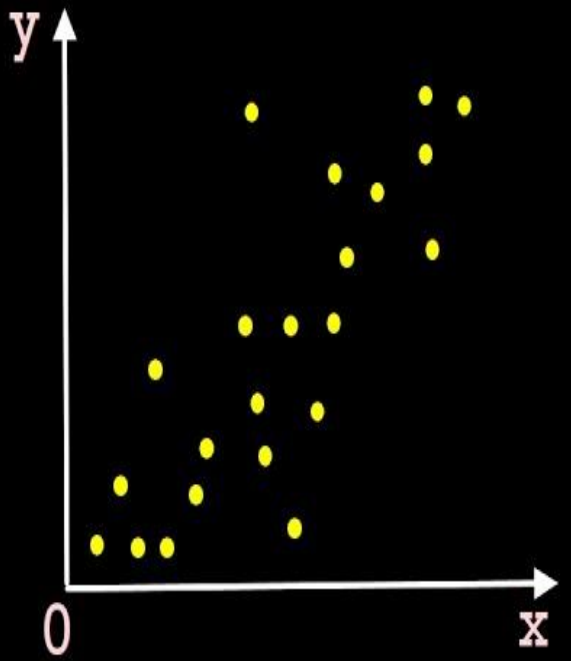
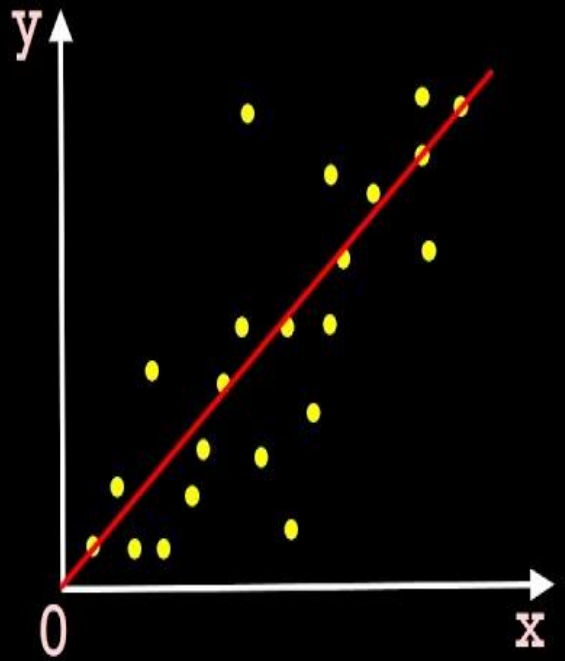
Correlation





Vs

Regression

Key Differences

Correlation	Regression
1. It indicates only the nature and extent of linear relationship	It is the study about the impact of the independent variable on the dependent variable. It is used for predictions.
2. If the linear correlation is coefficient is positive / negative , then the two variables are positively / or negatively correlated	The regression coefficient is positive, then for every unit increase in x, the corresponding average increase in y is b_{YX}. Similarly, if the regression coefficient is negative , then for every unit increase in x, the corresponding average decrease in y is b_{YX}.
3. One of the variables can be taken as x and the other one can be taken as the variable y.	Care must be taken for the choice of independent variable and dependent variable. We can not assign arbitrarily x as independent variable and y as dependent variable.
4. It is symmetric in x and y, *ie.*, $r_{XY}=r_{YX}$	It is not symmetric in x and y, that is, b_{XY} and b_{YX} have different meaning and interpretations.

Definition of Correlation -

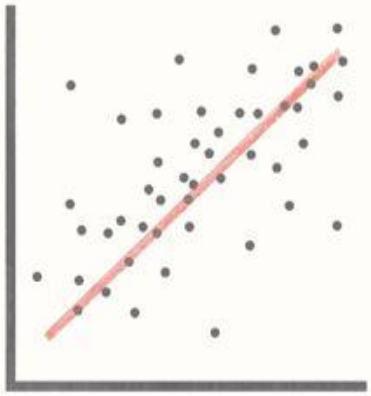
The term correlation is a combination of two words 'Co' (together) and the relation between two quantities.

Correlation can be either negative or positive.

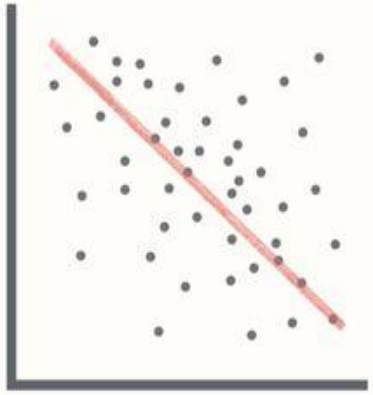
- **If the two variables move in the same direction, i.e. an increase in one variable results in the corresponding increase in another variable, and vice versa, then the variables are considered to be positively correlated. For example, Investment and profit.**

- **If the two variables move in different directions so that an increase in one variable leads to a decline in another variable and vice versa, this situation is known as a negative correlation. For example, Product price and demand.**

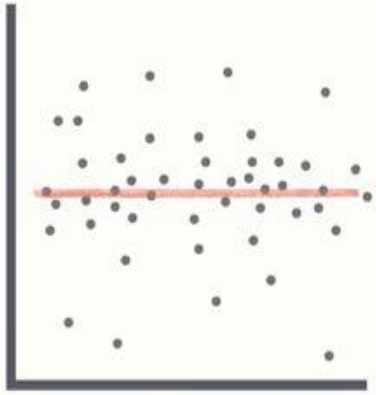
Correlation Coefficient



Positive Correlation



Negative Correlation



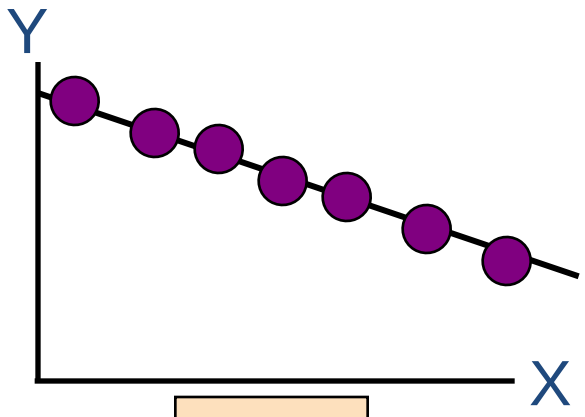
No Correlation

cov(X,Y) > 0 X and Y are positively correlated

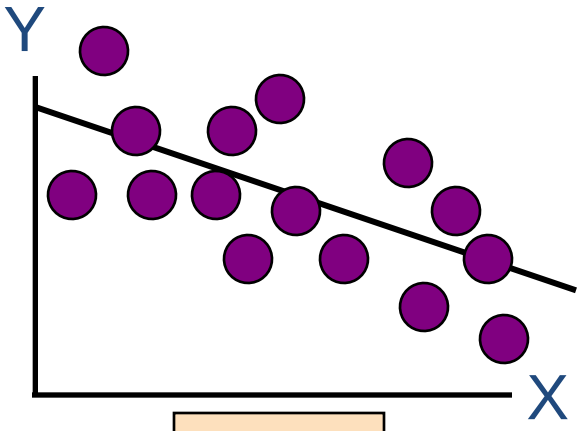
cov(X,Y) < 0 X and Y are inversely correlated

cov(X,Y) = 0 X and Y are independent

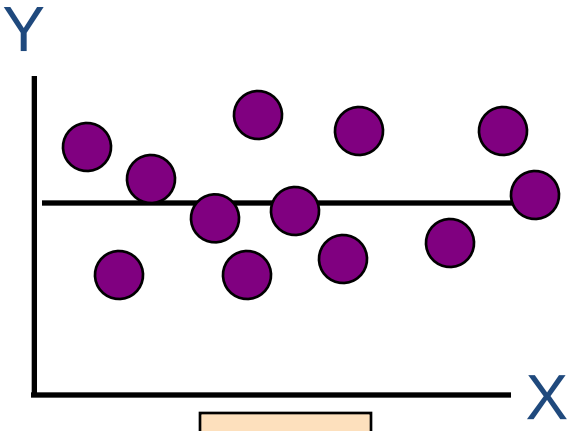
Scatter Plots of Data with Various Correlation Coefficients



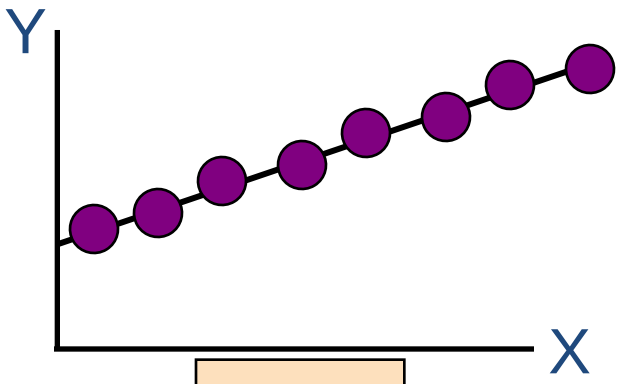
r = -1



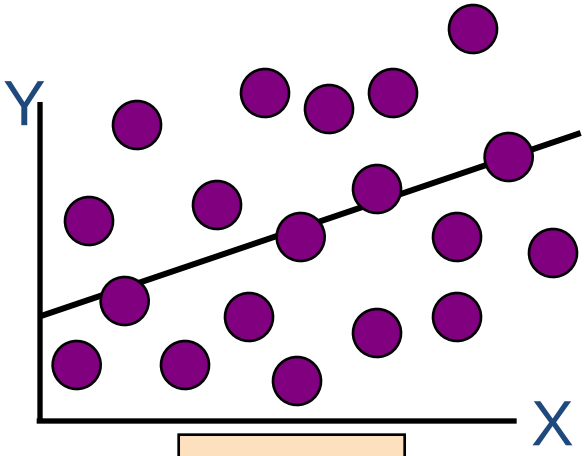
r = -.6



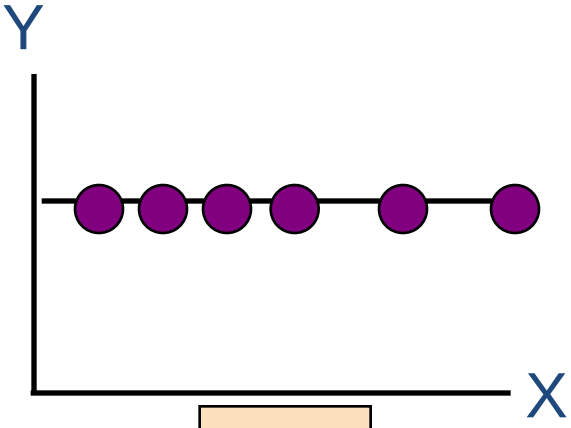
r = 0



r = +1



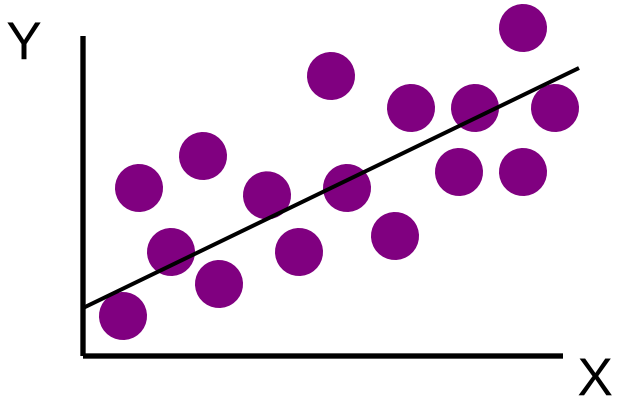
r = +.3

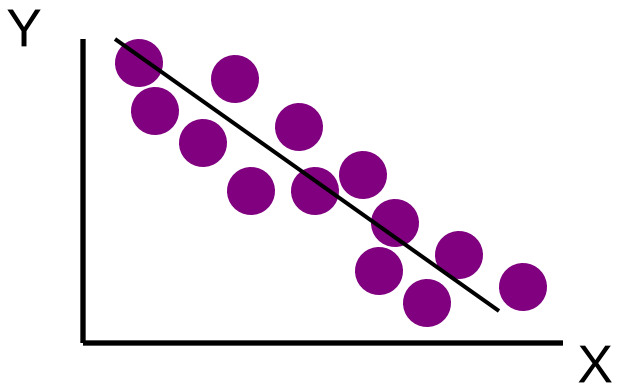


r = 0

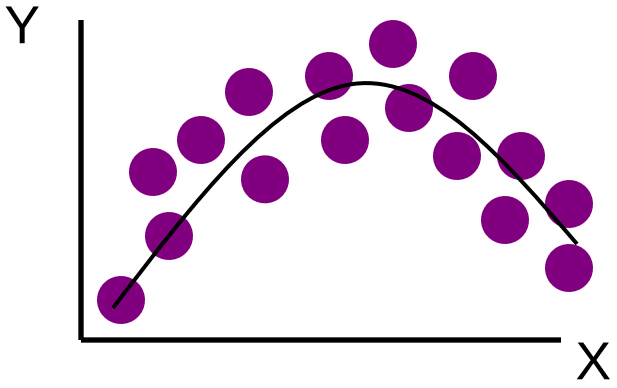
Linear Correlation

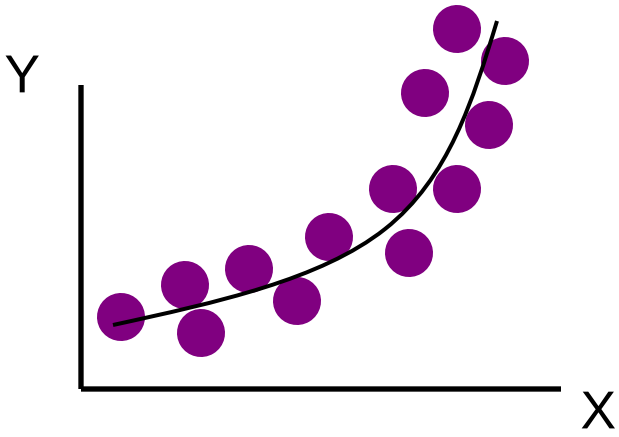
Linear relationships





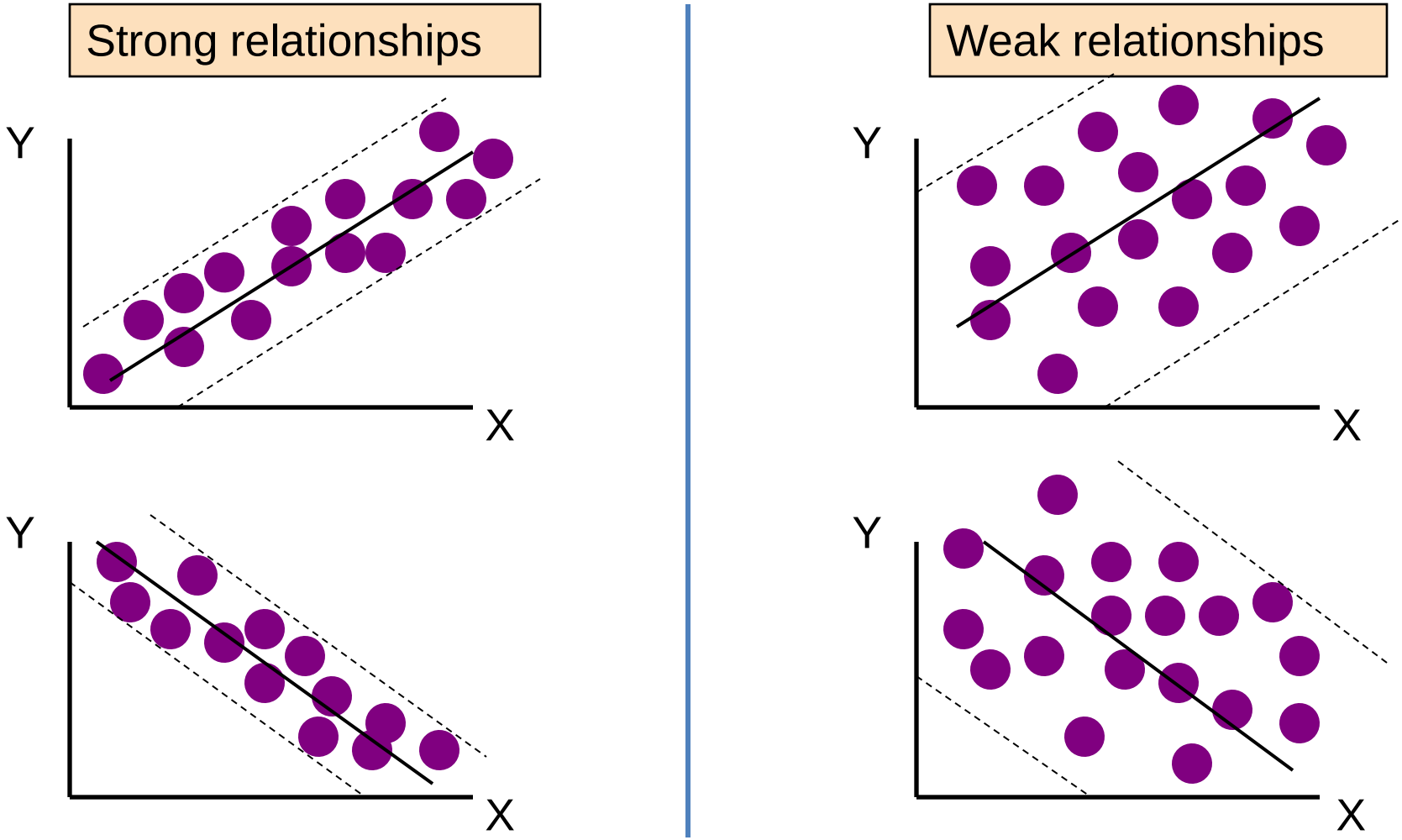
Curvilinear relationships





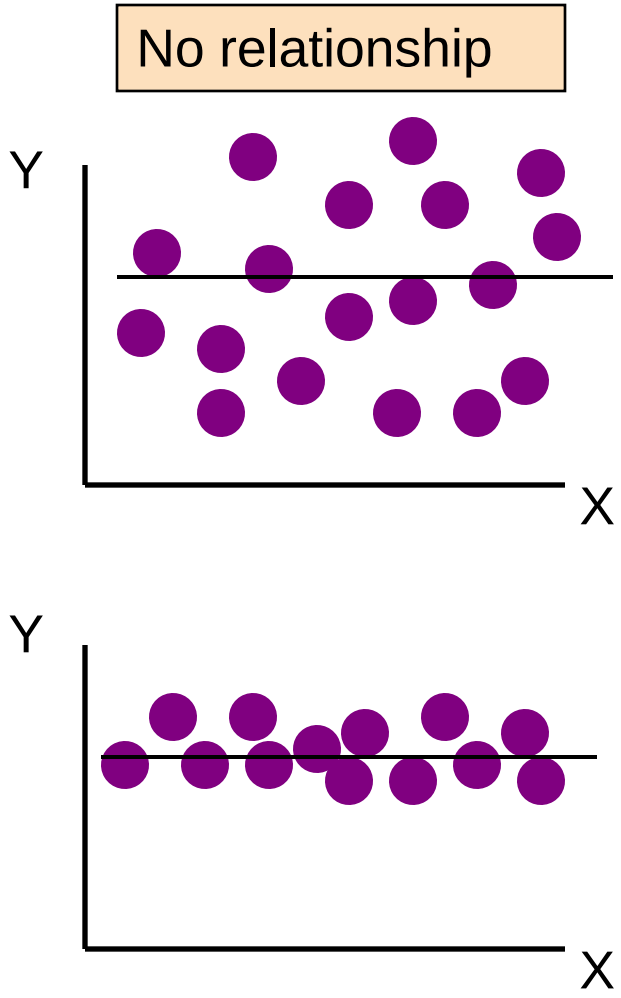
- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation



- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation



- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Calculating by hand…

$$\hat{r} = \frac{\mathrm{covariance}(x, y)}{\sqrt{\mathrm{var}\ x}\sqrt{\mathrm{var}\ y}} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

Simpler calculation formula…

$$\hat{r} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\cancel{n-1}}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\cancel{n-1}}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\cancel{n-1}}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

$$\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Karl Pearson Coefficient of Correlation.

As a measure of itensity or degree of linear relationship between two variables, Karl Pearson *(1867-1936). a British Biometrician.* deyeloped a.ormula called *Correlation Coefficient.*

Correlation coefficient between two random variables X and Y, usually denoted by

$$r(X, Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X \sigma_Y}$$

$$r(X, Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X\,\sigma_Y} = \frac{\frac{1}{n}\Sigma\,(x_i - \bar{x})\,(y_i - \bar{y})}{\left[\frac{1}{n}\Sigma\,(x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma\,(y_i - \bar{y})^2\right]^{1/2}},$$

Calculating the Correlation Coefficient

Sample correlation coefficient:

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

or the algebraic equivalent:

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

where:

r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

Type I : This method is used when given variables are small in magnitude.

Formula : $$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Example 1. Calculate Karl Pearson's coefficient of correlation between the age and weight of the children :

Age (years) :	1	2	3	4	5
Weight (kg.) :	3	4	6	7	12

Solution : $\Sigma X = 15$; $\Sigma Y = 32$; $\Sigma X^2 = 55$; $\Sigma Y^2 = 254$; $\Sigma XY = 117$

Age (X)	Weight (Y)	X^2	Y^2	XY
1	3	1	9	3
2	4	4	16	8
3	6	9	36	18
4	7	16	49	28
5	12	29	144	60
15	32	55	254	117

As $$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$\therefore\ r = \frac{5 \times 117 - 15 \times 32}{\sqrt{5 \times 55 - (15)^2}\ \sqrt{5 \times 254 - (32)^2}}$$

$$= \frac{585 - 480}{\sqrt{275 - 225}\ \sqrt{1270 - 1024}} = \frac{105}{\sqrt{50 \times 246}} = \frac{105}{\sqrt{12300}} = \frac{105}{110.90} = 0.9467$$ **Ans.**

Example 2. Calculate coefficient of correlation between death and birth rate for the following data.

Birth Rate	24	26	32	33	35	30
Death Rate	15	20	22	24	27	24

Solution

Birth Rate X	Death Rate Y	$(X-\bar{X}) = x$	$(Y-\bar{Y}) = y$	$(X-\bar{X})^2 = x^2$	$(Y-\bar{Y})^2 = y^2$	$(X-\bar{X})(Y-\bar{Y}) = xy$
24	15	–6	–7	36	49	42
26	20	–4	–2	16	4	8
32	22	2	0	4	0	0
33	24	3	2	9	4	6
35	27	5	5	25	25	25
30	24.	0	2	0	4	0
$\Sigma X=180$ $\bar{X} = \frac{180}{6} = 30$	$\Sigma Y=132$ $\bar{Y} = \frac{132}{6} = 22$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 90$	$\Sigma y^2 = 86$	$\Sigma xy = 81$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{(81)}{\sqrt{90 \times 86}} = \frac{81}{\sqrt{7740}} = \frac{81}{87.98} = .92$$

Example 10·2. *A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results :*

$$n = 25,\ \Sigma X = 125,\ \Sigma X^2 = 650,\ \Sigma Y = 100,\ \Sigma Y^2 = 460,\ \Sigma XY = 508$$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

Obtain the correct value of correlation coefficient.

(A) 0.67
(B) 1.37
(C) 0.03
(D) 2.37

Corrected $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\Sigma XY = 508 - 6\times14 - 8\times6 + 8\times12 + 6\times8 = 520$

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{1}{25}\times 125 = 5, \quad \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{25}\times 100 = 4$$

$$\text{Cov}(X,Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = \frac{1}{25}\times 520 - 5\times 4 = \frac{4}{5}$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \bar{X}^2 = \frac{1}{25}\times 650 - (5)^2 = 1$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \bar{Y}^2 = \frac{1}{25}\times 436 - 16 = \frac{36}{25}$$

$$\therefore \quad \text{Corrected } r(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{\frac{4}{5}}{1\times\frac{6}{5}} = \frac{2}{3} = 0.67$$

10·6·3. Limits for the Rank Correlation Coefficient. Spearman's rank correlation coefficient is given by

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

'ρ' is maximum, if $\sum_{i=1}^{n} d_i^2$ is minimum, *i.e.*, if each of the deviations d_i is minimum. But the minimum value of d_i is zero in the particular case $x_i = y_i$, *i.e.*, if the ranks of the ith individual in the two characteristic are equal. Hence the maximum value of ρ is + 1, *i.e.*, $\rho \leq 1$.

Grading Standards	Correlation Degree
$\rho = 0$	no correlation
$0 < \lvert\rho\rvert \leq 0.19$	very week
$0.20 \leq \lvert\rho\rvert \leq 0.39$	weak
$0.40 \leq \lvert\rho\rvert \leq 0.59$	moderate
$0.60 \leq \lvert\rho\rvert \leq 0.79$	strong
$0.80 \leq \lvert\rho\rvert \leq 1.00$	very strong
1.00	monotonic correlation

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

A 0.145

B. 0.245

c. 0.345

D None

	A	B	C	D
2				
3	IQ	Rock	Rank IQ	Rank Rock
4	99	2	4	2
5	120	0	9	1
6	98	25	3	9
7	102	45	5	10
8	123	14	10	4
9	105	20	6	7
10	85	15	1	5
11	110	19	7	6
12	117	22	8	8
13	90	4	2	3

	Marks									
English	56	75	45	71	61	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

(A)0.67
(B)0.77
(C)0.88
(D)0.99

English (mark)	Maths (mark)	Rank (English)	Rank (maths)
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
61	65	6.5	5
64	56	5	9
58	59	8	8
80	77	1	1
76	67	2	3
61	63	6.5	6

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

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

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

PROBABLE ERROR OF COEFFICIENT OF CORRELATION

- It is an measure of testing reliability of an observed value of coefficient of correlation. it depends on the condition of random sampling
- It is represented by "r"

$$P.E.r = 0.6745\ (1-r^2)/\sqrt{n}$$

r = coefficient of correlation.
n = number of pairs of observation.

Probable Error (P.E.)of r (contd…)

- P.E. (r) = $0.6745 \times \frac{1 - r^2}{n}$
- Limits of (population) correlation coefficient is defined as $r - P.E.(r) <= \rho <= r + P.E.(r)$ where ρ denotes correlation coefficient in population and r denotes correlation coefficient in sample
- By convention the rules are
  - If $|r| < 6 P.E.(r)$, then correlation is not significant and no correlation between 2 variables
  - If $|r| > 6 P.E.(r)$, then correlation is significant and this implies presence of strong correlation between 2 variables
  - If correlation coefficient is greater than 0.3 and probable error is relatively small, the correlation coefficient should be considered as significant

Regression is the determination of a statistical relationship between two or more variables

The basic relationship between *X* and *Y* is given by

$$\hat{Y} = a + bX$$

where the symbol $\hat{Y}$ denotes the estimated value of *Y* for a given value of *X*.

This equation is known as the regression equation of *Y* on *X* (also represents the regression line of *Y* on *X* when drawn on a graph) which means that each unit change in *X* produces a change of *b* in *Y*, which is positive for direct and negative for inverse relationships.

Regression Equations in Individual Series using Normal Equations

- This method is also called as Least Square Method.
- Under this method, regression equations can be calculated by solving two normal equations:
  - For regression equation Y on X: $Y = a + bX$
    - $\Sigma Y = Na + b\Sigma X$
    - $\Sigma XY = a\Sigma X + b\Sigma X^2$
  - Another Method
    - $b_{yx} = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2}$ & $a = \bar{Y} - b\bar{X}$
- Here a is the Y – intercept, indicates the minimum value of Y for X = 0
- & b is the slope of the line, indicates the absolute increase in Y for a unit increase in X.

Birinder Singh, Assistant Professor, PCTE

Regression Equations Using Regression Coefficients (Using Deviations from Actual Values)

- Regression Equation of Y on X
  - $Y - \bar{Y} = b_{yx}(X - \bar{X})$ where $b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$
- Regression Equation of X on Y
  - $X - \bar{X} = b_{xy}(Y - \bar{Y})$ where $b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

where ($\overline{X}$)is the mean of X series,

$\overline{Y}$ is the mean of Y series,

$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

Formula

for X on Y

$$X = a + bY$$

The normal equations are

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

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

Regression coefficient

$$b_{XY} = \frac{\sum XY}{\sum Y^2}$$

Regression equation

$$x = \bar{X} + b_{XY}(Y - \bar{Y})$$

for Y on X

$$Y = a + bX$$

The normal equations are

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

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

Regression coefficient

$$b_{YX} = \frac{\sum XY}{\sum X^2}$$

Regression equation

$$Y = \bar{Y} + b_{YX}(X - \bar{X})$$

Regression coefficients:

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

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

coefficient of correlation

$$r = \sqrt{(b_{yx} * b_{xy})}$$

Least Squares Fitting

A mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets ("the residuals") of the points from the curve.

Example: Using the method of least squares, find an equation of the form

$y = ax + b$ that fits the following data:

x	0	1	2	3	4
y	1	5	10	22	38

Solution: Consider the normal equations of least square fit of a straight line i.e

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \qquad (1)$$

$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i \qquad (2)$$

Here n =5.

Here n =5.

From the given data, we have,

x	y	xy	x^2
0	1	0	0
1	5	5	1
2	10	20	4
3	22	66	9
4	38	152	16
$\sum_i x_i = 10$	$\sum_i y_i = 76$	$\sum_i x_i y_i = 243$	$\sum_i x_i^2 = 30$

Therefore the normal equations are given by:

30a +10b =243 .....(3)

10a+5b=76.....(4)

On solving (3) and (4) we get

a = 9.1 , b= - 3 .....(5)

Hence the required fit for the given data is

y=9.1x - 3 ....(6)

Least Square Fit of a Straight Line

Suppose that we are given a data set $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\ldots, (x_n, y_n)$ of n observations from an experiment. Say that we are interested in fitting a straight line

$$y = ax + b$$

to the given data. Find the 'n' residuals e_i by:

$$e_i = y_i - (ax_i + b), \qquad i = 1, 2 \ldots\ldots n \qquad (2)$$

Now consider the sum of the squares of $e_i's$ i.e

$$E = \sum_{i=1}^{n} e_i^2$$

$$= \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 \qquad (3)$$

Note that E is a function of parameters a and b. We need to find a,b such that E is minimum. The necessary condition for E to be minimum is given by:

$$\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0 \qquad (4)$$

The condition $\frac{\partial E}{\partial a} = 0$ yields:

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{n} 2x_i[y_i - (ax_i + b)] = 0$$

i.e $$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \qquad (5)$$

Similarly the condition $\frac{\partial E}{\partial b} = 0$ yields

$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i \qquad (6)$$

Equations (5) and (6) are called as normal equations,which are to be solved to get desired values for a and b.

The expression for E i.e (3) can be re-written in a convenient way as follows:

$$E = \left(\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} y_i\right) \qquad (7)$$

1. Fit the straight line to the following data.

x	1	2	3	4	5
y	1	2	3	4	5

a) $y=x$

b) $y=x+1$

c) $y=2x$

d) $y=2x+1$