.

Sampling Distribution of a Statistic:

Consider a population of size N and let r samples be drawn, through random sampling, each of size n. Now we compute some statistic t, say the mean $\bar{x}$ or variance s^2 for each of the r samples. The values of the statistic t can be put in the frequency table. This data of values of t is called a **sampling distribution**. The standard deviation of this sampling distribution of this statistic is called the **standard error** of the statistic.

We are considering here the case of large samples. The sample is considered as **large** if the size of the sample $n \geq 30$. If the sample is large, following assumptions always hold good.

(*i*) The sampling distribution of a statistic is normal. Note that the distribution of the population may or may not be normal.

(*ii*) The sampling statistics can be taken as corresponding population parameters if they are not known.

CENTRAL LIMIT THEOREM:

If random samples of size n are drawn from a non-normal population with mean μ and standard deviation σ, n being large, then the sampling distribution of the sample mean $\bar{x}$ is normally distributed with mean μ and standard error = $\sigma/\sqrt{n}$ approximately. However, if the population itself is normal then the sampling distribution of $\bar{x}$ will be obviously normal.

EXAMPLE 33.1. Let a population consist of 5 numbers 3, 5, 7, 9, 11. If random samples each of size n (= 2) are selected without replacement then find the sampling distribution of the sample mean $\bar{x}$.

SOLUTION: There can be 10 equally likely random samples of size $n = 2$. The values of $\bar{x}$ are given in the following table.

Sample	*Sample units*	*Sample mean* $\bar{x}$
1	3, 5	4
2	3, 7	5
3	3, 9	6
4	3, 11	7
5	5, 7	6
6	5, 9	7
7	5, 11	8
8	7, 9	8
9	7, 11	9
10	9, 11	10

Therefore the sampling distribution of the sample $\bar{x}$, is

$\bar{x}$	4	5	6	7	8	9	10
$f(\bar{x})$	1	1	2	2	2	1	1

The population mean $\mu = \frac{3+5+7+9+11}{5} = 7$

and the population variance $\sigma^2 = \frac{1}{5}[(-4)^2 + (-2)^2 + 0 + 2^2 + 4^2] = 8.$

Now the mean of 'sample means'

$$= \frac{1}{10}(4 + 5 + 6 + 7 + 6 + 7 + 8 + 8 + 9 + 10) = 7.0$$

and the variance of "sample mean"

$$= (-3)^2 (0.1) + (-2)^2 (0.1) + (-1)^2 (0.2) + 0 + (1)^2 (0.2) + (2)^2 (0.1) + (3)^2 (0.1)$$
$$= 0.9 + 0.4 + 0.2 + 0.2 + 0.4 + 0.9 = 3.0.$$

Observe here that the mean of the 'sample means' is the same as the population mean but the variance of the sample mean is not the same as the population variance. The standard deviation of the sampling distribution of a statistic is called its **standard error** (S.E.).

In the present case the S.E. = $\sqrt{3}$.

TESTS OF SIGNIFICANCE FOR LARGE SAMPLES

In sampling theory it is very important to make decision about the parameter value. The tests of hypothesis enable us to decide whether the deviation between the observed and the theoretical value is significant or might be attributed to fluctuations of sampling. Since n is large the sampling distribution of the statistic under study, can be approximated to normal hence for large sample testing, normal distribution is applied.

Null and Alternative Hypotheses.

Given a population we want to have information about a characteristic of the population. We start with the assumption that there is no significant difference between the sample statistic and the corresponding population parameter or between two sample statistics. This assumption that there is no significan difference is called **a null hypothesis** and is denoted by H_0. A hypothesis that is different from th null hypothesis is called an **alternate hypothesis** and is denoted by H_1. There are **tests of hypothesi** which decide whether to accept or reject a null hypothesis or an **alternate hypothesis**.

For example, let the null hypothesis be defined as

H_0 : The population has an assumed value of mean μ_0, i.e., $\mu = \mu_0$.

The alternate hypothesis can be defined as any of the following.

(*i*) $H_1 : \mu \neq \mu_0$, that is, $\mu > \mu_0$ or $\mu < \mu_0$

(*ii*) $H_1 : \mu > \mu_0$

(*iii*) $H_1 : \mu < \mu_0$

The alternate hypothesis (*i*) is called a **two-tailed alternative** and is defined as two-tailed test.

(*ii*) is called the **right-tailed alternative** and is known as right-tailed test, and

(*iii*) is called the **left-tailed alternative** and is known as left-tailed test.

TESTING OF HYPOTHESIS AND TESTS OF SIGNIFICANCE

Suppose we have some information about a characteristic of the population and want to know whether this information can be accepted. We take a sample and obtain information about this very characteristic. On the basis of this sample information we decide whether the available information of the characteristic of the population can be accepted or rejected. We also want to know that if it can be accepted then to what degree of confidence it can be accepted.

Let θ be a parameter of the population and θ_0 be the corresponding sample statistic. Obviously there will be some difference between θ and θ_0. This may be due to the reason that the selection of the sample is not fully random. If this difference is large we say that the difference is significant. If θ_1 is the statistic obtained from a second random sample, we wish to know whether the difference between θ_0 and θ_1 is significant. The methods that are used to decide whether the difference is significant or not, are called **tests of significance** as discussed below.

Test Statistic:

In the large samples standard error (S.E.) forms the basis of the testing of hypothesis, If t is the statistic, it follows a normal distribution. The corresponding population parameter is the mean $E(t)$ and the standard deviation = S.E.(t). Thus, for the large samples,

$$Z = \frac{t - E(t)}{\text{S.E.}(t)} \sim N(0, 1)$$ and is called test statistic.

Different test statistics are chosen for various situations as follows.

I. **Test for difference between sample mean and population mean.**

If the population standard deviation σ is not known, then approximate σ by s, the standard deviation of sample and we take test statistic as

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

II. **Test for difference between means**

(*i*) If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, that is, two samples are drawn from populations with same standard deviation then the test statistic is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

(*ii*) If σ is not known and samples are large then we approximate σ^2 by

$$\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$ where s_1 and s_2 are s.d. of the samples.

Consider now a sample of size n drawn from a population whose mean is μ and variance σ^2. Let the sample observations be denoted by $x_1, x_2, ..., x_n$ which are independent and evenly distributed. Then $X_i \sim N(\mu, \sigma^2)$. Let us find the S.E. in the following cases.

I. **Sampling Distribution of the sample mean.**

"If a random sample of size n is selected from a population with mean μ and standard deviation σ then the sampling distribution of the sample mean $\bar{x}$ will have mean μ and standard error (S.E.) as $\sigma/\sqrt{n}$."

Proof: We know that sample mean $\bar{x} = \frac{1}{n}\sum x_i$ and sample variance $s^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$.

Now $$E(\bar{x}) = E\left(\frac{1}{n}\sum_1^n x_i\right) = \frac{1}{n}\sum_1^n E(x_i) = \frac{1}{n}\sum_1^n \mu = \mu$$

and $$\text{var}(\bar{x}) = \text{var}\left(\frac{x_1 + x_2 + ... + x_n}{n}\right) = \frac{1}{n^2}[\text{var}(x_1) + \text{var}(x_2) + ... + \text{var}(x_n)]$$

$$= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

Hence S.E. $(\bar{x}) = \sigma/\sqrt{n}$. Thus, $\bar{x}$ is distributed with mean μ and S.E $\sigma/\sqrt{n}$.

Note that if the population is normal, the distribution of $\bar{x}$ will be approximately normal for large n by the central limit theorem.

Thus the statistic Z is given by $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

II. **Sampling Distribution of the difference between the means of two samples**

Let n_1 and n_2 be the sizes of two independent random samples drawn from two different populations. Let $\bar{X}_1$ and $\bar{X}_2$ be the means of the two samples and let the two populations have same mean μ and variances σ_1^2 and σ_2^2 respectively.

Then $\bar{X}_1 = N\left(\mu, \frac{\sigma_1^2}{n_1}\right)$ and $\bar{X}_2 = N\left(\mu, \frac{\sigma_2^2}{n_2}\right)$

Clearly, $\bar{X}_1 - \bar{X}_2$ will also be normal and $E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu - \mu = 0$

and $$\text{Var}(\bar{X}_1 - \bar{X}_2) = \text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Here $$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

and the standard error (S.E.) $= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

$$\therefore \quad Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

In case σ_1^2 and σ_2^2 are not known, we take

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$ where s_1 and s_2 are the s.d.'s of two samples.

SAMPLING DISTRIBUTION OF SAMPLE PROPORTION:

In some practical situations we need to estimate the proportion 'p', for example, proportion of people in the population who have some characteristic say smoking or drinking. If x out of n sampled people have this characteristic, then the sample proportion $P\ (= x/n)$ can be taken as an estimate of the population proportion p. One can observe that the distribution of the random variable x is binomial with mean np and S.D. $\sqrt{npq}$ and hence $P = (= x/n)$ will be distributed like a binomial variate with mean as

$$E(P) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = p,$$

and variance as $\quad \text{Var}\,(P) = \text{Var}\left(\frac{x}{n}\right) = \frac{1}{n^2}\text{Var}\,(x) = \frac{npq}{n^2} = \frac{pq}{n}.$

and the standard error as S.E. $(P) = \sqrt{\frac{pq}{n}}$ where $q = 1 - p$.

Also, since the binomial distribution can be approximated to normal distribution for large n, the statistic Z is given by $Z = \frac{P - p}{\sqrt{pq/n}}$ and $Z \sim N(0, 1)$.

TEST OF DIFFERENCE BETWEEN PROPORTIONS FOR TWO SAMPLES FROM TWO POPULATIONS:

If two large samples are drawn from two populations, we may be interested to test the significance of the difference between two sample proportions p_1 and p_2. We take null hypothesis H_0 that there is no significant between the two sample proportions p_1 and p_2, then we consider the statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where} \quad P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} \quad \text{and} \quad Q = 1 - P.$$

CRITICAL VALUES AND LEVEL OF SIGNIFICANCE

Let the sample statistic t lie in a certain region R. If we decide that the difference between the parameter of the population and the sample statistic is significant, that is, the null hypothesis is rejected, then the region R is called **critical region or region of rejection**. The complementary region $\bar{R}$ is called the **region of acceptance**.

Following table gives list of test statistics.

Case	Test of significance of difference between	S.E. σ^*	Statistic Z
1.	Sample mean $\bar{X}$ and population mean μ	$\sigma/\sqrt{n}$	$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$
2.	Mean of two samples $\bar{X}_1$ and $\bar{X}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\frac{\bar{X}_1 - \bar{X}_2}{\sigma^*}$
3.	Sample S.D. $= s$, population S.D. $= \sigma$	$\frac{\sigma}{\sqrt{2n}}$	$\frac{s - \sigma}{\sigma/\sqrt{2n}}$
4.	Samples S.D. s_1 and s_2	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$	$\frac{s_1 - s_2}{\sigma^*}$
5.	Sample proportion p and population proportion P	$\sqrt{PQ/n}$	$\frac{p - P}{\sigma^*}$
6.	Two sample proportions p_1 and p_2	$\sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}$	$\frac{p_1 - p_2}{\sigma^*}$

Let t be the statistic against a random sample of size n and R be the critical region (region of rejection) and $\bar{R}$ the region of acceptance.

Also, define $P\left(t \in \frac{R}{H_0}\right) = \alpha, \quad P\left(t \in \frac{\bar{R}}{H_1}\right) = \beta.$

Then α is the probability that a random value of the statistic lies in the region R. We call α as the **level of significance**.

The area of the critical region is written as $\alpha\%$ level of significance.

The value of the test statistic Z which separates the region of rejection and the region of acceptance is called the **critical value** of Z. We denote it by Z_α where α is the level of significance.

For large samples the test statistic follows a normal distribution. $Z = \frac{t - E(t)}{\text{S.E.}\,(t)} \sim N(0, 1)$.

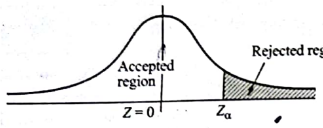
LEVEL OF SIGNIFICANCE

Suppose the level of significance α is given and Z_α is the critical value of the test statistic Z. Following are three tests.

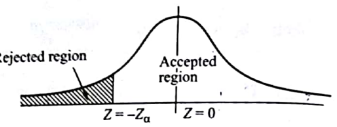
(*i*) **Right tailed test.** The critical value Z_α is obtained from the equation

$$P(Z > Z_\alpha) = \alpha.$$

The shaded area of the right tail of the probability curve, is the total area of the critical region. See the figure.



(Right tailed test)



(Left tailed test)

(*ii*) **Left tailed test.** The critical value Z_α is obtained from the equation

$$P(Z < -Z_\alpha) = \alpha.$$

The shaded area of the left tail of the probability curve, is the total area of the critical region.

(*iii*) **Two tailed Test.** The critical value Z_α is obtained from the equation

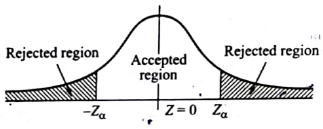
$$P(|Z| > Z_\alpha) = \alpha.$$

$$\therefore \quad P(|Z| > Z_\alpha) = P(Z > Z_\alpha) + P(Z < -Z_\alpha) = \alpha.$$

Since the probability curve is symmetrical about $Z = 0$, we have $P(Z > Z_\alpha) = P(Z < -Z_\alpha)$.

Therefore $P(Z > Z_\alpha) = \dfrac{\alpha}{2}$.

The acceptance region is given by $(-Z_\alpha, Z_\alpha)$. (See the adjoining figure).



The following table gives the critical values for the commonly used values of level of significance: 1%, 2%, 5%.

Test	C.V.	Level of Significance		
		1%	2%	5%
Two tailed	$\lvert Z_\alpha \rvert$	2.58	2.33	1.96
Right tailed	Z_α	2.33	2.055	1.645
Left tailed	Z_α	−2.33	−2.055	−1.645

Sometimes we may like to determine an interval in which the population parameter is supposed to lie. This interval is called **confidence interval** and its end points are called **confidence limits**.

For instance consider the case of $\alpha = 0.01$. We have

$$P[|Z| \leq 2.58] = 0.99 \quad \text{or} \quad P\left[\left|\frac{t - E(t)}{\text{S.E. } (t)}\right| \leq 2.58\right] = 0.99$$

or $$P[t - (2.58)\ \text{S.E. } (t) \leq E(t) \leq t + 2.58\ \text{S.E. } (t)] = 0.99$$

That is, the population parameter $E(t)$ will be in the interval $[t - 2.58\ \text{S.E.}(t),\ t + 2.58\ \text{S.E. } (t)]$ which is the confidence interval and the end points are the confidence limits.

EXAMPLE 33.2. (*a*) A coin is tossed 400 times and head turns up 216 times. Discuss whether the coin is biased or unbiased.

(*b*) A coin was tossed 400 times and head turned up 225 times. Test the hypothesis that the coin is unbiased at 5% level of significance.

[*GGSIPU IV Sem II Term 2015*]

SOLUTION: (*a*) Let us take the null hypothesis H_0: "The coin is unbiased".

Here the probability of success $p = \dfrac{1}{2}$ and hence $q = 1 - p = \dfrac{1}{2}$.

Standard deviation $= \sqrt{npq} = 10$.

Expected number of heads in 400 tosses $= 400\left(\dfrac{1}{2}\right) = 200$.

But the heads have actually appeared 216 times.

The difference between the expected number and observed ones $= 216 - 200 = 16$.

The deviation 16 is 1.6 times the standard derivation and $1.6 < 1.96$ (at 5% level), therefore the deviation is likely to appear as a result of fluctuations of simple sampling. Thus the hypothesis H_0 is accepted and the coin may be taken as unbiased. **Ans.**

(*b*) Here again $p = q = 1/2$, $s.d. = \sqrt{npq} = 10$, and the expected number of heads on 400 tosses = 200.

But the actual number of heads observed = 225, so the difference between the expected number and the observed one $= 225 - 200 = 25$.

This deviation of 25 is 2.5 times the standard deviation.

Since $2.5 > 1.96$ therefore the deviations is not likely to appear as a result of fluctuations. Therefore the hypothesis that the coin is unbiased, fails at 5% level of significance. **Ans.**

EXAMPLE 33.3. (*a*) A bag contains defective items also, the exact number being not known. A large sample of 100 items from the bag has 10 defective items. Find the 95% confidence limits for the proportion of defective items in the bag.

(*b*) From a large lot of apples a sample of 600 apples was drawn and 60 were found to be rotten. Obtain the standard error (S.E.) of the proportion of bad apples in the sample and find the 3σ limits for the percentage of bad apples in the lot.

SOLUTION: (*a*) Here $P = \dfrac{10}{100} = 0.1$, hence $Q = 1 - P = 0.9$. Since the required level of confidence is not given, let it be 95% hence 5 % level of confidence. The proportion of success of the population, p is not given in the problem. Therefore to obtain confidence limits we take P in place of p. At 5% level of significance $z_{0.5} = 1.96$.

$$\therefore \text{ Confidence limits are } P \pm z_{0.5}\sqrt{\frac{PQ}{n}} = 0.1 \pm 1.96\sqrt{\frac{0.1 \times 0.9}{100}} = 0.1 \pm 1.96\ (0.03)$$

$$= 0.1 \pm 0.0588 = \quad 0.1588, \quad 0.0412$$

Thus, the 95% confidence limits for defective items in the bag are (0.1588, 0.0412).

(*b*) Here $n = 600$. The number of bad apples in the sample is $x = 60$.

p = proportion of bad apples $= \frac{60}{600} = 0.1 \quad \therefore \quad q = 0.9$.

Since the proportion P of bad apples in the lot is not known, we assume $P = p = 0.1$.

Hence the limits for the proportion of bad apples in the lot are

$$p \pm 3\sqrt{\frac{PQ}{n}} = 0.1 \pm 3\,(0.0122) = 0.0634,\ 0.1366.$$

$\therefore$ Percentage of bad apples is in the range (0.0634, 0.1366). **Ans.**

EXAMPLE 33.4. (*a*) A firm produces electric bulbs that have burning life normally distributed with mean 800 hours and standard deviation 40 hrs. Find the probability that a random sample of 16 bulbs will have average burning life of less than 775 hours.

(*b*) In a sample of 1000 people, 800 are tea drinkers. Suddenly there was spurt in excise duty of tea and then 800 persons out of 1200 are observed to be taking tea. Can we conclude that the increase in duty has resulted in decrease in consumption of tea?

SOLUTION: (*a*) Let the average burning life of 16 bulbs be $\bar{x}$ which is a normal variate with mean 800 hours and standard error (S.E.) $= 40/\sqrt{16} = 10$.

$$\therefore \quad Z = \frac{\bar{x} - 800}{10} \sim N(0, 1) \quad \therefore \quad \bar{x} = 800 + 10Z$$

Therefore, $P(\bar{x} < 775) = P(Z < -2.5) = P(Z > 2.5)$

$= 0.5 - P(0 < Z < 2.5)$

$= 0.5 - 0.4938$ (see table)

Thus $P(\bar{x} < 775) = 0.0062$. **Ans.**

(*b*) For sample one, $n_1 = 800$, $p_1 = \frac{800}{1000} = 0.8$, For sample two, $n_2 = 1200$, $p_2 = \frac{800}{1200} = \frac{2}{3}$.

$$\therefore \quad P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{1000 \times \frac{4}{5} + 1200 \times \frac{2}{3}}{1000 + 1200} = \frac{8+8}{22} = \frac{8}{11} \quad \text{and} \quad Q = \frac{3}{11}.$$

We take null hypothesis here as $H_0 : p_1 = p_2$, that is, there is no significant difference in the consumption of tea before and after the increase in excise duty.

Let $H_1 : p_1 > p_2$ then $z = \dfrac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \dfrac{0.8 - 0.667}{\sqrt{\frac{8}{11} \cdot \frac{3}{11}\left(\frac{1}{1000} + \frac{1}{1200}\right)}} = 7$

The significance value of z at 5% level of significance is 1.65, and that at 1% level of significance is 2.33. The observed value of $z = 7$ which is much much greater than 2.33 and much greater than 1.65. Therefore the hypothesis H_0 is rejected and the conclude that there is significant decrease in the consumption of tea after the increase in exercise duty. **Ans.**

EXAMPLE 33.5. (*a*) In a factory producing nuts and bolts a machine produces both, a fraction of which are defective. A sample of 400 bolts is collected and of these 30 are found defective. The manufacture claims that it produces not more than 5% bolts defective. Find the 95% confidence limits of the proportion of the defective bolts.

(*b*) A die was thrown 9000 times and 3240 times 5 or 6 turned up. Does this data indicate that the die is unbiased?

SOLUTION: (*a*) Let us take the hypothesis H_0 that manufacturer's claim is true, Here $p = \frac{5}{100} = 0.05$

Observed proportion of the sample $= P = \frac{30}{400} = 0.075$ hence $Q = 1 - P = 0.925$.

The statistic $z = \dfrac{P - p}{\sqrt{\frac{PQ}{n}}} = \dfrac{0.075 - 0.05}{\sqrt{\frac{0.925 \times 0.075}{400}}} = 1.84$

We know that the tabular value of z at 5% level of significance is $z_{0.05} = 1.645$.

Here $1.84 > 1.645$.

Therefore H_0 is rejected at 5% level of significance, means the proportion of defective bolts is larger than what manufacturer claims. Confidence limits of the proportions are

$$p \pm z_{0.05}\sqrt{\frac{PQ}{n}} = 0.05 \pm 1.96\sqrt{\frac{0.925 \times 0.075}{400}} = 0.05 \pm 0.0258 = (0.0242, 0.0758) \quad \textbf{Ans.}$$

(*b*) Here the theoretical probability of success (getting 5 or 6) $= \frac{1}{3}$. $\therefore$ $P = \frac{1}{3}$, $Q = \frac{2}{3}$, and $n = 9000$. Observed proportion of success $p = \frac{3240}{9000} = 0.36$

Null hypothesis H_0: Die is unbiased, i.e., $p = \frac{1}{3}$.

Test statistic $z = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0.36 - 0.333}{\sqrt{(1/3)(2/3)/(9000)}} = \dfrac{0.027}{(1/90)\sqrt{(2/10)}} = 5.43.$

Since $|z| = 5.43$ which is much greater than 1.96, the hypothesis H_0 is rejected at 5% level of significance. We conclude that the die is biased. **Ans.**

EXAMPLE 33.6. (*a*) A normally distributed population has means 6.8 and standard deviation 1.5. A sample of size 400 has mean 6.75. Is the difference between the population mean and the sample mean significant?

(*b*) A random sample of 900 articles has mean 3.4 cm. Can it be reasonably regarded as a sample from a large populations with mean 3.2 cm and standard deviation 2.3 cm.

SOLUTION: (*a*) H_0 : there is no significant difference between $\bar{x}$ and μ.

Here $\mu = 6.8$, $\bar{x} = 6.75$, $\sigma = 1.5$ $n = 400$.

$$|z| = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} = \left|\frac{6.75 - 6.8}{1.5/\sqrt{400}}\right| = 0.67$$

The theoretical value of z at 5% level of significance is 1.96. Since the observed value of $z = 0.67 < 1.96$. The hypothesis H_0 is accepted, and we conclude that the sample mean is not significantly different than population mean. **Ans.**

(b) $\bar{x} = 3.4$, $\mu = 3.2$, $\sigma = 2.3$, $n = 900$.

The null hypothesis H_0 is that sample had been taken from the given population.

Under H_o, $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{3.4 - 3.2}{2.3/\sqrt{900}} = \dfrac{60}{23} = 2.61.$

Since the calculated value of $z = 2.61 > 1.96$ the theoretical value of z at 5% level of significance, H_0 is rejected at this level and we conclude that the sample can not be reasonably regarded as drawn from the given population. **Ans.**

EXAMPLE 33.7. **(a)** In a big city two samples of persons are drawn. In one sample of size 100 the average income of persons is Rs. 210, and s.d is 10 and in the other sample of size 150 persons the average income is Rs. 220 and s.d. is 12. Also given is that the s.d of the income of the people of the city is Rs. 11. Test if there is any significant difference between the two average incomes.

(b) In a school an I.Q test was given to a group of boys and to a group girls. The scores are as follows:

	Size	Mean Score	S.D
Boys	60	75	8
Girls	100	73	10

Examine if difference between the mean scores is significant.

SOLUTION: (a) The null hypothesis H_o is that the difference in the two sample means is not significant.

Here $n_1 = 100$, $\bar{x}_1 = 210$, $n_2 = 150$, $\bar{x}_2 = 220$, $s_1 = 10$ and $s_2 = 12$, $\sigma = 11$.

Also H_0: $\bar{X}_1 = \bar{X}_2$.

The test statistic $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{210 - 220}{\sqrt{\dfrac{10^2}{100} + \dfrac{12^2}{150}}} = \dfrac{-10}{\sqrt{1.0 + 0.96}} = \dfrac{-10}{1.4} = -7.1428$

Here $|z| = 7.1428$ which is much greater than 1.96 the table value of z at 5% level of significance. Therefore H_o is rejected and hence the two sample means differ significantly. **Ans.**

(b) Here we take H_0 as $\bar{x}_1 = \bar{x}_2$, that is, difference between the mean scores is insignificant.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{75 - 73}{\sqrt{\dfrac{8^2}{60} + \dfrac{10^2}{100}}} = \frac{2}{\sqrt{1 + 1.067}} = 1.3912$$

Since the computed value of the statistic $z = 1.3912$ is less than 1.96 the tabular value of z at 5% level of significance, there is no significant difference between the two means.

$\therefore$ H_0 is accepted. **Ans.**

EXAMPLE 33.8. **(a)** In a class of 100 students the average marks scored by 64 boys is 66 with s.d. 10, while the average marks scored by 36 girls is 70 with s.d. 8. Test at 1% level of significance whether the girls performed better than the boys.

(b) A tea company claims that its premium tea brand outsells its normal brand by 10%. If it is found that 46 out of a sample of 200 tea-users prefer premium brand and 19 out of another independent sample of 100 tea users prefer normal brand. Test the validity of the claims made by the company at 1% and 5% level of significance. *[GGSIPU IV Sem End Term 2015]*

SOLUTION: (a) Here $n_1 = 36$, $\bar{x}_1 = 70$, $\sigma_1 = 8$, $n_2 = 64$, $\bar{x}_2 = 66$, $\sigma_2 = 10$ and the level of significance = 1%.

Here the null hypothesis $H_0 : \mu_1 = \mu_2$, i.e., boys and girls performed equally well.

Alternate hypothesis $H_1 : \mu_1 \geq \mu_2$ (right tailed test, girls did better than boys).

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}} = \frac{70 - 66}{\sqrt{\left(\dfrac{64}{36} + \dfrac{100}{64}\right)}} = 2.189$$

Since $Z = 2.189 < 2.33$ we conclude that at 1% level of significance, the difference between μ_1 and μ_2 is not significant. Therefore the null hypothesis H_0 is accepted and H_1 is rejected. Which means that the girls performed better than boys. **Ans.**

(b) Let p_1 and p_2 be the true proportions of the premium and normal brands and let the null hypothesis H_0 be that the claim of the company is valid, that is, $H_0 : p_1 \sim p_2 = 0.1$ against the alternative $H_1 : p_1 \sim p_2 \neq 0.1$.

Here $n_1 = 200$, $x_1 = 46$, $\hat{p}_1 = \dfrac{x_1}{n_1} = \dfrac{46}{200} = 0.23$.

and $n_2 = 100$, $x_2 = 19$, $\hat{p}_2 = \dfrac{x_2}{n_2} = \dfrac{19}{100} = 0.19$

The test statistic is $z = \dfrac{|(\hat{p}_1 \sim \hat{p}_2) - (p_1 - p_2)|}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0, 1)$.

where $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \dfrac{46 + 19}{200 + 100} = \dfrac{65}{300} = 0.217$

$\therefore$ $\hat{q} = 1 - \hat{p} = 0.783$.

Therefore $|z| = \dfrac{|0.04 - 0.1|}{\sqrt{0.217\,(0.783)\left(\dfrac{1}{200} + \dfrac{1}{100}\right)}}$

$$= \frac{0.06}{\sqrt{0.15 \times 0.17}} = \frac{0.06}{0.0505} = 1.19$$

Since $|z| = 1.19 < 1.96$ claim is valid at 5% level of significance. **Ans.**

SMALL SAMPLES AND TESTS OF SIGNIFICANCE

Upto now we have discussed the large samples testing. All these tests were based on the central limit theorem to justify the normality of the test statistic. But, if we are not able to collect large sample, the test procedures discussed so far are of no use. Here we shall take up some equivalent procedures fit to be employed for small samples.

As per convention, a small sample is one whose size is less than 30 and in such a sample we cannot assume that a random sampling distribution of a statistic is normal distribution. Also, we cannot assume that the values given by the sample data are sufficiently close to the population value and cannot be used in their place for calculating the standard error of the estimate. However, we shall assume that the population (s) from which the samples are drawn are normal. In such cases we apply student's t-test and χ^2-test. Before these we should understand the concept of 'degrees of freedom'.

Degrees of Freedom (d.o.f.)

Suppose we are to choose four numbers whose sum is 40, we can choose any three numbers independently and the fourth will be 40 minus the sum of the first three. Thus, our choice is reduced to choosing three numbers because of one restraint on our freedom and we say that our degrees of freedom is $3(= 4 - 1)$. Same way, if two restrictions are imposed, our freedom to choose will be further curtailed and the degrees of freedom will be $2(= 4 - 2)$. Thus, in general, the d.o.f. is the total number of observations minus the number of restrictions. The d.o.f. is usually denoted by ν (a Greek letter pronounced as 'nu').

For small samples we use student's t-test, F-test and χ^2-test.

STUDENT'S t-DISTRIBUTION

We have seen that if the population is normal, the statistic $Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ has normal distribution whether the sample is large or small. However, if the population S.D. σ is not known and the sample size is small, we define an statistic 't' given by

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}} \quad \text{where} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

The distribution of the statistic t was given by the mathematician Gusset who worked under the pen-name 'student'.

The probability density function of the above 't' variate, is given by

$$f(t) = \frac{1}{\sqrt{\nu}\,\beta(\nu/2, \nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \quad -\infty < t < \infty \quad \text{where} \quad \nu = n - 1.$$

The probability curve is symmetrical about $t = 0$ and bell shaped. As ν increases the 't' distribution curve moves closer to the standard normal probability curve.

As a particular case, if $\nu = 1$, that is for $n = 2$, we get

$$f(t) = \frac{1}{\beta\left(\frac{1}{2}, \frac{1}{2}\right)} \cdot \frac{1}{1+t^2} = \frac{1}{\pi(1+t^2)}, \quad -\infty < t < \infty.$$

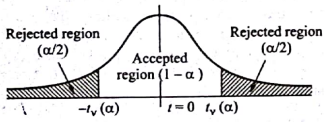
Critical or Significant Values of t

The critical (or significant) values of t at α-level of significance with ν-degrees of freedom, are given by

$$P[|t| < t_\nu(\alpha)] = \alpha \quad \text{for two tailed test}$$

$$P[t > t_\nu(\alpha)] = \alpha \quad \text{for the right tailed test}$$

and $$P[t < -t_\nu(\alpha)] = \alpha \quad \text{for the left tailed test.}$$



Since the distribution is symmetrical about $t = 0$, we have

$$P[t > t_\nu(\alpha)] + P[t < t_\nu(\alpha)] = \alpha$$

or $$2P[t < t_\nu(\alpha)] = \alpha \quad \text{or} \quad P[t > t_\nu(\alpha)] = \frac{\alpha}{2}$$

$$\Rightarrow \quad P[t > t_\nu(2\alpha)] = \alpha.$$

We conclude that the critical values for a single tailed test at α-level of significance is the same as for a two-tailed test at 2α-level of significance (for same degrees of freedom).

t-TEST FOR THE MEAN OF RANDOM SAMPLE

Here we test whether the mean of a sample drawn from a population deviates significantly from a stated value when variance of the population is unknown. In such cases we take the null hypothesis H_0 : The difference between the sample mean $\bar{x}$ and the population mean μ is not significant and we use the statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$ where $\bar{x}$ is the mean of the sample

and $s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$ and $(n - 1)$ is the degrees of freedom.

Critical value of t or significant value of t at the 'α' level of significance for r d.o.f., is denoted by $t_r(\alpha)$. If calculated value of t is such that $|t| < t_\alpha$, H_0 is accepted and if $|t| > t_\alpha$ then H_0 is rejected.

EXAMPLE 33.9. (*a*) A sample of 20 items has mean 42 units and standard deviation 5 units. Test the hypothesis that it is a random sample from a normal population with mean 45 units?

(*b*) Ten persons were chosen at random from a normal population and their heights were found to be 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71 inches. Test the hypothesis that the mean height of the population is 66 inches. Also find the 95% confidence limits for the true population mean.

SOLUTION: (**a**) Here $n = 20$, $\bar{X} = 42$, $S = 5$, $\nu = 19$ = *d.o.f*

The null hypothesis H_0: "No difference is there between the sample mean and the population mean."

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n-1} \quad \text{or} \quad s^2 = \frac{\Sigma(X - \bar{X})^2}{n} \cdot \frac{n}{n-1} = \frac{n}{n-1}S^2 \quad \text{where } S = \text{s.d. of the sample.}$$

$$\therefore \quad s^2 = \frac{n}{n-1}S^2 = \frac{20}{19} \cdot 5^2 = 26.31, \quad \text{hence} \quad s = 5.129.$$

Then $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{42 - 45}{5.129/\sqrt{20}} = -2.615 \quad \therefore \quad |t| = 2.615$$

From the t-table we have $t_{0.05}$ (for 19 *d.o.f.*) = 2.09.

Since $|t| > t_{0.05}$, H_0 is rejected, that is, there is significant difference between the sample mean, and population mean, in other words, sample could not have come from this population. **Ans.**

(*b*) Here $n = 10$, $\bar{x} = \frac{1}{10}(63 + 63 + 66 + 67 + 68 + 69 + 70 + 70 + 71 + 71) = 67.8$ inches

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{1}^{n}\left[d_i^2 - \frac{(\sum d_i)^2}{n}\right] \quad \text{where } d_i = x_i - 68.$$

$$= \frac{1}{9}\left[82 - \frac{4}{10}\right] = 9.067. \quad \text{Hence} \quad s = 3.011 \text{ inches.}$$

We test $H_0 : \mu = 66$ against two-tailed alternative $H_1 : \mu \neq 66$.

Under H_0, $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_\nu$ with $\nu = n - 1 = 9$ d.o.f.

$$\therefore \quad t = \frac{67.8 - 66}{3.011/\sqrt{10}} = 1.89.$$

From Table II $t_\nu(0.05) = 2.26$. Since t calculated is less than t tabulated hence the null hypothesis H_0 may be accepted at 5% level of significance.

Also we can write, the 95% confidence limits for the population mean μ as

$$\mu = \bar{x} \pm t_{\nu(0.05)}\,(s/\sqrt{n})$$

$$= 67.8 \pm (2.26)\,(3.011/\sqrt{10}) = 67.8 \pm 2.53 \quad \textbf{Ans.}$$

EXAMPLE 33.10. A sample of 10 boxes of chips is drawn in which the mean weight is 490 gr. and the standard derivation of weight is 9 gr, Can the sample be considered to have come from a population having mean weight 500 gr?

SOLUTION: Here $n = 10$, $\bar{x} = 490$, $S = 9$, $\mu = 500$

$$\therefore \quad s = S\sqrt{\frac{n}{n-1}} = 9\sqrt{\frac{10}{9}} = 9.486$$

Let the null hypothesis H_0 be that the difference between $\bar{x}$ and μ is not significant.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{490 - 500}{9.486/\sqrt{10}} = -0.33 \quad \text{But from the } t\text{-table } t_{0.05} = 2.26 \text{ for } 9 \text{ } d.o.f.$$

Since $|t| = 0.33 < t_{0.5} = 2.26$ the hypothesis H_0 is a accepted, and we conclude that the sample can be considered to have come from the population with mean 500 gr. **Ans.**

T-TEST FOR DIFFERENCE OF MEANS OF TWO SMALL SAMPLES

The t-test can also be applied to ascertain if two samples $(x_1, x_2, ..., x_{n1})$ and $(y_1, y_2, ..., y_{n2})$ of sizes n_1 and n_2 have been draw from two normal populations with means μ_1 and μ_2 but having same variance.

Here we take the null hypothesis H_0 that the samples have been drawn from normal populations with means μ_1 and μ_2. The samples have means $\bar{x}$ and $\bar{y}$.

Consider the statistic $t = \dfrac{\bar{x}-\bar{y}}{s\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ where $\bar{x} = \dfrac{1}{n_1}\sum_{n}^{n_1} x_i$, $\bar{y} = \dfrac{1}{n}\sum_{1}^{n_2} y_i$

and $s^2 = \dfrac{1}{n_1+n_2-2}\left[(n-1)s_1^2+(n_2-1)s_2^2\right] = \dfrac{1}{n_1+n_2-2}\left[\sum_{1}^{n_1}(x_i-\bar{x})^2+\sum_{1}^{n_2}(y_i-\bar{y})^2\right]$.

This t is to be compared with the table value of t at 5 % level of significance for $(n_1 + n_2 - 2)$ degrees of freedom.

EXAMPLE 33.11. In a school the heights of six randomly chosen girls are 63, 65, 68, 69, 71 and 72 inches and those of nine randomly chosen boys 61, 62, 65, 66, 69, 70, 71, 72, 73 inches. Test if the girls are taller than the boys.

SOLUTION: Let x_1 and x_2 denote the sample heights of girls and boys. Here $n_1 = 6$, $n_2 = 9$. The null hypothesis $H_0 : \mu_1 = \mu_2$, that is, means of both the samples are the same.

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = 68, \quad \bar{x}_2 = \frac{\sum x_2}{n_2} = 67.66 \quad \text{and} \quad \Sigma(x_1-\bar{x}_1)^2 = 60, \quad \Sigma(x_2-\bar{x}_2)^2 = 152.0002.$$

$$\therefore \quad s^2 = \frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{n_1+n_2-2} = \frac{60+152}{6+9-2} = 16.3077$$

or $s = \sqrt{16.3077} = 4.038.$

Therefore $t = \dfrac{\bar{x}_1-\bar{x}_2}{s\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \dfrac{68-67.66}{4.038\sqrt{\dfrac{1}{6}+\dfrac{1}{9}}} = 0.3031$

This is the calculated value of t for $n_1 + n_2 - 2$ (= 13) *d.o.f.* The tabular value of t for 13 *d.o.f.* at 5% level of significance is 1.77. Here $0.3031 < 1.77$ therefore H_0 is accepted, that is, there is no significant difference between the two means.

EXAMPLE 33.12. Two samples of C.F.L. of two brands were tested for length of life and results are given below:

	Size	Sample Mean	Sample S.D
Brand I	7	1036	40
Brand II	8	1234	36

Is the difference in two sample means significant to conclude that the Brand I has more life than Brand II?

SOLUTION: Here $n_1 = 7$, $\bar{x}_1 = 1036$, $s_1 = 40$.

$n_2 = 8$, $\bar{x}_2 = 1234$, $s_2 = 36$.

Let the null hypothesis H_0 be that the Brand I and II have same life span and in turn, H_1 be that Brand II is superior to Brand I.

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2} = \frac{8(36)^2+7(40)^2}{7+8-2} = 1659.07, \quad \text{hence} \quad s = 40.73.$$

Then $$t = \frac{|\bar{x}_1-\bar{x}_2|}{s\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{|1036-1234|}{40.73\sqrt{\dfrac{1}{8}+\dfrac{1}{7}}} = 18.148$$

From t-table $t_{0.05} = 1.77$ at 13 (= 7 + 8 – 2) d.o.f.. Since the calculated value $t > t_{0.05}$, the hypothesis H_0 is rejected and H_1 is accepted, that is, Brand II is superior to Brand I.

CHI-SQUARE VARIATE AND TEST FOR POPULATION VARIANCE

Earlier we have seen that an estimate of the population variance σ^2 usually required to make inferences about the population mean. But in some practical situations the knowledge of the variance of the sampled population may be more important than the population mean. For example, our concern may be to know the precision of a measuring instrument being used, or we may be much concerned about the variation of the water level at different points during flood.

Suppose we want to test if a random sample $x_1, x_2, ..., x_n$ has been drawn from a normal population with a specified variance σ^2. Then under the null hypothesis that the population variance is σ^2, the statistic χ^2 (called chi-square) variable is defined by

$$\chi^2 = \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sigma^2} = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{(n-1)S^2}{\sigma^2}$$

where $S^2 = \dfrac{1}{n-1}\sum_{n=1}^{n}(x_i-\bar{x})^2$ is an unbiased estimate of σ^2, follows sampling distribution with probability distribution given by

$$\frac{1}{2^{\nu/2}\Gamma(\nu/2)}\cdot\left[\exp\left(-\frac{1}{2}\chi^2\right)\right](\chi^2)^{\frac{\nu}{2}-1}, \quad 0 < \chi^2 < \infty.$$

The distribution defined above is called χ^2 probability distribution with $\nu = n - 1$ degrees of freedom (*df*). The probability curve for a chi-square distribution is shown in Fig. (*i*). The curve is skewed towards right and its shape varies with the degrees of freedom $\nu = (n - 1)$. The variate tends to standard normal variate as $n \to \infty$.

Critical values and test of significance. Let $\chi^2_{\nu[\alpha]}$ denote the value of chi-square variate for ν *df* such that the area to the right of this point α, that is, $P[\chi^2 > \chi^2_{\nu[\alpha]}] = \alpha$, as shown in Fig. (*i*). The Table (3) gives the critical values or significant values of $\chi^2_{\nu[\alpha]}$ for the right-tailed test for different degrees of freedom ν and decrease in α. At a specific level α and *df* ν, the null hypothesis H_0: $\sigma^2 = \sigma_0^2$ is rejected against the alternate hypothesis:

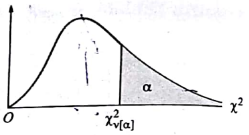


Fig. (*i*)

(*i*) $H_1 : \sigma^2 > \sigma_0^2$; if calculated $\chi^2 > \chi^2_{\nu[\alpha]}$, refer Fig. (*i*)

(*ii*) $H_1 : \sigma^2 < \sigma_0^2$; if calculated $\chi^2 < \chi^2_{\nu[1-\alpha]}$, refer Fig. (*ii*)

(*iii*) $H_1 : \sigma^2 \neq \sigma_0^2$; if calculated $\chi^2 > \chi^2_{\nu[\alpha/2]}$ or $\chi^2 - \chi^2_{\nu[1-\alpha/2]}$, refer Fig. (*iii*).

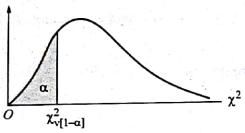


Fig. (*ii*)

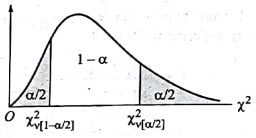


Fig. (*iii*)

Equal tails are used for the two-tailed χ^2 test as a matter of mathematical convenience only otherwise the chi-square distribution is not symmetric. However, normally in practice right-tailed test is applicable.

Example 33.13. A manufacturer of car batteries claims that the life of his batteries is approximately normally distributed with a S.D. of 0.9 years. If a random sample of 10 of these batteries has a S.D. of 1.2 years, do you think $\sigma > 0.9$ years at $\alpha = 0.05$?

Solution: We test $H_0 : \sigma^2 = 0.81$ against the right-tailed alternative $H_1 : \sigma^2 > 0.81$.

We have $n = 10$, $\sigma^2 = 0.81$, $S^2 = (1.2)^2 = 1.44$

Under H_0, the test statistic $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{9(1.44)}{0.81} = 16.0$

It follows χ^2-distribution with d.o.f. $= 9 (= 10 - 1)$.

From Table of χ^2, $\chi^2_9(0.05) = 16.92$. Since χ^2 calculated is less than χ^2-tabulated, hence the value is not significant and the hypothesis H_0 should be accepted at 5% level of significance.

Example 33.14. Following data gives the 11 measurements of the same object on the same instrument:
2.5, 2.3, 2.4, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, 2.5 and 2.3
At 1% level, test the hypothesis that the variance of the instrument is not more than 0.16.

Solution: We test the null hypothesis $H_0 : \sigma^2 = 0.16$, against alternative $H_1 : \sigma^2 > 0.16$.

For the given data we have $\bar{x} = \dfrac{27.6}{11} = 2.51$, $\sum (x - \bar{x})^2 = 0.1891$.

Under H_0, the test statistic χ^2 is given by

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum (x-\bar{x})^2}{\sigma^2} = \frac{0.1891}{0.16} = 1.182$$

It follows χ^2 distribution with degrees of freedom $\nu = n - 1 = 11 - 1 = 10$.

From Table III, $\chi^2_{10[0.1]} = 23.2$, and since the χ^2 calculated is less than the χ^2 tabulated, so hypothesis may be accepted at 1% level of significance.

χ^2-distribution can be applied in following two ways.

1. χ^2-distribution is used **to test the goodness of fit**. For example, suppose that we have fitted a binomial or a Poisson distribution to a given data of a sample. We use the χ^2-distribution to test whether this fitting of the Binomial or Poisson distribution to the data is acceptable or not.
2. χ^2-distribution is used **to test the independence of the attributes** of a population. For example, suppose that a population has two attributes or characteristics. The χ^2-distribution can be used to test whether the two attributes are dependent or independent, based on a random sample drawn from a population.

CHI-SQUARE TEST AS TEST OF GOODNESS OF FIT

Suppose that O_i and E_i, $i = 1, 2, \ldots, n$, are the observed and the expected theoretical frequencies of the i^{th} class with $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i$.

The expected frequencies are computed using the hypothesis assumed about the population. Then

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] \text{ with } \sum_1^n O_i = \sum_1^n E_i = N.$$

This follows a χ^2-distribution with $(n - 1)$ degrees of freedom. On simplification, we get

$$\chi^2 = \sum_1^n \frac{1}{E_i}(O_i^2 + E_i^2 - 2O_i E_i) = \sum_1^n \left[\frac{O_i^2}{E_i} + E_i - 2O_i\right]$$

$$= \sum_1^n \frac{O_i^2}{E_i} + \sum_1^n E_i - 2\sum_1^n O_i = \sum_1^n \frac{O_i^2}{E_i} - N.$$

Critical Values:

Let $\chi^2_n(\alpha)$ denote the values χ^2-distribution with n degrees of freedom at α-level of significance, then the critical values of χ^2 are given by

$$P[\chi^2 > \chi^2_n(\alpha)] = \alpha.$$

That is the area of the probability curve to the right of this point is α. The critical values of χ^2-distribution are available in Table 2 for different values of degrees of freedom η and levels of significance α.

EXAMPLE 33.15. (*a*) A die is thrown 276 times and results of thee throws are tabulated below.

Number on the die	1	2	3	4	5	6
Frequency	40	32	29	59	57	59

Test if the die is unbiased using χ^2-test.

(*b*) The demand for a particular spare part in a factory was found to vary from day-to-day. In a sample study the following information was obtained.

Days	*Mon*	*Tues*	*Wed*	*Thurs*	*Fri*	*Sat*
No. of Parts Demanded	1124	1125	1110	1120	1125	1116

Use chi-square to test the hypothesis that the no of parts demanded does not depend on the day of the week at 5% level of significance.

(*c*) The theory predicts the proportion of beans in the four groups G_1, G_2, G_3, G_4 should be in the ratio 9 : 3 : 3 : 1. In an experiment with 1600 beans the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory?

[*GGSIPU IV Sem II Test 2015*]

SOLUTION: (*a*) Let the null hypothesis H_0 be that the die is unbiased.

Under H_0, the expected frequency for each digit from 1 to 6 $= \frac{276}{6} = 46.$

Now to calculate the value of χ^2, we have

O_i	40	32	29	59	57	59
E_i	46	46	46	46	46	46
$(O_i - E_i)^2$	36	196	289	169	121	169
$\frac{1}{E_i}(O_i - E_i)^2$	0.782	4.261	6.282	3.674	2.63	3.674

$$\therefore \quad \chi^2 = \sum_1^6 \frac{(O_i - E_i)^2}{E_i} = 21.3.$$

The standard value of χ^2 from the table for 5% level of significance at 5 (= 6 − 1) d.o.f. is 11.07. Since the calculated value of $\chi^2 = 21.3$ is greater than the table value 11.07 the hypothesis H_0 is rejected and we conclude that coin is biased, die

(*b*) Let the null hypothesis H_0: The number of parts demanded does not depend on the day of the week.

Under the Hypothesis H_0, the expected frequency for each day from monday to saturday (i.e., for 1'st day to 6th day) $= \frac{6720}{6} = 1120$. Now to calculate χ^2, we have

Days	*Mon*	*Tues*	*Wed*	*Thurs*	*Fri*	*Sat*
No. of Parts Demanded O_i	1124	1125	1110	1120	1125	1116
Expected frequency E_i	1120	1120	1120	1120	1120	1120
$\frac{(O_i - E_i)^2}{E_i}$	0.01428	0.02232	0.08928	0	0.02232	0.01428

$$\therefore \quad \chi^2 = \sum_{i=1}^{6}\left[\frac{(O_i - E_i)^2}{E_i}\right] = 0.16248$$

The standard value of χ^2 from the table for 5% level of significance at 5(= 6 − 1) d.o.f. is 11.07.

Since the calculated value of $\chi^2 = 0.16248$ is less than the standard value (from table) therefore hypothesis H_0 is accepted.

(*c*) Let H_0: "The experimental result support the theory," i.e., there is no significant difference between the observed and expected frequency.

Expected frequency of group G_1 is $E(G_1) = \frac{1600\times 9}{16} = 900.$

Expected frequency of group G_2 is $E(G_2) = \frac{1600\times 3}{16} = 300.$

Expected frequency of group G_3 is $E(G_3) = \frac{1600\times 3}{16} = 300.$

Expected frequency of group G_4 is $E(G_4) = \frac{1600\times 1}{16} = 100.$

Observed frequency O_i	882	313	287	118
Expected frequency E_i	900	300	300	100
$\frac{(O_i - E_i)^2}{E_i}$	0.36	0.5633	0.5633	3.24

$$\chi^2 = \sum_{i=1}^{4}\frac{(O_i - E_i)^2}{E_i} = 4.7266$$

Conclusion: For degree of freedom 3(= 4 − 1), the tabular value of χ^2 at 5% level of significance is 7.815. Since the calculated value of χ^2 is less than the standard value of χ^2 hence H_0 is accepted.

EXAMPLE 33.16. (*a*) Following data gives the number of male and female births in 800 families having four children:

No. of males	0	1	2	3	4
No. females	4	3	2	1	0
No. of families	32	178	290	236	94

Test if the above data is consistent with the hypothesis that the binomial distribution holds and the probability of male birth is same as that of a female birth. [*GGSIPU IV Sem End Term 2015*]

(*b*) What is chi-square test? A random number table of 250 digits showed the following distribution of digits 0, 1, 2, ...9.

Digit	0	1	2	3	4	5	6	7	8	9
Observed frequency	17	31	29	18	14	20	35	30	20	36
Expected frequency	25	25	25	25	25	25	25	25	25	25

Does the observed distribution differ significantly from expected distribution using a significance level of 0.01? Given that $\chi^2_{0.99}$ for 9 degree of freedom is 21.7.

SOLUTION: (*a*) Hypothesis H_0 : Probability of male and female birth are equal $p = q = 1/2$, $N = 800$. Under the binomial distribution

$$P(x = r) = {}^nC_r\, p^r\, q^{n-r} = {}^4C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{4-r} = {}^4C_r \left(\frac{1}{2}\right)^4$$

$\therefore$ $N(r)$ = Number of families having r male children $= N \cdot {}^4C_r \left(\frac{1}{2}\right)^4$.

$\therefore$ $N(0) = 800\left(\frac{1}{2}\right)^4 = 50$, $N(1) = 800\ {}^4C_1\left(\frac{1}{2}\right)^4 = 200$,

$N(2) = 800 \times {}^4C_2\left(\frac{1}{2}\right)^4 = 300$, $N(3) = 200$ and $N(4) = 50$.

O_i	32	178	290	236	94
E_i	50	200	300	200	50
$\frac{(O_i - E_i)^2}{E_i}$	6.48	2.42	0.333	6.48	38.72

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 54.433$$

$\therefore$ Table value of χ^2 at 4 (5 – 1) d.o.f. = 9.49

The calculated value of χ^2 > table value of x^2, hence H_0 is rejected.

(*b*) Hypothesis H_0: The observed distribution does not differ significantly from expected distribution.

Digit	0	1	2	3	4	5	6	7	8	9
Observed frequency O_i	17	31	29	18	14	20	35	30	20	36
Expected frequency E_i	25	25	25	25	25	25	25	25	25	25
$\frac{(O_i - E_i)^2}{E_i}$	2.56	1.44	0.64	1.96	4.84	1	4	1	1	4.84

$$\chi^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} = 23.28$$

The standard value of χ^2 from the table given 21.7 as $\chi^2_{\text{cal.}} > \chi^2_{\text{stand value}}$ hence H_0 is rejected. **Ans.**

EXAMPLE 33.17. (*a*) Apply χ^2-test to ascertain if Poisson distribution can be assumed from the following data.

No, of defects	0	1	2	3	4	5
Frequency	6	13	13	8	4	3

(*b*) A survey of 320 families with 5 children each revealed the following distribution:

No of boys	5	4	3	2	1	0
No of girls	0	1	2	3	4	5
No. of families	14	56	110	88	40	12

Is the result consistent with the hypothesis that male and female births are equally probable? You may use the following table giving the value of chi-square:

Degrees of freedom "n"	χ^2-*values at levels*	
	0.05	0.01
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812

SOLUTION: (*a*) The null hypothesis H_0: Poisson distribution is a good fit to the data.

$$\text{Mean } m = \frac{\sum f_i x_i}{\sum f_i} = \frac{0+13+26+24+16+15}{47} = \frac{94}{47} = 2 = \text{one parameter of Poisson data} = m.$$

EXAMPLE 33.16. *(a)* Following data gives the number of male and female births in 800 families having four children:

No. of males	0	1	2	3	4
No. females	4	3	2	1	0
No. of families	32	178	290	236	94

Test if the above data is consistent with the hypothesis that the binomial distribution holds and the probability of male birth is same as that of a female birth. *[GGSIPU IV Sem End Term 2015]*

(b) What is chi-square test? A random number table of 250 digits showed the following distribution of digits 0, 1, 2, ...9.

Digit	0	1	2	3	4	5	6	7	8	9
Observed frequency	17	31	29	18	14	20	35	30	20	36
Expected frequency	25	25	25	25	25	25	25	25	25	25

Does the observed distribution differ significantly from expected distribution using a significance level of 0.01? Given that $\chi^2_{0.99}$ for 9 degree of freedom is 21.7.

SOLUTION: *(a)* Hypothesis H_0 : Probability of male and female birth are equal $p = q = 1/2$, $N = 800$. Under the binomial distribution

$$P(x = r) = {}^nC_r\, p^r\, q^{n-r} = {}^4C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{4-r} = {}^4C_r \left(\frac{1}{2}\right)^4$$

$\therefore$ $N(r)$ = Number of families having r male children $= N \cdot {}^4C_r \left(\frac{1}{2}\right)^4$.

$\therefore$ $N(0) = 800\left(\frac{1}{2}\right)^4 = 50$, $N(1) = 800\ {}^4C_1\left(\frac{1}{2}\right)^4 = 200$,

$N(2) = 800 \times {}^4C_2\left(\frac{1}{2}\right)^4 = 300$, $N(3) = 200$ and $N(4) = 50$.

O_i	32	178	290	236	94
E_i	50	200	300	200	50
$\frac{(O_i - E_i)^2}{E_i}$	6.48	2.42	0.333	6.48	38.72

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 54.433$$

$\therefore$ Table value of χ^2 at 4 (5 – 1) d.o.f. = 9.49

The calculated value of χ^2 > table value of x^2, hence H_0 is rejected.

(b) Hypothesis H_0: The observed distribution does not differ significantly from expected distribution.

Digit	0	1	2	3	4	5	6	7	8	9
Observed frequency O_i	17	31	29	18	14	20	35	30	20	36
Expected frequency E_i	25	25	25	25	25	25	25	25	25	25
$\frac{(O_i - E_i)^2}{E_i}$	2.56	1.44	0.64	1.96	4.84	1	4	1	1	4.84

$$\chi^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} = 23.28$$

The standard value of χ^2 from the table given 21.7 as $\chi^2_{cal.} > \chi^2_{stand\ value}$ hence H_0 is rejected. **Ans.**

EXAMPLE 33.17. *(a)* Apply χ^2-test to ascertain if Poisson distribution can be assumed from the following data.

No, of defects	0	1	2	3	4	5
Frequency	6	13	13	8	4	3

(b) A survey of 320 families with 5 children each revealed the following distribution:

No of boys	5	4	3	2	1	0
No of girls	0	1	2	3	4	5
No. of families	14	56	110	88	40	12

Is the result consistent with the hypothesis that male and female births are equally probable? You may use the following table giving the value of chi-square:

Degrees of freedom "n"	χ^2-*values at levels*	
	0.05	0.01
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812

SOLUTION: *(a)* The null hypothesis H_0: Poisson distribution is a good fit to the data.

Mean $m = \frac{\sum f_i x_i}{\sum f_i} = \frac{0+13+26+24+16+15}{47} = \frac{94}{47} = 2$ = one parameter of Poisson data = m.

Since probability of r successes $= e^{-m}\frac{m^r}{r!}$ hence the frequency with r successes

$$= Ne^{-m}\frac{m^r}{r!} = 47.\frac{e^{-2}.2^r}{r!} = N(r)$$

$\therefore$ $N(0) = 47e^{-2}\frac{2^4}{0!} = 6.36 \approx 6$, $N(1) = 47e^{-2}.2 = 12.72 \approx 13$, $N(2) = 47.e^{-2}\frac{2^2}{2} \approx 13$,

$N(3) = 47e^{-2}\frac{2^3}{6} = 8.48 \approx 9$, $N(4) = 47e^{-2}.\frac{2^4}{24} = 4.24 \approx 4$, $N(5) = 47e^{-2}.\frac{2^5}{120} = 1.696 \approx 2$.

X	0	1	2	3	4	5
O_i	6	13	13	8	4	2
E_i	6.36	12.72	12.72	8.48	4.24	1.696
$\frac{(O_i-E_i)^2}{E_i}$	0.2037	0.00616	0.00616	0.02716	0.0135	1.0026

$\therefore$ $$\chi^2 = \sum\frac{(O_i-E_i)^2}{E_i} = 1.2864$$

Tabulated value of χ^2 at 5% level of significance for d.o.f. $= \nu = 6 - 2$ is 9.49. Since the calculated value of χ^2 is less than 4 the tabulated value of χ^2, hence we can conclude that H_0 is accepted and Poisson is a good fit on the data.

where H_0: "Male and female births are equally probable: i.e., $p = 1/2 = q$.

(b) We use **binomial distribution** to calculate theoretical or expected frequency given by

$N(r) = \text{N.P}(r)$ where $P(r) = {}^nC_r\, p^r\, q^{n-r}$ and N = total frequency.

$N(r)$ = number of families with r male children

$P(r)$ = Probability of having r male children in a family out of n-chiledren p, q are probability of male and female births.

$N(0)$ = number of familes with 0 male child $= 320\,{}^5C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{5-0} = 320\cdot1\cdot\frac{1}{2^5} = 10$

$N(1)$ = number of familes with 1 male child $= 320\,{}^5C_1\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^4 = 320\cdot5\cdot\frac{1}{2^5} = 50$

$N(2)$ = number of familes with 2 male children $= 320\,{}^5C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = 320\cdot\frac{5\cdot4}{2!}\cdot\frac{1}{2^5} = 100$

$N(3)$ = number of familes with 3 male children $= 320\,{}^5C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = 320\cdot\frac{5\cdot4}{2}\cdot\frac{1}{2^5} = 100$

$N(4)$ = number of familes with 4 male children $= 320\,{}^5C_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) = 320\cdot5\cdot\frac{1}{1^5} = 50$ and

$N(5)$ = number of familes with 5 male children $= 320\,{}^5C_5\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = 320\cdot1\cdot\frac{1}{2^5} = 10$.

No of boys	0	1	2	3	4	5
No of girls	5	4	3	2	1	0
Observed frequency O_i	14	56	110	88	40	12
Expected frequency E_i	10	50	100	100	50	10
$\frac{(O_i-E_i)^2}{E_i}$	1.6	0.72	1	1.44	2	0.4

$\therefore$ $$\chi^2 = \sum_{i=1}^{6}\frac{(O_i-E_i)^2}{E_i} = 7.16$$

Conclusion: The calculated value of χ^2 is 7.16 for degree of freedom $6 - 1 = 5$. At 0.05 χ^2_{standard} is given 11.070 and hence $\chi^2_{\text{cal}} < \chi^2_{\text{standard value}}$ therefore H_0 is accepted.

At level 0.01 also $\chi^2_{\text{cal}} < \chi^2_{\text{standard value}}$ therefore H_0 is accepted.

i.e., The male and female births are equally probable.

χ^2-TEST AS TEST FOR INDEPENDENCE OF ATTRIBUTES

χ^2-test is also used to find whether two attributes are associated or not. For that we take the null hypothesis that the two attributes are not associated, that is, the two attributes are independent.

The sample data is written in two way table which is called **contingency table.** Let the two attributes be A and B. A divided into m classes $A_1, A_2, \ldots, A_m$ and B divided into n classes $B_1, B_2, \ldots, B_n$. Next $(A_i - B_j)$ represents the number of persons possessing the attributes A_i and B_j $(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$. We also have $\Sigma A_i = N = \Sigma B_j$ where N is the total frequency.

The contigency table is as follows:

A / B	A_1	A_2	A_3	...	A_m	Total
B_1	(A_1B_1)	(A_2B_1)	(A_3B_1)	...	(A_mB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	(A_3B_2)	...	(A_mB_2)	(B_2)
B_3	(A_1B_3)	(A_2B_3)	(A_3B_3)	...	(A_mB_3)	(B_3)
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
B_n	(A_1B_n)	(A_2B_m)	(A_3B_n)	...	(A_mB_n)	(B_n)
Total	(A_1)	(A_2)	(A_3)	...	(A_m)	N

$P(A_i)$ = Probability that a person possesses the attribute $A_i = \frac{(A_i)}{N}$, $i = 1, 2, \ldots, m$.

$P(B_j)$ = Probability that a person possesses the attribute $B_j = \frac{(B_j)}{N}$, $j = 1, 2, \ldots, n$.

If $(A_i B_j)$ is the expected number of persons possessing both the attributes A_i and B_j then

$$(A_i B_j)_e = N\,P(A_i B_j) = N\,P(Ai)\,P(Bj) = N.\frac{(A_i)}{N}.\frac{(B_j)}{N} = \frac{(A_i)(B_j)}{N}$$

(since A_i and B_j are independent attributes.)

$$\therefore \quad \chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{[(A_i B_j)_o - (A_i B_j)_e]^2}{(A_i B_j)_e}$$

which is distributed as a χ^2-variate with $(m-1)(n-1)$ degrees of freedom.

EXAMPLE 33.18. Following table gives the number of good and bad parts produced by each of the three shifts on a factory.

	Good parts (A_1)	Bad parts (A_2)	Total
(B_1) Day shift	960	40	1000
(B_2) Evening shift	940	50	990
(B_3) Night shift	950	45	995
Total	2850	135	2985

Test whether or not the production of bad parts is independent of the shift on which they were produced.

SOLUTION: Assume the null hypothesis H_0: The production of bad parts is independent of the shift on which they were produced.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{[(A_i B_j)_0 - (A_i B_j)_e]^2}{(A_i B_j)_e}$$

Expected frequencies

$$(A_1B_1) = \frac{A_1 B_1}{N} = \frac{1000\times 2850}{2985} = 954.77, \quad (A_2B_1) = \frac{A_2 B_1}{N} = \frac{135\times 1000}{2985} = 45.27$$

$$(A_1B_2) = \frac{A_1 B_2}{N} = \frac{990\times 2850}{2985} = 945.226, \quad (A_2B_2) = \frac{A_2 B_2}{N} = \frac{135\times 990}{2985} = 44.773$$

$$(A_1B_3) = \frac{A_1 B_3}{N} = \frac{2850\times 995}{2985} = 950, \quad (A_2B_3) = \frac{A_2 B_3}{N} = \frac{135\times 995}{2985} = 45$$

The χ^2 can be calculated from the following table

Class	O_i	E_i	$(O_i - E_i)^2/E_i$
(A_1B_1)	960	954.77	0.0286
(A_1B_2)	940	945.226	0.02889
(A_1B_3)	950	950	0
(A_2B_1)	40	45.27	0.61349
(A_2B_2)	50	44.773	0.61022
(A_2B_3)	45	45	0
		Total	1.28126 = χ^2

The tabulated value of χ^2 at 5% level of significance at $2(=(m-1)(n-1))$, d.o.f. is 5.991. Here the calculated value of χ^2 is less than the tabulated value, hence the hypothesis H_0 is accepted, i.e, the problem of bad parts has nothing, to do with shifts.

SNEDECOR'S F-DISTRIBUTION AND TEST FOR THE EQUALITY OF TWO POPULATION VARIANCES

Consider a situation where we want to compare the precisions of the two measuring instruments, means we are to compare two population variances. Suppose we want to test whether the two independent samples $x_1, x_2, ..., x_{n1}$ and $y_1, y_2, ..., y_{n2}$ have been drawn from the normal populations with the same variance s^2. Under the null hypothesis H_0 : that the population variances σ_x^2 and σ_y^2 are the same, that is, $\sigma_x^2 = \sigma_y^2 = \sigma^2$, we define the *variance ratio statistic* F, as

$$F = S_x^2/S_y^2,$$

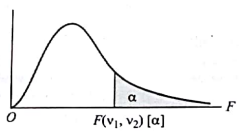
where $S_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$ and $S_y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2$.

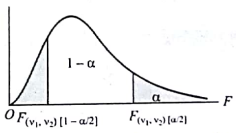
It follows sampling distribution with probability density function $f(F)$ given by

$$f(F) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{\beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}\cdot\frac{F^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, \quad 0 < F < \infty,$$

where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

The distribution is called *Snedecor's F-distribution with* (ν_1, ν_2) *degrees of freedom* and the variate F is denoted by $F_{(\nu_1, \nu_2)}$. Generally, the greater of the two variances S_x^2 and S_y^2 is taken as numerator and ν_1 corresponds to the variance in the numerator. The probability curve for the F-distribution is shown in the adjoining figure.





The curve is not symmetric and the shape depends on the degrees of freedom ν_1 and ν_2.

Critical values and test of significance. Let $F_{(\nu_1,\nu_2)[\alpha]}$ denote the value of F for (ν_1, ν_2) degrees of freedom such that the area to the right of this point is α, that is, $P[F > F_{(\nu_1,\nu_2)[\alpha]}] = \alpha$, as shown in the figure. The Tables at the end, gives critical values or significant values of $F_{(\nu_1,\nu_2)[\alpha]}$ for the right-tailed test for different df (ν_1, ν_2) and significant level $\alpha = 0.05$ and 0.01, respectively.

For an F-variate the reciprocal relation

$$F_{(\nu_1,\nu_2)[\alpha]} = \frac{1}{F_{(\nu_2,\nu_1)[1-\alpha]}}$$ also holds.

At a specific level α and degrees of freedom (ν_1, ν_2) the null hypothesis $H_0 : \sigma_x^2 = \sigma_y^2$ is rejected against the alternate hypothesis,

(*i*) $H_1 : \sigma_x^2 > \sigma_y^2$ and $F = \dfrac{S_x^2}{S_y^2}$, if $F > F_{(\nu_1, \nu_2)[\alpha]}$

(*ii*) $H_1 : \sigma_x^2 < \sigma_y^2$ and $F = \dfrac{S_y^2}{S_x^2}$, if $F > F_{(\nu_2, \nu_1)[\alpha]}$

(*iii*) $H_1 : \sigma_x^2 \neq \sigma_y^2$ and $F = \dfrac{S_x^2}{S_y^2}$, if $F > F_{(\nu_1, \nu_2)[\alpha/2]}$

EXAMPLE 33.19. There are two different choices to stimulate a certain chemical process. To test whether the variance of the yield is the same no matter which catalyst is used, a sample of 10 batches is produced using the first catalyst, and of 12 using the second. If the resulting data is $S_1^2 = 0.14$ and $S_2^2 = 0.28$, test the hypothesis of equal variance at 2% level.

SOLUTION: We have, $n_1 = 10$, $n_2 = 12$, $S_1^2 = 0.14$, $S_2^2 = 0.28$. We test $H_0 : \sigma_1^2 = \sigma_2^2$ against two-tail alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Under H_0, the test statistic F given by

$$F = \frac{S_2^2}{S_1^2} = \frac{0.28}{0.14} = 2 .$$

The statistic F follows F-distribution with (11, 9) degrees of freedom.

From the Table, $F_{(11, 9)[0.02]}$ two-tail alternative $= F_{(11, 9)[0.01]}$ at right-tail alternative = 4.67.

Since, F calculated is less than F-tabulated, it is not significant and hence hypothesis may be accepted at 2% level of significance.

EXAMPLE 33.20. Two random samples gave the following results:

Sample	*Size*	*Sample mean*	*Sum of squares of deviations from the mean*
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population at 5% level of significance.

SOLUTION: Since a normal population is specified by two parameters: mean μ and variance σ^2, thus to test that two independent samples have been drawn from the same population, we need to test (*i*) the equality of population means using t-test (*ii*) the equality of population variances, using F-test.

Since t-test is applied under the assumption that population variances are the same, so first we shall test for the equality of population variances.

Here, we have

$$n = 10,\ n_2 = 12,\ \bar{x} = 15,\ \bar{y} = 14,\ \sum (x_i - \bar{x})^2 = 90,\ \sum (y_i - \bar{y})^2 = 108,$$

$$S_1^2 = \frac{90}{9} = 10,\quad S_2^2 = \frac{108}{11},\quad \text{and}\quad S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2\right] = \frac{90 + 108}{20} = 9.9.$$

We test $H_0 : \sigma_1^2 = \sigma_2^2$ against the right-tailed alternative $H_1 : \sigma_1^2 > \sigma_2^2$. Under H_0, the statistic F is given by

$$F = \frac{S_1^2}{S_2^2} = \frac{10}{9.82} = 1.018 - F_{(9.11)}.$$

From the Table A, $F_{(9,\ 11)\ [0.05]} = 2.90$. Since F calculated is less than the F tabulated, hence H_0 is accepted.

Since H_0: $\sigma_1^2 = \sigma_2^2$ is established, we can now apply t test for testing H_0: $\mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$.

Under H_0, the statistic t is given by

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{15 - 14}{3.15\sqrt{\dfrac{1}{10} + \dfrac{1}{12}}} = \frac{1}{1.349} = 0.74 \sim t_{20}.$$

From Table II, $t_{20[0.05]} = 2.086$. Since t calculated is less than the t-tabulated hence the hypothesis H_0: $\mu_1 = \mu_2$ may be accepted.

Since both the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ and H_0: $\mu_1 = \mu_2$ are accepted so samples may be considered to come from the same normal population.

FISHER'S Z-DISTRIBUTION

In Snedecor's distribution with (ν_1, ν_2) degree of freedom, if we put

$$F = \exp(2Z) \qquad \text{or} \qquad Z = \frac{1}{2}\log F$$

the distribution of Z becomes

$$g(Z) = p(F)\frac{dF}{dZ} = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{\beta\left(\dfrac{\nu_1}{2}, \dfrac{\nu_2}{2}\right)} \cdot \frac{(e^{2z})^{(\nu_1/\nu_2 - 1)} 2e^{2Z}}{\left[1 + \dfrac{\nu_1}{\nu_2} e^{2Z}\right]^{(\nu_1 + \nu_2)^2}}$$

$$= 2\frac{(\nu_1/\nu_2)^{\nu_1/2}}{\beta\left(\dfrac{\nu_1}{2}, \dfrac{\nu_2}{2}\right)} \frac{e^{\nu_1 z}}{\left[1 + \dfrac{\nu_1}{\nu_2} e^{2Z}\right]^{(\nu_1 + \nu_2)/2}}$$

which is the probability function of Fisher's Z-distribution with (ν_1, ν_2) degrees of freedom.