Regression

1. From the following data, find (i) the two regression equations (ii) The coefficient of correlation between the marks in Economics and Statistics (iii) The most likely marks in statistics when marks in Economics are 30.

Marks in Economics	**25**	**28**	**35**	**32**	**31**	**36**	**29**	**38**	**34**	**32**
Marks in Statistics	**43**	**46**	**49**	**41**	**36**	**32**	**31**	**30**	**33**	**39**

Solution:

X	**Y**	$\boldsymbol{X-\overline{X}}$	$\boldsymbol{Y-\overline{Y}}$	$\boldsymbol{(X-\overline{X})^2}$	$\boldsymbol{(Y-\overline{Y})^2}$	$\boldsymbol{(X-\overline{X})(Y-\overline{Y})}$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21

38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
320	380	0	0	140	398	-93

Now $\bar{X} = \frac{\sum X}{n} = \frac{320}{10} = 32$

$\bar{Y} = \frac{\sum Y}{n} = \frac{380}{10} = 38$

Coefficient of regression of Y on X is $b_{YX} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$

$= -\frac{93}{140} = -0.6643$

Coefficient of regression of X on Y is $b_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2}$

$= -\frac{93}{398} = -0.2337$

Equation of the line of regression of X on Y is $x - \bar{x} = b_{XY}(y - \bar{y})$

$\Rightarrow x - 32 = -0.2337(y - 38)$

$\Rightarrow x = -0.2337y + 0.2337 \times 38 + 32$

$\Rightarrow x = -0.2337y + 40.8806$

Equation of the line of regression of Y on X is $y - \bar{y} = b_{YX}(x - \bar{x})$

$\Rightarrow y - 38 = -0.6642(y - 38)$

$\Rightarrow y = -0.6642x + 0.6642 \times 32 + 38$

$\Rightarrow y = -0.6642x + 59.2576$

Correlation of coefficient $r^2 = b_{YX} \times b_{XY}$

$$= -0.6643 \times (-0.2337)$$

$$= 0.1552$$

$$r = \pm\sqrt{0.1552}$$

$$= \pm 0.394$$

Now we have to find the most likely marks in statistics (Y) when marks in Economics (X) are 30. We use the line of regression of Y on X.

$\Rightarrow y = -0.6642x + 59.2576$

Put $x = 30$ we get

$\Rightarrow y = -0.6642 \times 30 + 59.2576$

$\Rightarrow y = 39.3286$

2. The two lines of regression are $8x - 10y + 66 = 0, 40x - 18y - 214 = 0$. The variance of X is 9. Find (i) the mean value of X and Y (ii) correlation coefficient between X and Y.

Solution:

Since both the lines of regression passes through the mean values $\bar{x}$ and $\bar{y}$, the point $(\bar{x}, \bar{y})$ must satisfy the two given regression lines.

$(1) \times 5 \Rightarrow 40\bar{x} - 50\bar{y} = -330$

$(2) \Rightarrow 40\bar{x} - 18\bar{y} = 214$

Subtracting (1) – (2) we get

$\Rightarrow 32\bar{y} = 544$

$\Rightarrow \bar{y} = 17$

Sub $\bar{y} = 17$ in (1) we get,

$(1) \Rightarrow 8\bar{x} - 10\bar{y} = -66$

$\Rightarrow 8\bar{x} - 10 \times 17 = -66$

$\Rightarrow 8\bar{x} = -66 + 170$

$\Rightarrow \bar{x} = 13$

Hence the mean value ae given by $\bar{x} = 13$ and $\bar{y} = 17$

(ii) Let us suppose that equation (A) is the equation of line of regression of Y on X and (B) is the equation of the line regression of X on Y, we get after rewriting (A) and (B)

$\Rightarrow 10y = 8x + 66$

$\Rightarrow y = \frac{8}{10}x + \frac{66}{10}$

$\Rightarrow b_{YX} = \frac{8}{10}$

$\Rightarrow 40x = 18y + 214$

$\Rightarrow x = \frac{18}{40}y + \frac{214}{40}$

$\Rightarrow b_{XY} = \frac{18}{40}$

Correlation of coefficient $r^2 = b_{YX} \times b_{XY}$

$$= \frac{8}{10} \times \frac{18}{40} = \frac{9}{25}$$

$$r = \pm\frac{3}{5}$$

$$= \pm 0.6$$

Since both the regression coefficients are positive, r must be positive.

Hence $r = 0.6$

Important note:

If we take equation (A) as the line of regression of X on Y we get,

$$\Rightarrow 8x = 10y - 66$$

$$\Rightarrow x = \frac{10}{8}y - \frac{66}{8}$$

$$\Rightarrow b_{XY} = \frac{10}{8}$$

$$\Rightarrow 18y = 40x - 214$$

$$\Rightarrow y = \frac{40}{18}x - \frac{214}{18}$$

$$\Rightarrow b_{YX} = \frac{40}{18}$$

Correlation of coefficient $r^2 = b_{YX} \times b_{XY}$

$$= \frac{10}{8} \times \frac{40}{8} = \frac{25}{9}$$

$$r = 2.78$$

But r^2 should always lies between 0 and 1. Hence our assumption that line (A) is line of regression of X on Y and the line (B) is line of regression of Y on X is wrong.

3. The two lines of regression are $\boldsymbol{4x - 5y + 33 = 0, 20x - 9y - 107 = 0}$. The variance of $\boldsymbol{X}$ is 25. Find (i) the mean value of X and Y (ii) correlation coefficient between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Solution:

Since both the lines of regression passes through the mean values $\bar{x}$ and $\bar{y}$, the point $(\bar{x}, \bar{y}\ )$ must satisfy the two given regression lines.

$(1) \Rightarrow 20\bar{x} - 9\bar{y} = 107$

$(2) \times 5 \Rightarrow 20\bar{x} - 25\bar{y} = -165$

Subtracting (1) – (2) we get

$\Rightarrow 16\bar{y} = 272$

$\Rightarrow \bar{y} = 17$

Sub $\bar{y} = 17$ in (1) we get,

$(2) \Rightarrow 4\bar{x} - 5\bar{y} = -33$

$\Rightarrow 4\bar{x} - 5 \times 17 = -33$

$\Rightarrow 4\bar{x} = -33 + 85$

$\Rightarrow \bar{x} = 13$

Hence the mean value ae given by $\bar{x} = 13$ and $\bar{y} = 17$

(ii) Let us suppose that equation (A) is the equation of line of regression of Y on X and (B) is the equation of the line regression of X on Y, we get after rewriting (A) and (B)

$\Rightarrow 5y = 4x + 33$

$\Rightarrow y = \frac{4}{5}x + \frac{33}{5}$

$\Rightarrow b_{YX} = \frac{4}{5}$

$\Rightarrow 20x = 9y + 107$

$\Rightarrow x = \frac{9}{20}y + \frac{107}{20}$

$\Rightarrow b_{XY} = \frac{9}{20}$

Correlation of coefficient $r^2 = b_{YX} \times b_{XY}$

$$= \frac{4}{5} \times \frac{9}{20} = \frac{3}{5}$$

$$r = \pm 0.6$$

4. Can $\boldsymbol{Y = 5 + 2.8X}$ and $\boldsymbol{X = 3 - 0.5Y}$ be the estimated regression equations of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{X}$ on $\boldsymbol{Y}$ respectively? Explain your answer.

Solution:

Given,

$\Rightarrow X = 3 - 0.5Y$

$\Rightarrow b_{XY} = -0.5$

$\Rightarrow Y = 5 + 2.8X$

$\Rightarrow b_{YX} = 2.8$

Correlation of coefficient $r^2 = b_{YX} \times b_{XY}$

$$= 2.8 \times (-0.5) = -1.4$$

$r = \sqrt{-1.4}$ which is imaginary quantity.

Here r cannot be imaginary.

Hence the given lines are not estimated as regression equations

Home Work

1. Out of the two lines of regression given by $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$, which one is the regression lines of $X\ on\ Y$? Use the equations to find the mean of $X\ and\ Y$. If the variance of X is 12, calculate the variance of Y.
2.

Method of least squares

This is the most widely used method of obtaining trend. If we fit a straight line using the method of least squares, then the sum of deviations of points on either side of the line is equal to zero (ie, sum of the deviations of the computed values from the actual values is zero). Also, the sum of the squares of these deviations will be least as compared to those obtained using other lines. Hence the name "Method of least squares". The straight line obtained using this method is called "the line of best fit".

Fitting a straight line using the method of least squares

The straight line trend is given by the equation $Y = a + bX$. Here a is the Y intercept and b is the slope of the trend line. To determine the values of a and b we solve the normal equations:

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

where n represents the number of years for which data are given.

To simplify calculations, we can take the midpoint in time as the origin for the variable X, so that $\sum X = 0$. If odd number of consecutive years are given, deviations are taken from the middle year. If even number of years are given, the origin is taken midway between the middle two years. Then we will have

$a = \frac{\sum Y}{n}$ and $b = \frac{\sum XY}{\sum X^2}$.

Fitting a Second degree parabola

The simplest form of the non-linear trend is the second degree parabola:

$$Y = a + bX + cX^2$$

where a is the Y' intercept, b is the slope of the curve at the origin and c is the rate of change in the slope.

The values of a, b, and c are obtained by solving the following normal equations:

$$\sum Y = na + b\sum X + c\sum X^2$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

If the origin of X is taken as the midpoint in time, then $\sum X = 0$ and $\sum X^3 = 0$ so that the normal equations are greatly simplified.

A second degree trend equation is an appropriate model for the secular trend of a time series when the data do not fall in a straight line.

Fitting an exponential curve

An exponential curve has equation $Y = ab^x$. Taking logarithm on both sides, we have,

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

Putting $Y = \log y$, $A = \log a$, $X = x$, $B = \log b$

the equation becomes $Y = A + BX$ and we proceed to find A and B using normal equations as in the case of straight line trend. Finally a and b are found by taking antilogs of A and B respectively.

The normal equations are

$$\sum Y = nA + B\sum X$$

$$\sum XY = A\sum X + B\sum X^2$$

Merits and Limitations of least square method.

Merits:

1. This is a mathematical method and is completely objective in character.

2. The straight line obtained by this method is the line of best fit.

3. This method gives trend values for the entire time period.

4. Since we get a functional relationship between X and Y, we can forecast future values.

Limitations

1. This method is tedious and time -- consuming

2. The type of curve to be fitted linear, parabolic etc - has to be selected carefully

3. The addition of even one more item necessitates fresh calculations. Hence it is not a flexible method.

Problems:

1. calculate the trend values by the method of least squares. Also calculate the sales for the years 1999 and 2000

Year	: 1991	1992	1993	1994	1995	1996	1997
Sales (in Lakhs)	: 125	128	133	135	140	141	143

Solution:

Let the trend equation be $Y = a + bX$. The normal equations are

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

The required sums are got from the following table

Year	Y (Sales)	X=Year-1994	X^2	XY	T
1991	125	-3	9	-375	125.67
1992	128	-2	4	-256	128.78
1993	133	-1	1	-133	131.89
1994	135	0	0	0	135
1995	140	1	1	140	138.11
1996	141	2	4	282	141.22

1997	143	3	9	429	144.33
Total	945	0	28	87	

$$\sum X = 0, \sum Y = 945, \sum X\,Y = 87, n = 7, \sum X^2 = 28$$

The normal equations are

$$\sum Y = na + b \sum X$$

$$945 = 7 * a + b * 0$$

$$\Rightarrow 7a = 945$$

$$\Rightarrow a = \frac{945}{7} = 135$$

$$\sum XY = a \sum X + b \sum X^2$$

$$87 = a(0) + b(28)$$

$$\Rightarrow 28b = 87$$

$$\Rightarrow b = \frac{87}{28} = 3.11$$

The straight line trend equation is $Y = a + bX$

$$Y = 135 + 3.11X$$

Trend value for $1991 = 135 + (3.11 * -3) = 125.67$

Trend value for $1992 = 135 + (3.11 * -2) = 128.78$

Trend value for $1993 = 135 + (3.11 * -1) = 131.89$

Trend value for $1994 = 135 + (3.11 * 0) = 135$

Trend value for $1995 = 135 + (3.11 * 1) = 138.11$

Trend value for $1996 = 135 + (3.11 * 2) = 141.22$

Trend value for $1997 = 135 + (3.11 * 3) = 144.33$

For the year $1999, X = 1999 - 1994 = 5$

Sales= $Y = 135 + (3.11 * 5) = 150.55\ lakhs$

For the year 2000, $X = 2000 - 1994 = 6$

Sales= $Y = 135 + (3.11 * 6) = 153.66\ lakhs$

2. Given below are the figures of production (in thousand quintals) of a sugar factory.

Year :	1974	1975	1976	1977	1978	1979	1980
Production :	77	88	94	85	91	98	90

Solution:

Let the trend equation be $Y = a + bX$. The normal equations are

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

The required sums are got from the following table

Year	Y (Sales)	X=Year-1997	X^2	XY	Trend Values (T)
1974	77	-3	9	-231	83
1975	88	-2	4	-176	85
1976	94	-1	1	-94	87
1977	85	0	0	0	89
1978	91	1	1	91	91
1979	98	2	4	196	93
1980	90	3	9	270	95
Total	623	0	28	56	

$$\sum X = 0, \sum Y = 623, \sum X\,Y = 56, n = 7, \sum X^2 = 28$$

The normal equations are

$$\sum Y = na + b\sum X$$

$$623 = 7 * a + b * 0$$

$$\Rightarrow 7a = 623$$

$$\Rightarrow a = \frac{623}{7} = 89$$

$$\sum XY = a\sum X + b\sum X^2$$

$$56 = a(0) + b(28)$$

$$\Rightarrow 28b = 56$$

$$\Rightarrow b = \frac{56}{28} = 2$$

The straight line trend equation is $Y = a + bX$

$$Y = 89 + 2X$$

Trend value for $1974 = 89 + 2(-3) = 83$

Trend value for $1975 = 89 + 2(-2) = 85$

Trend value for $1976 = 89 + 2(-1) = 87$

Trend value for $1977 = 89 + 2(0) = 89$

Trend value for $1978 = 89 + 2(1) = 91$

Trend value for $1979 = 189 + 2(2) = 93$

Trend value for $1980 = 89 + 2(3) = 95$

3.

Year	: 1989	1990	1991	1992	1993	1994	1995	1996
Sales(in Rs.Lakhs)	: 76	80	130	144	138	120	174	190

Calculate the trend values from 1989 to 1996 by the method of least squares and also calculate the predicted sales for the year 2001.

Solution:

Let the trend equation be $Y = a + bX$. The normal equations are

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Since there are even number of years, we take the origin to be between 1992 and 1993

(ie, 1992.5)

The required sums are got from the following table

Year	Y (Sales)	X=Year-1992.5	X^2	XY	Trend Values (T)
1989	76	-3.5	12.25	-266	80.155
1990	80	-2.5	6.25	-200	94.825
1991	130	-1.5	2.25	-195	109.495
1992	144	-0.5	0.25	-72	124.165
1993	138	0.5	0.25	69	138.835
1994	120	1.5	2.25	180	153.505
1995	174	2.5	6.25	435	168.175
1996	190	3.5	12.25	665	182.845
Total	1052	0	42	616	

$$\sum X = 0, \sum Y = 1052, \sum XY = 616, n = 8, \sum X^2 = 42$$

The normal equations are

$$\sum Y = na + b\sum X$$

$$1052 = 8 * a + b * 0$$

$$\Longrightarrow 8a = 1052$$

$$\Longrightarrow a = \frac{1052}{8} = 131.5$$

$$\sum XY = a\sum X + b\sum X^2$$

$$616 = a(0) + b(42)$$

$$\Longrightarrow 42b = 616$$

$$\Longrightarrow b = \frac{616}{42} = 14.67$$

The straight line trend equation is $Y = a + bX$

$$Y = 131.5 + 14.67X$$

Trend value for 1989 $= 131.5 + (14.67 * -3.5) = 80.155$

Trend value for 1990 $= 131.5 + (14.67 * -2.5) = 94.825$

Trend value for 1991 $= 131.5 + (14.67 * -1.5) = 109.495$

Trend value for 1992 $= 131.5 + (14.67 * -0.5) = 124.165$

Trend value for 1993 $= 131.5 + (14.67 * 0.5) = 138.835$

Trend value for 1994 $= 131.5 + (14.67 * 1.5) = 153.505$

Trend value for 1995 $= 131.5 + (14.67 * 2.5) = 168.175$

Trend value for 1996 $= 131.5 + (14.67 * 3.5) = 182.845$

For the year $2001, X = 2001 - 1992.5 = 8.5$

Predicted Sales for 2001= $Y = 131.5 + (14.67 * 8.5) = 256.195\ lakhs$

4. Fit a second degree polynomial equation for the following data

Year :	1976	1977	1978	1979	1980	1981	1982	1983	1984
Sales :	50	65	70	85	82	75	65	90	95

Solution:

The second degree polynomial

$$Y = a + bX + cX^2$$

The normal equations are:

$$\sum Y = na + b\sum X + c\sum X^2$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

$$we\ calculate\ the\ necessary\ sums$$

Year	Y	X (Year-1980)	X^2	X^3	X^4	XY	X^2Y
1976	50	-4	16	-64	256	-200	800
1977	65	-3	9	-27	81	-195	585
1978	70	-2	4	-8	16	-140	280
1979	85	-1	1	-1	1	-85	85
1980	82	0	0	0	0	0	0
1981	75	1	1	1	1	75	75
1982	65	2	4	8	16	130	260
1983	90	3	9	27	81	270	810
1984	95	4	16	64	256	380	1520
Total	677	0	60	0	708	235	4415

$$\sum X = 0, \sum X^2 = 60, \quad \sum X^3 = 0, \quad \sum X^4 = 708,$$

$$\sum Y = 677, \ \sum XY = 235, \quad \sum X^2Y = 4415, \ n = 9$$

The normal equations are:

$$\sum Y = na + b\sum X + c\sum X^2$$

$$677 = 9a + 60c$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

$$235 = 60\ b$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

$$4415 = 60a + 708\ c$$

Solving, we get The normal equations are:

$$a = 77.3509$$

$$b = 3.9167$$

$$c = -0.3193$$

Hence, the second degree polynomial trend equation is, The normal equations are:

$$Y = 77.3509 + 3.9167X - 0.3193X^2$$

5. The following table gives the profits of a concern for 5 years ending 1983.

Year	:	1979	1980	1981	1982	1983
Profit's(RS. '000)	:	1.6	4.5	13.8	40.2	125.0

Find out the trend values by using the equation $y_e = ab^x$

Solution:

The trend curve is $y = ab^x$

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

$$Y = \log y, \ A = \log a, \ X = x, \ B = \log b$$

$$Y = A + BX$$

The normal equations are

$$\sum Y = nA + B\sum X$$

$$\sum XY = A\sum X + B\sum X^2$$

Year x	Profits (y)	$X = x - 1981$	$\log y = Y$	X^2	XY	Trend Values
1979	1.6	-2	0.2041	4	-0.4082	1.56
1980	4.5	-1	0.6532	1	-0.6532	4.63
1981	13.8	0	1.1399	0	0	13.79
1982	40.2	1	1.6042	1	1.6042	41.05
1983	125.0	2	2.0969	4	4.1938	122.21
Total		0	5.6983	10	4.7366	

$$\sum X = 0, \sum Y = 5.6983, \sum X^2 = 10, \quad \sum XY = 4.7366, n = 5$$

$$\sum X = 0, the\ value\ of\ \ A = \frac{\sum Y}{n} = \frac{5.6983}{5} = 1.1397$$

$$a = Antilog\ A = 13.79$$

The Value of $B = \frac{\sum XY}{\sum X^2} = \frac{4.7366}{10} = 0.47366$

$$b = Antilog\ B = 2.977$$

The equation $y_e = ab^x$

$$y_e = 13.79 * (2.977)^x$$

Now when $x = -2$ the value of

$$y_e = 13.79 * (2.977)^{-2} = 1.56$$

Now when $x = -1$ the value of

$$y_e = 13.79 * (2.977)^{-1} = 4.63$$

Now when $x = 0$ the value of

$$y_e = 13.79 * (2.977)^0 = 13.79$$

Now when $x = 1$ the value of

$$y_e = 13.79 * (2.977)^1 = 41.05$$

Now when $x = 2$ the value of

$$y_e = 13.79 * (2.977)^2 = 122.21$$

Home Work

1. The following are the annual profits, in thousand rupees in a business

Year	: 1971	1972	1973	1974	1975	1976
Profit's (in Rs. '000)	: 83	92	71	90	169	191

Calculate the trend values by the method of least squares. Also estimate the profit for the year 1979.

2. The prices of a commodity during 1993-98 are given below. Fit a parabola $y = a + bx + cx^2$ to these data. Calculate the trend values. Estimate the price of the commodity for the year 1999.

Year :	1993	1994	1995	1996	1997	1998
Price :	100	107	128	140	181	192

3. You are given the population figures of India as follows:

Census Year (x) :	1911	1921	1931	1941	1951 n	1961	1971
Population (in crores (y)) :	25.0	25.1	27.9	31.9	36.1	43.9	54.7

Fit the exponential trend $y = ab^x$ to the above data by the method of least squares and find the trend values. Estimate the population in 1981.

Rank Correlation

Let $(x_i, y_i): i = 1,2,..,n$ be the ranks of $'n'$ individuals in the group for two characteristics A and B respectively. The correlation coefficient between the ranks $x_i \; and \; y_i$ is called the rank correlation between the two characteristics A and B for that group of individuals.

The Spearsman's coefficient of rank correlation is given by

$$\rho_s = 1 - \frac{6\sum_i {d_i}^2}{n(n^2-1)} \quad where \; d_i = x_i - y_i$$

Where ${d_i}^2$ is the square of the difference of corresponding ranks and n is the number of pairs of observations.

Note:

1. When the ranks are same $\rho = 1$,
2. The rank correlation co-efficient ρ lies between $-1 \; and \; 1$

$$-1 \leq \rho \leq 1$$

Repeated ranks

If there is more than one item with the same value in the series, then common ranks are given to the repeated items. As a result of this, the following adjustment (or) correction is made in the rank correlation formula.

In the rank correction coefficient formula, we add the correction factor $\frac{m(m^2-1)}{12}$ to $\sum_i {d_i}^2$, where $'m'$ is the number of items, an item is rtepeated. This correction factor is to be added for each repeated value.

$$\rho_s = 1 - \frac{6\left(\sum_i {d_i}^2 + correction\ factor\right)}{n(n^2-1)}$$

$$\rho_s = 1 - \frac{6\left(\sum_i {d_i}^2 + \frac{m(m^2-1)}{12}\right)}{n(n^2-1)}$$

1. Ten competition in a beauty contest are ranked by 3 judges in the following order:

A :	1	6	5	3	10	2	4	9	7	8
B :	3	5	8	4	7	10	2	1	6	9
C :	6	4	9	8	1	2	3	10	5	7

Find which pair of judges have the nearest approach to common taste of beauty.

Solution:

A	B	C	$d_1 = A - B$	$d_2 = B - C$	$d_3 = A - C$	${d_1}^2$	${d_2}^2$	${d_3}^2$
1	3	6	-2	-3	-5	4	9	2
6	5	4	1	1	2	1	1	4
5	8	9	-3	-1	-4	9	1	1
10	4	8	6	-4	2	36	16	4
3	7	1	-4	6	2	16	36	4
2	10	2	-8	8	0	64	64	0
4	2	3	2	-1	1	4	1	1
9	1	10	8	-9	-1	64	81	1
7	6	5	1	1	2	1	1	4
8	9	7	-1	2	-1	1	4	1
						$\sum {d_1}^2 = 200$	$\sum {d_2}^2 = 214$	$\sum {d_3}^2 = 60$

$$\rho_{AB} = 1 - \frac{6 * 200}{10(100-1)} = -0.212$$

$$\rho_{BC} = 1 - \frac{6 * 214}{10(100-1)} = -0.297$$

$$\rho_{AC} = 1 - \frac{6 * 60}{10(100-1)} = 0.636$$

Hence A and C have the nearest approach to common tastes of beauty.

2. Calculate Spearman's rank correlation coefficient for the following data:

X : 53 98 95 81 75 71 59 55

Y : 47 25 32 37 30 40 39 45

Solution

X	Y	Rank X	Rank Y	d	d^2
53	47	8	1	7	49
98	25	1	8	-7	49
95	32	2	6	-4	16
81	37	3	5	-2	4
75	30	4	7	-3	9
71	40	5	3	2	4
59	39	6	4	2	4
55	45	7	2	5	25
					$\sum d^2 = 160$

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

$$= 1 - \frac{6 * 160}{8(64 - 1)} = -0.9048$$

There is very high negative correlation between X and Y

3. Calculate the coefficient of rank correlation from the following data:

X : 48 34 40 12 16 16 66 25 16 57

Y : 15 15 24 8 13 6 20 9 9 15

Solution

X	Y	Rank X	Rank Y	d	d^2
48	15	8	7	1	1
34	15	6	7	-1	1
40	24	7	10	-3	1
12	8	1	2	-1	1
16	13	3	5	-2	4
16	6	3	1	2	4
66	20	10	9	1	1
25	9	5	3.5	1.5	2.25
16	9	3	3.5	-0.5	0.25
57	15	9	7	2	4
					$\sum d^2 = 27.5$

In X-series, the value 16 is repeated three times

In Y-series, the value 15 is repeated three times and the value is repeated two times and the value 9 is repeated two times.

$$\rho_s = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}({m_1}^3 - m_1) + \frac{1}{12}({m_2}^3 - m_2) + \frac{1}{12}({m_3}^3 - m_3)\right]}{n(n^2 - 1)}$$

$$\rho_s = 1 - \frac{6\left[27.5 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2)\right]}{10(100 - 1)}$$

$$= 1 - \frac{6\,[32]}{990} = 0.806$$

There is high positive correlation.

Home Work

1. The following are the ranks obtained by 10 students in statistics and mathematics. To what extent is knowledge of students I statistics related to knowledge in mathematics?

Statistics	:	1	2	3	4	5	6	7	8	9	10
Mathematics	:	2	4	1	5	3	9	7	10	6	8

2. The following data are a random sample of consumer's income and expenditure on certain luxury items. Compute the spearman rank correlation coefficient and test for the existence in the population.

Income: 23 17 34 56 49 31 28 80 65 40 26
Luxury items: 10 50 120 225 90 60 55 340 170 25 80

Test for Rank Correlation Coefficient

When n is greater than $30 (n > 30)$, the sampling distribution of r_s is approximately normal, with a mean of zero and a standard deviation of $\frac{1}{\sqrt{n-1}}$. Thus, under the null hypothesis, the test statistic is $Z = r_s(\sqrt{n-1})$

If $|Z| \leq Z_{\propto}$ we accept H_0 at a given level of significance α, otherwise we reject H_0

1. The following are the year of experience (X) and the average customer satisfac tion (Y) for 10 service providers. Is there a significant rank correlation between measures? Use the 0.05 level of significance.

X: 6.3 5.8 6.1 6.9 3.4 1.8 9.4 4.7 7.2 2.4

Y: 5.3 8.6 4.7 4.2 4.9 6.1 5.1 6.3 6.8 5.2

Solution
Null hypothesis
$H_0 : \rho_s = 0$, there is no significant rank correlation between the two measures.
Alternative hypothesis
$H_1 : \rho_s \neq 0$, i.e there is a significant rank correlation between the two measures.
Level of significance $\alpha = 5\%$

Test statistic

Let us find the ranks for X and Y.

X	Y	R_1	R_2	$d = R_1 - R_2$	d^2
6.3	5.3	4	5	-1	1
5.8	8.6	6	1	5	25
6.1	4.7	5	9	-4	16
69	4.2	3	10	-7	49
3.4	4.9	8	8	0	0
1.8	6.1	10	4	6	36
9.4	5.1	1	7	-6	36
4.7	6.3	7	3	4	16
7.2	6.8	2	2	0	0
2.4	5.2	9	6	3	9

					$\sum d^2 = 188$

Therefore the sample rank correlation coefficient

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

$$= 1 - \frac{6 * 188}{10(99 - 1)} = -0.139$$

The expected or critical value at 5% level of significance with $n = 10, is$ 0.6364.

Conclusion

$r_s < 0.6364$, we accept H_0 $and\ conclude\ that$ there is no significant rank correlation between the two measures.

2. Test the hypothesis that X and Y are independent against the alternative that they are dependent if for a sample of size $n = 50$ pairs of observations we find that $r_s = -0.29$.. Use $\alpha = 0.05$

Solution:
We are given, $n = 50$, $r_s = -0.29$
Null hypothesis

H_0: X and Y are independent.
Alternative hypothesis

H_1: X and Y are dependent. Level of significance $\alpha = 0.05$

Test statistic (for $n > 30$)

$$Z = r_s\left(\sqrt{n-1}\,\right) = -0.29\left(\sqrt{50-1}\right) = -2.03$$

The value of $Z_\propto$ at $\alpha = 0.05$ is 1.96

Conclusion

If $|Z| > Z_\propto$ we reject H_0 and conclude that X and Y are dependent.

3. A consumer panel tested 9 makes a microwave ovens for overall quality. The ranks assigned by the panel and the suggested retail prices were as follows

Manufactures :	1	2	3	4	5	6	7	8	9
Panel rating :	6	9	2	8	5	1	7	4	3

Suggested price : 480 395 575 550 510 545 400 465 420

Is there a significant relationship between the quality and the price of a microwave oven? Use $\alpha = 0.05$

Solution

Null hypothesis

$$H_0 : \rho_s = 0.$$

Alternative hypothesis

$H_1 : \rho_s \neq 0$.

Level of significance $\alpha = 5\% = 0.05$

Test statistic

Let us take panel rating rank= R_1, and suggested price rank= R_2

X	R_1	Y	R_2	$d = R_1 - R_2$	d^2
6	6	480	5	1	1
9	9	395	1	8	64
2	2	575	9	-7	49
8	8	550	8	0	0
5	5	510	6	-1	1
1	1	545	7	-6	36
7	7	400	2	5	25
4	4	465	4	0	0
3	3	420	5	0	0
					$\sum d^2 = 176$

The sample rank correlation coefficient

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

$$= 1 - \frac{6 * 176}{9(81 - 1)} = -0.4667$$

The expected or critical value at 5% level of significance with $n = 9, is$ 0.6833.

Conclusion

$|r_s| < 0.6833$, we accept H_0 *and conclude that* there is no significant relationship between the quality and the price of a microwave oven.

4. The following are ratings of aggressiveness (X) and amount of sales (Y) in the last year for eight salespeople. Is there a significant rank correlation between the two measures? Use the 0.10 significance level.

X: 30 17 35 28 42 25 19 29

Y: 35 31 43 46 50 32 33 42

Solution

Null hypothesis

$$H_0 : \rho_s = 0.$$

Alternative hypothesis

$H_1 : \rho_s \neq 0$.

Level of significance $\alpha = 0.10$

Test statistic

X	R_1	Y	R_2	$d = R_1 - R_2$	d^2
30	6	35	4	2	4
17	1	31	1	0	0
35	7	43	6	1	1
28	4	46	7	-3	9
42	8	50	8	0	0
25	3	32	2	1	1
19	2	33	3	-1	1
29	5	42	5	0	0
					$\sum d^2 = 16$

The sample rank correlation coefficient

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

$$= 1 - \frac{6 * 16}{8(64 - 1)} = 0.8095$$

The expected or critical value at 0.10 level of significance with

$n = 9, is\ 0.619.$

Conclusion

$|r_s| > 0.619$, we reject H_0 .

Home Work

1. Most people belive that managerial produces better interpersonal relationships between a manager and her employees. The quill corporation has the following data matching years of experience on the part of the manager with the number of grievances filed last year by the employees reporting to that manager. At the 0.05 level of significance, does the rank correlation between these two suggest that experience improves relationships?

Age of manager :	32	43	42	29	56	62	45	39	40	35
No Of grievances :	5	2	4	4	3	2	4	5	4	6